UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

FACULDADE DE MEDICINA

PROGRAMA DE PÓS-GRADUAÇÃO EM MEDICINA: CIÊNCIAS MÉDICAS

LIPOCALINA ASSOCIADA COM GELATINASE DE NEUTRÓFILOS HUMANOS (NGAL) PARA O DIAGNÓSTICO DA INJÚRIA RENAL AGUDA APÓS TRANSPLANTE DE RIM

GISELE RODRIGUES LOBATO

PORTO ALEGRE

2010

LIPOCALINA ASSOCIADA COM GELATINASE DE NEUTRÓFILOS HUMANOS (NGAL) PARA O DIAGNÓSTICO DA INJÚRIA RENAL AGUDA APÓS TRANSPLANTE DE RIM

GISELE RODRIGUES LOBATO

A apresentação desta tese de doutorado é exigência do Programa de Pós-Graduação em Medicina: Ciências Médicas, Universidade Federal do Rio Grande do Sul, para a obtenção do título de Mestre em Medicina: Ciências Médicas.

.

Orientadora: Profa. Dra. Maria Isabel Albano Edelweiss
Co-orientadora: Profa. Dra. Lídia Rosi Medeiros

PORTO ALEGRE

2010

L884g Lobato, Gisele

LIPOCALINA ASSOCIADA COM GELATINASE DE NEUTRÓFILOS HUMANOS (NGAL) PARA O DIAGNÓSTICO DA INJÚRIA RENAL AGUDA APÓS TRANSPLANTE DE RIM/ Gisele Lobatto ; orient. Maria Isabel Albano Edelweiss ; co-orient. Lidia Rosi Medeiros. – 2010.

266 f. : il.

Dissertação (Mestrado) - Universidade Federal do Rio Grande do Sul. Faculdade de Medicina. Programa de Pós-Graduação em Medicina: Ciências Médicas. Porto Alegre, BR-RS, 2010.

1. Injuria Renal Aguda 2. Revisão sistemática 3. Transplante de rim 4. Lipocalina I. Edelweiss, Maria Isabel Albano II. Medeiros, Lidia Rosi III. Título.

NLM: WJ 342

Catalogação Biblioteca FAMED/HCPA

AGRADECIMENTOS

Agradeço, com muito, reconhecimento, a todos os que, de uma maneira ou outra, direta ou indiretamente, possibilitaram que este trabalho fosse realizado. De maneira particular sou grata

à Professora Doutora Lídia Rosi de Freitas Medeiros, Mestre e Doutora em Epidemiologia pela UFRGS , por sua amizade, disponibilidade, paciência, compreensão, incentivo e que, por seu entusiasmo à ciência e à pesquisa, me auxiliou na elaboração desta dissertação;

à professora Maria Isabel Albano Edelweiss, Mestre em Nefrologia pela UFRGS e Doutora em Patologia pela USP, que despertou meu interesse pela pesquisa e incentivou meu ingresso no Curso de Pós-Graduação, por sua amizade, carinho e pela oportunidade de aprimoramento de crescimento científico, pelas palavras e por ter confiado e acreditado;

ao Doutor César Pereira Lima, meu grande amigo e com seu exemplo de luta pela vida me transmitiu que o importante é buscar seus objetivos, apesar de todos os revezes que a vida nos apresenta;

ao Doutor Theobaldo de Oliveira Thomaz pela sua competência profissional e pela ajuda que me deu nos momentos tão difíceis que tive que enfrentar;

à Mariane Rodrigues Lobato, Mestre em Biologia, minha irmã, que foi desde o início a grande colaboradora para a realização desta dissertação;

ao Dirceu Schimidlin Condessa Júnior pela sua paciência, compreensão e parceria nos momentos de dedicação a este trabalho;

aos meus colegas do Hospital Ernesto Dornelles, em especial à Dra Cinthia Vieira, Roberta Venturella e Bartolomé Diano, pelo convívio, incentivo e apoio;

aos professores do Curso de Pós-Graduação em Clínica Médica, pelos conhecimentos transmitidos, fundamentais para o constante aperfeiçoamento profissional e pessoal;

aos funcionários da Secretaria do Curso de Pós-Graduação em Clínica Médica, pela eficiência no atendimento e pela ajuda na resolução de problemas;

ao Lucas Garcia, responsável pela editoração do texto, feita com inegável capacidade, pela disponibilidade e carinho;

À minha família, alicerce primordial de tudo o que sou e realizei.

A minha mãe Rita e ao meu pai, Oly pelos exemplos de trabalho, dedicação e com quem aprendi que, para indivíduos bem preparados, os limites da vida s ão relativos

Aos meus filhos, Helena e Bruno, que por existirem me tornaram uma pessoa melhor.

Ao querido Antonio Carlos Herz da Jornada, nosso Toto, que há muito não está mais conosco, mas, que sei o quanto vibraria com todas as nossas conquistas.

RESUMO

Introdução: Na Nefrologia, e em particular nos transplantes renais, houve uma substancial evolução nos últimos anos. Atualmente a Lipocalina Associada à Gelatinase dos Neutrófilos Humanos (NGAL) é considerado um biomarcador precoce de injúria renal nos pacientes transplantados de rim. **Metodologia:** A revisão sistemática com enfoque diagnóstico – NGAL urinário e sérico *versus* creatinina para injúria renal aguda em rim transplantado pela AKIN (*Acute Kidney Injury Network*), foram realizadas nos seguintes bancos de dados: MEDLINE (Pubmed and OVID interface), Cochrane Central de registros de Estudos Controlados (CENTRAL), EMBASE, LILACS, SCOPUS, Web of Science, BIOSIS, GOOGLE ACADÊMICO, entre janeiro de 2000 até 20/07/2010. Foram procuradas também publicações através de busca manual, em periódicos médicos relevantes até maio de 2010 e listas de referência de artigos bem como resumos de conferências médicas em Congressos. Estudos comparando NGAL urinário e sérico e creatinina no diagnóstico da injúria renal aguda em rim transplantado, começaram a ser realizadas a partir de 2000. Os dados foram extraídos de forma independente por dois revisores que aferiam a qualidade do estudo e a qualidade dos dados extraídos. A extração dos dados inclui avaliar as características clínicas de cada estudo, como tipo de participantes, intervenção e desfechos e tipos de exames realizados (o teste padrão-creatinina sérica e o teste diagnóstico NGAL urinário e sérico). A qualidade dos estudos de foi apurada por um tipo de instrumento de avaliação metodológica: QUADAS (Qualidade dos estudos de Acurácia Diagnóstica incluídos numa Revisão Sistemática). **Casuística:** Foram revisados 147 estudos publicados, de janeiro de 2000 até maio de 2010, sobre NGAL e IRA, sendo 17 escolhidos por trazer a maioria dos dados completos e destes, 7 foram selecionados para serem submetidos à análise crítica da validade metodológica, pois, preenchiam os critérios exigidos. **Resultados:** A revisão sistemática com enfoque diagnóstico – NGAL urinário e sérico *versus* creatinina sérica na avaliação de injúria renal aguda em rim transplantado, somente encontrou quatro estudos de coorte e três transversais, mas com pobre qualidade metodológica. Portanto, tornou-se impossível realizar a meta-análise devido à diferença entre os estudos. Devido à presença de uma heterogeneidade devido a aspectos metodológicos, clínicos ou /e estatística, não se realizou os cálculos de uma meta-análise. **Conclusão:** A revisão sistemática com enfoque diagnóstico não encontrou evidências que apóiem o uso da dosagem urinária e sérica de NGAL como uma ferramenta diagnóstica na avaliação da injúria aguda pós-transplante renal. Verifica-se a necessidade de estudos com boa qualidade metodológica comparando o NGAL urinário e sérico e creatinina sérica no manejo da injúria renal que pode acometer os rins transplantados. Apesar de todas as nossas tentativas, através de uma revisão minuciosa da literatura médica em diversos bancos de dados e na Grey Literature (British Library e Google Acadêmico) da última década, de podermos eleger o NGAL como um instrumento capaz de nos dar uma indicação precoce sobre os eventos agudos de perda de função renal, não conseguimos demonstrar nosso objetivo, pois, não obtivemos fundamentos que pudessem alicerçar esta afirmação. Desta forma depreendemos que há a necessidade para o prosseguimento desta investigação, uma vez que ao longo da revisão bibliográfica, este teste se mostrou de extrema importância no manejo precoce da injúria renal aguda que pode acometer os rins transplantados no pós-operatório imediato. Este é o teste em que não há qualquer tipo de intervenção no paciente se comparado ao que o que temos hoje ainda é a

biópsia renal, que é um método extremamente invasivo, agredindo um órgão com um universo de respostas imunológicas desordenadas e com potencial de fracasso através desta abordagem diagnóstica mais agressiva. Seria interessante dar prosseguimento a estes trabalhos elaborando estudos prospectivos com enfoque diagnóstico, aprimorados metodologicamente, evitando determinados vieses detectados nessa revisão. Nosso trabalho de Doutorado pretende contemplar este aspecto trazendo uma contribuição mais sustentável para o uso do NGAL na prática Nefrológica dos Transplantes.

ABSTRACTS

Introduction: In Nephrology, and in particular in the renal transplants, there was a substantial evolution in the last years. Nowadays Lipocalin Associate to Gelatinase of Neutrophil Human (NGAL) is considered a biomarker precocious of renal offense in the patients transplanted of kidney. **Methodology:** The systematic review with focus diagnosis – serum and urinary NGAL *versus* creatinine for acute renal offense in kidney transplanted by AKIN (*Acute Kidney Injury Network*), was accomplished in the next databases: MEDLINE (Pubmed and OVID interface), Cochrane records Central of Controlled Studies (CENTRAL), BASE, LILACS, Scopus, Web of Science, BIOSIS, GOOGLE ACADEMIC, between January 2000 up to 20/07/2010. Also were demanded publications through seeks manual, in important medical periodicals until May 2010 and goods reference lists as well as summaries of medical conferences in Congresses. Studies comparing serum and urinary NGAL and creatinina in the diagnosis of the acute renal offense in transplanted kidney, they started to accomplished being starting from 2000. The data were extracted of independent form for two reviewers who checked the study quality and the quality of the extracted data. The data extraction includes evaluate the clinical characteristics of each study, like kind of participants, intervention and outcomes and kinds of accomplished exams (the test standard- serum creatinine and the test diagnosis serum and urinary NGAL). Studies quality was select for a kind of instrument of methodological evaluation: QUADAS (Studies quality of accuracy diagnostic included in a Systematic Revision). **Casuistically:** They were revised 147 published studies, of January 2000 until May 2010, about NGAL and IRA, being 17 chosen to for bring most complete data and of these, 7 were selected to be submitted to the critical analysis of the methodological validity, because, performed the demanded criteria. **Results:** The systematic revision with focus diagnosis – serum and urinary NGAL *versus* creatinine in the evaluation of acute renal offense in transplanted kidney, only found four cohort studies and three cross-sectional, but with poor methodological quality. Therefore, it became impossible to accomplish the goal-analysis due to the difference among studies. Due to the presence of a heterogeneity due to methodological, clinical aspects or /e statistical, did not accomplish the calculations of a goal-analysis. **Conclusion:** The systematic revision with focus diagnosis did not find evidences that support the use of the urinary and serum dosage and NGAL as a tool diagnostic in the evaluation of the acute offense post-transplant renal. It verifies-if the need to studies with good methodological quality comparing NGAL urinary and sérico and serum creatinine in the handling of the renal offense that can attack the transplanted kidneys. Despite all our attempts, through a meticulous revision of the medical literature in several databases and in Grey Literature (British Library and Google Academic) of the last decade, of can choose NGAL as an instrument capable of give us a precocious indication about the acute events of loss of renal function, do not manage to demonstrate our goal, because, did not obtain foundations that could sustained this affirmation. Thus deduce that there is the need for the prosecution of this investigation, once along the bibliographical revision, this test showed of extreme importance in the precocious handling of the acute renal offense that can attack the kidneys transplanted in immediate postoperative. This is the test in which there is not any kind of intervention in the patient if compared to that what we have today still is the renal biopsy, that is an extremely invasive method, attacking an

organ with an immunologic answers universe disordered and with failure potential through of this approach diagnostic more aggressive. It would be interesting to give prosecution to these jobs elaborating prospective studies with focus diagnosis, improved methodological, avoiding determined bias detected in this revision. Our Doctorate work intends to contemplate this aspect bringing a more sustainable contribution for NGAL's Use in the practice Nephrologists of the Transplants.

LISTA DE ABREVIATURAS

AKIN	*Acute Kidney Injury Network*
AUC	Área sob a curva
BIOSIS	*Research database with the most current sources of life science information*
CC quimiocinas	Citocinas que funcionam como mediadores ou reguladores de inflama ção. São divididas em 4 sub-famílias dependendo da posição dos resí duos cisteína
CCL 17	Citocina, gene no cromossomo 16; induz quimiotaxia celular
CCL 2	Também conhecida como MIP-1, proteína inflamatória dos macrófagos
CCL 22	Citocina, gene no cromossomo 22; induz quimiotaxia celular
CCL 3	MIP-1α
CCL 4	MIP-1β
CCR 3 / 4 / 5 / 8	receptor de CCL 3 / 4 / 5 / 8
CD	Células Dendríticas
CD 8 / 26 / 30 / 62 L	Genes de ativação linfocitária
CXC	Citocina que apresenta duas cópias de cisteína
CXCL 8 / 10 / 12	Proteína 8 induzida pela citocina-interferon
CyC	Cistatina C
EMBASE	Excerpta *Medica Database*
HLA	Antígeno de Histocompatibilidade
IBECS	Índice Bibliográfico Espanõl en Ciencias de la Salud
IFKF	International Foundation of Kidney Failure
IGFBP-1 / 2 / 6	Proteína Ligadora do Fator de Crescimento *"Insulina-Like"*
IL-10	Interleucina 10 participa no controle de reações imunitárias inatas e celulares.
IL-12	Interleucina 12, chamada de fator de maturação de linfócito citotóxico
IL-14	Interleucina 14, mediadora entre as células plasmáticas e o citoplasma
IL-4	Interleucina 4, relacionada com a proliferação de linfócito B
IP-10	O mesmo que CLCX-10
IRA	Injúria Renal Aguda
IRC	Insuficiência Renal Crônica
ISN	*International Society of Nephrology*
KDa	Kilodaltons
KIM	Molécula da Injúria Renal - 1
LAG 3	Gene de ação linfocitária 3
L-FABP	Proteína ligadora dos ácidos graxos (do fígado)
LILACS	Literatura Latino-Americana em Ciências da Saúde
MEDLINE	Medical Literature Retrieval System Online
NAG	n-acetil-D Glicosaminidase
NGAL	Lipocalina Associada com Gelatinase de Neutrófilos Humanos
OR	*Odds ratio*

PUBMED	Banco de dados de informações bibliográficas na área da ciência Desenvolvido pela *National Center for Biotechnology Information (NCBI) at the National Library of Medicine*
QUADAS	*Quality assessment of studies of diagnostic accuracy included in systematic reviews*
RANTES	*Regulated on activation normal T cell expressed and secreted*
RIFLE	Renal / Injury / Failure / Loss / End Stage
RNAm	Ácido Ribonucléico mensageiro
ROC	*Receiver operator characteristics*
RS	Revisão Sistemática
S	Sensibilidade
SBN	Sociedade Brasileira de Nefrologia
SCOPUS	*Database of abstracts and citations for scholarly journal articles and peer reviewed literature*
TIMs	Inibidores teciduais da Metaloproteinase
TNFα	Fator de Necrose Tumoral
TR	Transplante Renal
VEGF	Fator de Crescimento do Endotélio Vascular
WEB OF SCIENCE	Banco de Dados de referências bibliográficas do *Institute for Scientific Information (ISI)*, que contém informações sobre A produção científica produzida no mundo a partir de 1974

SUMÁRIO

INTRODUÇÃO

1. INTRODUÇÃO

Dentro da estatística mundial as doenças renais vêm crescendo de forma epidêmica (1). O rim assim como vários outros órgãos que não tinham a grandeza do coração, órgão sempre entendido como o mais vital de todos, ficou silenciosamente escondido da visão de todos inclusive dos médicos.

Em 1941, na Holanda ocupada pelos nazistas um médico, que administrava um dos primeiros Bancos de Sangue da Europa, ofereceu tratamento hemodialítico através do chamado “rim artificial” para alguns soldados vitimados por queimaduras extensas, e que, em conseqüência destas desenvolveram insuficiência renal (2). Este médico que se chamava Willem Johan Kolff (1911-2009), holandês naturalizado americano trouxe, através desta iniciativa pioneira, um reconhecimento sobre a importância dos rins que começaram a ter seu lugar de destaque na lista dos órgãos imprescindíveis para a manutenção da vida (2).

Hoje, graças a todos os avanços diagnósticos e terapêuticos, podemos reverter situações médicas que até então seriam impossíveis de acontecer, e dentro destas está a insuficiência renal nos seus mais variados graus de severidade. Os maiores recursos foram desenvolvidos, inicialmente, no aprimoramento de máquinas de hemodiálise que tivessem a capacidade de trazer o mais fiel possível às condições homeostáticas vitais. Entretanto, a insuficiência renal não é um fim em si, e todos os nossos esforços estão voltados para a recuperação destes rins e se tal não for possível, para o transplante deste órgão.

E é aqui que nos deparamos com uma outra situação, extremamente grave que é possibilidade de que tal insuficiência se faça no novo rim, colocando por terra todo o

nosso empenho em trazer as condições idéias para nosso paciente, fazendo-o retornar ao tratamento hemodialítico. Nosso trabalho visa, fundamentalmente, avaliar através de uma revisão sistemática, a relevância de uma nova ferramenta não-invasiva que possibilitará um diagnóstico precoce da injúria renal aguda no rim transplantado, nos possibilitando uma abordagem mais focada e recuperação mais efetiva desta agressão ao rim, diminuindo assim a exposição de nossos pacientes a condutas invasivas e com drogas imunossupressoras capazes de aumentar sua morbi-mortalidade.

Para este propósito fizemos um estudo de Revisão Sistemática nesta área, buscando as evidências que pudessem fortalecer nosso objetivo principal que é o de determinar a acurácia diagnóstica do NGAL urinário e sérico no diagnóstico da injúria renal aguda em pacientes transplantados de rim (ANEXO 1).

REVISÃO DA LITERATURA

2. REVISÃO DA LITERATURA

De acordo com as estatísticas da Sociedade Internacional de Nefrologia (ISN) (1) e da Fundação Internacional de Insuficiência Renal (IFKF) (2) há ,em todo o mundo, 500 milhões de pessoas que sofrem de problemas renais e 1,5 milhão delas estão em diálise. As estatísticas revelam também que uma em cada dez pessoas no mundo sofre de doença renal crônica e o custo cumulativo global para diálise e transplantes para a próxima década deve exceder US$ 1 trilhão (3). Mais de 80% dos pacientes que fazem hemodiálise estão nos países desenvolvidos (2). Na Índia e Paquistão, por exemplo, menos de 10% das pessoas que precisam recebem algum tipo de terapia e na África quase não há acesso ao tratamento e o desfecho é sempre fatal (3). Cerca de 2 milhões de brasileiros sofrem de doenças renais e cerca de 60 % não sabem que tem o problema (4).

O crescente número de doentes renais no Brasil já o tornou o terceiro maior mercado de hemodiálise do mundo. De 2000 a 2008, o número de pacientes que fazem diálise no Brasil cresceu 84% e estima-se que neste ano, de 2010, o número de pacientes seja de 125 mil (4).

Segundo a Sociedade Brasileira de Nefrologia (4), em 2005, foram 32.329 novos pacientes. A taxa de aumento de 2005 para 2006 foi estimada em 8,8% (4). Segundo as informações desta Sociedade, dos 120 mil brasileiros que precisam fazer hemodiálise, apenas 70 mil estão em tratamento. Os números apontam ainda que 47% dos pacientes em diálise estão na fila do transplante renal e 25% dos pacientes em tratamento, são

diabéticos. Estima-se que em 2010 o número de pessoas em diálise no Brasil seja de 125 mil (4).

Conforme os dados apontados neste último censo da SBN a insuficiência renal crônica (IRC) representa um problema clínico comum e devastador, com uma taxa permanentemente alta de mortalidade e morbidade (4). O custo, não só de vidas, mas, também econômico muitas vezes é de tal monta que nos perguntamos, freqüentemente, de que forma podemos evitar ou se não corrigir no menor tempo possível este tipo de situação, que é em princípio passível de ser revertida.

Transplante de Rim (TR) é o tratamento de escolha para a insuficiência renal estágio 5 (5). Hoje, no Brasil, aproximadamente 35.000 estão em fila para transplante, mas, somente 3.000 conseguem ser transplantados, ou seja, apenas 10% da população necessitada (4). Se somarmos os pacientes transplantados (10%) aos que morrem em hemodiálise (15 a 25%) restam, anualmente 65 a 75% dos pacientes em lista de espera. A esse grupo devem-se somar os novos renais crônicos que surgem todo o ano, em torno de 35 a 50 / milhão de habitantes (4) e assim teremos esta realidade arrasadora, que é a da população sem tratamento adequado para sua patologia renal, apesar de todos os esforços realizados até hoje para mudar este cenário. Além disso, tais pacientes estão mais idosos e tem comorbidades adicionais que representam uma alta taxa de mortalidade após o transplante (6). Na realidade, há uma concordância que o baixo prognóstico de sucesso, nesta população de pacientes, é devido à interação entre a alta prevalência dos clássicos fatores de risco e condições inerentes ao TR (3). A melhoria na sobrevida dos transplantes de rim apresenta um salto de qualidade neste último ano, incomparável ao que até agora vinha se obtendo se pensarmos nas últimas três décadas (7). A escolha de doador, o preparo do receptor que se submete a uma ampla investigação multissistêmica com o objetivo de diminuir as comorbidades, o

aperfeiçoamento das técnicas cirúrgicas com um menor tempo de isquemia-reperfusão, o progresso nas drogas imunossupressoras, a qual nos possibilita a recuperação de um rim transplantado sem que tornemos esta terapêutica de tal forma agressiva que venhamos a salvar o rim e perder o paciente. Todos estes progressos têm nos amparado quando da decisão de submetermos um paciente ao transplante de rim. Entretanto, de todo estes progressos ainda têm uma situação, cujo empenho para solucioná-la, tem sido alvo de um esforço importante da comunidade científica, se incluído aqui os laboratórios que investigam marcadores imunológicos.

A injúria renal aguda resultante do atraso na recuperação da função renal complica de 4-10% dos rins transplantados de doadores vivos e de 5-50 % de todos os transplantes efetuados com rim de cadáver (8). Entende-se por injúria renal aguda (IRA), no transplante renal, a necessidade de procedimento dialítico dentro da primeira semana pós-cirurgia (6) . Com a melhoria das estratégias para o aumento do pool de doadores usando-se “doador com critério-extendido” e rins “doados pós-morte cardíaca” a chance de aumento na incidência de IRA é muito maior (6) .

O transplante renal pode ser realizado com três tipos de doadores: o doador vivo relacionado, o doador vivo não relacionado e o doador cadáver. O doador vivo relacionado inclui somente até o 4º grau de parentesco e o doador não relacionado, com exceção de cônjuge, necessita de autorização judicial (4).

A nefropatia crônica do transplante assim como a IRA é decisiva na evolução do transplante renal e, esta última que é indutora de disfunção do enxerto renal é dependente da atividade transcriptional das células infiltrativas (5-11). Este conceito é consistente nas pesquisas feitas por Grimm *et a*l (10) onde a rejeição clínica é distinguível da subclínica apenas pela presença de uma população ativada, de macrófagos (5, 10).

Embora não seja nossa intenção no presente trabalho, discorrer sobre o transplante renal propriamente dito achamos pertinente discutir alguns aspectos sobre a fisiopatologia da rejeição.

A maioria das causas de perda do rim transplantado são, sabidamente, imuno-mediadas e classificadas como devida à rejeição aguda (11,7%), fibrose intersticial e atrofia tubular (IF/TA) (30,7%) com metade destes relacionados à nefropatia poliômica (7,1%) e rejeição mediada por anticorpos ou celular (8,5%), ou relacionada à doença glomerular (36,6%--doença recurrente (15%), *de novo* (6,5%) e glomerulopatia do transplante (TG 15%) (11) .

A rejeição do enxerto é classificada segundo características histopatológicas, usando-se a classificação por critérios adotados após reunião em BANFF (Canadá), ou do tempo em curso podendo ser hiper-aguda, aguda e crônica (12, 13). A rejeição hiper-aguda é caracterizada pela oclusão trombótica da vasculatura do enxerto que se iniciam minutos a horas após a anastomose vascular do enxerto no receptor. É mediada por anticorpos pré-existentes na circulação do hospedeiro que se liga a antígenos endoteliais do doador (10).

A rejeição aguda ocorre nos primeiros seis meses pós-transplante e é caracterizada por um aumento abrupto dos níveis de uréia e creatinina associado à redução do volume urinário (5) . Desde a introdução da ciclosporina, há aproximadamente 20 anos, dor à palpação do enxerto e febre frequentemente observados no passado, não são mais comuns (14).

A rejeição crônica denominada hoje como nefropatia crônica do enxerto, apresenta-se de forma insidiosa, com aumento lento dos níveis de uréia e creatinina. Para o seu diagnóstico é importante a biópsia renal, devendo ser diferenciada das glomerulopatias recidivadas, "*de novo*", infecções oportunistas e rejeições sub-clínicas.

Estas últimas são caracterizadas pela presença histológica de rejeição aguda sem alteração da função renal (15).

Baseado nos critérios de BANFF, a biópsia do rim transplantado pode ser classificada em normal, rejeição mediada por anticorpos, rejeição *borderline*, rejeição celular aguda e nefropatia crônica do enxerto (12, 13).

A rejeição mediada por anticorpos é caracterizada pela presença de pelo menos três dentre quatro alterações; C4d no enxerto renal, presença de anticorpo antidoador (anti-HLA), características tissulares de rejeição aguda e piora de função renal (13, 14). A *borderline* é caracterizada pela suspeita clínica de rejeição celular aguda com focos e tubulite. A rejeição celular aguda é caracterizada histologicamente pela infiltração intersticial de células mononucleares e ocasionalmente eosinófilos e rompimento da membrana basal tubular (tubulite) por infiltração celular. A nefropatia crônica tem como característica a presença de fibrose intersticial e atrofia tubular (13, 16).

A rejeição aguda é um processo de lesão vascular e parenquimatosa mediada por células T e anticorpos, e que geralmente se inicia após a primeira semana do transplante. As células T ativadas, em resposta principalmente a moléculas de antígeno de histocompatibilidade humanos (HLA) (16) em células parenquimatosa e endoteliais, destroem células do enxerto ou produzem citocinas que recrutam células inflamatórias com consequente lesão do enxerto. As células endoteliais são os alvos mais precoces a serem lesados nos enxertos renais. Endotelite microvascular é um achado inicial frequente em episódios de rejeição aguda. Endotelite ou arterite intimal em artérias de tamanho médio (16) também ocorrem em um estágio inicial de rejeição aguda e é indicativa de rejeição grave. As Células T CD4 e CD8 contribuem com a rejeição aguda. Várias evidências sugerem que o reconhecimento e morte de células do enxerto por linfócitos T CD8 são importantes mecanismos de rejeição aguda. As células T CD4,

secretando citocinas e induzindo reações tipo hipersensibilidades no enxerto, parecem ser também importantes na rejeição aguda. Os anticorpos podem também mediar rejeição aguda, levando à necrose transmural da parede que ocorre na rejeição hiperaguda (16).

A rejeição crônica, também denominada de nefropatia crônica do enxerto, é caracterizada pela fibrose e distúrbios vasculares com perda da função do enxerto, ocorrendo por um período prolongado (16). Embora seja a principal causa de perda do enxerto sua patogênese é menos conhecida que a rejeição aguda e hiperaguda. A fibrose pode ser o resultado de reações imunes ou de produção de citocinas que estimulam fibroblastos ou representa cicatriz após a necrose celular da rejeição aguda (16).

Após esta pequena revisão, relataremos agora nosso trabalho sobre biomarcadores de função renal no transplante de Rim, no tocante à IRA.

Entendemos que o primeiro conceito que devemos traçar, para dar seguimento a nossa revisão, diz respeito aos critérios da injúria renal aguda. Uma definição precisa e operacional tem permanecido sujeita a mais de uma interpretação. Poderíamos aqui citar todos os conceitos que vem sendo sugeridos ao longo dos anos, mas, nosso objetivo é usar o consenso mundial.

A IRA é reconhecidamente encontrada em todos os campos da prática médica. Infelizmente, esta síndrome tem sido difundida através de definições inconsistentes, esquemas fisiopatológicos simplistas, e ferramentas diagnósticas insensíveis. Recentes avanços na definição de IRA, entendendo sua fisiopatologia, e aumentando a acurácia diagnóstica de ferramentas testes podem, eventualmente, ter impacto sobre o manejo desta doença e seu desfecho clínico (17).

Este conceito se baseia numa proposta da AKIN (Acute Kidney Injury Network), que vem ganhando aceitação clínica, e que usa duas opções para a medida da redução aguda da função renal (< 48 horas) (17):

1. Aumento nos níveis séricos de creatinina (absoluto, > 0,3 mg/dl) ou aumento >150 a 200 % (1,5-2 X) do valor basal;
2. Oligúria (<0,5 ml /Kg/h por mais de 6 horas); e aqui se entende como um volume de urina < 400mL /6 h.

O termo IRA é sugerido para enfatizar a natureza reversível da condição que acomete o rim (17), particularmente quando nosso raciocínio é em relação aos transplantes renais, pois, é uma situação passível de melhora e cura, se usada terapêutica correta e no tempo certo. [18,21] Os estudos que procuraram investigar a epidemiologia, o tratamento e a prevenção da insuficiência renal utilizaram definições laboratoriais heterogêneas, o que prejudicou a análise dos trabalhos e motivou um grupo de médicos intensivistas e nefrologistas a se reunirem no grupo ADQI (*Acute Dialysis Quality Initiative)* para o desenvolvimento de diretrizes e consensos baseados em evidências para o tratamento e prevenção da IRA, com uma definição específica e universalizada de lesão renal aguda, chamada de critério RIFLE (22, 23). Este critério, porém, já sofreu modificações, propostas pela *Acute Kidney Injury Network*, e será descrito a seguir.

Graças ao critério RIFLE (18) o entendimento médico da IRA vem sendo bem utilizado. Este acrômio define IRA baseado em três graus de aumento de severidade (Risco, Injuria, Insuficiência-F*ailure)* e dois tipos de desfecho (Perda-Loss, Estágio final). Este sistema, complementar àquele definido pela AKIN, descreve a severidade da

disfunção renal baseada no aumento dos níveis de creatinina e declínio no débito urinário. HOSTE *et al* (18) observou que os três graus que descrevem o RIFLE-risco, injúria, perda, estão associados com níveis de 8,8%, 11,4% e 26,3%, respectivamente em pacientes internados em hospital.

O diagnóstico precoce de IRA é tão difícil, se não impossível. Tentativa com modelos experimentais de IRA não tem se mostrado adequada e tem sido pouco conclusiva, enquanto marcador molecular de detecção de prejuízo da função renal, em tempo real, ainda não foi confirmado cientificamente (18), ou se o foram, não se somaram às ferramentas de diagnóstico e prognóstico da IRA no transplante renal.

Estudos pré-clínicos da fisiopatologia da IRA têm contribuído imensamente para o desenvolvimento de biomarcadores para esta condição de injúria aguda. Esta fisiopatologia envolve fatores tubulares, inflamatórios, e vasculares (19). Os marcadores séricos e urinários tradicionais (creatinina, uréia, cilindrúria, fração de excreção do sódio, habilidade de concentração urinária) que vem sendo usados por décadas nos estudos clínicos para diagnóstico e prognósticos da IRA são insensíveis e inespecíficos e não refletem diretamente a injúria das células renais. Muitos destes marcadores representam, na realidade, conseqüências funcionais da lesão (20).

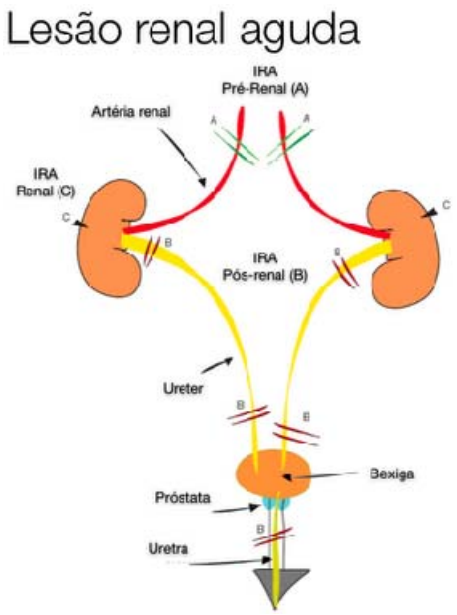


Figura 1: Lesão Renal Aguda (21)

2.1 Critério RIFLE

Consiste em 3 níveis graduados de lesão renal (RIFLE=RISK, INJURY AND FAILURE) e 2 medidas de desfecho (LOSS and END STAGE RENAL FAILURE) :

1) RISK (Risco) = aumento em 1,5x na creatinina sérica ou decréscimo de 25% na taxa de filtração glomerular ou débito urinário menor que 0,5 ml/Kg/h por 6 horas; (Risco de morte 2,4x maior).

2) INJURY (Lesão) = aumento em 2,0x na creatinina sérica ou decréscimo de 50% na taxa de filtração glomerular ou débito urinário menor que 0,5 mL/kg/h por 12 horas; (Risco de morte 4,15x maior).

3) FAILURE (Insuficiência) = aumento em 3x na creatinina sérica ou decréscimo de 75% na taxa de filtração glomerular ou débito urinário menor que 0,5 mL/kg/h por 24 h ou anúria por 12 h; (Risco de morte 6,17x maior);

4) LOSS (Perda de função) = Perda completa da função renal (ie,: necessidade de terapia de reposição renal) por mais de 4 semanas;

5) END STAGE RENAL DISEASE (doença renal em estágio terminal): Perda completa da função renal (ie,: necessidade de terapia de reposição renal) por mais de 3 meses.

Problemas com os critérios RIFLE:

1) Os níveis de creatinina, TFG e débito urinário são “aleatórios”, não baseados em evidências. Em um estudo, a creatinina era um bom preditor de mortalidade na UTI, porém isso não se repetia quando se analisava o débito urinário.

2) As mudanças na creatinina sérica durante um episódio de IRA não se correlacionam bem com a taxa de filtração glomerular.

3) É impossível calcular o risco do paciente, se não houver uma medida basal de creatinina prévia.

2.2 CRITÉRIOS AKIN (*ACUTE KIDNEY INJURY NETWORK*)

Uma modificação foi proposta em alguns dos critérios RIFLE pela rede AKIN em 2007 no periódico *Critical Care* e é demonstrada na tabela abaixo:

Critérios da Acute Kidney Injury Network (AKIN) para a Lesão Renal Aguda.

Estágio	Critério de creatinina sérica	Critério de débito urinário
1	Aumento na creatinina sérica maior ou igual a 0,3 mg/dL ou aumento maior que 150 a 200%(1,5-2x) do valor basal	Menos de 0,5 mL/kg/h por mais de 6 horas
2	Aumento na creatinina sérica maior ou igual a 200 a 300% (2-3x) do basal	Menos de 0,5 mL/kg/h por mais de 12 horas
3	Aumento na creatinina sérica maior que 300% (>3x) do valor basal ou creatinina sérica maior ou igual a 4,0 mg/dL com um aumento agudo de, pelo menos, 0,5 mg/dL	Menos de 0,3 mL/kg/h por 24 horas ou anúria por 12 horas

* Modificado do critério *RIFLE* (Risk, Injury, Failure, Loss, and End-stage kidney disease). O estadiamento proposto tem alta sensibilidade e é baseado em dados recentes que mostram que pequenos aumentos na creatinina sérica resultam em grandes influências sobre desfechos clínicos. O critério só deve ser aplicado após a otimização da hidratação e da hemodinâmica do paciente. Somente um critério é necessário para a qualificação no estágio analisado.
* Dada a grande variabilidade nas indicações e no tempo de início da terapia de reposição renal (TRR), indivíduos em diálise são considerados como tendo alcançado o estágio 3 independentemente do estágio em que estavam quando foi indicada a TRR.

Assim, além de ser capaz de fazer um diagnóstico precoce e ter valor prognóstico o biomarcador desejado deveria, também, possuir a capacidade de diferenciar que subtipo de IRA, identificando a etiologia e predizendo o desfecho, permitindo que sejamos capazes de fazer uma estratificação de risco e monitorizar a resposta às intervenções (21, 22). O diagnóstico precoce da IRA é freqüentemente problemático, devido à baixa aptidão de biomarcadores precoces entre dano e função renal (24). Usando tecnologias genômicas e de microarranjos protéicos, uma série de moléculas tem sido identificada como potenciais marcadores para IRA (24).

Ao longo dos últimos anos, vários biomarcadores vem sendo incluídos nesta pesquisa são eles: a Interleucina 18 (IL-18), Molécula 1 da Injúria Renal (KIM), Cistatina C (Cys), Proteína Ligadora dos Ácidos Graxos tipo hepáticos (L-FABP), n-acetil-D glicosaminidase (NAG), Fator de Crescimento do Endotélio Vascular (VEGF), Proteína-10 induzida pela Quimiocina-Interferon (CXCL10) e Lipocalina associada a

gelatinase nos neutrofílos humanos (NGAL); apenas um parece estar sendo capaz de nos responder a grande maioria das perguntas, embora não todas (25). Como já comentamos, anteriormente, o diagnóstico precoce da IRA é freqüentemente problemático, devido à baixa aptidão de biomarcadores capazes da diagnosticar de forma antecipada a diferença entre dano e função renal (26).

A urina pode conter marcadores sensíveis e específicos do dano renal e que estão presentes tanto devido ao prejuízo na capacidade de reabsorção tubular e catabolismo das moléculas filtradas, quanto da liberação das células protéicas tubulares em resposta às injúrias isquêmicas ou nefrotóxicas (7). No nosso caso, em especial toda a nossa atenção fica voltada ao padrão isquemia-reperfusão que acontece no transplante renal. O objetivo dos biomarcadores é o diagnóstico precoce da rejeição, determinar prognóstico e adequar a terapêutica imunossupressiva a uma forma não-invasiva e custo-efetivo. Os biomarcadores pesquisados têm sido focados nas áreas primárias da injúria renal, túbulos e células que os infiltram (27). A dosagem da creatinina sérica e a biópsia de rim permanecem sendo o “padrão ouro” para a avaliação dos enxertos renais. Estes testes têm limitações significativas quanto à informação de que tal ou qual paciente desenvolverá imuno-tolerância ou perda do enxerto mediada pela resposta imune, ou se haverá necessidade de imunossupressão por longo tempo. Além do que a creatinina sérica não nos dá a medida exata da agressão ao rim, particularmente em relação ao tempo, pois, ela só aumenta quando uma fração significativa da função renal foi perdida (24). Ao longo destes anos se elegeu como marcador para avaliar esta situação aguda, a dosagem da creatinina sérica e o volume de diurese nas primeiras 24 horas pós-transplante e como marcador definitivo, em relação à histologia da rejeição, a punção biópsia do rim transplantado (28). Diz o bom senso, e aqui não precisamos de qualquer fundamento científico, que a agressão advinda da introdução de uma agulha num órgão

recém transplantado, não se justifica como rotina.

Por tal razão, vários são os estudos que vem sendo desenvolvidos pelos mais diferentes pesquisadores, sempre com a intenção de abreviar uma ação que deva ser tomada quando da IRA em um pós-transplante. Conforme já comentamos há uma idéia bem estabelecida de que só um marcador, destes todos até então estudados, não fornece todas as informações necessárias para que, com segurança, possamos tomar uma atitude terapêutica que realmente seja capaz de lidar com o dano ao rim. Entretanto, um destes biomarcadores tem se diferenciado, nos múltiplos trabalhos em que vem sendo usado, como uma ferramenta útil no diagnóstico precoce da IRA no TR.

Quando se fala em biomarcador espera-se de um que ele seja capaz de fazer o diagnóstico da lesão, de uma forma rápida, segura, de baixo impacto econômico, de pouca ou nenhuma agressão ao paciente de forma que possam ser repetidas tantas vezes quantas forem necessárias, e que preceda a alteração clínica.

Conforme Sachin Soni *et al* (20) como no caso de qualquer insulto celular, na IRA inclusive, a lesão começa pela indução de modificações moleculares que evoluem para o dano celular. As células iniciariam a produção de marcadores da lesão e a síndrome clínica se desenvolveria posteriormente. É postulado, então, que o relógio biológico (expressão do biomarcador) deveria sempre preceder o relógio clínico (20). Assim, a detecção do biomarcador específico daria uma janela, entre a manifestação molecular e a clínica, dando tempo de uma intervenção direta sobre a agressão ao rim e com isso estaríamos impedindo a evolução deste quadro, no nosso interesse em particular, que é de tanto não perder o enxerto precocemente, quanto o paciente por uso excessivo, e não justificado de imunossupressores em dose desproporcional ao evento que o causou.

Ainda, segundo Sachin Soni *et al* (20), as características ideais de um

biomarcador para a Injúria Renal Aguda (IRA) deveriam ser:

1. Não-invasivo
2. Facilmente detectável em amostras acessíveis como soro ou urina.
3. Altamente sensível e específico para IRA.
4. De mensuração rápida e confiável.
5. Capaz de detecção precoce.
6. Com características que possam informar a etiologia, a natureza e a duração da lesão.
7. Ser um marcador de injúria e de função.
8. Predizer a severidade e a reversibilidade da IRA.
9. Ser de grande auxílio na monitorização e na resposta às intervenções.
10. Não ser afetado por outras variáveis biológicas.
11. Ser de baixo custo.

Usando tecnologias genômicas e de microarranjos protéicos, uma série de moléculas tem sido identificada como potenciais marcadores para IRA (20). As quimiocinas são citocinas que apresentam papel central na fisiologia leucocitária ao controlar o tráfego basal e inflamatório (29). São polipeptídios de 8 a 12 kD com duas alças internas de dissulfeto. É classificado em famílias baseadas no número e localização de resíduos de cisteínas N-terminais. As duas principais famílias são as CC quimiocinas, com resíduos de cisteína adjacentes, e família CXC, onde estes resíduos são separados por um aminoácido (16). As quimiocinas podem ser categorizadas por qualquer estímulo que altera a homeostase celular, e o RNAm das quimiocinas podem aumentar mais de 300 vezes em poucas horas de ativação (30). As constitutivas são responsáveis pelo tráfego leucocitário basal e pela formação da arquitetura de órgãos

linfóide secundários (30).

Segundo Cardoni *et al* (31), na etapa de indução da resposta inflamatória do transplante renal, células apresentadoras de antígenos, linfócitos T ativados e células natural killers, na presença de Interleucina 12 e 18 (IL-12 e IL-18) e Fator de Necrose Tumoral alfa (TNFα) levam ao aumento de produção de células T helper 1.

Conseqüentemente, há um aumento da expressão de genes de CD-26 e LAG-3 (genes de ativação linfocitária), e de receptores CCR5 (receptores de RANTES/CCL5, MIP-1α/CCL3 e MIP-1β/CCL4) e CXCR3 (receptor de IP-10/CXCL10) e vários outros (27).

No caso de estarem presentes as IL-4, IL-10, IL-13, ocorre aumento da produção de células T helper e estímulo de expressão CD30, CD62L (expressão em células ativas leva a entrada de estruturas linfóides periféricos), CCR3 (14) (receptor de eotaxina em eosinófilos), CCR4 (receptor de TARC/CCL17, quimiocinas regulada por ativação e MDC/CCL22 (quimiocinas derivadas de macrófagos), expresso em células T, killers e células dendríticas (CD), CCR8 (receptor de I-309/CCL1, derivado do gene 3 de ativação celular, que é expresso em células T e neutrófilos) e CXCR4 (receptor de SDF-1α/β/CXCL12, fator derivado do estroma, que é expresso em T-CD4 ativada, CD e eosinófilos) (23).

De acordo com Hu e Knechtle (32), as quimiocinas podem influenciar pelo menos três aspectos da biologia do enxerto renal: 1- a restauração do fluxo sanguíneo no enxerto pode levar à lesão de isquemia reperfusão, onde quimiocinas recrutam leucócitos; 2- respostas do receptor à infecção durante a supressão imune envolvem as quimiocinas; 3- os componentes inflamatórios na rejeição aguda e na nefropatia crônica do enxerto são controlados por quimiocinas.

Estes autores (Hu e Knechle) (32), avaliando as quimiocinas e citocinas urinárias

em transplantados renais as divide em três grupos:

1- Aquelas que estão aumentadas em receptores com Rejeição Aguda, Necrose Tubular Aguda, Nefropatia Crônica do Enxerto, função normal do enxerto, assim como indivíduos hígidos: angiogenina, TIMs (*tissue inhibitors of metalloproteinase*), receptor 2 de TNF solúvel, ligante de receptor 3 indutor de apoptose relacionado ao TNF);

2- Aquelas que têm baixa expressão em receptores de transplante renal e indivíduos hígidos: IL-1beta, IL-2, IL-6, MIP-1α/CCL3, MIP-1β/CCL4, MIP-3α/CCL20, IL-18 e TNFα;

3- Aquelas que estão mais aumentadas em receptores com rejeição aguda, necrose tubular aguda, nefropatia crônica do enxerto, do que naqueles com função normal do enxerto e indivíduos hígidos: Adiponectina, IGFBP-1 (proteína ligadora de fator de crescimento “insulina-like”-1), IGFBP-2, IGFBP-6, IL-8/CXCL8, leptina, MCP-1/ CCL2, MIP-1δ, sTNFR1, osteoprotegerina e receptor ativador de urokinase plasminogênio (33).

Neste sentido, Li *et al* (34) mostraram um método para diagnóstico de rejeição aguda em enxertos renais mensurando RNAm de perfurina e granzima B em células de sedimento urinário com sensibilidade e especificidade maior que 80% (31). Devarajan (35) demonstrou promissores biomarcadores de injúria renal aguda (IRA), compostos de NGAL (lipocalina associada com gelatinase de neutrófilos humanos) e cistatina C para um painel plasmático e NGAL, IL-18 (citocina induzida e clivada em túbulo proximal) e KIM-1 (molécula de injúria renal 1, expressa em células tubulares renais proximais) para um painel urinário (31).

Recentemente, uma lipocalina produzida no nefrón distal, chamada Lipocalina

Gelatinase-associada Neutrofílica (NGAL) está emergindo como um novo biomarcador da IRA.

As lipocalinas são uma família de mais de 20 proteínas com a habilidade de se ligar a uma grande variedade de moléculas. NGAL, também conhecida como lipocalina 2, é membro desta família, e foi a primeira a ser isolada nos neutrofílos humanos e está expressa em um grande número de tecidos humanos incluindo trato gastrintestinal, respiratório e urinário (25). Aumento na expressão sistêmica e tecidual de NGAL tem sido bem documentado em muitas condições caracterizadas por infecção ou inflamação, incluindo diverticulite, apendicite, doença Intestinal inflamatória ou infecção urinária, mas, na prática, nosso interesse está focado na IRA (25).

Há fortes evidências que baseiam a estreita correlação entre a IRA por diversas etiologias (nefropatia do contraste, transplante de rim, cirurgia cardíaca, preeclampia e sepse) e o aumento nos níveis plasmáticos e urinários da NGAL (26).

Na era da medicina baseada em evidências, com revisões sistemáticas como a pedra angular, os instrumentos para avaliação da qualidade devem estar disponíveis (36).

As revisões sistemáticas têm por objetivo identificar e avaliar todas as evidências disponíveis relacionáveis a um objetivo particular. Acesso à qualidade é uma parte integrante de qualquer revisão sistemática. Se os resultados dos estudos individuais derem margem por serem imprecisos então o trabalho como um todo estará também com potencialidade para não fornecer dados que possam ser usados em protocolos para diagnóstico (36). É, portanto, essencial que a qualidade individual dos estudos incluídos em uma revisão sistemática seja passível de avaliação sob vários aspectos, entre eles: potencial para polarização, ausência de aplicabilidade e, inevitavelmente sobre a qualidade da informação obtida.

O estudo da qualidade (QUADAS-*Quality assessment of studies of diagnostic accuracy included in systematic reviews*) (37) é tão importante nas revisões sistemáticas dos estudos de acurácia diagnóstica quanto para outras revisões. Entretanto, os estudos de acurácia diagnóstica têm uma série de características únicas em termos de estrutura que diferem das avaliações clássicas de intervenção. Ele ajuda a determinar o quão bom é um teste, em particular, na detecção do que se deseja provar (37).

2.3 A Biologia do NGAL

NGAL humana foi originalmente identificada como uma proteína de 25-kDa ligada covalentemente à gelatinase dos neutrófilos (38, 39). Como outras lipocalinas, NGAL forma uma estrutura terciária em forma de barreira com uma porção hidrofóbica que se liga às moléculas lipofílicas (40). Os maiores ligantes à NGAL são os sideróforos, pequenas moléculas que são sintetizadas pelas bactérias para adquirir ferro, e a NGAL exerce um efeito bacteriostático pela depleção destes sideróforos (41). Por outro lado, os sideróforos produzidos pelos eucarióticos participam da bomba de ferro mediada pela NGAL que é crítica a várias respostas celulares tais como proliferação e diferenciação (40). Embora a NGAL seja expressa só em níveis muito baixos nos vários tecidos humanos, ela pode aumentar marcadamente quando há injúria às células epiteliais, incluindo às renais (24). A região promotora do gene do NGAL contém locais de ligação para um grande número de fatores de transcrição incluindo NF-kB (29, 38). NF-kB é conhecido por ser rapidamente ativado nas células tubulares renais após injúrias agudas (28) e desempenha um papel central no controle da sobrevida e proliferação celular (24). Estes achados nos indicam que há um mecanismo molecular em potencial desempenhado pelo NGAL nos fenótipos epiteliais, tanto durante o

desenvolvimento do rim quanto ao que se segue a IRA (24).

Estudos transcriptônicos pré-clínicos identificaram NGAL (também chamada de lipocalina 2 ou *lcn*2) como sendo um dos genes renais mais regulados no rim imediatamente após a injúria renal aguda em modelos animais (19, 42). Análises proteômicas eletroforéticas também revelam que o NGAL é a proteína mais facilmente induzida no rim após IRA nefrotóxica ou isquêmica, em modelos animais (43, 44).

Um número de estudos tem implicado, agora, um papel importante da NGAL, como um biomarcador diagnóstico precoce para a IRA em várias situações clínicas comuns. Em estudos prospectivos em crianças, com função renal normal e sem qualquer comorbidade, que se submeteram à cirurgia cardíaca, IRA (definida como um aumento de 50% na creatinina sérica) ocorreu em, aproximadamente, 30% dos pacientes, 2-3 dias após a intervenção (39, 45). Ao mesmo tempo, a medida de NGAL pela técnica de Enzima Imuno-Ensaio (ELISA), nestas crianças, revelou um aumento de 10 vezes ou mais no plasma e na urina, em apenas 2-6 horas naqueles que desenvolveram IRA como conseqüência da cirurgia (35). Tanto o NGAL plasmático quanto urinário se mostraram excelentes marcadores precoces, independentes, da IRA, com uma área sob a curva (AUC) de > 0,9 para as medidas de NGAL tanto na urina quanto no plasma (39, 45). Estes achados têm sido confirmados através de estudos prospectivos de adultos que desenvolveram IRA após cirurgia cardíaca, nos quais o NGAL urinário estava significativamente elevado após as primeiras 3 horas pós-cirurgia (46, 47). IRA, definida nos padrões clássicos que é um aumento da creatinina em >50% só veio a acontecer 2 a 3 dias depois do evento, mostrando com isso que a NGAL pode ser um instrumento de diagnóstico precoce na IRA (48).

NGAL também tem sido avaliada como um biomarcador em transplante renal. Nos protocolos de biópsia de rim obtida, após 1 hora da anastomose vascular, revelou-

se uma correlação significativa entre a intensidade da coloração do tecido renal e o subseqüente desenvolvimento de retardo da resposta do enxerto (49). Em um estudo prospectivo multicêntrico com crianças e adultos, o nível de NGAL em amostras urinárias coletadas no primeiro dia do transplante identificou aqueles que subsequentemente desenvolveriam função retardada do enxerto renal (a qual se desenvolveu 2-4 dias após), com uma AUC de 0,9 (50). As medidas plasmáticas de NGAL também vêm sendo correlacionáveis com enxertos renais que tem sua função retardada em pacientes que receberam transplante de rim de doadores após morte cardíaca (47).

Nosso trabalho tem como objetivo principal, fazer uma revisão sistemática das pesquisas desenvolvidas na última década, sobre um biomarcador, Lipocalina Associada à Gelatinase dos Neutrófilos Humanos (NGAL), que tem sido considerado como um grande avanço no diagnóstico das rejeições de rim transplantado, assegurando uma evolução em termos diagnósticos sem prejuízo à integridade de nossos pacientes.

OBJETIVO

3-OBJETIVO

3.1 OBJETIVO PRINCIPAL:

- Proceder a um estudo de Revisão Sistemática das pesquisas desenvolvidas na última década, sobre um biomarcador sérico e urinário, Lipocalina Associada à Gelatinase dos Neutrófilos Humanos (NGAL).

REFERÊNCIAS DA REVISÃO DE LITERATURA

4. REFERÊNCIAS BIBLIOGRÁFICAS

1. International Society of Nephrology. Acessado em 22 de Maio de 2010: Disponível em: <http://www.isn-online.org/isn/index.html>; 2010.

2. Tuoto EA. Biografias Médicas. Acessado em 28 de Maio de 2010. Disponível em: <http://medbiography.blogspot.com/>; 2010.

3. NEFROCARE Clínica de Diálise. Campinas.: Acessado em 30 de Maio de 2010. Disponível em: <http://www.nefrocare.com.br/noticias>; 2010.

4. Sociedade Brasileira de Nefrologia. Acessado em 30 de Maio de 2010. Disponível em: <http://www.sbn.org.br/>; 2010.

5. Nickerson P. Post-transplant monitoring of renal allografts: are we there yet? Curr Opin Immunol. 2009 Oct;21(5):563-8.

6. Waikar SS, Bonventre JV. Biomarkers for the diagnosis of acute kidney injury. Curr Opin Nephrol Hypertens. 2007 Nov;16(6):557-64.

7. Meldrum KK, Hile K, Meldrum DR, Crone JA, Gearhart JP, Burnett AL. Simulated ischemia induces renal tubular cell apoptosis through a nuclear factor-kappaB dependent mechanism. J Urol. 2002 Jul;168(1):248-52.

8. Wagener G, Jan M, Kim M, Mori K, Barasch JM, Sladen RN, et al. Association between increases in urinary neutrophil gelatinase-associated lipocalin and acute renal dysfunction after adult cardiac surgery. Anesthesiology. 2006 Sep;105(3):485-91.

9. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA. 1999 Sep 15;282(11):1061-6.

10. Grimm PC, McKenna R, Nickerson P, Russell ME, Gough J, Gospodarek E, et al. Clinical rejection is distinguished from subclinical rejection by increased infiltration by a population of activated macrophages. J Am Soc Nephrol. 1999 Jul;10(7):1582-9.

11. Hall IE, Yarlagadda SG, Coca SG, Wang Z, Doshi M, Devarajan P, et al. IL-18 and urinary NGAL predict dialysis and graft recovery after kidney transplantation. J Am Soc Nephrol. Jan;21(1):189-97.

12. Racusen LC, Halloran PF, Solez K. Banff 2003 meeting report: new diagnostic insights and standards. Am J Transplant. 2004 Oct;4(10):1562-6.

13. Racusen LC, Solez K, Colvin RB, Bonsib SM, Castro MC, Cavallo T, et al. The Banff 97 working classification of renal allograft pathology. Kidney Int. 1999 Feb;55(2):713-23.

14. Pereira AB. Quimiocinas e citocinas urinárias no transplante de rim.

(Dissertação de Mestrado). Programa de Pós-Graduação em Medicina, Universidade Federal de Minas Gerais2008.

15. Nankivell BJ, Borrows RJ, Fung CL, O'Connell PJ, Allen RD, Chapman JR. The natural history of chronic allograft nephropathy. N Engl J Med. 2003 Dec 11;349(24):2326-33.

16. Portilla D, Dent C, Sugaya T, Nagothu KK, Kundi I, Moore P, et al. Liver fatty acid-binding protein as a biomarker of acute kidney injury after cardiac surgery. Kidney Int. 2008 Feb;73(4):465-72.

17. Lattanzio MR, Kopyt NP. Acute kidney injury: new concepts in definition, diagnosis, pathophysiology, and treatment. J Am Osteopath Assoc. 2009 Jan;109(1):13-9.

18. Hoste EA, Clermont G, Kersten A, Venkataraman R, Angus DC, De Bacquer D, et al. RIFLE criteria for acute kidney injury are associated with hospital mortality in critically ill patients: a cohort analysis. Crit Care. 2006;10(3):R73.

19. Vaidya VS, Waikar SS, Ferguson MA, Collings FB, Sunderland K, Gioules C, et al. Urinary biomarkers for sensitive and specific detection of acute kidney injury in humans. Clin Transl Sci. 2008 Dec;1(3):200-8.

20. Soni SS, Ronco C, Katz N, Cruz DN. Early diagnosis of acute kidney injury: the promise of novel biomarkers. Blood Purif. 2009;28(3):165-74.

21. Medicinidade. Novos critérios de insuficiências renal. Disponível em: <http://pedrorpb.blogspot.com/2009/07/novos-criterios-de-insuficiencia-renal.html>; 2009.

22. Mehta RL, Kellum JA, Shah SV, Molitoris BA, Ronco C, Warnock DG, et al. Acute Kidney Injury Network: report of an initiative to improve outcomes in acute kidney injury. Crit Care. 2007;11(2):R31.

23. Ronco C. N-GAL: diagnosing AKI as soon as possible. Crit Care. 2007;11(6):173.

24. Mishra J, Mori K, Ma Q, Kelly C, Yang J, Mitsnefes M, et al. Amelioration of ischemic acute renal injury by neutrophil gelatinase-associated lipocalin. J Am Soc Nephrol. 2004 Dec;15(12):3073-82.

25. Devarajan P. Neutrophil gelatinase-associated lipocalin--an emerging troponin for kidney injury. Nephrol Dial Transplant. 2008 Dec;23(12):3737-43.

26. Di Grande A, Giuffrida C, Carpinteri G, Narbone G, Pirrone G, Di Mauro A, et al. Neutrophil gelatinase-associated lipocalin: a novel biomarker for the early diagnosis of acute kidney injury in the emergency department. Eur Rev Med Pharmacol Sci. 2009 May-Jun;13(3):197-200.

27. Gillespie A, Lee IJ. Biomarkers in renal transplantation. Biomark Med. 2008

Dec;2(6):603-12.

28. Haussler U, von Wichert G, Schmid RM, Keller F, Schneider G. Epidermal growth factor activates nuclear factor-kappaB in human proximal tubule cells. Am J Physiol Renal Physiol. 2005 Oct;289(4):F808-15.

29. Cowland JB, Sorensen OE, Sehested M, Borregaard N. Neutrophil gelatinase-associated lipocalin is up-regulated in human epithelial cells by IL-1 beta, but not by TNF-alpha. J Immunol. 2003 Dec 15;171(12):6630-9.

30. Gerard C, Rollins BJ. Chemokines and disease. Nat Immunol. 2001 Feb;2(2):108-15.

31. Cardoni RL, Prigoshin N, Tambutti ML, Ferraris JR. [Regulatory cytokines in the response to the allogeneic renal transplant]. Medicina (B Aires). 2005;65(1):54-62.

32. Hu H, Knechtle SJ. Elevation of multiple cytokines/chemokines in urine f human renal transplant recipients with acute and chronic injuries: potencial usage for diagnosis and monitoring. Transplantation Review. 2006;20:165-71.

33. Afzali B, Taylor AL, Goldsmith DJ. What we CAN do about chronic allograft nephropathy: role of immunosuppressive modulations. Kidney Int. 2005 Dec;68(6):2429-43.

34. Li B, Hartono C, Ding R, Sharma VK, Ramaswamy R, Qian B, et al. Noninvasive diagnosis of renal-allograft rejection by measurement of messenger RNA for perforin and granzyme B in urine. N Engl J Med. 2001 Mar 29;344(13):947-54.

35. Devarajan P. Emerging biomarkers of acute kidney injury. Contrib Nephrol. 2007;156:203-12.

36. Jaeschke R, Guyatt G, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? Evidence-Based Medicine Working Group. JAMA. 1994 Feb 2;271(5):389-91.

37. Whiting PF, Weswood ME, Rutjes AW, Reitsma JB, Bossuyt PN, Kleijnen J. Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. BMC Med Res Methodol. 2006;6:9.

38. Cowland JB, Borregaard N. Molecular characterization and pattern of tissue expression of the gene for neutrophil gelatinase-associated lipocalin from humans. Genomics. 1997 Oct 1;45(1):17-23.

39. Mishra J, Ma Q, Kelly C, Mitsnefes M, Mori K, Barasch J, et al. Kidney NGAL is a novel early marker of acute injury following transplantation. Pediatr Nephrol. 2006 Jun;21(6):856-63.

40. Supavekin S, Zhang W, Kucherlapati R, Kaskel FJ, Moore LC, Devarajan P. Differential gene expression following early renal ischemia/reperfusion. Kidney Int. 2003 May;63(5):1714-24.

41. Mori K, Lee HT, Rapoport D, Drexler IR, Foster K, Yang J, et al. Endocytic delivery of lipocalin-siderophore-iron complex rescues the kidney from ischemia-reperfusion injury. J Clin Invest. 2005 Mar;115(3):610-21.

42. Kieran NE, Doran PP, Connolly SB, Greenan MC, Higgins DF, Leonard M, et al. Modification of the transcriptomic response to renal ischemia/reperfusion injury by lipoxin analog. Kidney Int. 2003 Aug;64(2):480-92.

43. Mishra J, Ma Q, Prada A, Mitsnefes M, Zahedi K, Yang J, et al. Identification of neutrophil gelatinase-associated lipocalin as a novel early urinary biomarker for ischemic renal injury. J Am Soc Nephrol. 2003 Oct;14(10):2534-43.

44. Mishra J, Mori K, Ma Q, Kelly C, Barasch J, Devarajan P. Neutrophil gelatinase-associated lipocalin: a novel early urinary biomarker for cisplatin nephrotoxicity. Am J Nephrol. 2004 May-Jun;24(3):307-15.

45. Mishra J, Dent C, Tarabishi R, Mitsnefes MM, Ma Q, Kelly C, et al. Neutrophil gelatinase-associated lipocalin (NGAL) as a biomarker for acute renal injury after cardiac surgery. Lancet. 2005 Apr 2-8;365(9466):1231-8.

46. Koyner JL, Bennett MR, Worcester EM, Ma Q, Raman J, Jeevanandam V, et al. Urinary cystatin C as an early biomarker of acute kidney injury following adult cardiothoracic surgery. Kidney Int. 2008 Oct;74(8):1059-69.

47. Kusaka M, Kuroyanagi Y, Mori T, Nagaoka K, Sasaki H, Maruyama T, et al. Serum neutrophil gelatinase-associated lipocalin as a predictor of organ recovery from delayed graft function after kidney transplantation from donors after cardiac death. Cell Transplant. 2008;17(1-2):129-34.

48. Parikh CR, Mishra J, Thiessen-Philbrook H, Dursun B, Ma Q, Kelly C, et al. Urinary IL-18 is an early predictive biomarker of acute kidney injury after cardiac surgery. Kidney Int. 2006 Jul;70(1):199-203.

49. Parikh CR, Jani A, Mishra J, Ma Q, Kelly C, Barasch J, et al. Urine NGAL and IL-18 are predictive biomarkers for delayed graft function following kidney transplantation. Am J Transplant. 2006 Jul;6(7):1639-45.

50. Hernandez D, Rufino M, Gonzalez-Posada JM, Estupinan S, Perez G, Marrero-Miranda D, et al. Predicting delayed graft function and mortality in kidney transplantation. Transplant Rev (Orlando). 2008 Jan;22(1):21-6.

ARTIGO ORIGINAL EM INGLÊS

5. ARTIGO ORIGINAL EM INGLÊS

Neutrophil Gelatinase Associated Lipocalin (NGAL) for Diagnose of acute Renal Injury in Kidney Transplantation

LOBATO GR, MEDEIROS LR, EDELWEIS MI

ABSTRACT

Background: Neutrophil gelatinase-associated lipocalin (NGAL) appears to be a promising biomarker for the early diagnosis of acute kidney injury (AKI). ***Purpose:*** To determine the accuracy of NGAL compared with serum creatinine in patients with acute renal failure after kidney transplantation. ***Study Design***: Systematic review. ***Data Source***: MEDLINE (through Pub Med and OVID interface), EMBASE, LILACS, BIOSIS, Cochrane Central Register of Controlled Trials, Web of Science, IBECS, SCOPUS and grey literature was searched for studies reporting the value of NGAL to predict AKI in KT and renal replacement therapy (RRT) from January 2000 to July 2010. ***Study selection:*** Two reviewers scrutinized abstract and examined potentially inclusion and independently extracted study data. Methodological quality was assessed by using the QUADAS instrument. ***Results***: We found 7 trials (4 cohort and 3 cross-sectional). The statistical heterogeneity was present, because the differences in methodological, clinics and statistical aspects, we realized a systematic review and not a meta-analysis, concordant with Cochrane Library and PRISMA's methodological orientations. ***Limitations:*** Studies used different index test (serum NGAL and urinary NGAL) and had heterogeneous findings. ***Conclusion***: Available evidence is inconclusive but suggests that NGAL could be superior to the serum creatinine for identifying acute renal injury after kidney transplantation. High quality studies on their role in the diagnostic investigation of AKI are urgently needed in KT.

Key-Words: Neutrophil gelatinase-associated lipocalin (NGAL); urine NGAL; qualitative systematic review; acute kidney injury (AKI); kidney transplantation (KT).

INTRODUCTION

Renal ischemia-reperfusion injury is the leading cause of ARI in the native as well as the transplanted kidney (1).

Acute kidney injury (AKI) represents a common and potentially devasting problem in clinical medicine, with a persistently high mortality and morbidity (2, 3).

Measurement of serum creatinine and biopsy remains the current gold standards for the evaluation of renal allograft. These tests have significant limitations in predicting which patients are destined for immune tolerance or immune-mediated graft loss, and aiding in the management of long-term immunosuppressant (4).

Indeed, understanding the early stress response of the kidney to acute injuries has revealed a number of potential biomarkers (5). The bench-to-bedside journey of neutrophil gelatinase-associated lipocalin (NGAL), arguably the most promising novel AKI biomarker, is explained in turn.

NGAL has also been evaluated as a biomarker of AKI in kidney transplantation. Protocol biopsies of kidneys obtained 1 h after vascular anastomosis revealed a significant correlation between NGAL staining intensity and the subsequent development of delayed graft function (6).

The purpose of this systematic review was to determine diagnostic performance of NGAL compared with serum creatinine for diagnosis of acute kidney injury after renal transplantation.

METHODS

Findings are reported according with to the STARD (Standards for Reporting of Diagnostic Accuracy), PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-analysis), GRADE system (Grading of recommendations Assessment,

Development and Evaluation) and according with Guidance on the Conduct of Narrative Synthesis in Systematic reviews (7-10).

Search Strategy

We searched MEDLINE (PubMed and OVID interface), EMBASE, LILACS (trough Scielo interface), Cochrane Central Register of Controlled Trials, IBECS, BIOSIS , Web of Science, Congress Abstracts to identify potentially relevant articles or abstracts, and Grey literature (Google scholar; British Library), from January 2000 to July 2010. were searched using the following terms, both as text words and, as appropriate, Medical Subjects Heading (MeSH) or equivalent subject heading /thesaurus terms: *"neutrophil gelatinase-associated lipocalin"*, *"NGAL"*, and *"Kidney transplantation*" were combined with the MeSH term *"diagnosis"*(sensitivity, specificity, false positive, false negative, predicted value, reference value, ROC, likelihood ratio, accuracy). This sensitive filter was created by combining three filters for the identification for diagnostic studies via Boolean operator "OR" and "AND". The search was limited to human studies but had no languages restrictions. Reference lists of all available primary studies were reviewed to identify additional relevant citations. Additionally, we checked reference of relevant reviews, meta-analysis, guidelines, and commentaries identified in MEDLINE and EMBASE. Some articles seem to us unclear and we try to contact the authors, but we can not get answer and this study was not included. The complete search strategy is available on request.

Study Selection

Two reviewers (G.R.L. and M.I.E.) independently screened studies identified for inclusion and determined study eligibility. Disagreements were resolved by a third

opinion (L.F.M.). Selection was restricted to published prospective cohort studies and cross-sectional studies of humans investigating the diagnostic and prognostic accuracy of NGAL level to predict graft loss after kidney transplantation. Trials enrolling patients in interventional studies and those exclusively with AKI for other reasons were excluded.

Articles were selected according to the following criteria:

Population: patients that were submitted kidney transplantation

Index test: Serum e urinary NGAL

Reference Standard: Serum creatinine

Target condition: grading of renal graft loss

Outcomes: Early diagnosis of the delayed graft function within 1 wk of transplantation

Data Extraction

Data extraction was performed by on reviewer (G.R.L) and checked by a second reviewer (M.I.E). The following information was extracted for all studies when reported: study details (identifier author, setting, year, and design), participants' details (age, sex, comorbidities, inclusion criteria, and donor cardiac death), test index details, reference standard details). Acute kidney injury was described which an increase in serum creatinine of more than or equal to 0,3mg/dL or increase to more than or equal 150 to 200 percent from baseline or less than 0,5mL/kg per hour the urine for more than 6 hours to the diagnose according RIFLE criteria (Table 1) (11).

Quality assessment

All articles meeting the eligibility criteria were assessed for their methodological quality. Quality was assessment of Diagnostic Accuracy Studies Tool (12) recommends

by the Cochrane Collaboration) (13). The modified version consists of 11 items on study characteristics with the potential to introduce bias. Items were scored as positive (no bias), negative (potential bias), or insufficient information (inadequate - ?).

Data synthesis and statistical analysis

Qualitative analysis was performed because of the heterogeneity of the studies with regard to populations, interventions and outcome measured, we refrained from statistical pooling. Results were considered contradictory if the overall conclusions regarding effectiveness in different studies on the same indication were contradictory or inconsistent (8, 9).

RESULTS

Study identification and eligibility

The process of study selections is summarized in figure 1. Our initial search identified 157 potentially relevant articles. We excluded 140 published studies after reviewing their titles and abstracts, because the three independent reviewers considered that they did not relate to the question under review. Seventeen full-text articles were retrieved and 10 were excluded after scrutiny. A complete list of excluded studies is available from the authors. Seven primary studies included 408 patients, met the criteria for inclusion and were analyzed (Table 2) (14-17)**.**

Interrater agreement for study eligibility and methodological quality was 79% (κ = 0,64), indicating good agreement (18). Disagreement between reviewers occurred during analysis of the 17 studies and they were related to inclusion or exclusion criteria (characteristics and scoring of study quality), but it was solved by consensus. Because the strong heterogeneity that appears in this studies that we can diagnose for different

aspects in clinical, methodological and statistics (9), we would be realized a qualitative systematic review and do not do the meta-analysis. This conduct is in conformity with the methodological orientations of the Cochrane Handbook 5 (13) and PRISMA (7, 8, 10).

Study Description

Details of the participants, interventions, and quality assessments of the studies selected for meta-analysis are summarized in Table 3. The range age of the participants across studies was 4-61. The prevalence of delayed graft function was 14.4% (57 patients).

All studies were not-blind; four were cohort and three were cross-sectional studies from a narrow population, but included insufficient experimental details and the proper diagnostic tests and diagnostic reference standards. Of the 7 included studies, there was one study with high methodological quality (at least eight of 11 qualitative items). The results assessments of the included studies are summarized in the Table 3 using QUADAS. Figure 2 summarizes the methodological quality of the included studies. In seven studies, the sample spectrum was representative of the examined population. Most studies described the tests in insufficient detail to permit their replication. However, these studies were conducted within a spectrum of patients typical (renal transplantation). None studies performed well, receiving a positive assessment of at least 8 of 11 items (14-17, 19-21).

Sensitivity Analysis

Sensitivity analysis was not performed because we performed qualitative systematic review.

DISCUSSION

We preformed a systematic review to clarify the predictive value of NGAL for the early diagnosis of AKI in patients with transplanted kidney.

However, there are limitations to this study that apply to the field of biomarkers in AKI in general. In spite of this affirmation after the evaluation of the studies our conclusion was: 1. NGAL level to be a useful early predictor of AKI, both overall and across a range of clinical settings; 2. The performance of NGAL level improved when standardized clinical laboratory platforms with a cutoff NGAL concentration > 150 ng / mL were used, in comparison to research-based assays; 3. NGAL level had prognostic value for clinical outcomes, such as diminish of imunnossupression in the transplanted patients.

In the literature, different definitions of AKI, various settings of AKI, and varying timings of NGAL measurement with regard to a renal insult have been used to assess the predictive value of NGAL, thus creating effective modifiers of NGAL usefulness as a biomarker (22).

Although serum creatinine is typically used for diagnosis of AKI, it is an insensitive and unreliable biomarker during acute changes in kidney function (23). The serum creatinine concentration does not increase until about half of the kidney function is lost. The lack of an early biomarker has been a major impediment in developing newer preventive strategies for AKI (24).

The need for a simple, accurate, and minimally invasive marker of AKI has been limiting factor in clinical nephrology research and practice. More accurate methods would be necessary to ensure that this biomarker can be used like a relevant aspect in the early diagnose of AKI in transplanted patients.

In the clinical routine, on the other hand, only seldom are we able to grasp

precisely when the kidney insult took place and implement adequate countermeasures (25).

Looking to the natural evolution of AKI, we can identify different milestones along the timeline of the syndrome. The injury begins by inducing molecular modifications that subsequently evolve into cellular damage. The cells start to produce biomarkers of injury and only subsequently does the clinical picture of the syndrome develop with the typical signs and symptoms. Therefore, it can be posited that the molecular and cellular clocks always anticipate the clinical clock, which always runs late. The biological clock provided by biomarkers display an intermediate time into the progression, but it is most certainly reflective of an earlier stage when compared to the clinical clock (25).

Nowadays our understanding of the AKI syndrome in general, and specifically in KT, suggests that early biomarkers represent a unique possibility for timely diagnosis and intervention to protect the kidney from further insults and to prevent tissue damage from existing risk factors. These biomarkers might very well revolutionize the outcomes of AKI in the near future; new preventive measures, new protective drugs or old simple strategic approaches not adequately tested in the right patients at the right times in the past may find interesting new applications in AKI patients, thereby modifying the natural course of the syndrome and changing the pattern of patient response to therapy and now, again, we empathize the importance of all of this conducts in a kidney transplant patient because, in this particular case we do not have chance to commit a wrong and latter therapy because this attitude can evaluate to the loose of the kidney and sometimes the dead of our patient.

Strengths and weakness of our review

The limitations that we can see in these studies and that will be unable for us to make a meta-analysis was: first, the majority of reports represent a relatively small number of patients and the results will need to be prospectively validated in a large population. Second, the possible confounding effects of comorbid conditions on NGAL excretion are unknown. Third, it will be important to the partner with industry to design point-of-care kits and platforms for biomarkers panels that *(1)* are easy to perform at the bedside or in a standard clinical laboratory, using easily accessible samples such as blood or urine; *(2)* are highly sensitive to facilitate early detection and with a wide dynamic range and cutoff values that allow for risk stratification; *(3)* are highly specific for AKI and enable the identification of AKI subtypes, etiologies and duration; and *(4)* exhibit strong biomarker properties on receiver operating characteristics curves. The availability of such tools will enable the design of rational interventional studies that are initiated early in the course of AKI.

However, we performed this systematic review according to the most recent guidelines for conducting a diagnostic review as described in the Cochrane Handbook (26). We use an extensive search strategy with different database, but using a methodological filter. By reference checking we tried to track down those publications that our search strategy might have failed to identify. We not use of a language restriction during selection phase. Anticipating poor agreement on some items of the QUADAS (12), two reviewers independently assessed all papers for methodological quality and reached consensus by discussing disagreements on individual scores. We performed this systematic review according grading quality of evidence strength of recommendations for diagnostic tests and strategies (8, 10).

RECOMENDATIONS

Diagnostic tests as first line investigation in care after kidney transplantation need to be valid, easy to perform, well tolerated by patients, and sensitive, especially in case of serious disease with kidney transplantation.

WHAT WE ALREADY KNOW ABOUT THIS TOPIC?

Our systematic review show us that NGAL in the AKIN's diagnosis in patients with renal transplantation might prove to be such tests. We therefore urgently need high quality diagnostic cohort studies enrolling consecutive patients after renal transplantation. In future research, NGAL should be an important factor in the analysis, especially as tests that are able to diagnose early stages of AKIN and is important tools to reduce the gravity which case.

"We can lose a kidney but we can make sure that the patients will survive"!!

REFERENCES

1. International Society of Nephrology. Acessado em 22 de Maio de 2010: Disponível em: <http://www.isn-online.org/isn/index.html>; 2010.

2. Sociedade Brasileira de Nefrologia. Acessado em 30 de Maio de 2010. Disponível em: <http://www.sbn.org.br/>; 2010.

3. Nankivell BJ, Borrows RJ, Fung CL, O'Connell PJ, Allen RD, Chapman JR. The natural history of chronic allograft nephropathy. N Engl J Med. 2003 Dec 11;349(24):2326-33.

4. Devarajan P. Emerging biomarkers of acute kidney injury. Contrib Nephrol. 2007;156:203-12.

5. Lattanzio MR, Kopyt NP. Acute kidney injury: new concepts in definition, diagnosis, pathophysiology, and treatment. J Am Osteopath Assoc. 2009 Jan;109(1):13-9.

6. Waikar SS, Bonventre JV. Biomarkers for the diagnosis of acute kidney injury. Curr Opin Nephrol Hypertens. 2007 Nov;16(6):557-64.

7. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. Ann Intern Med. 2003 Jan 7;138(1):40-4.

8. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JP, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ. 2009;339:b2700.

9. Popay J, Roberts H, Sowden A, Petticrew M, Arai L, Rodgers M, et al. Guidance on the Conduct of Narrative Synthesis in Systematic Reviews A Product from the ESRC Methods Programme. Swindon: ESRC; 2006.

10. Schunemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ. 2008 May 17;336(7653):1106-10.

11. Hoste EA, Clermont G, Kersten A, Venkataraman R, Angus DC, De Bacquer D, et al. RIFLE criteria for acute kidney injury are associated with hospital mortality in critically ill patients: a cohort analysis. Crit Care. 2006;10(3):R73.

12. Whiting PF, Weswood ME, Rutjes AW, Reitsma JB, Bossuyt PN, Kleijnen J. Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. BMC Med Res Methodol. 2006;6:9.

13. Smidt N, Deeks J, Bossuy P, Gatsonis C. Guide to the contents of a Cochrane review and protocol. Cochrane Handbook for Systematic Review of Systematic review of Diagnostic Test Accuracy: Cochrane Colaboration. Avaliable in:

<http://srdta.cochrane.org/handbook-dta-reviews>; 2009.

14. Hall IE, Yarlagadda SG, Coca SG, Wang Z, Doshi M, Devarajan P, et al. IL-18 and urinary NGAL predict dialysis and graft recovery after kidney transplantation. J Am Soc Nephrol. Jan;21(1):189-97.

15. Kusaka M, Kuroyanagi Y, Mori T, Nagaoka K, Sasaki H, Maruyama T, et al. Serum neutrophil gelatinase-associated lipocalin as a predictor of organ recovery from delayed graft function after kidney transplantation from donors after cardiac death. Cell Transplant. 2008;17(1-2):129-34.

16. Lebkowska U, Malyszko J, Lebkowska A, Koc-Zorawska E, Lebkowski W, Malyszko JS, et al. Neutrophil gelatinase-associated lipocalin and cystatin C could predict renal outcome in patients undergoing kidney allograft transplantation: a prospective study. Transplant Proc. 2009 Jan-Feb;41(1):154-7.

17. Malyszko J, Malyszko JS, Mysliwiec M. Serum neutrophil gelatinase-associated lipocalin correlates with kidney function in renal allograft recipients. Clin Transplant. 2009 Sep-Oct;23(5):681-6.

18. Altman DG. Some common problems in medical research. Practical statistics for medical research. 9 ed. London: Chapman 1999. p. 403-9.

19. Malyszko J, Malyszko JS, Bachorzewska-Gajewska H, Poniatowski B, Dobrzycki S, Mysliwiec M. Neutrophil gelatinase-associated lipocalin is a new and sensitive marker of kidney function in chronic kidney disease patients and renal allograft recipients. Transplant Proc. 2009 Jan-Feb;41(1):158-61.

20. Mishra J, Ma Q, Kelly C, Mitsnefes M, Mori K, Barasch J, et al. Kidney NGAL is a novel early marker of acute injury following transplantation. Pediatr Nephrol. 2006 Jun;21(6):856-63.

21. Parikh CR, Jani A, Mishra J, Ma Q, Kelly C, Barasch J, et al. Urine NGAL and IL-18 are predictive biomarkers for delayed graft function following kidney transplantation. Am J Transplant. 2006 Jul;6(7):1639-45.

22. Ronco C. NGAL: an emerging biomarker of acute kidney injury. Int J Artif Organs. 2008 Mar;31(3):199-200.

23. Mishra J, Mori K, Ma Q, Kelly C, Barasch J, Devarajan P. Neutrophil gelatinase-associated lipocalin: a novel early urinary biomarker for cisplatin nephrotoxicity. Am J Nephrol. 2004 May-Jun;24(3):307-15.

24. Vaidya VS, Waikar SS, Ferguson MA, Collings FB, Sunderland K, Gioules C, et al. Urinary biomarkers for sensitive and specific detection of acute kidney injury in humans. Clin Transl Sci. 2008 Dec;1(3):200-8.

25. Mishra J, Ma Q, Prada A, Mitsnefes M, Zahedi K, Yang J, et al. Identification of neutrophil gelatinase-associated lipocalin as a novel early urinary biomarker for ischemic renal injury. J Am Soc Nephrol. 2003 Oct;14(10):2534-43.

26. Cochrane Collaboration. Cochrane Diagnostic Reviwe. Chichester: John Wiley & Sons, Inc; 2005.

Table 1. Risk, Injury, Failure, Loss, and End-stage Kidney (RIFLE) classification

Risk, Injury, Failure, Loss, and End-stage Kidney (RIFLE) classification

Class	Glomerular filtration rate criteria	Urine output criteria
Risk	Serum creatinine × 1.5	< 0.5 ml/kg/hour × 6 hours
Injury	Serum creatinine × 2	< 0.5 ml/kg/hour × 12 hours
Failure	Serum creatinine × 3, or serum creatinine ≥ 4 mg/dl with an acute rise > 0.5 mg/dl	< 0.3 ml/kg/hour × 24 hours, or anuria × 12 hours
Loss	Persistent acute renal failure = complete loss of kidney function > 4 weeks	
End-stage kidney disease	End-stage kidney disease > 3 months	

For conversion of creatinine expressed in conventional units to SI units, multiply by 88.4. RIFLE class is determined based on the worst of either glomerular filtration criteria or urine output criteria. Glomerular filtration criteria are calculated as an increase of serum creatinine above the baseline serum creatinine level. Acute kidney injury should be both abrupt (within 1–7 days) and sustained (more than 24 hours). When the baseline serum creatinine is not known and patients are without a history of chronic kidney insufficiency, it is recommend to calculate a baseline serum creatinine using the Modification of Diet in Renal Disease equation for assessment of kidney function, assuming a glomerular filtration rate of 75 ml/min/1.73 m^2. When the baseline serum creatinine is elevated, an abrupt rise of at least 0.5 mg/dl to more than 4 mg/dl is all that is required to achieve class Failure.

Table 2. Characteristics of included studies

Study	Clinical features and settings	Participants	Study design	Target condition and reference standard text	Index and comparator test	Follow up	Notes
Hall et al., 2010	Patients were submitted kidney transplantation. Multicenter study (Connecticut, Kansas, Michigan,Ohio Pennsylvania) in the USA.	91 patients were submitted kidney transplantation. (34 women and 57 men). Age mean 51 (SD±11.9). 11 patients were submitted previous transplant. 73 were submitted hemodialysis and 18 were submitted peritoneal dialysis. Cause of end-stage renal disease: 30 had hypertension, 23 had diabetes and 8 had polycystic kidney disease and 29 had another cause. 34 patients were submitted dialysis within 1 wk of kidney transplantation. There were no clinically suspected or biopsy-confirmed cases of acute	Prospective cohort study. Patients were separated in 3 groups: Delayed graft function (DGF) patients need dialysis within 1 wk of transplantation); Slow graph function (SGF) and immediate graft function (IGF).	Patients were submitted kidney transplantation. Reference standard was serum creatinine (mg/dl) in the first day: DGF 1.9(1.2-2.7) SGF:1.6 (1.2-2.3) IGF: 1.5 (1.0-2.0).	Urine NGAL, urine IL-18 and urine KIM-1 (*kidney injury molecule*) 0h,6h, 12h, 18h after surgery and first and second postoperative day. Receiver operating characteristic (ROC) curve showed NGAL and IL-18 were moderately accurate in predicting dialysis, whereas KIM-1 was not. There were no significant changes in areas under curve (AUCs) for NGAL, IL-18 or KIM-1 after normalization for urine creatinine. The value NGAL (ng/ml) were in the first day: DGF:1035 (95-3143) SGF: 248 (22-756.1) IGF: 60.5 (15.3-249.2).	3 months.	Impossible to construction contingency table 2x2 to compared serum creatinine (<1.5mg/dl>) with urine NGAL (<150 ng/ml>) in the loss evaluation of the renal function.

		rejection during the 7-d study period.					
Kusaka et al., 2008	Patients were submitted kidney transplantation from the Department of Urology in Fujita Health University Scholl of Medicine, Aichi, Japan.	16 patients. 11 patients who received kidney from living donors and 5 patients from donor with cardiac death. Not describe clinics date in first 11 patients. 5 patients received kidney from cardiac death had mean age (30-60 years), 2 were women and 3 were men. Time of hemodialysis duration were 10-15 years.	Prospective cohort study. The primary outcome variable was the development of delayed graft function (DGF), defined as the need for hemodialysis within the first few weeks after transplantation.	Patients were submitted kidney transplantation. Not describe reference standard text.	Serum NGAL (ng/ml) ELISA. Only two patients from 11 required hemodialysis in the first week. They had serum NGAL (459±27 ng/ml). All patients the received kidney from cardiac death had necessity hemodialysis in the first month (5-22 days). But study not describe mean serum NGAL in this group.	1 month (?)	Impossible to construction contingency table 2x2 to compared serum creatinine (<1.5mg/dl>) with serum NGAL (<150 ng/ml>) in the loss evaluation of the renal function. Not describe serum NGAL in the group the five patients that received kidney from cardiac death. Only describe that value is high. Not describe follow up in adequate form.
Lebkowska, 2009	Patients were submitted kidney transplantation from Department Vascular Transplantation surgery, Medical University, Bialystok, Poland.	41 patients who received kidney from donor with cardiac death. Age range (17-69), mean age (45.2±14.1). 17 were women and 24 were men. Before transplantation, each subject was treated with	Cohort study. Not informed if prospective or retrospective study. The primary outcome variable was the development of delayed graft function (DGF).	Patients were submitted kidney transplantation. Only describe serum creatinine on the day 1,3,6,10 after surgery: Day 1 (5.86±1.43) Day 3 (5.56±1.3) Day 6 (4.01±1.37) Day 10 (3.12±1.59) But not describe	Serum NGAL (ng/ml). Study describe NGAL delayed graft function on the day 1,3,6,10 after surgery: Day 1 (307.54±103.1) Day 3 (279.5±106.5) Day 6 (241.9±108.6) Day 10 (206.1±99.2).	Not report.	Impossible to construction contingency table 2x2 to compared serum creatinine (<1.5mg/dl>) with serum NGAL (<150 ng/ml>) in the loss evaluation of the renal function.

		hemodialysis for a mean time of 34.5 ±27.6 months. After transplantation only four patients with DGF required hemodialysis for 3 to 5 days.		witch standard test.			
Malyszko, 2009	Patients were submitted kidney transplantation from Department of Nephrology, Medical University, Bialystok, Poland.	100 patients were submitted kidney transplantation. Mean age (45.1±14.1). 48 were women and 52 were men.	Cross-sectional study.	Patients were submitted kidney transplantation. Serum creatinine was measured by the standard laboratory method (Jaffe) in the central laboratory. Serum creatinine (mg/dl) was measured of the group kidney transplantation: Mean (1.53±0.56).	Serum NGAL (ng/ml). Study describe NGAL delayed graft function, but not specific the time this result after transplantation: NGAL (120.44±73.07)	Not reported (cross-sectional study).	Impossible to construction contingency table 2x2 to compared serum creatinine (<1.5mg/dl>) with serum NGAL (<150 ng/ml>) in the loss evaluation of the renal function.
Malyszko, 2009[]	Patients were submitted kidney transplantation from Department of Nephrology and Transplantology, Medical University, Bialystok, Poland.	80 patients were submitted kidney transplantation. Mean age (55.1±13.5).	Coss-sectional study.	Patients were submitted kidney transplantation. Serum creatinine (mg/dl) was measured of the group kidney transplantation: Mean (1.43±0.59).	Serum NGAL (ng/ml). Study describe NGAL delayed graft function, but not specific the time this result after transplantation: NGAL (129.4±27.6)	Not reported (cross-sectional study).	Impossible to construction contingency table 2x2 to compared serum creatinine (<1.5mg/dl>) with serum NGAL (<150 ng/ml>) in the loss evaluation of the renal function.
Mishra, 2006	Patients were submitted kidney transplantation from	25 patients were submitted kidney	Cross-sectional study.	Patients were submitted kidney	NGAL antibody in the paraffin-	Not reported (cross-sectional	Impossible to construction

	Nephrology and Hypertension, University of Cincinnati College of Medicine, Cincinnati, OH, USA.	transplantation. Age (years) range 7 to 19 years. 13 were male and 12 were female.		transplantation. Not describe reference standard text. Only give information about serum creatinine on the day 0, 1, 2 and 3 before and after surgery in the group of patients the received the kidney from cadaveric donors and living-related donor. Four patients from group cadaveric donors need dialysis (serum creatinine between 10.2 to 14.1).	embedded protocol biopsy specimens obtained at approximately one hour or reperfusion after transplantation. All samples were viewed under the same intensity, and images captured under identical conditions. Used an arbitrary scoring system of 0 (none), 1 (mild), 2 (moderate), and 3 (intense) to NGAL antibody in the histological sample. The staining intensity was correlated with cold ischemia time, peak serum creatinine and dialysis requirement. NGAL expression was significant increased in the group that receive the kidney from cadaveric donor (46% had intensity antibody NGAL). Over the side, 58% that received kidney from living donor not showed image capture of NGAL	study).	contingency table 2x2 to compared serum creatinine and NGAL.

					antibody.		
Parikh, 2006	Patients were submitted kidney transplantation. Multicenter study (Connecticut, Colorado, Ohio, New York) in the USA.	53 patients were submitted kidney transplantation (25 women and 28 men). Age range between 4 to 61 old years.	Prospective cohort study. Patients were separated in 3 groups: living related donor (LRD) kidney transplantation and prompt graft function; deceased donor (CAD) kidney transplantation and prompt graph function and deceases donor (CAD) kidney transplantation with delayed graft function (DGF) patients that need dialysis within 1 wk of transplantation.	Patients were submitted kidney transplantation. Only describe serum creatinine on the day 0, 1,3,4 after surgery. Patients in the group deceased donor (CAD) kidney transplantation have delayed graft function and serum creatinine range between 3.5 until 14.2. However, serum creatinine in the group the patients that received the kidney from living donor had serum creatinine between 0.3 to 1.4.	Urine NGAL, urine IL-18 in the day 0, day 1, day 2, day 3, day 4 after surgery. In the group that received kidney from deceased donor NGAL in the day 0 had mean value 3306 ng/mg (range 17.6-5850). However, the group that received the kidney from living donor had NGAL between 35 to 2500.	Not reported.	Not describe follow up. It is possible to draw contingency table 2x2 to compared serum creatinine (<1.5mg/dl>) with urine NGAL (<150 ng/ml>) in the loss evaluation of the renal function.

Table 3. Results of risk assessment per study according to items on checklist for the quality assessment of diagnostic accuracy studies (QUADAS)[12].

	1	2	3	4	5	6	7	8	9	10	11
Hall, 2009[14]	+	-	+	+	+	+	?	?	+	+	?
Kusaka, 2008[15]	+	-	+	?	-	?	+	-	-	-	-
Lebkowska, 2009[16]	+	+	+	+	+	+	?	?	-	-	-
Malyszko, 2009[17]	+	-	-	?	?	?	?	?	?	-	-
Malyszko, 2009[18]	+	?	?	+	+	+	?	?	-	-	-
Mishra, 2006[19]	+	+	+	+	+	+	?	?	+	-	-
Parikh, 2006[20]	+	+	+	+	+	+	?	?	+	-	-

(+)= no bias ;(−) =potential bias; (?)3=bias unclear.
*1=valid selection, representative patients, 2=Acceptable reference standard, 3=Acceptable delay between tests, 4=Partial verification avoided, 5=Differential verification avoided, 6=incorporation avoided, 7= Reference standard results blinded, 8= Index test results blinded, 9= relevant clinical information, 10=Uninterruptible results reported 11= Withdrawals explained

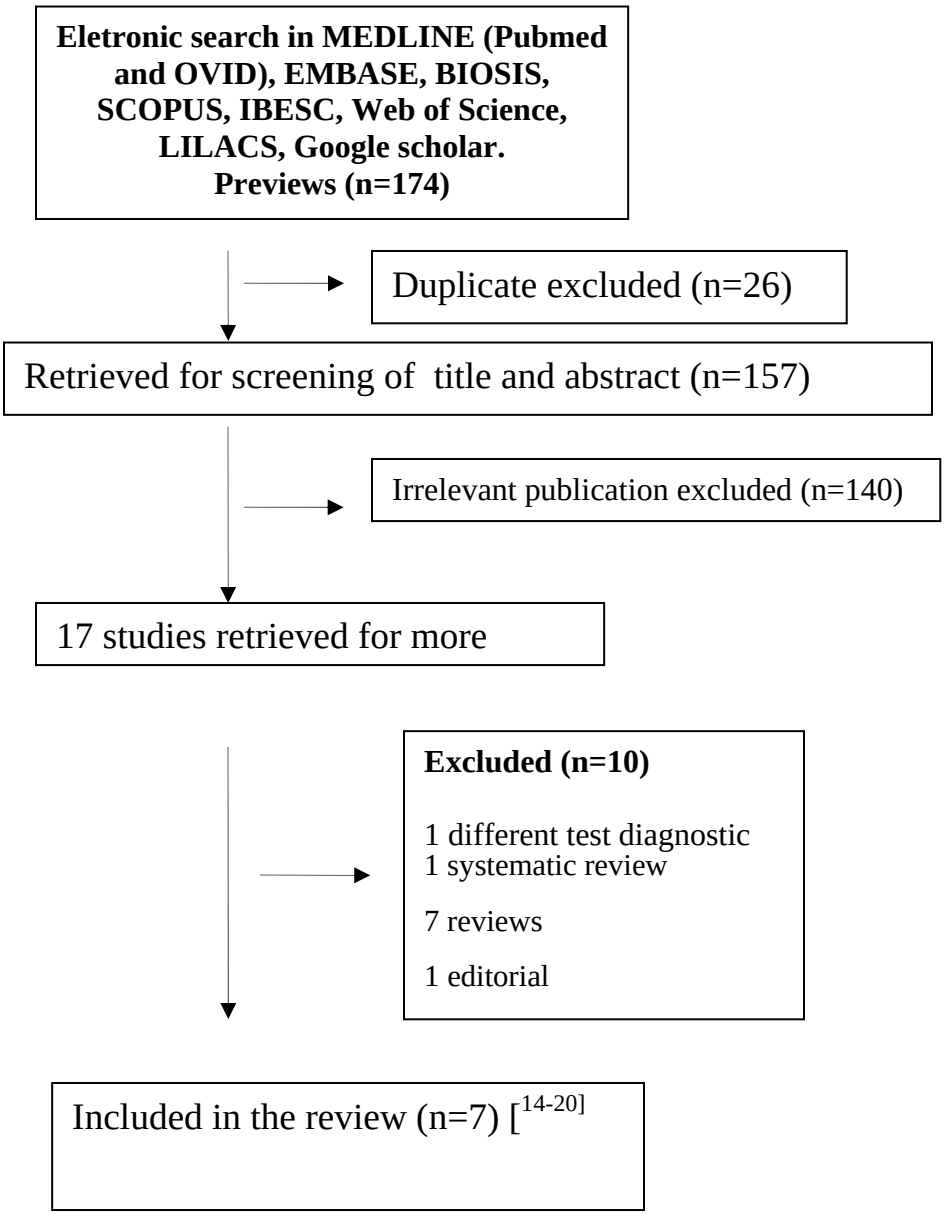


Figure 1. Study flow diagram

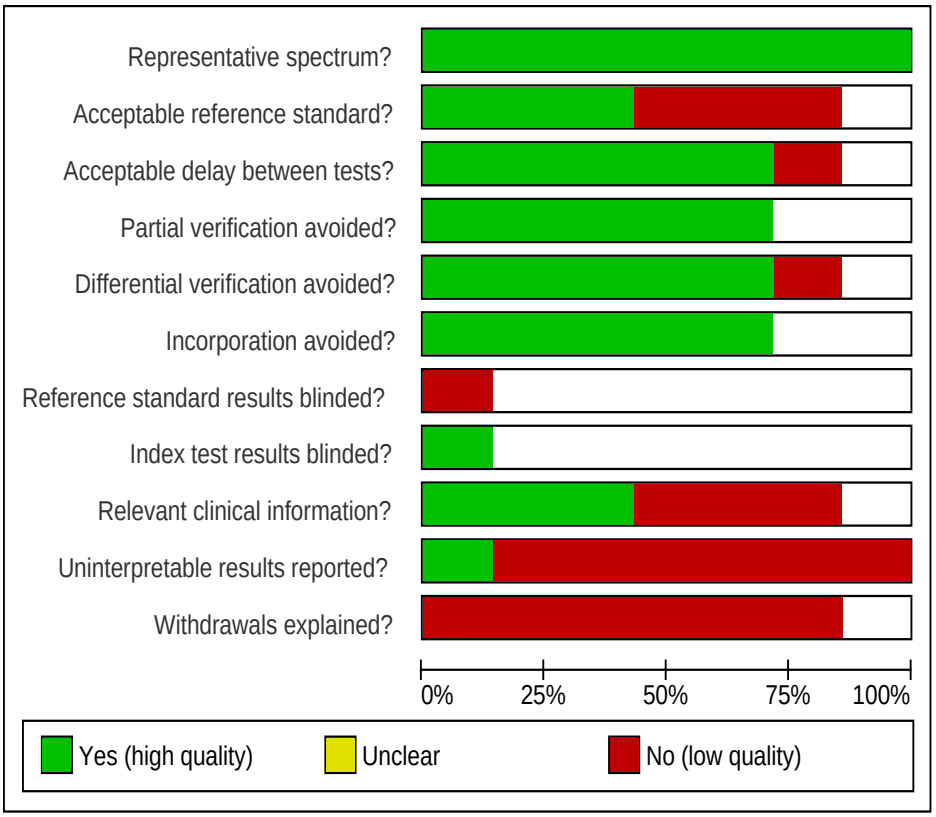


Figure 2. Reviewer judgments about methodological quality items, reported as percentages across all included studies

CONCLUSÕES E CONSIDERAÇÕES FINAIS

6. CONCLUSÕES E CONSIDERAÇÕES FINAIS

Nossa revisão com enfoque diagnóstico não encontrou evidências que apóiem o uso da dosagem urinária e sérica de NGAL como uma ferramenta diagnóstica na avaliação da injúria aguda pós-transplante renal. Verifica-se a necessidade de estudos com boa qualidade metodológica comparando o NGAL urinário e sérico e creatinina sérica no manejo da injúria renal que pode acometer os rins transplantados. Apesar de todas as nossas tentativas, através de uma revisão minuciosa da literatura médica em diversos bancos de dados e na Grey Literature (British Library e Google Acadêmico) da última década, de podermos eleger o NGAL como um instrumento capaz de nos dar uma indicação precoce sobre os eventos agudos de perda de função renal, não conseguimos demonstrar nosso objetivo, pois, não obtivemos fundamentos que pudessem sustentar esta afirmação. Desta forma entendemos que há a necessidade do prosseguimento desta investigação, uma vez que ao longo da revisão bibliográfica, este teste se mostrou de extrema importância no manejo precoce da injúria renal aguda que pode acontecer com os rins transplantados.

Este é o teste em que não há qualquer tipo de intervenção no paciente se comparado ao que o que temos hoje ainda é a biópsia renal, que é um método invasivo. Seria interessante dar prosseguimento a estes trabalhos elaborando estudos prospectivos com enfoque diagnóstico, aprimorados metodologicamente, evitando determinados vieses detectados nessa revisão. Nosso trabalho de Doutorado pretende contemplar este aspecto trazendo uma contribuição mais sustentável para o uso do NGAL na prática Nefrológica dos Transplantes.

Entendemos que é de grande relevância o estudo sobre este potencial biomarcador, pois, se não temos este tipo de parâmetro imediato após um transplante de rim, corremos o grande risco de SALVAR UM RIM, MAS, DE PERDERMOS O PACIENTE.

ANEXOS

ANEXO 1: PROJETO DE PESQUISA

ANEXO 1: PROJETO DE PESQUISA

1. Introdução

Conforme os últimos censos das Sociedades Nefrológica Mundial e Brasileira, o número de pacientes com insuficiência renal vem crescendo muito (1, 2). Sabemos, também, que a maior chance de recuperação destes pacientes e seu retorno à vida é através de um transplante de rim. Grandes avanços nesta área, não só sob o ponto de vista técnico-cirúrgico, quanto de entendimento do comportamento imunológico desencadeado por este órgão "estranho", vem acontecendo nos últimos anos.

Entretanto, há vários pontos que permanecem sem resposta e um deles, ao qual dedicaremos esta revisão sistemática diz respeito a métodos diagnósticos não-invasivos e precoces de injúria renal aguda neste rim transplantado. O surgimento de um biomarcador conhecido como Lipocalina Associada à Gelatinase dos Neutrófilos Humanos (NGAL), tem despertado uma grande expectativa nos pesquisadores, pois, talvez possa ser a ferramenta que todos estavam esperando (3).

O diagnóstico precoce da injúria renal aguda que acontece, também, nos rins transplantados tem sua importância mais do que fundamentada através do questionamento que se apresenta quando da necessidade de aumentarmos ou diminuirmos as drogas imunossupressoras (4).

Entendemos que este estudo se justifica por tentar mostrar os mecanismos que podemos

ter a nosso alcance quando uma situação de tal monta se apresenta nesta população de pacientes, extremamente vulnerável a qualquer intercorrência.

2 - Objetivos

2.1. Geral:

- Avaliar a acurácia diagnóstica do NGAL urinário com a creatinina sérica após o transplante renal, mostrando que os valores da Lipocalina traduzem, de forma mais precoce, a injúria renal aguda que pode acontecer nos transplantes renais, obrigando-nos a uma atitude imediata.

2.2. Específicos:

- Proceder uma revisão sistemática de estudos com enfoque diagnóstico, que avaliem as medidas urinárias e séricas do NGAL e creatinina, respectivamente, associada à injúria renal aguda pós-transplante renal.

- Avaliar a prevalência de injúria renal em pacientes transplantados de rim.

- Avaliar a sensibilidade de NGAL no diagnóstico da injúria renal em pacientes pós-transplante de rim.

- Avaliar a razão de verossimilhança positiva e negativa do NGAL no diagnóstico de injúria renal em pacientes pós-transplante de rim.

- Avaliar a probabilidade pós-teste positiva e negativa do NGAL no diagnóstico da injúria renal pós-transplante de rim.

3 - Método

3.3.1 – Critérios para Considerar Estudos da Revisão Sistemática

3.3.1.1 – Tipo de Delineamento dos Estudos

Serão incluídos todos os estudos de coorte e transversais que considerem os dois testes diagnósticos – NGAL urinário e creatinina sérica - para a abordagem da injúria renal aguda nos transplantes renais.

3.3.1.2 Critérios para inclusão dos estudos:

- Estudos comparativos entre a dosagem urinária de NGAL dentro das primeiras 6-48 horas desde o início do transplante de rim com o aumento da creatinina sérica de mais ou igual a 0,3 mg/dL ou aumento igual ou maior à 150 a 200 vezes o valor do momento zero do transplante.

3.3.1.3 Critérios para exclusão dos estudos:

- estudos que difiram significativamente na investigação clínica;
- estudos de revisão;
- o desfecho de interesse não foi estudado;
- estudos que não informam a média, o desvio padrão ou o intervalo de

confiança de 95% para as variáveis quantitativas;

- testes diagnósticos que não foram descritos de forma adequada

3.3.1.4 – Tipo de Participantes

Critérios de inclusão:

- Pacientes transplantados de rim com acompanhamento laboratorial pós-operatório imediato, sem alterações nas dosagens clássicas de injúria renal aguda.

Critérios de exclusão:

- Pacientes submetidos à Biópsia Renal no pós-operatório imediato.

3.3.2 – Tipos de testes diagnósticos avaliados

3.3.2.1 Teste diagnóstico padrão: A técnica laboratorial usada nas dosagens de creatinina sérica deverá ser a mesma em todos os trabalhos. A creatinina é obtida através da técnica do picrato alcalino (Método de Jaffé) clássico. Valores considerados alterados no teste diagnóstico da creatinina sérica para diagnóstico de injuria renal:

Segundo David S. Jacobs et al (5) insuficiência renal é definida por uma concentração de creatinina sérica > 1,5 mg/dL.

3.3.2.1 Teste diagnóstico: a técnica laboratorial deverá ser a mesma em todos os trabalhos separando-se apenas os que usaram NGAL urinária e sérica. Serão medidas através do kit de ensaio imuno-enzimático (enzyme-linked immunoassay - ELISA) comercializado por

BIOPORTO DIAGNOSTICS A/S, Gentofte, Denmark. Este protocolo utiliza um anticorpo monoclonal específico para determinação da quimiocina estudada, fornecido pelo fabricante.Valores considerados alterados no teste diagnóstico para injuria renal o valor de *cut-off* de NGAL considerado ótimo para predizer precocemente a injúria aguda renal no rim transplantado é, aproximadamente, 150 ng/mL (6).

3.3.3 - Estratégias de Busca dos estudos da Revisão Sistemática

As publicações com enfoque na injúria renal aguda pós-transplante de rim foram pesquisadas no banco de dados do ***PUBMED***, ***EMBASE CENTRAL***, ***MEDLINE-OVID***, ***LILACS***, biblioteca da ***COCHRANE LIBRARY***, ***IBECS***, ***BIOSIS***, em referências dos próprios artigos e em literatura menos científica ***(GREY LITERATURE-Google scholar; ISI-Web of science; British Library;)*** de janeiro de 2000 até final desta revisão**.** Sem restrição de língua ou qualquer outro tipo de limitação.

3.3.1. Estratégias de busca no MEDLINE-PUBMED (de maio de 2000 até final da revisão, 2010)

1. "sensitivity and specificity" [all fields]
2. "sensitivity and specificity/standards" [all fields]
3. "specificity" [all fields]
4. "screening" [all fields]
5. "false positive" [all fields]
6. "false negative" [all fields]

7. #1 OR #2 OR #3 OR #4 OR #5 OR #6
8. “accuracy” [all fields]
9. “predictive value” [all fields]
10. “predictive value of tests” [all fields]
11. “reference value” [all fields]
12. “reference values” [all fields]
13. “reference standards” [all fields]
14. #8 OR #9 OR #10 OR #11 OR #12 OR #13
15. “roc” [all fields]
16. “roc analysis” [all fields]
17. “roc and” [all fields]
18. “roc area” [all fields]
19. “roc auc” [all fields]
20. “roc characteristics” [all fields]
21. “roc curve” [all fields]
22. “roc curve method” [all fields]
23. “roc curves” [all fields]
24. “roc estimated” [all fields]
25. “roc evaluation” [all fields]
26. “likelihood ratio” [all fields]
27. #16 OR #17 OR #18 OR #19 OR #20 OR #21 OR #22 OR #23 OR #24 OR #25 OR #26
28. #7 OR #14 OR #27

29. animals[mh] not human [mh]
30. #28 NOT #29
31. Renal transplantation [mh]
32. Kidney transplantation [mh]
33. Kidney disease [mh]
34. Kidney*[tw] AND injury*[tw]
35. Renal*[tw] AND transplant*[tw]
36. #31 OR #31 OR #33 OR #34 OR #35
37. Biomarkers [tw] AND renal [tw]
38. Biomarkers [tw] AND kidney [tw]
39. Biomarkers [tw] AND renal [tw] AND transplant*[tw]
40. Biomarkers [tw] AND kidney [tw] AND transplant*[tw]
41. #37 OR #38 OR #39 OR #40
42. Acute renal disease
43. Acute kidney disease
44. Acute*[tw] AND renal [tw] AND disease [tw]
45. Acute*[tw] AND kidney [tw] AND disease [tw]
46. Acute*[tw] AND renal [tw] AND injury*[tw]
47. Acute*[tw] AND kidney [tw] AND injury*[tw]
48. #42 OR #43 OR #44 OR #45 OR #46 OR #47
49. #36 OR #41 OR #48
50. NGAL kidney
51. NGAL acute kidney injury

52. NGAL AND kidney
53. NGAL AND renal
54. #50 OR #51 OR #52 OR #53
55. #49 AND #54
56. #30 AND #55

3.3.2 -Estratégia de busca no EMBASE: (de 2000 até final da revisão, 2010)

1. sensitivity and specificity
2. sensitivity and specificity$ AND standards
3. specificity
4. screening
5. false ADJ positive
6. false ADJ negative
7. (#1 OR #2 OR #3 OR #4 OR #5 OR #6)
8. accuracy diagnostic
9. (predictive ADJ value)
10. (predictive ADJ value ADJ tests)
11. (reference ADJ value)
12. (reference ADJ values)
13. (reference ADJ standards)
14. (#8 OR #9 OR #10 OR #11 OR #12 OR #13)
15. Roc

16. (roc ADJ analysis)
17. (roc ADJ area)
18. (roc ADJ auc)
19. (roc ADJ characteristics)
20. (roc ADJ curve)
21. (roc ADJ curve ADJ method)
22. (roc ADJ curves)
23. (roc ADJ estimated)
24. (roc ADJ evaluation)
25. (likelihood ADJ ratio)
26. (#15 OR #16 OR #17 OR #18 OR #19 OR #20 OR #25)
27. (#7 OR #14 OR #26)
28. Nohuman
29. (animal not human)
30. (#28 OR #29)
31. (#27 NOT #30)
32. Renal AND transplantation
33. Kidney AND transplantation
34. (kidney ADJ disease)
35. (kidney ADJ injury)
36. (renal ADJ disease ADJ transplant$)
37. (renal ADJ injury ADJ transplant$)
38. (kidney ADJ disease ADJ transplant$)

39. (kidney ADJ injury ADJ transplant$)
40. #32 OR #33 OR #34 OR #35 OR #36 OR #37 OR #38 OR #39
41. (Biomarkers ADJ renal)
42. (Biomarkers ADJ kidney)
43. (Biomarkers ADJ renal ADJ transplant$)
44. (Biomarkers ADJ kidney ADJ transplant$)
45. (biomarkers ADJ renal)
46. (biomarkers ADJ AND kidney)
47. #41 OR #42 OR #43 OR #44 OR #45 OR #46
48. acute AND “renal disease”
49. acute AND “kidney disease”
50. (acute ADJ renal ADJ disease)
51. (acute ADJ kidney ADJ disease)
52. (acute ADJ renal ADJ injury)
53. (acute ADJ kidney ADJ injury)
54. #48 OR #49 OR #50 OR #51 OR #52 OR #53
55. #40 OR #47 OR #54
56. NGAL AND kidney
57. NGAL AND renal
58. #56 OR #57
59. #55 AND #58
60. #31 AND #59

3.3.3. Estratégias de busca do OVID (de 2000 até final da revisão, 2010):

1. sensitivity and specificity/
2. (sensitivity and specificity).tw.
3. specificity/
4. screening/
5. (false positive).mp.
6. (false negative).mp.
7. OR/1-6
8. (accuracy).mp.
9. (predictive value).mp.
10. (predictive value tests).mp.
11. (reference value).mp.
12. (reference values).mp.
13. (reference standards).mp.
14. OR/8-13
15. (Roc).mp.
16. (roc analysis).mp.
17. (roc area).mp.
18. (roc auc).mp.
19. (roc characteristics).mp.
20. (roc curve).mp.
21. (roc curve method).mp.

22. (roc curves).mp.

23. (roc estimated).mp.

24. (roc evaluation).mp.

25. (likelihood ratio).mp.

26. OR/15-25

27. (#7 OR #14 OR #26)

28. Nohuman

29. (animal not human)

30. (#28 OR #29)

31. (#27 NOT #30)

32. Exp Renal transplantation/

33. (Kidney$ adj transplantat$).tw.

34. (kidney adj disease).tw.

35. (kidney adj injury).tw.

36. (renal adj disease adj transplant$).tw.

37. (renal adj injury adj transplant$).tw.

38. (kidney adj disease adj transplant$).tw.

39. (kidney adj injury adj transplant$).tw.

40. OR /32-39

41. (Biomarkers adj renal).tw.

42. (Biomarkers adj kidney).tw.

43. (Biomarkers adj renal adj transplant$).tw.

44. (Biomarkers adj kidney adj transplant$).tw.

45. (biomarkers adj renal).tw.
46. (biomarkers adj kidney).tw.
47. OR/41-46
48. (acute adj renal adj disease).tw.
49. (acute adj kidney adj disease).tw.
50. (acute adj renal adj disease).tw.

3.3.4. Estratégia de busca na *COCHRANE LIBRARY-Cochrane Controlled Trial* (CCTR), *National Research Register* (NNR), *LILACS, BIOSIS, GREY LITERATURE* and *Clinical Trials* (de 2000 até final da revisão, 2010).

- Acute Kidney Injury (and/or)
- Kidney Transplantation (and/or)
- NGAL (and/or)
- Biomarker (and/or)

3.3.5 – Método da Revisão

Avaliação da Qualidade Metodológica e Extração dos Dados

Os estudos selecionados serão identificados de forma independente por dois revisores (G.R.L. e M.I.E.). A inclusão ou exclusão será feita usando a tabela QUADAS (7). Discordâncias sobre inclusão ou exclusão serão feitas inicialmente resolvidas através de

consenso, e quando isto não for possível, elas serão resolvidas por um revisor (L.R.M.). A concordância estatística entre os revisores será computada.

Todos os artigos encontrados pelo critério de elegibilidade terão suas qualidades metodológicas avaliadas. Essa avaliação envolve escrutinar o desenho clínico do estudo e dados relevantes sobre a população em estudo, teste diagnóstico e teste de referência (8, 9). Estas características incluem os métodos como os dados foram coletados, como foi realizada a seleção dos pacientes, a descrição do método laboratorial empregado, e a presença de viés (10, 11). Os resultados da avaliação da qualidade dos artigos serão sumarizados através do **QUADAS (*Quality assessment of studies of diagnostic accuracy included in systematic reviews*** (7)**.** Serão avaliadas determinadas características dos estudo, segundo **QUADAS**, e as informações foram categorizadas como: presente (+), ausente (-) e duvidosa (?) com as seguintes indagações:

1. O espectro de pacientes foi representativo dos pacientes que receberam o teste diagnóstico (NGAL)?
2. O teste de referencia padrão (creatinina sérica) classifica corretamente a condição de doença do paciente?
3. O tempo entre a execução do teste de referencia padrão e o teste diagnóstico (NGAL) foi suficiente para que não houvesse modificação na condições do paciente e na modificação dos dois testes?
4. Toda amostra recebe a avaliação do teste de referencia padrão (creatinina sérica)?
5. Todos pacientes receberam o mesmo teste de referência padrão (creatinina sérica) independente do resultado do teste diagnóstico (NGAL)?
6. Teste de referência padrão (creatinina urinária) era independe do teste

diagnóstico (NAGL), isto é o teste diagnóstico não faz parte do teste de referência padrão?

7. Foi o teste diagnóstico (NGAL) interpretado sem o conhecimento do resultado do teste de referência padrão (creatinina sérica)?

8. Foi o teste de referência padrão (creatinina sérica) interpretado sem o conhecimento dos resultados do teste diagnóstico (NGAL)?

9. Estavam os dados clínicos disponíveis quando os resultados dos testes forem interpretados e podem estar disponíveis quando o teste for usado na prática?

10. Os resultados que não forem interpretados, ou que forem inconclusivos, foram reportados?

11. Se houve descrição das razões para retirada de participantes do estudo?

3.3.6 - TIPOS DE DESFECHOS ANALISADOS

- correlação entre o aumento de NGAL urinária e sérica com aumento da creatinina sérica, no pós-transplante imediato (6-48h);
- correlação entre o aumento de NGAL urinária e sérica com uma diminuição do volume urinário de 0,5 mL/kg por hora por mais de 6-48h);
- correlação entre o aumento de NGAL urinária e sérica com uma diminuição do volume urinário de 0,5 mL/kg por hora por mais de 6 horas;

3.3.7 - Síntese dos Dados e Análise Estatística

Para avaliar a concordância entre a eligibilidade e a qualidade metodológica assim como a concordância entre os resultados do NGAL urinário e a creatinina sérica, serão calculadas as percentagens de concordância e o coeficiente κ (8). Em cada estudo será construída uma tabela 2 x 2 na qual os resultados do NGAL e da creatinina serão classificadas como valores < 150 > ng/mL e creatinina sérica < 1,5 > mg/dL, e serão calculados os índices de valores positivo-verdadeiros (sensibilidade) e índice falso-positivo (1-especificidade), as razões de verossimilhança positiva ou negativa e a probabilidade pós-teste para cada estudo com seus respectivos intervalos de confiança (ICs). Quando em alguma tabela 2 x 2 houver o valor de 0, será acrescentado o valor de 0,5, que será também adicionado aos outros valores das demais caselas.

A associação entre sensibilidade e especificidade será calculada pelo teste de Sperman de correlação entre duas variáveis contínuas para os três tipos de resultados:1. negativo: (NGAL > 150ng/mL e creatinina > 1,5 mg/dL); 2. positivo: (NGAL < 150 ng/mL e creatinina < 1,5 mg/dL); 3. Inconclusivo: (NGAL < 150 ng/mL e creatinina > 1,5 mg/dL) (8). Se não ocorrer correlação positiva, serão estimativas a sensibilidade e a especificidade porque se terá dois tipos de categoria – os resultados com desfecho positivo ou negativo do teste – e não existe variabilidade no limiar do teste diagnóstico (12). Razões de verossimilhança poderão ser estimadas a partir do sumário das estimativas de sensibilidade e especificidade usando-se as seguintes fórmulas: para razão de verossimilhança positiva, sensibilidade/1-especificidade e para

razão de verossimilhança negativa, 1-sensibilidade/1-especificidade (13).

Razão de verossimilhança indica que quanto maior o valor dado para a dosagem urinária

de

NGAL maior ou menor é a probabilidade do resultado final de o diagnóstico ser positivo, negativo ou inconclusivo para injúria renal aguda no rim transplantado (13).

Em adição, será calculado o somatório da probabilidade pré-teste (prevalência) usando o número de verdadeiros positivos de injúria renal aguda mais falso negativo de injúria renal divididos pelo número total de pacientes submetidos transplante renal. A heterogeneidade da sensibilidade e especificidade será calculada usando-se o teste estatístico *Cochran's Q* onde o valor do P será obtido comparando-o com o teste estatístico de distribuição do qui-quadrado com N-1 graus de liberdade (12).Caso a sensibilidade e a especificidade forem homogêneas, serão usados modelos de efeitos fixos para os cálculos do somatório geral, também com intervalo de confiança de 95% (12-14). A análise estatística será realizada com *software* Meta-Disc® (versão 1.4) (15). Em caso de ocorrer importante heterogeneidade e inconsistência dos dados, em decorrência de aspectos metodológicos, clínicos e estatísticos, realizar-se-á uma revisão sistemática narrativa sem realizar a metanálise estando nossa conduta de acordo com as orientações metodológicas do Handbook Cochrane 5 (16) (ANEXO 7) e do Guidance on the Conduct of Narrative Synthesis in Systematic Reviews (ANEXO 6) (17) e do PRISMA (ANEXO 8) (18).

A análise de subgrupo será feita separando os artigos que avaliarem o NGAL sérico e o NGAL urinário comparando com a creatinina sérica. E também avaliando as doações de doador vivo e doador cadáver quanto à perda da função renal e os resultados do NGAL urinário e serico e a cretinina sérica nestes grupos.

Análise Sensibilidade

Para avaliarmos a robustez dos nossos dados iremos realizar análise de sensibilidade elegendo os artigos de melhor qualidade metodológica como aqueles que tiverem pelo menos oito avaliações positivas pelo QUADAS (7).

4. REFERÊNCIAS BIBLIOGRÁFICAS

1. International Society of Nephrology. Acessado em 22 de Maio de 2010: Disponível em: <http://www.isn-online.org/isn/index.html>; 2010.

2. Sociedade Brasileira de Nefrologia. Acessado em 30 de Maio de 2010. Disponível em: <http://www.sbn.org.br/>; 2010.

3. Devarajan P. Emerging biomarkers of acute kidney injury. Contrib Nephrol. 2007;156:203-12.

4. Nickerson P. Post-transplant monitoring of renal allografts: are we there yet? Curr Opin Immunol. 2009 Oct;21(5):563-8.

5. Jacobs DS, Oxley DK, DeMott WR. Laboratory Test Handbook. 3rd ed. Hudson (Cleveland): Lexi-Comp; 2004.

6. Clinical Laboratory Internacional (CLI). NGAL ELISA Kits. Gentofte, Denmark: BIOPORTO DIAGNOSTICS A/S; 2010.

7. Whiting PF, Weswood ME, Rutjes AW, Reitsma JB, Bossuyt PN, Kleijnen J. Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. BMC Med Res Methodol. 2006;6:9.

8. Altman DG. Some common problems in medical research. Practical statistics for medical research. 9th ed. London: Chapman 1999. p. 396-439.

9. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. Ann Intern Med. 2003 Jan 7;138(1):40-4.

10. Jaeschke R, Guyatt G, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? Evidence-Based Medicine Working Group. JAMA. 1994 Feb 2;271(5):389-91.

11. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA. 1999 Sep 15;282(11):1061-6.

12. Deeks JJ. Systematic reviews of evaluation of diagnostic and screening test. In: Egger M, Smith GD, Altman DG, editors. Systematic reviews in Health care: Meat-analysis in context. 2nd ed. London. England: BMJ Piblishing; 2001. p. 248-82.

13. Irwig L, Tosteson AN, Gatsonis C, Lau J, Colditz G, Chalmers TC, et al. Guidelines for

meta-analyses evaluating diagnostic tests. Ann Intern Med. 1994 Apr 15;120(8):667-76.

14. Sutton AJ. Random affects methods for combining study estimates. In: Sutton A, J, Abrams RK, Jones DR, Sheldon TA, Song F, editors. Methods for Meta-analysis in Medical Research. Chichester, England: John Wiley; 2000. p. 73-86.

15. Zamora J, Abraira V, Muriel A, Khan K, Coomarasamy A. Meta-DiSc: a software for meta-analysis of test accuracy data. BMC Med Res Methodol. 2006;6:31.

16. Higgins JPT, Green S. Cochrane Handbook for Systematic Reviews of Interventions 5.0.2 [updated September 2009]: Available from: <http:// www.cochrane.org/training/cochrane-handbook> acess in 22/06/2010; 2009.

17. Popay J, Roberts H, Sowden A, Petticrew M, Arai L, Rodgers M, et al. Guidance on the Conduct of Narrative Synthesis in Systematic Reviews A Product from the ESRC Methods Programme. Swindon: ESRC; 2006.

18. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JP, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ. 2009;339:b2700.

5. CRONOGRAMA

	Janeiro de 2010	Fevereiro de 2010	Março de 2010	Abril de 2010	Maio de 2010	Ju
Revisão da literatura	X	X				
Compra do material	X					
Avaliação de estudos selecionados	X	X	X			
Análise estatística		X	X			
Análise de resultados para publicação				X		
Discussão				X	X	
Elaboração dos manuscritos				X	X	
Publicação e Apresentação do trabalho						

ANEXO 2: RIFLE

Research

Open Access

RIFLE criteria for acute kidney injury are associated with hospital mortality in critically ill patients: a cohort analysis

Eric AJ Hoste[1,2], Gilles Clermont[1], Alexander Kersten[1], Ramesh Venkataraman[1], Derek C Angus[1], Dirk De Bacquer[3] and John A Kellum[1]

[1]The Clinical Research, Investigation, and Systems Modeling of Acute Illness (CRISMA) Laboratory, Department of Critical Care Medicine, University of Pittsburgh, School of Medicine, Pittsburgh, Pennsylvania, USA
[2]Intensive Care Unit, Ghent University Hospital, Gent, Belgium
[3]Department of Public Health, Ghent University, Gent, Belgium

Corresponding author: John A Kellum, kellumja@ccm.upmc.edu

Received: 10 Mar 2006 Revisions requested: 27 Mar 2006 Revisions received: 1 Apr 2006 Accepted: 10 Apr 2006 Published: 12 May 2006

Critical Care 2006, **10**:R73 (doi:10.1186/cc4915)
This article is online at: http://ccforum.com/content/10/3/R73

Abstract

Introduction The lack of a standard definition for acute kidney injury has resulted in a large variation in the reported incidence and associated mortality. RIFLE, a newly developed international consensus classification for acute kidney injury, defines three grades of severity – risk (class R), injury (class I) and failure (class F) – but has not yet been evaluated in a clinical series.

Methods We performed a retrospective cohort study, in seven intensive care units in a single tertiary care academic center, on 5,383 patients admitted during a one year period (1 July 2000–30 June 2001).

Results Acute kidney injury occurred in 67% of intensive care unit admissions, with maximum RIFLE class R, class I and class F in 12%, 27% and 28%, respectively. Of the 1,510 patients (28%) that reached a level of risk, 840 (56%) progressed. Patients with maximum RIFLE class R, class I and class F had hospital mortality rates of 8.8%, 11.4% and 26.3%, respectively, compared with 5.5% for patients without acute kidney injury. Additionally, acute kidney injury (hazard ratio, 1.7; 95% confidence interval, 1.28–2.13; $P < 0.001$) and maximum RIFLE class I (hazard ratio, 1.4; 95% confidence interval, 1.02–1.88; $P = 0.037$) and class F (hazard ratio, 2.7; 95% confidence interval, 2.03–3.55; $P < 0.001$) were associated with hospital mortality after adjusting for multiple covariates.

Conclusion In this general intensive care unit population, acute kidney 'risk, injury, failure', as defined by the newly developed RIFLE classification, is associated with increased hospital mortality and resource use. Patients with RIFLE class R are indeed at high risk of progression to class I or class F. Patients with RIFLE class I or class F incur a significantly increased length of stay and an increased risk of inhospital mortality compared with those who do not progress past class R or those who never develop acute kidney injury, even after adjusting for baseline severity of illness, case mix, race, gender and age.

Introduction

Acute kidney injury is well recognized for its impact on the outcome of patients admitted to the intensive care unit (ICU). Illness severity scores such as the Acute Physiology and Chronic Health Evaluation version III (APACHE III) scoring system [1] and the Sequential Organ Failure Assessment score (SOFA) [2] both weight kidney dysfunction heavily (20% and 16.6% of the total scores for acute physiology). Yet there is no consensus on the amount of dysfunction that defines acute kidney injury, with more than 30 definitions in use in the literature today [3]. The variety of definitions used in clinical studies may be partly responsible for the large variations in the reported incidence (1–31%) [4-6] and the associated mortality (19–83%) [3,6-9] of acute kidney injury. Indeed, the lack of a uniform definition for acute kidney injury is believed to be a major impediment to research in the field [10]. Acute kidney injury is generally defined as 'an abrupt and sustained decrease in kidney function'. Until recently there has not been

APACHE III = Acute Physiology and Chronic Health Evaluation, version III; class F = failure, according to the RIFLE classification; class I = injury, according to the RIFLE classification; class R = risk, according to the RIFLE classification; Cr_{MDRD} = serum creatinine based upon the MDRD equation; ICU = intensive care unit; MDRD = Modification of Diet in Renal Disease; RIFLE = Risk, Injury, Failure, Loss, and End-stage Kidney; SOFA = Sequential Organ Failure Assessment score; $SOFA_{nonrenal}$ = SOFA score without points for renal insufficiency.

a consensus on how best to assess kidney function; namely, what markers best reflect kidney function, and what values of those markers discriminate normal from abnormal kidney function.

To establish a uniform definition for acute kidney injury, the Acute Dialysis Quality Initiative formulated the Risk, Injury, Failure, Loss, and End-stage Kidney (RIFLE) classification [11]. RIFLE defines three grades of increasing severity of acute kidney injury – risk (class R), injury (class I) and failure (class F) – and two outcome classes (loss and end-stage kidney disease) (see Table 1). A unique feature of the RIFLE classification is that it provides three grades of severity for acute kidney injury based on changes in either serum creatinine or urine output from the baseline condition. This allows classification of patients with acute kidney injury into one of the three RIFLE severity classes (Table 1).

RIFLE represents a new classification system issued from a process of formal evidence appraisal and expert opinion [11,12]. Three studies were recently published that used the RIFLE classification to evaluate the occurrence rate and/or outcome of acute kidney injury in two relatively small cohorts (207 ICU patients treated with renal replacement therapy and 183 ICU patients with acute kidney injury) and one larger cohort (813 patients after cardiac surgery) [13-15]. The clinical characteristics and predictive ability of this classification have not, however, been clinically validated in a large general ICU population. The aims of this study were therefore to characterize acute kidney injury defined by the maximum RIFLE classification, to examine the progression between stages of the classification, and to relate this classification to the length of stay and mortality in a large cohort of critically ill patients.

Patients and methods

Study population

We constructed a retrospective cohort of all adult hospitalizations during a 12 month period (1 July 2000–30 June 2001) at the University of Pittsburgh Medical Center that were admitted to one of its seven ICUs during their hospital stay. We excluded patients receiving chronic hemodialysis ($n = 146$) from the study cohort, and we only considered the first admission for patients who were readmitted to the ICU during the study period ($n = 327$). The University of Pittsburgh Medical Center is a tertiary care academic medical center with seven ICUs and more than 120 ICU beds serving medical, surgical, neurological, trauma and solid organ transplant patients.

Data collection

The study was approved by the Institutional Review Board of the University of Pittsburgh Medical Center. Data from different sources were merged by a non-investigator data manager (such as, an honest broker) and were stripped of all identifying information to preserve patient anonymity and to comply with local and federal regulations. Demographic data were retrieved from the electronic hospital database, laboratory data were retrieved from the laboratory database, and patient data were retrieved from the electronic hospital records. After merging data from the different sources, we performed automated and manual data verification. The patient data included demographic, administrative, physiologic, laboratory and hospital outcome information. Ethnicity, reported as white, black or other, was reported by the admitting nurse, and was used to calculate the glomerular filtration rate assessed by the Modification of Diet in Renal Disease (MDRD) equation [16]. High-density (every two hours) physiologic data were only available while patients were in the ICU, while other data sources covered the entire hospitalization. Urine output was recorded at least once every two hours, and serum creatinine was measured at least once daily.

RIFLE criteria

We classified patients according to the maximum RIFLE class (class R, class I or class F) reached during their hospital stay. The RIFLE class was determined based on the worst of either glomerular filtration rate criteria or urine output criteria. We used the change in serum creatinine level and urine output to classify patients according to the RIFLE criteria.

Patients who met any of the criteria of the RIFLE classification were classified as acute kidney injury patients. For patients without chronic kidney insufficiency as reported in the medical history, we calculated a serum creatinine level using the MDRD equation [16] (Cr_{MDRD}) as recommended by the Acute Dialysis Quality Initiative, by solving the MDRD equation for serum creatinine assuming a glomerular filtration rate of 75 ml/minute/1.73 m^2. We then used the lowest creatinine value among the hospital admission creatinine, the ICU admission creatinine or the Cr_{MDRD} creatinine as the baseline value. Approximately one-half of patients were classified using the Cr_{MDRD} as a baseline. None of these values differed by very much, however (mean difference between creatinine on admission and $Cr_{MDRD} = 0$; interquartile range -0.3 to 0.3), and our results are not qualitatively different regardless of which baseline is used. For patients with a history of kidney insufficiency (but not on chronic dialysis) we used their hospital admission creatinine as their baseline. We did not evaluate the outcome classes of RIFLE (loss and end-stage kidney disease criteria) in this study.

Severity of illness

The APACHE III [1] and the SOFA [2] scores were calculated based on the worst variables recorded during the first 24 hours of ICU admission. The nonrenal total SOFA score was calculated from the total SOFA score minus the points for kidney insufficiency. In addition, we calculated the SOFA score and the nonrenal SOFA score on basis of the worst variables recorded during the 24 hours preceding the maximum RIFLE class.

Table 1

Risk, Injury, Failure, Loss, and End-stage Kidney (RIFLE) classification

Class	Glomerular filtration rate criteria	Urine output criteria
Risk	Serum creatinine × 1.5	< 0.5 ml/kg/hour × 6 hours
Injury	Serum creatinine × 2	< 0.5 ml/kg/hour × 12 hours
Failure	Serum creatinine × 3, or serum creatinine ≥ 4 mg/dl with an acute rise > 0.5 mg/dl	< 0.3 ml/kg/hour × 24 hours, or anuria × 12 hours
Loss	Persistent acute renal failure = complete loss of kidney function > 4 weeks	
End-stage kidney disease	End-stage kidney disease > 3 months	

For conversion of creatinine expressed in conventional units to SI units, multiply by 88.4. RIFLE class is determined based on the worst of either glomerular filtration criteria or urine output criteria. Glomerular filtration criteria are calculated as an increase of serum creatinine above the baseline serum creatinine level. Acute kidney injury should be both abrupt (within 1–7 days) and sustained (more than 24 hours). When the baseline serum creatinine is not known and patients are without a history of chronic kidney insufficiency, it is recommend to calculate a baseline serum creatinine using the Modification of Diet in Renal Disease equation for assessment of kidney function, assuming a glomerular filtration rate of 75 ml/min/1.73 m^2. When the baseline serum creatinine is elevated, an abrupt rise of at least 0.5 mg/dl to more than 4 mg/dl is all that is required to achieve class Failure.

Statistical analysis

The central tendency for continuous data is expressed as the mean ± standard deviation or the median (interquartile range). We tested continuous variables for normality by distribution plots. We compared means using the Student's *t* test when normally distributed, and the Mann-Whitney *U* test when not. Comparisons across multiple groups were performed using the *F* test, with Bonferroni correction for multiple comparisons [17]. When data were not normally distributed, we used the Kruskal-Wallis *H* analysis of variance test; significant changes over the observation period were tested with the Mann-Whitney *U* test.

We performed univariate and multivariable logistic regression to assess the impact of different baseline characteristics found to be significantly different over the four groups, on the occurrence of acute kidney injury and on maximum RIFLE class F. In the multivariable model, all covariates were entered simultaneously (enter method). We analyzed for collinearity by assessing correlation between covariates. For continuous variables we analyzed the relationship between the outcome and the variable with locally weighted scatterplot smoothing in order to assess whether categorization was necessary. Finally, the goodness of fit of the model was tested by means of the Hosmer-Lemeshow statistic.

We analyzed hospital survival across groups using the chi-square and the Kaplan-Meier methods, and we tested differences between groups using the log-rank test. Patients alive at hospital discharge were censored. We performed a Cox proportional hazards regression analysis to examine whether the maximum RIFLE class and the incidence of acute kidney injury (defined as patients who fulfilled one of the RIFLE classes) were associated with mortality.

To correct for differences in patient characteristics, we included simultaneously age, gender, race, the main reason for ICU admission, the medical or surgical admission category, and the nonrenal SOFA score on ICU admission or at the maximum RIFLE class in the model (enter method). The nonrenal SOFA score was chosen as a covariate to control for multicollinearity between the RIFLE classification and scoring systems that include points for kidney insufficiency such as the APACHE III and SOFA scores. Interactions between the 'main reason for admission' and the maximum RIFLE class were explored, and were found not to be significant.

We tested whether it was appropriate to treat continuous variables as continuous by a residuals plot. We tested the assumption of proportionality of hazards by plotting hazard rates against time for the four different categories, as well as by the numerical method proposed by Lin and colleagues [18] derived from cumulative sums of martingale residuals. We found no evidence of violating the proportional hazards assumption. Finally, we tested the qualitative goodness of fit of the model with residual plots. A double-sided *P* value less than 0.05 was considered significant. Analysis was performed with the statistical software package SPSS 11.0.1 (SPSS Inc., Chicago, IL, USA).

Results

Characteristics of patients with acute kidney injury

A total of 5,383 patients was evaluated. The baseline characteristics of the patient cohort are presented according to the maximum RIFLE class in Table 2. The four groups differed in age, race, pre-existing chronic kidney insufficiency, admission type, severity of illness on admission and at the time of maximum RIFLE class, APACHE III score, SOFA score and the nonrenal SOFA score, and the proportion of patients already admitted inhospital to another non-ICU ward.

Results of the regression analyses examining the impact of the different baseline characteristics on the appearance of acute kidney injury and maximum RIFLE class F are presented in Table 3. Increasing age, greater severity of illness (APACHE III, SOFA and nonrenal SOFA scores), pre-existing chronic

Table 2

Baseline characteristics of patients classified according to the maximum Risk, Injury, Failure, Loss, and End-stage Kidney (RIFLE) class

	No acute kidney injury	Risk	Injury	Failure
n	1,766 (32.8%)	670 (12.4%)	1,436 (26.7%)	1,511 (28.1%)
Sex (male)	988 (55.9%)	372 (55.5%)	841 (58.6%)	570 (57.0%)
Age (years)*	56.6 (18.2)	63.4 (17.0)	62.6 (16.6)	62.1 (16.4)
Race ($n = 5,101$)**				
White	1,491 (89.5%)	587 (91.3%)	1,237 (90.7%)	1,255 (87.9%)
Black	146 (8.8%)	45 (7.0%)	111 (8.1%)	156 (10.9%)
Other	29 (1.7%)	11 (1.7%)	16 (1.2%)	17 (1.2%)
Chronic kidney insufficiency*	17 (1.0%)	4 (0.6%)	17 (1.2%)	121 (8.0%)
Admission type* ($n = 5,375$)				
Medical	778 (44.2%)	277 (41.4%)	566 (39.4%)	545 (36.1%)
Surgical	984 (55.8%)	392 (58.6%)	870 (60.6%)	963 (63.9%)
Reason for admission according to organ system* ($n = 5,375$)				
Cardiovascular disease	446 (25.3%)	236 (35.3%)	446 (31.1%)	513 (34.0%)
Neurological disease	311 (17.7%)	109 (16.3%)	314 (21.9%)	327 (21.7%)
Pulmonary disease/infection	226 (12.8%)	101 (15.1%)	232 (16.2%)	291 (19.3%)
Trauma	317 (18.0%)	85 (12.7%)	203 (14.1%)	163 (10.8%)
Malignancy	117 (6.6%)	34 (5.1%)	38 (2.6%)	35 (2.3%)
Gastrointestinal disease	67 (3.8%)	22 (3.3%)	32 (2.2%)	26 (1.7%)
Other	278 (15.8%)	82 (12.3%)	171 (11.9%)	153 (10.1%)
APACHE III score* ($n = 3,400$)	36 (26–47)	46 (35–57)	46 (36–58)	56 (41–73)
SOFA score*	5.1 (3.3)	6.3 (4.0)	6.8 (4.1)	7.8 (4.5)
$SOFA_{nonrenal}$ score*	4.5 (3.1)	5.3 (3.5)	5.6 (3.5)	5.9 (3.9)
SOFA $RIFLE_{max}$ score* ($n = 4,994$)	3.7 (3.2)	5.3 (3.8)	5.9 (3.8)	6.7 (4.3)
$SOFA_{nonrenal}$ $RIFLE_{max}$ score* ($n = 4,994$)	3.2 (2.8)	4.5 (3.5)	5.0 (3.4)	5.0 (3.7)
Inhospital before ICU admission*	527 (29.8%)	243 (36.3%)	476 (33.1%)	592 (39.2%)
Pre-ICU LOS* (days)	1 (1–4)	2 (1–4)	2 (1–5)	2 (1–6)
Time to $RIFLE_{max}$* (days)		2 (1–3)	2 (1–4)	2 (1–7)
RIFLE class on glomerular filtration rate criteria*		463 (69.1%)	929 (64.7%)	1,110 (73.5%)

Continuous variables are presented as the mean (standard deviation) when normally distributed or as the median (interquartile interval) when not normally distributed. Categorical variables are presented as percentages. No acute kidney injury is those patients without any occurrence of RIFLE criteria; APACHE III, Acute Physiology and Chronic Health Evaluation, version III; $SOFA_{nonrenal}$, Sequential Organ Failure Assessment score without points for renal insufficiency; SOFA $RIFLE_{max}$, Sequential Organ Failure Assessment score at the time of maximum RIFLE class; pre-ICU LOS, length of hospital stay before intensive care unit admission (only for patients who were in hospital before intensive care unit admission). *$P < 0.001$, **$P = 0.035$.

kidney insufficiency and a preceding admission to a non-ICU ward in the hospital were associated with increased risk for occurrence of acute kidney injury and maximum RIFLE class F. In addition, black patients had increased risk for development of maximum RIFLE class F. Medical admissions were less likely to result in acute kidney injury or RIFLE class F compared with surgical admissions.

Progression of acute kidney injury to maximum RIFLE class

The progression of acute kidney injury during the ICU stay to the maximum RIFLE class is shown in Figure 1. On the first day of ICU admission, 1,182 patients (21.9%) already had acute kidney injury, defined by the RIFLE criteria. During the entire ICU stay, 3,617 patients (67.2%) had an episode of acute kid-

Table 3

Impact of baseline characteristics on the occurrence of acute kidney injury (multivariate logistic regression analysis)

Characteristic	Covariates associated with occurrence of acute kidney injury		Covariates associated with occurrence of maximum RIFLE class failure	
	Odds ratio (95% confidence interval)	*P*	Odds ratio (95% confidence interval)	*P*
Age (per year older)	1.02 (1.02–1.03)	< 0.001	1.01 (1.00–1.01)	0.001
Race (reference white)		0.130		0.001
Black	1.20 (0.96–1.50)	0.111	1.50 (1.21–1.86)	< 0.001
Other	0.73 (0.44–1.23)	0.237	0.78 (0.41–1.38)	0.397
Chronic kidney insufficiency	4.19 (2.48–7.10)	< 0.001	8.86 (6.01–13.05)	< 0.001
Medical admission (reference surgical)	0.79 (0.69–0.90)	< 0.001	0.76 (0.66–0.87)	< 0.001
Reason for admission according to organ system (reference cardiovascular disease)		< 0.001		< 0.001
Trauma	0.64 (0.53–0.79)	< 0.001	0.64 (0.52–0.80)	< 0.001
Neurological disease	0.93 (0.78–1.13)	0.481	1.02 (0.85–1.2)	0.830
Pulmonary disease and infection	1.08 (0.88–1.32)	0.461	1.16 (0.96–1.40)	0.120
Gastrointestinal disease	0.51 (0.35–0.73)	< 0.001	0.51 (0.32–0.66)	0.004
Malignancy	0.36 (0.27–0.49)	< 0.001	0.45 (0.31–0.66)	< 0.001
Other	0.57 (0.47–0.70)	< 0.001	0.60 (0.48–0.74)	< 0.001
$SOFA_{nonrenal}$ (per point greater)	1.19 (1.16–1.21)	< 0.001	1.08 (1.06–1.10)	< 0.001
In hospital before ICU admission	1.18 (1.03–1.36)	0.015	1.19 (1.04–1.36)	0.012

$SOFA_{nonrenal}$, Sequential Organ Failure Assessment score without points for kidney insufficiency; ICU, intensive care unit. The odds ratios were calculated with logistic regression analysis. The goodness of fit of the multivariable regression model was tested by the Hosmer-Lemeshow statistic: $P = 0.080$ for the model with acute kidney injury as the endpoint, and $P = 0.019$ for the model with maximum Risk, Injury, Failure, Loss, and End-stage Kidney (RIFLE) class failure as the endpoint.

ney injury defined by RIFLE criteria. One-half of the patients reached the maximum RIFLE class within 2 days after ICU admission (Table 2).

More than 50% of the patients with RIFLE class R progressed to RIFLE class I or class F, and more than one-third of the patients with RIFLE class I progressed to class F. The time to progress to class I was 1 (0.5–3.6) days and the time to progress to class F was 4 (1.4–9.5) days. Patients who progressed to a higher RIFLE class during the ICU stay were older compared to patients whose renal function did not deteriorate, (62.4 ± 16.6 years versus 58.5 ± 18.0 years, $P < 0.001$) and were already more severely ill on admission, as illustrated by the greater APACHE III score (47 (36–62) versus 41 (29–56), $P < 0.001$) and the greater nonrenal SOFA score (5.7 (3.6) versus 4.8 (3.4), $P < 0.001$).

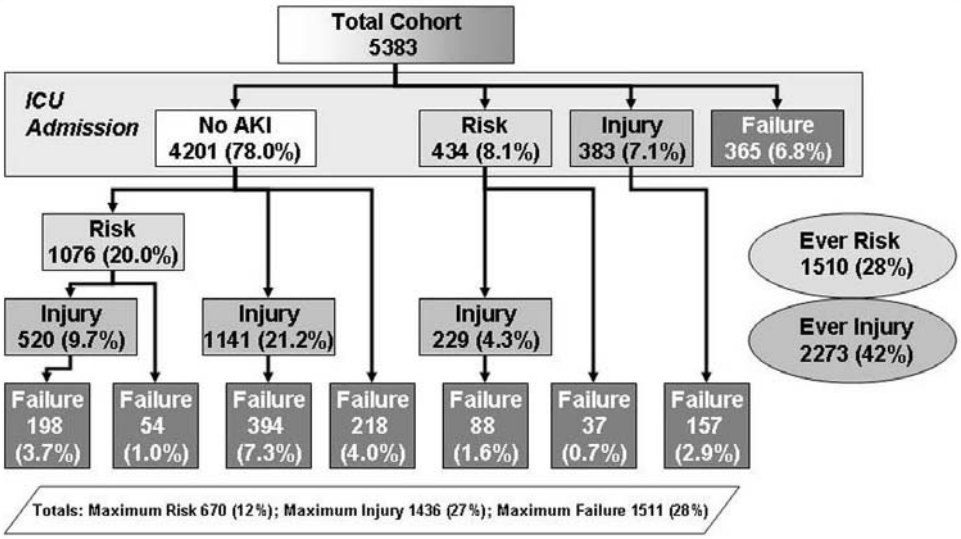
Mortality, length of stay and renal replacement therapy

Less than 1% of patients with maximum RIFLE class I and 14.2% of patients with maximum RIFLE class F received renal replacement therapy (Table 4). Increasing severity of acute kidney injury was associated with an increasing length of ICU stay and hospital stay, and higher mortality (Table 5 and Figure 2). Patients with maximum RIFLE class R, class I and class F had hospital mortality rates of 8.8%, 11.4% and 26.3%, respectively, compared with 5.5% for patients without acute kidney injury. Patients with maximum RIFLE class F based on glomerular filtration rate criteria had a somewhat higher inhospital mortality compared with patients who had a maximum RIFLE class F on urine output criteria (27.9% versus 21.9%, $P = 0.020$). The unadjusted hazard ratios (95% confidence interval) for hospital mortality for acute kidney injury and RIFLE class R, class I and class F were, respectively, 2.1 (1.67–2.57, $P < 0.001$), 1.3 (0.91–1.93, $P = 0.142$), 1.9 (1.45–2.48, $P < 0.001$) and 3.4 (2.64–4.29, $P < 0.001$). After adjustment for covariates, acute kidney injury was still associated with an almost twofold increased hazard for hospital mortality (Table 5, panel A). Maximum RIFLE class I and class F were both associated with mortality in the covariate-adjusted Cox regression model (Table 5, panel B). These results were unchanged when the nonrenal SOFA at the time of maximum RIFLE class was substituted for the nonrenal SOFA at ICU admission.

Discussion

We found that acute kidney injury, defined by the RIFLE classification, had a high incidence (67.2%) and was associated with an increased risk for hospital mortality compared with those who never developed acute kidney injury. The incidence of almost 70% may appear at odds with the existing literature [5]. Even when limiting cases to those with RIFLE class F (28%) we found a higher rate for ICU patients than typically reported. Fourteen percent of class F patients received renal

Figure 1



Flow chart of the clinical course of patients until the maximum RIFLE class. Data expressed as patient numbers who were identified at each level, and the percentage of the total number of patients. Patients who appear to skip a grade (class risk or class injury) do so because they did not remain at a transition state for at least 24 hours. 'Ever Risk' and 'Ever Failure' refers to the number of patients who could be identified at this stage. AKI, acute kidney injury; ICU, intensive care unit; RIFLE, Risk, Injury, Failure, Loss, and End-stage Kidney Disease.

replacement therapy, however, leading to a rate of 4–5% among ICU patients, consistent with previous reports [9,19]. Indeed, our study highlights the potential for under-reporting when renal replacement therapy is used to 'define' acute kidney injury. Importantly, even milder degrees of kidney dysfunction, RIFLE class R or class I, were still associated with excess mortality compared with patients who maintained normal function. RIFLE provided a well-balanced classification system for determination of patients with different severity of acute kidney injury, at least as far as risk of mortality or need for renal replacement therapy is concerned.

Not surprisingly, the occurrence of acute kidney injury and maximum RIFLE class F were associated with increased baseline severity of illness and older age (Tables 2 and 3). Patients developing acute kidney injury were slightly older and had higher APACHE III and SOFA scores, even when kidney dysfunction was not counted. However, the severity within acute kidney injury was not so affected by these factors. Patients progressing to RIFLE class I and class F were no older and their nonrenal SOFA scores no greater than patients remaining in RIFLE class R. Although, Herget-Rosenthal and colleagues have also described the progression of acute kidney injury in a selected cohort of 85 ICU patients [20], this is to our knowledge the first time that the progression of acute kidney injury has been examined in a large dataset of general ICU patients.

RIFLE class R would appear to be aptly named. More than one-half of the patients of class R progressed to more severe RIFLE classes, yet those that did not were not at increased risk of hospital mortality. Future studies could target this population for prevention. RIFLE class I may also have been fortuitously named, for this is the stage at which risk for hospital mortality increases even after controlling for covariates. It was commonly held until fairly recently that patients die 'with, and not of, acute renal failure'. Medication (for example, erythropoietin and diuretics) and renal replacement therapy were thought to 'replace' the loss of kidney function. It has already been demonstrated in critically ill patients that severe acute renal failure, defined as the need for renal replacement therapy or oliguria, is independently associated with mortality [4,5,19,21]. In addition, in a cardiothoracic surgery population and in a cohort of hospitalized patients, both with a lower baseline mortality compared with general ICU patients, small changes in serum creatinine were associated with a worse outcome [22,23]. In the present study we confirm the association of acute kidney injury with increased hospital mortality in a general ICU population. This is a remarkable finding considering how common this condition appears to be – 55% of all

Table 4

Outcomes for all patients and for patients classified according to the maximum Risk, Injury, Failure, Loss, and End-stage Kidney (RIFLE) class

	No acute kidney injury ($n = 1,766$)	Risk ($n = 670$)	Injury ($n = 1,436$)	Failure ($n = 1,511$)	All injury ($n = 5,383$)
Renal replacement therapy*	1 (0.1%)	0 (0%)	4 (0.3%)	214 (14.2%)	219 (4.1%)
Hospital LOS after reaching maximum RIFLE class (days)*	5 (3–10)	5 (3–10)	7 (4–14)	11 (5–23)	7 (3–14)
ICU LOS (days)*	3 (2–4)	3 (2–6)	5 (3–10)	9 (4–21)	4 (2–9)
Hospital LOS (days)*	6 (4–10)	8 (5–14)	10 (6–19)	16 (9–31)	9 (5–19)
Hospital mortality*	97 (5.5%)	59 (8.8%)	163 (11.4%)	398 (26.3%)	717 (13.3%)

Continuous variables presented as the median (interquartile interval) and categorical variables presented as the percentage. LOS, length of stay; ICU, intensive care unit. *$P = 0.001$ between the four subgroups no acute kidney injury, RIFLE class risk, RIFLE class injury and RIFLE class failure.

patients had RIFLE class I or class F. Furthermore, there was increasing mortality risk over RIFLE classes, despite the fact that these ICU patients had similar comorbidity, as reflected by the nonrenal SOFA score.

The finding that moderate degrees of kidney dysfunction pose a significant risk of death is particularly notable given that we know very little of why this should be. Acute kidney injury may simply be colinear with unmeasured elements of comorbidity, or it may be causally related to the increased mortality. Future studies should consider exploring whether alternative management of patients with mild degrees of kidney dysfunction could change the outcome. If the problem is actually the kidney, then possible mechanisms underlying the excess mortality associated with acute kidney injury are likely to be found in the pathophysiologic changes resulting from kidney insufficiency and adverse effects of renal replacement therapy [24,25]. Salt and water retention resulting in volume overload, hyperkalemia and acid-base derangements [26], perhaps leading to decreased blood pressure, cardiac output, hepatic and renal blood flow [27], to insulin resistance and protein breakdown, and even to alterations in innate immunity [28], all may contribute to the excess mortality in this group of patients. Furthermore, patients with acute kidney injury have a high incidence of infectious complications [29-31] and frequently develop anemia. Finally, acute kidney injury itself can lead to a non-infectious, proinflammatory response with activation of leukocytes, secretion of proinflammatory cytokines and recruitment of neutrophils and macrophages with resultant lung injury, as has been demonstrated in animal models of ischemia-reperfusion-induced acute renal failure [32,33]. All these changes may occur prior to, or even in, patients never receiving renal replacement therapy. These same mechanisms, however, may explain why patients who are treated with a lower dose of renal replacement therapy have a worse survival [34-36].

Our study has certain limitations. First, we did not attempt to compare RIFLE with other classification systems; nor did we compare urine output and creatinine criteria, but rather used the criteria as proposed by the Acute Dialysis Quality Initiative workgroup, as the worst classification by each criterion. It is possible that urine output and creatinine criteria provide complementary information, which is lost when these criteria are combined.

The Acute Dialysis Quality Initiative recommended the use of a baseline serum creatinine, yet a true baseline is often unknown for patients admitted to the ICU. Several possible baseline values existed for our patients (hospital admission, ICU admission, or a calculated baseline from the MDRD equation). Our use of the lowest of these values for any given patient may have lead to a higher estimate of change and therefore a higher estimate of the incidence of acute kidney injury. Although the MDRD equation was developed and validated on a large number of patients, conflicting results have been published regarding the validation of this equation in different patient populations. We acknowledge that this equation is only a substitute for the actual glomerular filtration rate, but validation of this equation or developing an alternative for the MDRD-derived baseline creatinine was beyond the scope of this study.

We also acknowledge that some members of our research group have contributed to the consensus process by which RIFLE was developed and by which MDRD recommendations were made. In addition, patient follow-up in our study was limited to hospital discharge information.

Some patients may have died shortly after hospital discharge. As shown in Figure 2, the curves continue to separate, particularly for those in the class F group. Longer follow-up would also be required to examine the RIFLE endpoints 'loss' and 'end-stage disease'. Early renal replacement therapy may theoretically influence the criteria, and patients that would have

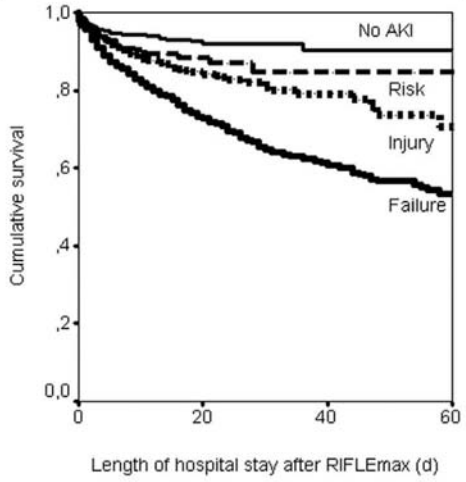
Table 5

Association of Risk, Injury, Failure, Loss, and End-stage Kidney (RIFLE) criteria with mortality

	Hazard ratio (95% confidence interval)	*P*
Panel A: association of AKI, defined by meeting any RIFLE criteria, with mortality		
AKI (reference no AKI)	1.7 (1.28–2.13)	< 0.001
$SOFA_{nonrenal}$ score (per point)	1.1 (1.10–1.15)	< 0.001
Age (per 10 years)	1.1 (1.07–1.18)	< 0.001
Sex (female)	0.9 (0.78–1.07)	0.252
Race (reference white)		0.528
Black	1.1 (0.88–1.45)	0.324
Other	0.8 (0.41–1.68)	0.608
Medical admission (reference surgical admission)	2.9 (2.44–3.40)	< 0.001
Main reason for admission (reference cardiovascular disease)		< 0.001
Trauma	0.9 (0.66–1.09)	0.196
Neurological disease	1.3 (1.03–1.52)	0.026
Pulmonary disease, infection	0.8 (0.65–1.03)	0.087
Gastrointestinal disease	0.5 (0.21–1.25)	0.140
Malignancy	0.2 (0.04–0.63)	0.009
Other	0.4 (0.22–0.62)	< 0.001
Panel B: association of maximum RIFLE class with mortality		
$RIFLE_{max}$ (reference no AKI)		< 0.001
Risk	1.0 (0.68–1.56)	0.896
Injury	1.4 (1.02–1.88)	0.037
Failure	2.7 (2.03–3.55)	< 0.001
$SOFA_{nonrenal}$ score (per point)	1.1 (1.10–1.14)	< 0.001
Age (per 10 years)	1.1 (1.08–1.20)	< 0.001
Sex (female)	0.9 (0.76–1.06)	0.190
Race (reference white)		0.673
Black	1.1 (0.87–1.45)	0.389
Other	0.9 (0.44–1.97)	0.844
Medical admission (reference surgical admission)	2.7 (2.27–3.21)	< 0.001
Main reason for admission (reference cardiovascular disease)		< 0.001
Trauma	0.9 (0.69–1.18)	0.463
Neurological disease	1.3 (1.06–1.60)	0.011
Pulmonary disease, infection	0.8 (0.66–1.07)	0.154
Gastrointestinal disease	0.6 (0.24–1.41)	0.230
Malignancy	0.2 (0.04–0.69)	0.013
Other	0.3 (0.20–0.60	< 0.001

Covariate-adjusted Cox proportional hazard regression analysis. AKI, acute kidney injury, patients meeting at least one of the RIFLE criteria; $RIFLE_{max}$, maximum RIFLE class; $SOFA_{nonrenal}$, Sequential Organ Failure Assessment score without points for kidney failure, determined on data from the first 24 hours of admission.

Figure 2



Kaplan-Meier curves for survival (inhospital) by maximum RIFLE class. Patients discharged alive were censored. Log-rank statistic, $P < 0.001$. AKI, acute kidney injury; $RIFLE_{max}$, maximum Risk, Injury, Failure, Loss, and End-stage Kidney Disease (RIFLE) class during the intensive care unit stay (days).

reached class F could be classified in our study as class R or class I. Only four class I patients were treated with renal replacement therapy, however, and reclassification of these patients to class F does not influence our results.

Although our study is relatively large and included seven ICUs, it was conducted at a single medical center whose case mix and referral patterns may not be representative of other centers. The case mix of this study cohort could have hindered the detection of specific conditions that influence the development of acute kidney injury. Finally, our retrospective study design, using existing medical records, limited our ability to look outside the ICU and to collect information on potential mechanisms of injury. Our design also prohibited the use of more sophisticated measures of kidney function. Indeed, our assessment of time to progression of acute kidney injury may have been artificially lengthened due to daily measurement of creatinine – some patients appeared to skip class R or class I because of this limitation.

Conclusion

In this general ICU population, acute kidney 'risk, injury, failure' as defined by the newly developed RIFLE classification is associated with increased hospital mortality and resource use. Patients with RIFLE class R are indeed at high risk of progression to class I or class F. Patients with RIFLE class I or class F incur a significantly increased length of stay and an increased risk of inhospital mortality compared with those who do not progress past class R or those who never develop acute kidney injury, even after adjusting for baseline severity of illness, case mix, race, gender and age.

Key messages

- The RIFLE classification is a very sensitive definition of acute kidney injury: acute kidney injury defined by the RIFLE classification occurred in two thirds of general ICU patients.
- RIFLE classes injury and failure are independently associated with increased risk for in-hospital dead.
- Patients who meet the very sensitive RIFLE "risk" criteria, are at significant risk for progression to injury or failure, and therefore in-hospital dead.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

EAJH designed the study, analyzed the raw data set, performed the statistical analysis and contributed to writing of the paper. GC set up the raw data set, helped to design the study and contributed to writing of the paper. AK helped analyze the raw data set and helped in the design of the study. RV and DCA helped to design the study and contributed to writing the paper. DDB helped with the statistical analysis. JAK designed the study, analyzed the data and contributed to writing the paper. All authors read, edited and ultimately approved the final manuscript.

Acknowledgements

Part of this work has been presented in abstract form at the 24th International Symposium on Intensive Care and Emergency Medicine, Brussels, Belgium, 2004. This study was conducted without external financial support.

References

1. Knaus WA, Wagner DP, Draper EA, Zimmerman JE, Bergner M, Bastos PG, Sirio CA, Murphy DJ, Lotring T, Damiano A, *et al.*: **The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults.** *Chest* 1991, **100:**1619-1636.
2. Vincent J-L, Moreno R, Takala J, Willatts S, de Mendonça A, Bruining H, Reinhart CK, Suter PeterM, Thijs LG: **The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine.** *Intensive Care Med* 1996, **22:**707-710.
3. Kellum JA, Levin N, Bouman C, Lameire N: **Developing a consensus classification system for acute renal failure.** *Curr Opin Crit Care* 2002, **8:**509-514.
4. Chertow GM, Levy EM, Hammermeister KM, Grover F, Daley J: **Independent association between acute renal failure and mortality following cardiac surgery.** *Am J Med* 1998, **104:**343-348.
5. de Mendonca A, Vincent JL, Suter PM, Moreno R, Dearden NM, Antonelli M, Takala J, Sprung C, Cantraine F: **Acute renal failure in the ICU: risk factors and outcome evaluated by the SOFA score.** *Intensive Care Med* 2000, **26:**915-921.

6. Vivino G, Antonelli M, Moro M, Cottini F, Conti G, Bufi M, Cannata F, Gasparetto A: **Risk factors for acute renal failure in trauma patients.** *Intensive Care Med* 1998, **24**:808-814.
7. Nash K, Hafeez A, Hou S: **Hospital-acquired renal insufficiency.** *Am J Kidney Dis* 2002, **39**:930-936.
8. Mora Mangano C, Diamondstone LS, Ramsay JG, Aggarwal A, Herskowitz A, Mangano DT, for the Multicenter Study of Perioperative Ischemia Research Group: **Renal dysfunction after myocardial revascularization: risk factors, adverse outcomes, and hospital resource utilization.** *Ann Intern Med* 1998, **128**:194-203.
9. Guerin C, Girard R, Selli JM, Perdrix JP, Ayzac L: **Initial versus delayed acute renal failure in the intensive care unit. A multicenter prospective epidemiological study. Rhone-Alpes Area Study Group on Acute Renal Failure.** *Am J Respir Crit Care Med* 2000, **161**:872-879.
10. Bellomo R, Kellum J, Ronco C: **Acute renal failure: time for consensus.** *Intensive Care Med* 2001, **27**:1685-1688.
11. Bellomo R, Ronco C, Kellum JA, Mehta RL, Palevsky P, the ADQI workgroup: **Acute renal failure – definition, outcome measures, animal models, fluid therapy and information technology needs: the Second International Consensus Conference of the Acute Dialysis Quality Initiative (ADQI) Group.** *Crit Care* 2004, **8**:R204-R212.
12. Kellum JA: **The Acute Dialysis Quality Initiative: methodology.** *Adv Ren Replace Ther* 2002, **9**:245-247.
13. Bell M, Liljestam E, Granath F, Fryckstedt J, Ekbom A, Martling CR: **Optimal follow-up time after continuous renal replacement therapy in actual renal failure patients stratified with the RIFLE criteria.** *Nephrol Dial Transplant* 2005, **20**:354-360.
14. Abosaif NY, Tolba YA, Heap M, Russell J, El Nahas AM: **The outcome of acute renal failure in the intensive care unit according to RIFLE: model application, sensitivity, and predictability.** *Am J Kidney Dis* 2005, **46**:1038-1048.
15. Kuitunen A, Vento A, Suojaranta-Ylinen R, Pettila V: **Acute renal failure after cardiac surgery: evaluation of the RIFLE classification.** *Ann Thorac Surg* 2006, **81**:542-546.
16. **K/DOQI clinical practice guidelines for chronic kidney disease: evaluation, classification, and stratification.** *Am J Kidney Dis* 2002, **39**:S1-266.
17. Bland JM, Altman DG: **Multiple significance tests: the Bonferroni method.** *BMJ* 1995, **310**:170.
18. Lin DY, Wei LJ, Ying Z: **Checking the Cox model with cumulative sums of Martingale-based residuals.** *Biometrika* 1993, **80**:557-572.
19. Metnitz PG, Krenn CG, Steltzer H, Lang T, Ploder J, Lenz K, Le Gall JR, Druml W: **Effect of acute renal failure requiring renal replacement therapy on outcome in critically ill patients.** *Crit Care Med* 2002, **30**:2051-2058.
20. Herget-Rosenthal S, Marggraf G, Husing J, Goring F, Pietruck F, Janssen O, Philipp T, Kribben A: **Early detection of acute renal failure by serum cystatin C.** *Kidney Int* 2004, **66**:1115-1122.
21. Hoste EA, Lameire NH, Vanholder RC, Benoit DD, Decruyenaere JM, Colardyn FA: **Acute renal failure in patients with sepsis in a surgical ICU: predictive factors, incidence, comorbidity, and outcome.** *J Am Soc Nephrol* 2003, **14**:1022-1030.
22. Lassnigg A, Schmidlin D, Mouhieddine M, Bachmann LM, Druml W, Bauer P, Hiesmayr M: **Minimal changes of serum creatinine predict prognosis in patients after cardiothoracic surgery: a prospective cohort study.** *J Am Soc Nephrol* 2004, **15**:1597-1605.
23. Chertow GM, Burdick E, Honour M, Bonventre JV, Bates DW: **Acute kidney injury, mortality, length of stay, and costs in hospitalized patients.** *J Am Soc Nephrol* 2005, **16**:3365-3370.
24. Klahr S, Miller SB: **Acute oliguria.** *N Engl J Med* 1998, **338**:671-675.
25. Hoste EA, Kellum JA: **ARF in the critically ill: impact on morbidity and mortality.** *Contrib Nephrol* 2004, **144**:1-11.
26. Rocktaeschel J, Morimatsu H, Uchino S, Goldsmith D, Poustie S, Story D, Gutteridge G, Bellomo R: **Acid-base status of critically ill patients with acute renal failure: analysis based on Stewart-Figge methodology.** *Crit Care* 2003, **7**:R60-R66.
27. Gunnerson KJ, Song M, Kellum JA: **Influence of acid-base balance in patients with sepsis.** In *Yearbook of Intensive Care and Emergency Medicine 2004* Edited by: Vincent JL. Berlin: Springer-Verlag; 2004:58-67.
28. Kellum JA, Song M, Li J: **Lactic and hydrochloric acids induce different patterns of inflammatory response in LPS-stimulated RAW 264.7 cells.** *Am J Physiol Regul Integr Comp Physiol* 2004, **286**:R686-R692.
29. Hoste EA, Blot SI, Lameire NH, Vanholder RC, De Bacquer D, Colardyn FA: **Effect of nosocomial bloodstream infection on the outcome of critically ill patients with acute renal failure treated with renal replacement therapy.** *J Am Soc Nephrol* 2004, **15**:454-462.
30. Woodrow G, Turney JH: **Cause of death in acute renal failure.** *Nephrol Dial Transplant* 1992, **7**:230-234.
31. Thakar CV, Yared JP, Worley S, Cotman K, Paganini EP: **Renal dysfunction and serious infections after open-heart surgery.** *Kidney Int* 2003, **64**:239-246.
32. Kramer AA, Postler G, Salhab KF, Mendez C, Carey LC, Rabb H: **Renal ischemia/reperfusion leads to macrophage-mediated increase in pulmonary vascular permeability.** *Kidney Int* 1999, **55**:2362-2367.
33. Donnahoo KK, Shames BD, Harken AH, Meldrum DR: **Review article: the role of tumor necrosis factor in renal ischemia-reperfusion injury.** *J Urol* 1999, **162**:196-203.
34. Ronco C, Bellomo R, Homel P, Brendolan A, Dan M, Piccinni P, La Greca G: **Effect of different doses in continuous veno-venous haemofiltration on outcomes of acute renal failure: a prospective randomised trial.** *Lancet* 2000, **356**:26-30.
35. Schiffl H, Lang SM, Fischer R: **Daily hemodialysis and the outcome of acute renal failure.** *N Engl J Med* 2002, **346**:305-310.
36. Phu NH, Hien TT, Mai NT, Chau TT, Chuong LV, Loc PP, Winearls C, Farrar J, White N, Day N: **Hemofiltration and peritoneal dialysis in infection-associated acute renal failure in Vietnam.** *N Engl J Med* 2002, **347**:895-902.

ANEXO 3: AKIN

Research

Open Access

Acute Kidney Injury Network: report of an initiative to improve outcomes in acute kidney injury

Ravindra L Mehta[1], John A Kellum[2], Sudhir V Shah[3], Bruce A Molitoris[4], Claudio Ronco[5], David G Warnock[6], Adeera Levin[7] and the Acute Kidney Injury Network

[1]Department of Medicine, University of California San Diego Medical Center 8342, 200 W. Arbor Drive, San Diego, CA 92103, USA
[2]Department of Critical Care Medicine, University of Pittsburgh, 3550 Terrace Street, 608 Scaife Hall, Pittsburgh, PA 15261, USA
[3]Division of Nephrology, UAMS College of Medicine, 4301 West Markham, Slot 501, Little Rock, AR 72205, USA
[4]Department of Medicine, Indiana University, Indianapolis, IN, USA
[5]Department of Nephrology Dialysis & Transplantation, San Bortolo Hospital, Vicenza, Italy
[6]Department of Medicine, University of Alabama, 1900 University Blvd, Birmingham, AL, USA
[7]Department of Medicine, University of British Columbia, St Pauls Hospital, 1160 Burrard St, Vancouver BC, V6Z1Y8, Canada

Corresponding author: Ravindra L Mehta, rmehta@ucsd.edu

Received: 8 Dec 2006 Revisions requested: 28 Dec 2006 Revisions received: 9 Feb 2007 Accepted: 1 Mar 2007 Published: 1 Mar 2007

Critical Care 2007, **11**:R31 (doi:10.1186/cc5713)
This article is online at: http://ccforum.com/content/11/2/R31

Acute Kidney Injury Network participants: Arvind Bagga, Aysin Bakkaloglu, Joseph V Bonventre, Emmanuel A Burdmann, Yipu Chen, Prasad Devarajan, Vince D'Intini, Geoff Dobb, Charles G Durbin Jr., Kai-Uwe Eckardt, Claude Guerin, Stefan Herget-Rosenthal, Eric Hoste, Michael Joannidis, John A Kellum, Ashok Kirpalani, Andrea Lassnigg, Jean-Roger Le Gall, Adeera Levin, Raul Lombardi, William Macias, Constantine Manthous, Ravindra L Mehta, Bruce A Molitoris, Claudio Ronco, Miet Schetz, Frederique Schortgen, Sudhir V Shah, Patrick SK Tan, Haiyan Wang, David G Warnock and Steve Webb.

Abstract

Introduction Acute kidney injury (AKI) is a complex disorder for which currently there is no accepted definition. Having a uniform standard for diagnosing and classifying AKI would enhance our ability to manage these patients. Future clinical and translational research in AKI will require collaborative networks of investigators drawn from various disciplines, dissemination of information via multidisciplinary joint conferences and publications, and improved translation of knowledge from preclinical research. We describe an initiative to develop uniform standards for defining and classifying AKI and to establish a forum for multidisciplinary interaction to improve care for patients with or at risk for AKI.

Methods Members representing key societies in critical care and nephrology along with additional experts in adult and pediatric AKI participated in a two day conference in Amsterdam, The Netherlands, in September 2005 and were assigned to one of three workgroups. Each group's discussions formed the basis for draft recommendations that were later refined and improved during discussion with the larger group. Dissenting opinions were also noted. The final draft recommendations were circulated to all participants and subsequently agreed upon as the consensus recommendations for this report. Participating societies endorsed the recommendations and agreed to help disseminate the results.

Results The term AKI is proposed to represent the entire spectrum of acute renal failure. Diagnostic criteria for AKI are proposed based on acute alterations in serum creatinine or urine output. A staging system for AKI which reflects quantitative changes in serum creatinine and urine output has been developed.

Conclusion We describe the formation of a multidisciplinary collaborative network focused on AKI. We have proposed uniform standards for diagnosing and classifying AKI which will need to be validated in future studies. The Acute Kidney Injury Network offers a mechanism for proceeding with efforts to improve patient outcomes.

ADQI = Acute Dialysis Quality Initiative; AKI = acute kidney injury; AKIN = Acute Kidney Injury Network; ARF = acute renal failure; ASN = American Society of Nephrology; CKD = chronic kidney disease; GFR = glomerular filtration rate; ISN = International Society of Nephrology; NKF = National Kidney Foundation; RIFLE = Risk, Injury, Failure, Loss, and End-stage kidney disease; RRT = renal replacement therapy.

Introduction

Acute renal failure (ARF) is a complex disorder that occurs in a variety of settings with clinical manifestations ranging from a minimal elevation in serum creatinine to anuric renal failure. It is often under-recognized and is associated with severe consequences [1-4]. Recent epidemiological studies demonstrate the wide variation in etiologies and risk factors [1,5-7], describe the increased mortality associated with this disease (particularly when dialysis is required) [1,4,6,8,9], and suggest a relationship to the subsequent development of chronic kidney disease (CKD) and progression to dialysis dependency [1,4,8,10-12]. Emerging evidence suggests that even minor changes in serum creatinine are associated with increased inpatient mortality [13-20]. ARF has been the focus of extensive clinical and basic research efforts over the last decades. The lack of a universally recognized definition of ARF has posed a significant limitation. Despite the significant progress made in understanding the biology and mechanism of ARF in animal models, translation of this knowledge into improved management and outcomes for patients has been limited.

During the last five years, several groups have recognized these limitations and have worked to identify the knowledge gaps and define the necessary steps to correct these deficiencies. These efforts have included consensus conferences and publications from the Acute Dialysis Quality Initiative (ADQI) group [19,21-25], the American Society of Nephrology (ASN) ARF Advisory group [26], the International Society of Nephrology (ISN), and the National Kidney Foundation (NKF) and KDIGO (Kidney Disease: Improving Global Outcomes) groups [27]. Additionally, the critical care societies have developed formal intersociety collaborations such as the International Consensus Conferences in Critical Care [28]. Recognizing that future clinical and translational research in ARF will require multidisciplinary collaborative networks, the ADQI group and representatives from three nephrology societies (ASN, ISN, and NKF) and the European Society of Intensive Care Medicine met in Vicenza, Italy, in September 2004. They proposed the term acute kidney injury (AKI) to reflect the entire spectrum of ARF, recognizing that an acute decline in kidney function is often secondary to an injury that causes functional or structural changes in the kidneys. The group established the Acute Kidney Injury Network (AKIN) as an independent collaborative network comprised of experts selected by the participating societies to represent both their area of expertise and their sponsoring organization. AKIN is intended to facilitate international, interdisciplinary, and intersocietal collaborations to ensure progress in the field of AKI and obtain the best outcomes for patients with or at risk for AKI.

This report describes an interim definition and staging system for AKI and a plan for further activities of the collaborative network which were developed at the first AKIN conference held in Amsterdam, The Netherlands, in September 2005.

Materials and methods

Representatives of the major critical care and nephrology societies and associations and invited content experts were assigned to workgroups to consider three topics: (a) the development of uniform standards for definition and classification of AKI, (b) joint conference topics, and (c) the interdisciplinary collaborative research network. Each workgroup had an assigned chair and co-chair to facilitate the discussion and develop summary recommendations of the workgroup. The draft recommendations were then refined and improved during discussion with the larger group. Key points and issues were noted and then discussed a second time if no resolution was reached initially. When a majority view was not evident or when the area was felt to be of extreme importance, votes were tallied. Dissenting opinions were also noted. The final recommendations were circulated to all participants and subsequently agreed upon as the consensus recommendations for this report. After an iterative process of revisions, the final manuscript was presented to each of the respective societies for endorsement. Societies were asked to facilitate dissemination of the findings to their membership through presentations in society conferences and publication of summary reports in society journals, Web sites, and other forms of communication.

Results

1. Proposal for uniform standards for definition and classification of AKI

Definition and diagnostic criteria of AKI

For any condition, the clinician needs to know whether the disease is present and, if so, where and when the patient falls in the natural history of the disease. The former facilitates recognition whereas the latter defines time points for intervention. Unfortunately, there has been no uniformly accepted definition of AKI. Studies describe ARF or AKI based on serum creatinine changes, absolute levels of serum creatinine, changes in blood urea nitrogen or urine output, or the need for dialysis [1,11,20,29-36]. The wide variation in definitions has made it difficult to compare information across studies and populations [37].

Diagnostic criteria

Recognition of AKI requires the delineation of easily measured criteria that can be widely applied. Serum creatinine levels and changes in urine output are the most commonly applied measures of renal function; however, they are each influenced by factors other than the glomerular filtration rate (GFR) and do not provide any information about the nature or site of kidney injury. The proposed diagnostic criteria (Table 1) were based on consideration of the following concepts:

1. The definition needs to be broad enough to accommodate variations in clinical presentation over age groups, locations, and clinical situations.

Table 1

Diagnostic criteria for acute kidney injury

An abrupt (within 48 hours) reduction in kidney function currently defined as an absolute increase in serum creatinine of more than or equal to 0.3 mg/dl (≥ 26.4 μmol/l), a percentage increase in serum creatinine of more than or equal to 50% (1.5-fold from baseline), or a reduction in urine output (documented oliguria of less than 0.5 ml/kg per hour for more than six hours).
The above criteria include both an absolute and a percentage change in creatinine to accommodate variations related to age, gender, and body mass index and to reduce the need for a baseline creatinine but do require at least two creatinine values within 48 hours. The urine output criterion was included based on the predictive importance of this measure but with the awareness that urine outputs may not be measured routinely in non-intensive care unit settings. It is assumed that the diagnosis based on the urine output criterion alone will require exclusion of urinary tract obstructions that reduce urine output or of other easily reversible causes of reduced urine output. The above criteria should be used in the context of the clinical presentation and following adequate fluid resuscitation when applicable. Note: Many acute kidney diseases exist, and some (but not all) of them may result in acute kidney injury (AKI). Because diagnostic criteria are not documented, some cases of AKI may not be diagnosed. Furthermore, AKI may be superimposed on or lead to chronic kidney disease.

2. Sensitive and specific markers for kidney injury are not currently available in clinical practice. Several groups are working on developing and validating biomarkers of kidney injury and GFR which may be used in the future for diagnosis and prognosis.

3. There is accumulating evidence that small increments in serum creatinine are associated, in a variety of settings, with adverse outcomes [13-20] that are manifest in short-term morbidity and mortality and in longer-term outcomes, including 1-year mortality [15-17]. Current clinical practice does not focus much attention on small increments in serum creatinine, which are often attributed to lab variations. However, the coefficient of variation of serum creatinine with modern analyzers is relatively small and therefore increments of 0.3 mg/dl (25 μmol/l) are unlikely to be due to assay variation [38]. Changes in volume status can influence serum creatinine levels [39]. Because the amount of fluid resuscitation depends on the underlying clinical situation [40], the group agreed that application of the diagnostic criteria would be used only after an optimal state of hydration had been achieved.

4. A time constraint of 48 hours for diagnosis was selected based on the evidence that adverse outcomes with small changes in creatinine were observed when the creatinine elevation occurred within 24 to 48 hours [15,16] and to ensure that the process was acute and representative of events within a clinically relevant time period. In the two aforementioned studies, there was no distinction of underlying CKD or *de novo* AKI. However, in the study by Chertow and colleagues [13], the odds ratio for mortality with a change in creatinine of 0.3 mg/dl (25 μmol/l) was 4.1 (confidence interval 3.1 to 5.5) adjusting for CKD. There is no requirement to wait 48 hours to diagnose AKI or initiate appropriate measures to treat AKI. Instead, the time period is designed to eliminate situations in which the increase in serum creatinine by 0.3 is very slow and thus is not 'acute.'

5. It was recognized that AKI is often superimposed on pre-existing CKD. Further validation will be required to determine whether the criterion of a creatinine elevation of 0.3 mg/dl (25 μmol/l) is applicable to these patients (that is, whether a creatinine increase of more than 0.3 mg/dl from an elevated baseline represents AKI and has the same risks as a creatinine increase from a normal baseline).

6. The need for including urine output as a diagnostic criterion is based on the knowledge of critically ill patients in whom this parameter often heralds renal dysfunction before serum creatinine increases.

A minority of group members, both intensivists and nephrologists, felt that a urine output reduction of less than 0.5 ml/kg per hour over the span of six hours was not specific enough to lead confidently to the designation of AKI. It was recognized that the hydration state, use of diuretics, and presence of obstruction could influence the urine volume, hence the need to consider the clinical context. Additionally, accurate measurements of urine output may not be easily available in all cases, particularly in patients in non-intensive care unit settings. Despite these limitations, it was felt that the use of changes in urine offers a sensitive and easily discernible means of identifying patients, but its value as an independent criterion for diagnosis of AKI will need to be validated.

The proposed diagnostic criteria for AKI are designed to facilitate acquisition of new knowledge and validate the emerging concept that small alterations in kidney function may contribute to adverse outcomes. The goal of adopting these explicit diagnostic criteria is to increase the clinical awareness and diagnosis of AKI. It is recognized that there may be an increase in false-positives, so that some patients labeled with AKI will not have the condition. There was consensus that adopting the more inclusive criteria is preferable to the current situation, in which the condition is under-recognized and many people are identified late in the course of their illness and potentially miss the opportunity for prevention or application of strategies to minimize further kidney damage.

Staging/classification

The goal of a staging system is to classify the course of a disease in a reproducible manner that supports accurate identification and prognostication and informs diagnostic or therapeutic interventions. The group recognized that a number of systems for staging and classifying AKI are currently in use or have been proposed [41]. The RIFLE (Risk, Injury, Failure,

Table 2

Classification/staging system for acute kidney injury[a]

Stage	Serum creatinine criteria	Urine output criteria
1	Increase in serum creatinine of more than or equal to 0.3 mg/dl (≥ 26.4 μmol/l) or increase to more than or equal to 150% to 200% (1.5- to 2-fold) from baseline	Less than 0.5 ml/kg per hour for more than 6 hours
2[b]	Increase in serum creatinine to more than 200% to 300% (> 2- to 3-fold) from baseline	Less than 0.5 ml/kg per hour for more than 12 hours
3[c]	Increase in serum creatinine to more than 300% (> 3-fold) from baseline (or serum creatinine of more than or equal to 4.0 mg/dl [≥ 354 μmol/l] with an acute increase of at least 0.5 mg/dl [44 μmol/l])	Less than 0.3 ml/kg per hour for 24 hours or anuria for 12 hours

[a]Modified from RIFLE (Risk, Injury, Failure, Loss, and End-stage kidney disease) criteria [26]. The staging system proposed is a highly sensitive interim staging system and is based on recent data indicating that a small change in serum creatinine influences outcome. Only one criterion (creatinine or urine output) has to be fulfilled to qualify for a stage. [b]200% to 300% increase = 2- to 3-fold increase. [c]Given wide variation in indications and timing of initiation of renal replacement therapy (RRT), individuals who receive RRT are considered to have met the criteria for stage 3 irrespective of the stage they are in at the time of RRT.

Loss, and End-stage kidney disease) criteria [25] proposed by the ADQI group were developed by an interdisciplinary, international consensus process and are now being validated by different groups worldwide [36,37]. However, according to data that have emerged since then, smaller changes in serum creatinine than those considered in the RIFLE criteria might be associated with adverse outcomes [13-18]. Additionally, given the consensus definition for AKI (Table 1), RIFLE criteria have been modified so that patients meeting the definition for AKI could be staged (Table 2). The proposed staging system retains the emphasis on changes in serum creatinine and urine output but includes the following principles:

1. Although diagnosis of AKI is based on changes over the course of 48 hours, staging occurs over a slightly longer time frame. One week was proposed by the ADQI group in the original RIFLE criteria [25].

2. There was a conscious decision not to include the therapy for AKI (that is, renal replacement therapy [RRT]) as a distinct stage because this constitutes an outcome of AKI.

3. The new staging system maps to the RIFLE stages as follows:

3a. RIFLE 'Risk' category should have the same criteria as for

Table 3

Potential topics identified for future consensus conferences

	Subject	Topics
1.	Epidemiology of AKI	What is a 'nomenclature' that is based on simple, universally available data and that can identify all patients globally with AKI irrespective of location and age? What are the data to help determine etiology once AKI is identified? What are the correlates of AKI in regard to pathology/physiology? Is there a validated method for assessing severity of AKI separate from multiple organ failure? What is the relationship between degree of severity and outcomes?
2.	Outcomes from AKI	What are the clinically meaningful outcomes that are important in clinical studies of AKI?
3.	Strategies to change outcomes	Prevention Treatment Non-dialytic Dialysis Timing of initiation Modality selection (CRRT, IHD, PD) Intensity of therapy (dose) Cessation of renal replacement therapy
4.	Data needed to advance knowledge in AKI	Datasets collected at contact with health care system Intensive care unit admission Biological sample repositories
5	Process outcomes	Measures of effectiveness of current processes for changing behavior/attitude of caregivers and ultimately patient outcomes from AKI.

AKI, acute kidney injury; CRRT, continuous renal replacement therapy; IHD, intermittent hemodialysis; PD, peritoneal dialysis.

Table 4

Recommendations for establishing a collaborative network for acute kidney injury (AKI) research

	Component	Principles and approach
1.	Identify the key roles of the participating groups	a. The collaborative effort should be inclusive and open to all interested societies/ organizations. b. Participation in the collaborative organization will require commitment of time, expertise, and/or resources as appropriate to the specific initiative and in accordance with the means of the organization/group. c. An organizational structure will be required to coordinate the activities. d. Work products from the collaborative effort will require a mechanism for recognizing the contributions of each group.
2.	Scope of collaborations	a. Identify topics in AKI areas of mutual interest and of wide application. b. Develop consensus statements for best practice where there is limited or no evidence and where, due to accepted practices, it will be difficult to get evidence. c. Develop tools to standardize the management of AKI. d. Develop evidence through clinical research where feasible. e. Develop practice recommendations/guidelines. f. Implement guidelines.
3.	Define infrastructure needs	a. Identify key components needed (for example, database, protocols for Web-based information transfer). b. Establish the requirements for sharing information with regulatory agencies. c. Define training needs for developing researchers and the resources that are required and define what hurdles will need to be overcome. d. International collaboration will require identification of peer-reviewed, public, and commercial sources of financial support. e. Develop an inventory of current collaborative efforts and establish relationships with these existing networks.
4.	Identify common unifying principles that would form the basis of ongoing collaboration	a. Establish protocols for consistent data entry that allows benchmarking of participating units. b. Identify questions that interest the majority of the participants. c. Initiate a short-term collaborative project to validate proposed AKI definition as an initial project.

the diagnosis of stage 1 AKI.

3b. Those who are classified as having 'Injury' and 'Failure' categories map to stages 2 and 3 of AKI.

3c. The 'Loss' and 'End-stage kidney disease' categories were removed from the staging system and remain outcomes.

3d. Given the variability inherent in commencing RRT and due to variability in resources in different populations and countries, patients receiving RRT are to be included in stage 3 (analogous to stage 5 CKD, GFR of less than 15, or dialysis).

2. Future joint conference topics and key collaborative research questions

There is a need to ensure that collaborative and integrated joint conferences are planned to facilitate the dissemination of knowledge, clarify clinical practice, and enhance research. Many organizations are currently in the process of planning meetings on ARF/AKI. These meeting take various forms: knowledge exchange/scientific meetings, consensus controversies, and research initiatives. The group described five key topics that should be addressed by any of the professional communities involved in the care of patients with AKI. The particular venue and the process and products of these conferences were not discussed in detail. An overview of the topics and issues that would be well served by a multidisciplinary consensus or controversies conference is presented in Table 3. These topics reflect important areas in which there is a need for ongoing research to develop evidence. A key step for future conferences will be to determine which research questions are most important and pressing to advance the field and improve outcomes from AKI.

3. Need for an international collaborative network

AKI is a global problem with varying etiologies and manifestations, but the outcomes are similar [1-4,6]. Given the wide global variation in the natural history and management of AKI, it is essential that mechanisms for sharing information and for collaboration among centers be developed. It was felt that the establishment of an international collaborative research effort for AKI would contribute to international research and education about AKI. The group proposed four major topics that would need to be addressed by this initiative (Table 4).

Conclusion

AKI is a complex disorder for which there is no currently accepted uniform definition. Having a standard for diagnosing and classifying AKI would enhance our ability to improve the management of these patients. We have described the formation of a multidisciplinary collaborative network focused on AKI and have proposed uniform standards for diagnosing and classifying AKI. The proposed standards will need to be validated in future studies. These standards build upon existing knowl-

Table 5

Acute Kidney Injury Network summit meeting participants and workgroups

Name	Representation	Joint conference	Interdisciplinary collaborative research network	Interim proposals for terminology, diagnosis, classification, and staging
Miet Schetz	Acute Dialysis Quality Initiative		X	
Sudhir V Shah	ASN	X (co-chair)		
Bruce A Molitoris	ASN		X	
Aysin Bakkaloglu	IPNA	X		
Arvind Bagga	IPNA		X	
Prasad Devarajan	American Society of Pediatric Nephrologists			X
Raul Lombardi	SLANH		X	
Emmanuel A Burdmann	SLANH	X		
Kai-Uwe Eckardt	European Dialysis and Transplant Association-European Renal Association		X (co-chair)	
Claudio Ronco	International Society of Nephrology	X		
Ravindra L Mehta	International Society of Nephrology			X (co-chair)
Adeera Levin	NKF	X		
David G Warnock	NKF	X		
Ashok Kirpalani	Indian Society of Nephrology			X
Haiyan Wang	CSN			X
Yipu Chen	CSN	X		
Vince D'Intini	Asian Pacific Society of Nephrology		X	
Michael Joannidis	European Society of Intensive Care Medicine		X	
Charles G Durbin Jr.	Society of Critical Care Medicine		X (co-chair)	
Patrick SK Tan	Asia Pacific Association of Critical Care Medicine		X	
Constantine Manthous	American Thoracic Society	X (co-chair)		
Claude Guerin	French Society			X
Frederique Schortgen	French Society		X	
John A Kellum	American College of Chest Physicians			X (co-chair)
Steve Webb	ANZICS	X		
Geoff Dobb	ANZICS		X	
Jean-Roger Le Gall	Expert			X
Eric Hoste	Expert		X	
Andrea Lassnigg	Expert			X
William Macias	Expert			X
Stefan Herget-Rosenthal	Expert			X
Joseph V Bonventre	Expert			X

ANZICS, Australian and New Zealand Intensive Care Society; ASN, American Society of Nephrology; CSN, Chinese Society of Nephrology; IPNA, International Pediatric Nephrology Association; NKF, National Kidney Foundation; SLANH, Sociedade Latino-Americana de Nefrologia e Hipertensão.

edge and permit individuals using current staging systems (for example, RIFLE) to transition to the new system without loss of comparability. These recommendations have been endorsed by the participating societies, which represent the majority of the critical care and nephrology societies worldwide and which have been asked to disseminate the results via presentations at the national and regional society conferences and through publication of summary reports in society journals (see Table 5 for society endorsement details). We believe that these recommendations provide a stepping stone to standard-

izing the care of patients with AKI and will greatly enhance our ability to design prospective studies to evaluate potential prevention and treatment strategies. One of the limitations of consensus recommendations is that they are often not adopted. We anticipate that the broad support and commitment obtained through society involvement will significantly enhance the ability to disseminate the results to the worldwide community and to address this limitation. Future clinical and translational research in AKI will require the development of collaborative networks of investigators drawn from various disciplines to facilitate the acquisition of evidence through well-designed and well-conducted clinical trials, dissemination of information via multidisciplinary joint conferences and publications, and improvement of the translation of knowledge from pre-clinical research. We anticipate that the AKIN will provide an effective mechanism for facilitating efforts to improve patient outcomes.

Key messages

- AKI is a complex disorder, and we have proposed uniform standards for diagnosing and classifying AKI on the basis of existing systems (that is, RIFLE). These proposals will require validation.
- Our recommendations have been endorsed by participating societies that represent the majority of critical care and nephrology societies worldwide.
- These recommendations provide a stepping stone to standardizing the care of patients with AKI and will greatly enhance our ability to design prospective studies to evaluate potential prevention and treatment strategies.
- Future clinical and translational research in AKI will require the development of collaborative networks. The AKIN was formed to provide an effective mechanism for facilitating such efforts.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

All authors participated in the AKIN conference workgroups, development of the summary statement, and review of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We would like to recognize the financial support provided by the ASN, ISN, and NKF for the AKIN conference. We are grateful for the logistical support provided by the NKF and the special efforts of Sue Levey for the meeting arrangements. Members from participating societies were supported by their respective societies (see Table 5).

Society and organization endorsements

Acute Dialysis Quality Initiative

Nephrology

American Society of Nephrology, American Society of Pediatric Nephrologists, Asian Pacific Society of nephrology, Chinese Society of Nephrology, European Dialysis and Transplant Association-European Renal Association, Indian Society of Nephrology, International Pediatric Nephrology Association, International Society of Nephrology, National Kidney Foundation, and Sociedade Latino-Americana de Nefrologia e Hipertensão.

Critical care

American College of Chest Physicians, American Thoracic Society, Asia Pacific Association of Critical Care Medicine, Australian and New Zealand Intensive Care Society, European Society of Intensive Care Medicine, Société de Réanimation de Langue Française, and Society of Critical Care Medicine.

References

1. Mehta RL, Pascual MT, Soroko S, Savage BR, Himmelfarb J, Ikizler TA, Paganini EP, Chertow GM: **Spectrum of acute renal failure in the intensive care unit: the PICARD experience.** *Kidney Int* 2004, **66**:1613-1621.
2. Palevsky PM: **Epidemiology of acute renal failure: the tip of the iceberg.** *Clin J Am Soc Nephrol* 2006, **1**:6-7.
3. Ympa YP, Sakr Y, Reinhart K, Vincent JL: **Has mortality from acute renal failure decreased? A systematic review of the literature.** *Am J Med* 2005, **118**:827-832.
4. Metnitz PG, Krenn CG, Steltzer H, Lang T, Ploder J, Lenz K, Le Gall JR, Druml W: **Effect of acute renal failure requiring renal replacement therapy on outcome in critically ill patients.** *Crit Care Med* 2002, **30**:2051-2058.
5. Waikar SS, Curhan GC, Wald R, McCarthy EP, Chertow GM: **Declining mortality in patients with acute renal failure, 1988 to 2002.** *J Am Soc Nephrol* 2006, **17**:1143-1150.
6. Uchino S, Kellum JA, Bellomo R, Doig GS, Morimatsu H, Morgera S, Schetz M, Tan I, Bouman C, Macedo E, *et al.*: **Acute renal failure in critically ill patients: a multinational, multicenter study.** *JAMA* 2005, **294**:813-818.
7. Liangos O, Wald R, O'Bell JW, Price L, Pereira BJ, Jaber BL: **Epidemiology and outcomes of acute renal failure in hospitalized patients: a national survey.** *Clin J Am Soc Nephrol* 2006, **1**:43-51.
8. Clermont G, Acker CG, Angus DC, Sirio CA, Pinsky MR, Johnson JP: **Renal failure in the ICU: comparison of the impact of acute renal failure and end-stage renal disease on ICU outcomes.** *Kidney Int* 2002, **62**:986-996.
9. Thakar CV, Worley S, Arrigain S, Yared JP, Paganini EP: **Influence of renal dysfunction on mortality after cardiac surgery: modifying effect of preoperative renal function.** *Kidney Int* 2005, **67**:1112-1119.
10. Druml W: **Long term prognosis of patients with acute renal failure: is intensive care worth it?** *Intensive Care Med* 2005, **31**:1145-1147.
11. Liano F, Junco E, Pascual J, Madero R, Verde E: **The spectrum of acute renal failure in the intensive care unit compared with that seen in other settings. The Madrid Acute Renal Failure Study Group.** *Kidney Int Suppl* 1998, **66**:S16-24.
12. Mehta RL, Pascual MT, Soroko S, Chertow GM: **Diuretics, mortality, and nonrecovery of renal function in acute renal failure.** *JAMA* 2002, **288**:2547-2553.
13. Chertow GM, Burdick E, Honour M, Bonventre JV, Bates DW: **Acute kidney injury, mortality, length of stay, and costs in hospitalized patients.** *J Am Soc Nephrol* 2005, **16**:3365-3370.
14. Gruberg L, Mintz GS, Mehran R, Gangas G, Lansky AJ, Kent KM, Pichard AD, Satler LF, Leon MB: **The prognostic implications of further renal function deterioration within 48 h of interventional coronary procedures in patients with pre-existent chronic renal insufficiency.** *J Am Coll Cardiol* 2000, **36**:1542-1548.

15. Lassnigg A, Schmidlin D, Mouhieddine M, Bachmann LM, Druml W, Bauer P, Hiesmayr M: **Minimal changes of serum creatinine predict prognosis in patients after cardiothoracic surgery: a prospective cohort study.** *J Am Soc Nephrol* 2004, **15**:1597-1605.
16. Levy MM, Macias WL, Vincent JL, Russell JA, Silva E, Trzaskoma B, Williams MD: **Early changes in organ function predict eventual survival in severe sepsis.** *Crit Care Med* 2005, **33**:2194-2201.
17. McCullough PA, Soman SS: **Contrast-induced nephropathy.** *Crit Care Clin* 2005, **21**:261-280.
18. Praught ML, Shlipak MG: **Are small changes in serum creatinine an important risk factor?** *Curr Opin Nephrol Hypertens* 2005, **14**:265-270.
19. Bellomo R, Ronco C, Kellum JA, Mehta RL, Palevsky P: **Acute renal failure – definition, outcome measures, animal models, fluid therapy and information technology needs: the Second International Consensus Conference of the Acute Dialysis Quality Initiative (ADQI) Group.** *Crit Care* 2004, **8**:R204-212.
20. Hoste EAJ, Clermont G, Kersten A, Venkataraman R, Angus DC, De Bacquer D, Kellum JA: **RIFLE criteria for acute kidney injury is associated with hospital mortality in critically ill patients: a cohort analysis.** *Crit Care* 2006, **10**:R73-82.
21. Palevsky PM, Metnitz PG, Piccinni P, Vinsonneau C: **Selection of endpoints for clinical trials of acute renal failure in critically ill patients.** *Curr Opin Crit Care* 2002, **8**:515-518.
22. Kellum JA, Leblanc M, Gibney RT, Tumlin J, Lieberthal W, Ronco C: **Primary prevention of acute renal failure in the critically ill.** *Curr Opin Crit Care* 2005, **11**:537-541.
23. Kellum JA, Ronco C, Mehta R, Bellomo R: **Consensus development in acute renal failure: The Acute Dialysis Quality Initiative.** *Curr Opin Crit Care* 2005, **11**:527-532.
24. Leblanc M, Kellum JA, Gibney RT, Lieberthal W, Tumlin J, Mehta R: **Risk factors for acute renal failure: inherent and modifiable risks.** *Curr Opin Crit Care* 2005, **11**:533-536.
25. Kellum JA, Mehta RL, Angus DC, Palevsky P, Ronco C: **The first international consensus conference on continuous renal replacement therapy.** *Kidney Int* 2002, **62**:1855-1863.
26. **American Society of Nephrology Renal Research Report.** *J Am Soc Nephrol* 2005, **16**:1886-1903.
27. Eknoyan G, Lameire N, Barsoum R, Eckardt KU, Levin A, Levin N, Locatelli F, MacLeod A, Vanholder R, Walker R, *et al.*: **The burden of kidney disease: improving global outcomes.** *Kidney Int* 2004, **66**:1310-1314.
28. Thompson BT, Cox PN, Antonelli M, Carlet JM, Cassell J, Hill NS, Hinds CJ, Pimentel JM, Reinhart K, Thijs LG: **Challenges in end-of-life care in the ICU: statement of the 5th International Consensus Conference in Critical Care: Brussels, Belgium, April 2003: executive summary.** *Crit Care Med* 2004, **32**:1781-1784.
29. Lameire N, Van Biesen W, Vanholder R: **Acute renal failure.** *Lancet* 2005, **365**:417-430.
30. Brivet FG, Kleinknecht DJ, Loirat P, Landais PJ: **Acute renal failure in intensive care units – causes, outcome, and prognostic factors of hospital mortality; a prospective, multicenter study. French Study Group on Acute Renal Failure.** *Crit Care Med* 1996, **24**:192-198.
31. Chertow GM, Lazarus JM, Christiansen CL, Cook EF, Hammermeister KE, Grover F, Daley J: **Preoperative renal risk stratification.** *Circulation* 1997, **95**:878-884.
32. Mehta RL, McDonald B, Gabbai F, Pahl M, Farkas A, Pascual MT, Zhuang S, Kaplan RM, Chertow GM: **Nephrology consultation in acute renal failure: does timing matter?** *Am J Med* 2002, **113**:456-461.
33. Mehta RL, Pascual MT, Gruta CG, Zhuang S, Chertow GM: **Refining predictive models in critically ill patients with acute renal failure.** *J Am Soc Nephrol* 2002, **13**:1350-1357.
34. Vincent JL: **Incidence of acute renal failure in the intensive care unit.** *Contrib Nephrol* 2001:1-6.
35. Hoste EA, Lameire NH, Vanholder RC, Benoit DD, Decruyenaere JM, Colardyn FA: **Acute renal failure in patients with sepsis in a surgical ICU: predictive factors, incidence, comorbidity, and outcome.** *J Am Soc Nephrol* 2003, **14**:1022-1030.
36. Uchino S, Bellomo R, Goldsmith D, Bates S, Ronco C: **An assessment of the RIFLE criteria for acute renal failure in hospitalized patients.** *Crit Care Med* 2006, **34**:1913-1917.
37. Bellomo R, Kellum JA, Ronco C: **Defining acute renal failure: physiological principles.** *Intensive Care Med* 2004, **30**:33-37.
38. Perrone RD, Madias NE, Levey AS: **Serum creatinine as an index of renal function: new insights into old concepts.** *Clin Chem* 1992, **38**:1933-1953.
39. Moran SM, Myers BD: **Course of acute renal failure studied by a model of creatinine kinetics.** *Kidney Int* 1985, **27**:928-937.
40. Rivers EP: **Early goal-directed therapy in severe sepsis and septic shock: converting science to reality.** *Chest* 2006, **129**:217-218.
41. Mehta RL, Chertow GM: **Acute renal failure definitions and classification: time for change?** *J Am Soc Nephrol* 2003, **14**:2178-2187.

ANEXO 4: QUADAS

Table 1: Questionnaire for evaluation of QUADAS

a) Review details
b) Content of the tool:
• Did QUADAS cover all important items?
• Were any QUADAS items omitted, added or modified?
c) Background document:
• Was the background document easy to understand?
• Were scoring instructions understandable?
• Should any items have been scored differently?
d) Technical points
• How long did it take to complete QUADAS?
• Was inter-rater reliability assessed?
e) Overall conclusions
• Reviewers were asked to rate coverage, ease of use, clarity of instructions, and validity (whether QUADAS helped to distinguish between studies of different qualities) on a five point scale
• Would you us QUADAS again?
f) Additional questions
• How were the results of quality incorporated into the review?
• Was a training session organised to ensure reviewers applied the tool consistently?
• Reviewer details, including age, experience, professional background
• Have you previously been involved in the quality assessment of studies included in a systematic review?
g) Final comments

evidence and a Delphi procedure involving a panel of experts in diagnostic research [1]. Like all quality assessment tools, QUADAS is a measurement, implying that its characteristics have to be evaluated: does it measure what it aims to measure, how well does it do this, and are results reproducible between different observers [2]? The objective of this study was to evaluate QUADAS by determining agreement between reviewers and the consensus rating and variability among raters, and gathering feedback on reviewers' experiences of using QUADAS.

Methods

Assessment of the consistency and reliability of QUADAS

Three reviewers were asked to use QUADAS to independently rate the quality of 30 studies as part of a systematic review on the diagnosis of peripheral arterial disease. One QUADAS item, the use of an appropriate reference standard, was not assessed as studies were only included in the review if they used a specified reference standard.

The three reviewers had different backgrounds and levels of experience. Reviewer 1 had previously carried out several diagnostic systematic reviews and had used QUADAS; she also had a background in primary diagnostics. Reviewer 2 was a new reviewer – this was the first review that she had worked on, but she had previously worked in primary diagnostics. Reviewer 3 was an experienced reviewer who had worked on a number of systematic reviews. This combination of reviewers with was chosen to reflect the spectrum of likely QUADAS users.

A limited amount of information specific to the diagnosis of peripheral arterial disease was provided to help with the scoring of QUADAS, this applied to items 1 (spectrum composition), 4 (disease progression bias), and 12 (availability of clinical information). For all other items, the guidelines on scoring provided in the QUADAS background document were briefly summarised [3]. Although reviewers did have access to the background document they were not specifically requested to read this or use it when assessing study quality.

Our main interest was in the amount of agreement between the rating of each reviewer and the consensus rating, calculated as the proportion of studies for which each reviewer agreed with the consensus rating. In addition, we also examined inter-observer variability by calculating the kappa statistic. Both analyses were carried out for all QUADAS items combined and for each individual item. We chose to focus on the proportion of agreements between reviewers and the final consensus, as kappas can be misleading in certain circumstances [4].

Piloting QUADAS in ongoing reviews

Reviewers who had used QUADAS in their reviews completed a short structured questionnaire asking how they used QUADAS and what their opinions of its usefulness were. Details of the questionnaire are provided in Table 1. A narrative synthesis was used to summarise results.

Results

Assessment of the consistency and reliability of QUADAS

Table 2 summarises the agreement between reviewers. Agreement between reviewers 1 and 2 and the final consensus rating was very good at 91 and 90%, and was

Table 2: Overall agreement between reviewers and agreement with consensus for each of the QUADAS items and for all items combined

QUADAS item		Agreement with consensus diagnosis (%) (95% confidence interval)			Reviewer variability (κ) (95% confidence interval)
		1	2	3	
All items combined		91 (88–94)	90 (86–93)	85 (81–89)	0.66 (0.63 to 0.67)
1	Was the spectrum of patients representative of the patients who will receive the test in practice? (spectrum composition)*	90 (73–98)	87 (69–96)	83 (65–94)	0.73 (0.60 to 0.76)
2	Were selection criteria clearly described? (selection criteria)	90 (73–98)	83 (65–94)	73 (54–88)	0.55 (0.33 to 0.61)
3	*Is the reference standard likely to correctly classify the target condition? (reference standard)**				
4	Is the time period between reference standard and index test short enough to be reasonably sure that the target condition did not change between the two tests? (disease progression bias)*	87 (69–96)	90 (73–98)	83 (65–94)	0.68 (0.63 to 0.86)
5	Did the whole sample or a random selection of the sample, receive verification using a reference standard of diagnosis? (partial verification)	87 (69–96)	90 (73–98)	93 (78–99)	0.27(-0.06 to 0.39)
6	Did patients receive the same reference standard regardless of the index test result? (differential verification)	97 (83–100)	97 (83–100)	97 (83–100)	0.31 (-0.01 to 0.46)
7	Was the reference standard independent of the index test (i.e. the index test did not form part of the reference standard)? (incorporation bias)	100 (88–100)	100 (88–100)	93 (78–99)	-0.02 (-0.03 to -0.01)
8	Was the execution of the index test described in sufficient detail to permit replication of the test? (index test execution)	97 (83–100)	100 (88–100)	87 (69–96)	0.60 (0.33 to 0.73)
9	Was the execution of the reference standard described in sufficient detail to permit its replication? (reference standard execution)	93 (78–99)	93 (78–99)	93 (78–99)	0.81 (0.60 to 0.87)
10	Were the index test results interpreted without knowledge of the results of the reference standard? (test review bias)	90 (73–98)	87 (69–96)	97 (83–100)	0.55 (-0.04 to 0.75)
11	Were the reference standard results interpreted without knowledge of the results of the index test? (reference standard review bias)	93 (78–99)	93 (78–99)	93 (78–99)	0.68 (0.46 to 0.76)
12	Were the same clinical data available when test results were interpreted as would be available when the test is used in practice? (clinical review bias)*	90 (73–98)	93 (78–99)	50 (31–69)	0.18 (-0.13 to 0.36)
13	Were uninterpretable/ intermediate test results reported? (uninterpretable test results)	83 (65–94)	70 (50–85)	87 (69–96)	0.32 (0.18 to 0.44)
14	Were withdrawals from the study explained? (withdrawals)	90 (73–98)	83 (65–94)	80 (61–92)	0.38 (0.33 to 0.51)

* Items for which review specific details were added to QUADAS
*The item relating to reference standard was not assessed for this review

slightly lower (85%) for reviewer 3. Overall reviewer variability was good [5] with a kappa of 0.65.

Agreement between reviewers and the final consensus rating was over 80% for all but four items: selection criteria, availability of clinical information, uninterpretable test results and withdrawals. The poor agreement for the availability of clinical information was related to reviewer 3 who had a very poor level of agreement (50%) with the final consensus rating; the other reviewers showed over 90% agreement with the final consensus. This suggests that reviewer 3 was interpreting this item differently to the other reviewers. The other three items, selection criteria, uninterpretable results and withdrawals, showed moderate agreement between each reviewer and the consensus rating suggesting that there may be difficulties in applying these items.

Piloting QUADAS in ongoing reviews

Twenty reviewers used QUADAS in their reviews and provided feedback via the structured questionnaire (Table 3). Fifteen reviewers came from the UK, two from Australia, two from the Netherlands, and one from Switzerland. Of those from the UK, seven were employees of the Centre for Reviews and Dissemination (CRD), which is where some of the researchers who developed QUADAS were based. The topics covered by the reviews included the diagnosis of: tuberculosis, urinary tract infection in children, haematuria, Dengue fever, prostate cancer, shoulder pain, epilepsy seizure focus, angina and myocardial infarction, infected diabetic foot ulcers, bacterial infections, lumbar fusion, multiple sclerosis, and osteoporosis. Diagnostic tests under evaluation included laboratory tests, imaging and physical examination. The number of studies included in the reviews ranged from 1 to 208 (median 28).

Content of tool

The feedback from 20 reviewers on the content of the tool was generally positive: eighteen reviewers thought that QUADAS covered all important items, seventeen did not omit any items, sixteen did not add any items, and nineteen did not modify any items.

Two reviewers thought that QUADAS did not cover all important items, one felt that it did not adequately cover population characteristics (description of spectrum, age, setting, prevalence), that questions regarding therapy, the positivity threshold of test results, and study design should have been included as separate items. These comments were mainly related to the desire to have information on these items so that they could be explored in subgroup analysis. The other reviewer thought that the tool should cover whether data could be extracted into a 2 × 2 table.

Three reviewers omitted items from QUADAS. One stated, "on occasions there were no withdrawals". One reviewer omitted items on: reference standard, disease progression bias, partial verification bias, differential verification bias and incorporation bias as these were not applicable to the topic area because there was no reference standard (the review was on prostate biopsies). The other reviewer omitted the item relating to disease progression bias as this did not apply to studies included in their review. Another reviewer stated that they did not omit any items but that as most of the studies included in their review were diagnostic case control studies, items on the availability of clinical information and withdrawals were difficult to answer, and in most cases the issue of follow-up was not relevant.

Four reviewers added items to QUADAS: one added clinically relevant items specific to their review, one added "Do you have plans to characterise data which are unsuitable for primary analysis?", one added "Was the raw data available?" and one added a number of items relating to the availability of 2 × 2 data, confidence intervals, a description of the index and reference tests and a description of the test threshold.

One reviewer modified the items on uninterpretable results and withdrawals to add a "not appropriate" response. She stated that if there were no uninterpretable test results it was unclear how to rate this item.

Background document

All but one reviewer found the background document easy to understand, two did not understand the scoring guidelines, and one reviewer thought that the items concerning differential and partial verification bias should have been scored differently. One reviewer found the item on disease progression bias difficult to understand. However, this difficulty appeared to be related to how to score this item specifically for their review rather than a problem with the instructions provided in the background document. Two reviewers stated that they added topic specific information to the background document to help determine exactly how to score items for their review.

Despite efforts to keep the wording of QUADAS simple to increase international applicability, two non-native English speakers had some difficulty in understanding the QUADAS background document. They found the item on the availability of clinical information difficult to understand and did not know what was meant by uninterpretable or indeterminate data or results, and felt that the background document did not clarify this. In future revisions, clarity of phrasing will be a key consideration.

The reviewer who thought items should have been scored differently felt that the items relating to verification bias should have been formulated differently and suggested "was verification bias avoided? (i.e. did the whole sample or a random selection of the sample receive verification using a reference standard)".

Technical points

The time taken to complete QUADAS ranged from less than 10 minutes to over an hour. Five reviewers reported that it took them <10 minutes, five that it took 10–15 minutes, seven that it took 15 to 30 minutes, two that it took 30 to 60 minutes and one that it took more than an hour. Some of the reviewers included the time to read the whole paper and carry out data extraction and completing QUADAS in this time, whereas others only included the time taken to complete QUADAS. None of the reviewers assessed inter-rater reliability.

Overall conclusions

Reviewers' ratings of QUADAS for coverage, ease of use, clarity of instructions and validity were generally good, especially for coverage, which was rated as good or very good by all reviewers, and ease of use, which was rated as at least average by all reviewers. One reviewer rated the clarity of instructions and the validity of QUADAS as being poor; she had earlier stated that she did not understand the instructions for scoring QUADAS. She also felt the studies in her review were of fairly poor quality but still fulfilled at least half the QUADAS items. All reviewers stated that they would use QUADAS again, although one stated that she may not use all 14 items next time and another stated that this was because there is currently no better tool available.

Additional comments

A major theme in reviewers' additional comments related to the poor quality of reporting of primary studies and the fact that this often limits the quality assessment. Another theme was that it is important to have an understanding of the clinical context while scoring some of the items. One reviewer suggested that it might be helpful to group the questionnaire using subheadings such as "general", "reference standard", and "index test". Another comment was that initial training on how to use the tool would be helpful.

Discussion

Principal findings

This evaluation has shown good agreement between reviewers and the final consensus rating for most QUADAS items and very positive feedback from reviewers who have used QUADAS. Two items, uninterpretable results and withdrawals, were found to be problematic. There was poorer agreement among reviewers and between reviewers and consensus for these items than for other items; feedback from reviewers also suggested problems with these items. One reviewer suggested that this might be because it is difficult to know what to do it if it is unclear if there are any uninterpretable results or withdrawals. Our own use of QUADAS supports this: we have found it very difficult to know how to score this item if the study does not report whether there were any uninterpretable results/withdrawals, and if all patients who entered the study appear to be accounted for. In such situations it is often unclear whether the study authors simply excluded uninterpretable results or withdrawals from their reports, or if there truly were no uninterpretable results or withdrawals. We have handled this problem by giving more explicit instructions for scoring these QUADAS items: we have stated that they should be scored as yes if it appears that all patients who were entered into the study completed the study.

The assessment of inter-rater reliability also highlighted possible problems with the items on the availability of clinical information and selection criteria. The item on clinical information is very specific to each review and it is therefore essential that clear guidelines on scoring this item be provided, outlining exactly what information should be available to the person interpreting the results of the index test. This definition should be agreed a priori. This was done for the review used for this evaluation and is reflected in the very high levels of agreement between two of the reviewers and the final consensus. It is unclear why the third reviewer showed much poorer agreement (50%) with the final consensus rating. It is unclear why the item on selection criteria showed poorer agreement with the consensus rating. This item was not highlighted as problematic in the feedback from reviewers. It may be related to the fact that no review specific information was provided for this item.

All additional items suggested for inclusion in QUADAS were considered as part of the development of QUADAS but were items that were not selected by the panel of experts for inclusion in the final tool. One of the items suggested for inclusion, the item relating to the threshold for the index test could be covered as part of item 8 (description of index test details). This is something to consider including in the guidelines for scoring this item when making guidelines specific to your review.

There was substantial variation in the time taken to complete QUADAS, ranging from less than 10 minutes to over 1 hour. This may be explained by the fact that some reviewers counted the time taken for the whole process of data extraction, including reading the paper, whereas others only counted the time taken to complete QUADAS. Despite this, half the reviewers took less than 15 minutes

Table 3: Summary of responses to the questionnaire on reviewers' experience of using QUADAS

Reviewer Number	Coverage?	Omit items?	Add items?	Modify item?	Easy to understand?	Scoring?	Different scoring?	Time to complete (minutes)	Inter-rater reliability ?	Coverage*	Ease of use*	Clarity of instruction*	Validity*	Use again?*
1	+	-	+	-	+	+	-	30–60	-	5	3	5	5	+
2	+	-	-	+	+	+	-	<10	-	4	4	3	4	+
3	+	+	-	-	+	+	-	<10	-	4	4	5	3	+
4	+	-	-	-	+	+	-	30–60	-	4	4	4	4	+
5	+	-	-	-	+	+	-	15–30	-	5	4	5	5	+
6	+	-	-	-	+	+	-	15–30	-	5	3	4	4	+
7	+	-	-	-	+	-	-	15–30	+	4	4	2	2	+
8	+	-	-	-	+	-	-	15–30	+	4	4	3	3	+
9	-	-	+	-	+	+	+	>60	-	4	5	5	4	+
10	+	-	-	-	+	+	-	<10	-	5	5	5	5	+
11	+	-	+	-	+	+	-	15–30	-	4	3	4	3	+
12	-	-	-	-	+	+	-	15–30	-	5	3	5	3	+
13	+	+	+	-	-	-	-	10–15	-	4	3	3	5	+
14	+	+	-	-	+	+	-	10–15	-	4	4	4	4	+
15	+	-	-	-	+	+	-	10–15	-	5	4	4	4	+
16	+	-	-	-	+	+	-	10–15	-	4	3	4	4	+
17	+	-	-	-	+	+	-	<10	-	4	4	4	4	+
18	+	-	-	-	+	+	-	<10	-	5	5	5	3	+
19	+	-	-	-	+	+	-	10–15	-	4	4	4	4	+
20	+	-	-	-	+	+	-	15–30	-	5	4	5	4	+

*Each of these items were rated from 1–5 where 1 is strongly disagree (very poor) and 5 is strongly agree (excellent)

Table 4: Proposed modifications to the QUADAS background document*

13. Were uninterpretable/intermediate test results reported?

c. How to score this item

If it is clear that all test results, including uninterpretable/indeterminate/intermediate are reported then this item should be scored as "yes". **If the authors do not report any uninterpretable/indeterminate/intermediate results, and if results are reported for all patients who were described as having been entered into the study then this item should also be scored as "yes".** If you think that such results occurred but have not been reported then this item should be scored as "no". If it is not clear whether all study results have been reported then this item should be scored as "unclear".

14. Were withdrawals from the study explained?

c. How to score this item

If it is clear what happened to all patients who entered the study, for example if a flow diagram of study participants is reported, then this item should be scored as "yes". **If the authors do not report any withdrawals and if results are available for all patients who were reported to have been entered into the study then this item should also be scored as "yes".** If it appears that some of the participants who entered the study did not complete the study, i.e. did not receive both the index test and reference standard, and these patients were not accounted for then this item should be scored as "no". If it is not clear whether all patients who entered the study were accounted for then this item should be scored as "unclear".

* Proposed changes are highlighted in bold

*Each of these items were rated from 1–5 where 1 is strongly disagree (very poor) and 5 is strongly agree (excellent)

and 17/20 took less than half an hour to complete QUADAS suggesting that QUADAS is relatively quick to complete.

Strengths and weaknesses of the study

The major strength of this study is that we carried out a detailed evaluation of QUADAS, which specifically included the views and experience of users. We are unaware of any other quality assessment tools for diagnostic accuracy studies that have undergone any process of evaluation.

Ideally, we would have liked to assess the "construct validity" of the tool – "the degree to which a test measures what it claims, or purports, to be measuring" [6]. As QUADAS aims to provide an indication of the quality of a study one way to assess this would be to take a set of "high" quality studies and a set of "low quality" studies and determine whether QUADAS can distinguish between these. This is known as "extreme groups" [6]. The problem with this process is determining which studies are high quality and which are low quality: there is no objective way of doing this. In addition, a systematic review is likely to include studies covering a range of quality. A quality assessment tool needs to be able to distinguish subtle differences across this full range of study quality, not just the extremes. We therefore decided against this method of evaluation.

Unanswered questions and future research

We originally proposed to carry out a meta-epidemiological regression analysis to investigate the association of individual QUADAS items with estimates of test performance. However, due to limited time and resources such an evaluation was not feasible. This is an area where future research would be beneficial. The Cochrane Collaboration is planning to extend its database to include diagnostic test accuracy reviews and is in the process of producing a handbook providing guidelines for the conduct of such reviews. The recommendations on quality assessment include a modified version of QUADAS (items 2, 8 and 9, the items relating to reporting rather than quality have been removed), and this will be built into the new Cochrane software. All diagnostic reviews included in the new Cochrane Database will therefore include an assessment of QUADAS with the results entered into the Review Manager Software in a structured way. In the future, once a number of Cochrane Test Accuracy Reviews have been completed, a meta-epidemiological regression analysis can be pursued.

Conclusions – Suggestions for modifications to QUADAS

We do not feel that major modifications to the content of QUADAS itself, in terms of items included, are necessary. However, the evaluation highlighted particular difficulties in scoring the items on uninterpretable results and withdrawals. We therefore recommend that the guidelines for scoring these items in the QUADAS background document be modified as shown in Table 4. In addition, we would like to highlight the importance of tailoring the guidelines for scoring items to each particular review, and of ensuring that all reviewers are clear on how studies should be scored for each of the items. It is not possible to provide a generic description of what should be considered an "appropriate patient spectrum", or what should be considered an "appropriate reference standard". It is therefore essential that all reviewers using QUADAS carefully consider how each individual item should be applied to their review and adapt the background document to make the guidelines for scoring specific to their review. This should be done in close collaboration with a clinical expert in the area of the review. Reviewers should also carefully consider whether all QUADAS items are relevant to their review, and also whether there are additional quality items not included in QUADAS which may

be of importance to their topic area and which they should assess as part of their review. Consensus should be established on all of these issues before starting the quality assessment. Lastly, an improvement in the quality of reporting, by endorsing the standards for reporting of diagnostic accuracy studies, the STARD initiative [7], should occur. This will allow reviewers to assess study quality rather than the quality of reporting.

Competing interests

The author(s) declare that they have no competing interests.

Authors' contributions

PW, JK and PB conceived the study. All authors contributed to the design of the study. PW and MW collected the data. PW carried out the analysis and drafted the paper. All authors commented on drafts of the manuscript and read and approved the final manuscript.

Acknowledgements

No financial or material support was provided for this study. We would like to thank the reviewers who participated in the assessment of inter-rater reliability and those who completed the questionnaire to provide feedback on their use of QUADAS.

References

1. Whiting P, Rutjes AWS, Reitsma JB, Bossuyt PM, Kleijnen J: **The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews.** *BMC Medical Research Methodology* 2003, **3**:25.
2. Bland JM, Altman DG: **Statistics Notes: Validating scales and indexes.** *BMJ* 2002, **324**:606-607.
3. Whiting P, Rutjes AW, Dinnes J, Reitsma J, Bossuyt PM, Kleijnen J: **Development and validation of methods for assessing the quality of diagnostic accuracy studies.** *Health Technol Assess* 2004, **8**:1-234.
4. Lantz CA, Nebenzahl E: **Behavior and interpretation of the [kappa] statistic: Resolution of the two paradoxes.** *Journal of Clinical Epidemiology* 1996, **49**:431-434.
5. Altman DG: **14.3 Inter-rater agreement.** In *Practical Statistics for Medical Research* First edition edition. London, Chapman & Hall; 1999:403-408.
6. Brown JD: *Testing in language problems* Upper Saddle River, NJ, Prentice Hall Regents; 1996.
7. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Lijmer JG, Moher D, Rennie D, de Vet HC: **Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative.** *Ann Intern Med* 2003, **138**:40-44.

Pre-publication history

The pre-publication history for this paper can be accessed here:

http://www.biomedcentral.com/1471-2288/6/9/prepub

ANEXO 5: ARTIGOS EXCLUÍDOS

NOME DO ESTUDO	AUTORES	REVISTA/ANO DA PUBLICAÇÃO	MOTIVO DA EXCLUSÃO
I. Detection of Subclinical Tubular Injury After Renal Transplantation: Comparison Urine Protein Analysis with Allograft Histopathology	Stefan Schaub, Michael Mayr, Gideon Hönger et al.	Transplantation Volume 84, Number 1, July 15, 2007	Different Methodology
II. Accuracy of Neutrophil Gelatinase-Associated Lipocalin(NGAL) in Diagnosis and Prognosis in Acute Kidney Injury: A Systematic Review and Meta-analysis	Michael Haase, Rinaldo Bellomo, Prased Devarajan, Peter Schlattmann et al.	American Journal of Kidney Disease, Vol 54, No 6 (December),2009:pp 1012-1024	Systematic Review
III. Acute Kidney Injury: New Concepts in Definition, Diagnosis, Pathophysiology, and Treatment	Michael R. Lattanzio Nelson P. Kopyt	JAOA. Vol 109. No 1. January 2009	Review
IV. NGAL: An emerging biomarker of acute kidney injury	Claudio Ronco	The International Journal of Artificial Organs / Vol 31 / no. 3, 2008 / p. 199-200	Editorial
V. Neutrophil gelatinase-associated lipocalin: a novel biomarker for the early diagnosis of acute kidney injury in the Emergency Department	A. Di Grande, C. Giuffrida, G. Carpinteri, G. Narbone, G. Pirrone, A. Di Mauro, S. Calandra, P. Noto, C. Le Moli, et al	European Review for Medical and Pharmacological Science 2009; 13: 197-200	Review
VI. Early Diagnosis of Acute Kidney Injury: The Promise of Novel Biomarkers	Sachin S. Soni, Claudio Ronco, Nevin Katz, Dinna N. Cruz	Blood Purif 2009; 28: 165-174	Review

VII. *Biomarkers - A Potential Route for Improved Diagnosis and Manegement of Ongoing Renal Damage*	R. Oberbauer	Transplantation Proceedings, 40, S44-S47 (2008)	Review
VIII. *Pos- transplant monitoring of renal allograft: are we there yet?*	Peter Nickerson	Current Opinion in Immunology 2009, 21:563-568	Review
IX. *Neutrophil gelatinase-associated lipocalin- an emerging troponin for kidney injury*	Prasad Devarajan	Nephrol Dial Transplantation (2008) 23; 3737-3743	Review
X. *Predicting delayed graft function and mortality in kidney transplantation*	Domingo Hernández, Margarita Rufino, José Manuel González-Posada et al	Transplantion Reviews 22 (2008) 21-26	Review

ANEXO 6: GUIDANCE ON THE CONDUCT OF NARRATIVE SYNTHESIS IN SYSTEMATIC REVIEWS

Guidance on the Conduct of Narrative Synthesis in Systematic Reviews

A Product from the ESRC Methods Programme

Jennie Popay[1]
Helen Roberts[2]
Amanda Sowden[3]
Mark Petticrew[4]
Lisa Arai[2]
Mark Rodgers[3]
Nicky Britten[5]
with
Katrina Roen and Steven Duffy

[1]Institute for Health Research, Lancaster University

[2]Child Health Research and Policy Unit, City University

[3]Centre for Reviews and Dissemination, University of York

[4]MRC Social and Public Health Sciences Unit, University of Glasgow: Mark Petticrew is funded by the Chief Scientist Office of the Scottish Executive Department of Health

[5]Peninsula Medical School, Universities of Exeter and Plymouth

Version 1: April 2006

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

The research reported here was funded by the ESRC (Grant reference number H33250019) within the ESRC Methods Programme. There are many people who have contributed to this work along the way to whom we would like to extend our thanks. We are grateful to Professor Angela Dale, the Director of the programme for her support throughout the project. We are also grateful to the international panel of experts in research synthesis including in particular: Professor David DuBois, University of Illinois; Dr Jeanne Daly, Mother and Child Health Research, La Trobe University; Professor Mike Fisher, Social Care Institute for Excellence; Angela Harden, The EPPI Centre, Institute of Education, London; Professor Cindy Mulrow, University of Texas; Dr Pierre Pluye, McGill University; Professor Helen Thomas, McMaster University; Dr Carl Thompson, University of York; Dr Heather McIntosh, NHS Quality Improvement Scotland; and Ros Collins, Dr Catriona McDaid and Dr Nerys Woolacott at CRD University of York. The extensive and invaluable comments from panel members helped us to improve on earlier versions of this guidance but the responsibility for the final product is, of course, entirely our own. Finally, we would like to thank the individuals in our various administrative teams and research support offices who have supported the work in many ways. Thanks are due in particular to Erja Nikander, Yvonne Moorhouse, and Vicki Bell in Lancaster and Mel Bryan at City who helped at various points with producing documents, travel arrangements, organising meetings, managing budgets and to Lisa Stirk at CRD in York and Vicki Bell at Lancaster who developed and maintain our website.

We would also like to pay tribute to Professor Sally Baldwin who died as a result of an accident in October 2003. Sally developed the proposal for this work with us. There have been many tributes to her extraordinary qualities as a researcher, teacher, colleague and friend. We will not repeat these here – suffice it to say that her stimulating and good humoured contributions to our intellectual endeavours have been badly missed.

CHAPTER 1: ABOUT THE GUIDANCE

Do domestic smoke alarms save lives? Can young offenders be 'scared straight' through tough penal measures? What factors should be considered when designing and implementing a multi-sectoral injury prevention programme in a local area? Making sense of large bodies of evidence drawn from research using a range of methods is a challenge. Ensuring that the product of this synthesis process can be trusted is important for policy makers, for practitioners and for the people research is intended to benefit. There are a number of ways in which research evidence can be brought together to give an overall picture of current knowledge that can be used to inform policy and practice decisions. However, the trustworthiness of some of these methods remains problematic.

The guidance we set out here focuses on a particular approach - **narrative synthesis.** Variants of this approach are widely used in work on evidence synthesis, including Cochrane reviews, but there is currently no consensus on the constituent elements of narrative synthesis and the conditions for establishing trustworthiness – notably a systematic and transparent approach to the synthesis process with safeguards in place to avoid bias resulting from the undue emphasis on one study relative to another – are frequently absent. This guidance therefore aims to contribute to improving the quality of narrative approaches to evidence synthesis.

1.1 Telling stories – the nature of narrative synthesis

Narrative synthesis is sometimes viewed as a 'second best' approach for the synthesis of findings from multiple studies, only to be used when statistical meta-analysis or another specialist form of synthesis (such as meta-ethnography for qualitative studies) is not feasible. In fact, even when specialist methods are used to synthesise findings from multiple studies, those who want to increase the chances of a scientific synthesis being used in policy and practice are likely to find a narrative synthesis helpful in the initial stages of a review. Recognising this, the guidance on undertaking systematic reviews produced by The Centre for Reviews and Dissemination at the University of York suggests that reviewers should first undertake a narrative synthesis of the results of the included studies to help them decide what other methods are appropriate.[1]

Narrative synthesis is a form of story telling. We are part of a story telling culture, and bringing together evidence in a way that tells a convincing story of why something needs to be done, or needs to be stopped, or why we have no idea whether a long established policy or practice makes a positive difference is one of the ways in which the gap between research, policy and practice can start to be bridged. Telling a trustworthy story is at the heart of narrative synthesis.

1.2 Narrative synthesis, narrative reviews and evidence synthesis

'Narrative' synthesis' refers to an approach to the systematic review and synthesis of findings from multiple studies that relies primarily on the use of words and text to summarise and explain the findings of the synthesis. Whilst narrative synthesis can involve the manipulation of statistical data, the defining characteristic is that it adopts a textual approach to the process of synthesis to 'tell the story' of the findings from the included studies. As used here 'narrative synthesis' refers to a process of synthesis that can be used in systematic reviews focusing on a wide range of questions, not only those relating to the effectiveness of a particular intervention.

Narrative review is a phrase some commentators have used to describe more traditional literature reviews and they are typically not systematic or transparent in their approach to synthesis.[2] Narrative synthesis - the focus of this guidance - in contrast, is part of a larger review process that includes a systematic approach to searching for and quality appraising research based evidence as well as the synthesis of this evidence. A narrative review can also be another name for a description, and is used in fields as diverse as performance review of staff[3] to assessing familial patterns in colorectal cancer.[4] Narrative reviews in the sense of traditional literature reviews can be distinguished from narrative synthesis as the latter refers specifically to a specific approach to that part of a systematic review process concerned with combining the findings of multiple studies.

Evidence synthesis includes, but is not restricted to, systematic reviews. Findings from research using a wide range of designs including randomised controlled trials, observational studies, designs

that produce economic and qualitative data may all need to be combined to inform judgements on the effectiveness, cost-effectiveness, appropriateness and feasibility of a wide range of interventions and policies. Evidence syntheses may also addresses many other types of questions including, for example, questions about the current state of knowledge on the causes of particular health or social problems. They are also undertaken in diverse fields from health services research and sociology to engineering and urban planning.

1.3 Why this guidance has been produced?

The Cochrane Collaboration, established in 1993, is an international non-profit and independent organisation, dedicated to making up-to-date, accurate information about the effects of healthcare readily available worldwide. It produces and disseminates systematic reviews of healthcare interventions and promotes the search for evidence in the form of clinical trials and other studies of interventions.

Since its inception, there have been major developments in methods for the systematic review of research evidence which have increased the reliability of the evidence about ***effectiveness*** available to decision makers by combining findings from good quality studies which evaluate policies, specific interventions or professional practices. However, even in reviews focusing on effectiveness, meta-analysis is often an inappropriate approach to synthesis. Additionally, there has been increasing recognition of the need for review and synthesis of evidence to answer questions other than those focusing on effectiveness, particularly those relating to the local implementation of interventions shown to be effective in experimental contexts. Methods for the synthesis of evidence on effectiveness when meta-analysis is not appropriate or for the synthesis of more diverse evidence are, however, not well developed.

Unlike meta-analysis, ***narrative synthesis*** does not rest on an authoritative body of knowledge or on reliable and rigorous techniques developed and tested over time. In the absence of such a body of knowledge there is, as the Cochrane handbook argues[5]

'a possibility that systematic reviews adopting a narrative approach to synthesis will be prone to bias, and may generate unsound conclusions leading to harmful decisions'

This problem is not confined to narrative synthesis - statistical techniques have produced misleading results in the past (and continue to do so from time to time). However, given the widespread use of narrative synthesis in systematic reviews there is a pressing need for the methodological foundation of this approach to be strengthened, if systematic reviews produced to inform the choice and implementation of interventions are to be credible. This is the aim of this guidance.

1.4 What the guidance is about

The guidance provides advice on the conduct of ***narrative synthesis*** in the context of systematic reviews of research evidence and describes some specific tools and techniques that can be used in the synthesis. The (synthesis) product, *at a minimum*, is a summary of the current state of knowledge in relation to a particular review question. This question might relate to effectiveness or cost effectiveness, to issues of efficacy, appropriateness (to need), feasibility of implementation, or to some or all of these.

We recognise that narrative synthesis can be utilised in reviews addressing a wide range of questions. However, for practical reasons, we have focused this guidance on the conduct of the narrative synthesis of ***research*** evidence in the context of two types of systematic review which have particular salience for those who want their work to inform policy and practice: those addressing questions concerned with the ***effects*** of interventions and those concerned with the ***implementation*** of interventions shown to be effective in experimental settings.

1.5 Who the guidance is for

The guidance is intended to be accessible to a range of people involved in systematic reviewing. However, whilst users of the guidance will not need to be systematic review experts, they will need a

reasonable level of research literacy and we would advise anybody without experience of systematic review work to collaborate with more experienced colleagues.

The phrase *evidence synthesis* can be used to mean many different things. At its most simple, synthesis will involve the juxtaposition of findings from multiple studies, perhaps with some analysis of common themes or findings across studies. More sophisticated approaches to synthesis involve the integration or interpretation of results from multiple studies, with the aim of producing new knowledge/findings. It has been suggested[2] that different types of evidence synthesis can be located along a continuum from quantitative approaches, which involve the pooling of findings from multiple studies (e.g. meta-analysis), to qualitative approaches, which involve an interpretative approach (e.g. meta-ethnography). The guidance provided here lies between these two. Narrative synthesis will always involve the 'simple' juxtaposition of findings from the studies that have been included in the review. However, it may also involve some element of integration and/or interpretation, depending on the type of evidence included. These methods necessarily require some familiarity with research processes if they are to be done well.

1.6 When might the guidance be used?

The process of evidence synthesis is not linear, so reviewers may use a number of different approaches to synthesis in an iterative way. Narrative synthesis might be used:

- Before undertaking a specialist synthesis approach such as statistical meta-analysis or meta-ethnography
- Instead of a specialist synthesis approach because the studies included are insufficiently similar to allow for this
- Where the review question dictates the inclusion of a wide range of research designs, producing qualitative and/or quantitative findings for which other approaches to synthesis are inappropriate.

1.7 Developing the guidance

The methods used in the development of the guidance are described in detail in the appendix and summarised here. The process began with a systematic search of the methodological literature in an attempt to identify existing guidance on the conduct of narrative synthesis and any specific tools and techniques that could potentially be used in the narrative synthesis process. The search process and results are shown in Figure 1.

The search included three elements: i) a database search, ii) a search of internet sites and iii) identification of relevant text by members of the research team. This generated 1,309 items. On the basis of an initial review of titles and, where available, abstracts by at least two members of the research team 264 of these items were retrieved and read in full by at least two members of the research team. This process resulted in 69 articles, reports and/or books being included in the methodological review. None specifically related to narrative synthesis although some elements of guidance on established methodologies such as meta-ethnography and 'case survey' method, for example, were judged relevant to the conduct of narrative synthesis.

Methodological guidance on the conduct of various different approaches to review and synthesis were used to identify common generic elements of an evidence synthesis process. Other text provided 'tips' on aspects of the evidence review process in general, such as how to structure results and/or present data and described a number of specific tools and techniques for the management, manipulation and presentation of quantitative and/or qualitative data. This material formed the basis of an initial draft of the guidance on narrative synthesis. The guidance was then applied to two 'demonstration' syntheses: one focusing on the effectiveness of intervention(s); the other on the implementation of intervention(s). These demonstration syntheses have been incorporated into the final version of the guidance to illustrate how the guidance may be used to inform decisions about which specific tools and techniques to use in the context of a particular review.

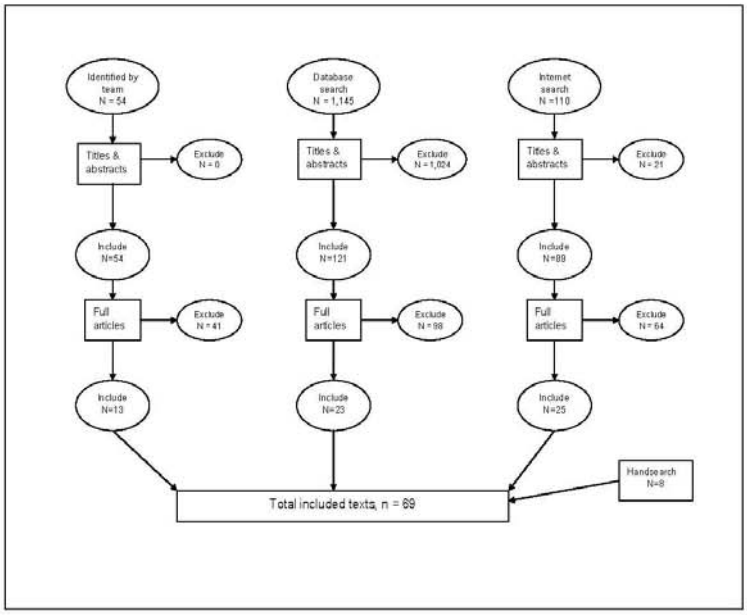


Figure 1. Search process and results

1.8 What the guidance does not do

The guidance does not describe a new approach to the synthesis of qualitative or mixed method research. Instead this guidance seeks to provide an over-arching framework to guide the conduct of a narrative synthesis and suggests ways in which current approaches to narrative synthesis may be further enhanced and developed. Similarly, the guidance is not intended as a source of detailed methodological advice on the systematic review process as a whole. Whilst there is some limited discussion, for example, of search strategies and study quality appraisal, the guidance does not provide details of specific methods for these. We include references to detailed methodological advice in these and other areas in Appendix 2.

CHAPTER 2: THE SYSTEMATIC REVIEW PROCESS – AN OVERVIEW

The process of undertaking a systematic review has been well documented and there is broad agreement about the main elements involved. Six main elements are identified here including the process of synthesis, the focus of this guidance. The other five elements of a systematic review are not described in detail. References to detailed methodological advice on systematic reviewing are included in Appendix 2. This chapter provides a framework to aid understanding of where the synthesis occurs in the systematic review process.

2.1 Identifying the review focus, searching for and mapping the available evidence

Getting the question(s) 'right' is critical to the success of the systematic review process overall. The review question has to be both ***relevant*** to potential users of the review and in theory at least ***answerable***. In some instances the question is clearly formulated at an early stage. More often, however, whilst an initial focus for the review is identified, a '***mapping***' of the available relevant evidence needs to be carried out before the specific question(s) for the review can be clearly specified.[6]

The mapping exercise can be used to assess the need for a systematic review and/or to guide and refine the scope of the review. It is especially useful in situations where a broad question is of interest, such as "how effective are interventions to prevent unintentional injuries?" By mapping the available literature addressing this topic it is possible to:

- Describe the types of interventions that have been evaluated
- Describe the sorts of study designs used in these evaluations and
- Assess the volume of potentially relevant literature.

Based on this initial mapping the scope of the review can be refined, so that the questions to be addressed are both answerable and relevant. The search for studies should be comprehensive and appropriate to the question posed so a mapping exercise may also help to refine a search strategy.

2.2 Specifying the review question

It will take time to get the review question right. In the context of reviews of the effectiveness of interventions, there is general agreement that a well-formulated question involves three key components: the people (or participants) who are the focus of the interventions, the interventions, and the outcomes. Sometimes a fourth component that relates to type of study design is also included. If the review intends to focus on the factors shaping the implementation of an intervention then the question will also have to include components related to this, such as aspects of the context in which the intervention was implemented.

2.3 Identifying studies to include in the review

Once the precise review question has been agreed, the key components of the question form the basis of specific selection criteria, each of which any given study must meet in order to be included in the review. It is usually necessary to elaborate on the key components of the review question so as to aid process of identifying studies to include in the review and make sure that decisions made are transparent to users of the review. These might include, for example, being more precise about the age groups of participants to be included in the review or about aspects of the intervention design.

2.4 Data extraction and study quality appraisal

Once studies are selected for inclusion a process of study quality appraisal and data extraction takes place. Decisions about which data should be extracted from individual studies should also be guided by the review question. In the context of a systematic review addressing a question about the effect of a particular intervention, for example, the data to be extracted should include details of: the

participants, the interventions, the outcomes and, where used, the study design. For reviews focusing on implementation, it would be important to extract detailed data on the design of the intervention, the context in which it was introduced and on the factors and/or processes identified as impacting on implementation. The specific data and/or information to be extracted and recorded are usually those which could affect the interpretation of the study results or which may be helpful in assessing how applicable the results are to different population groups or other settings. This may be referred to as applicability, generalisability or external validity.

Study appraisal - also called validity assessment, assessment of study quality and critical appraisal - usually refers to a process of assessing the methodological quality of individual studies. This is important as it may affect both the results of the individual studies and ultimately the conclusions reached from the body of studies - although 'quality' in general and validity in particular are defined differently in relation to different types of study designs. In the context of effectiveness reviews study quality is often used as a criterion on which to base decisions about including or excluding particular studies, although this does depend on the approach taken by the reviewers. Whatever the focus of the review, reviewers may choose to exclude studies from the synthesis on grounds of methodological quality; others may opt to include all studies, but in this case it is important to differentiate clearly between more and less robust studies. There are many different appraisal tools available for use in relation to both quantitative and qualitative study designs and details of how to get information about some of these are provided in Appendix 2.

2.5 The synthesis

The key element of a systematic review is the synthesis: that is the process that brings together the findings from the set of included studies in order to draw conclusions based on the body of evidence. The two main approaches are quantitative (statistical pooling) and narrative, and sometimes both approaches are used to synthesise the same set of data. One approach - narrative synthesis - is the focus of detailed attention in this guidance.

2.6 Reporting the results of the review and dissemination

Once the review is complete the findings need to be disseminated to potential users, although communication needs to be considered from the start often with the involvement of policy, practice and end point users and throughout the review process. We have included some useful references to the 'art' of dissemination - an often neglected component of the systematic review process in Appendix 2.

CHAPTER 3: GUIDANCE ON NARRATIVE SYNTHESIS – AN OVERVIEW

As we have noted this guidance focuses on the conduct of narrative synthesis in systematic reviews of research-based evidence on:

- The ***effects*** of interventions and/or
- The factors shaping the ***implementation*** of interventions.

Although we have restricted our focus in this way, the guidance may also be helpful for people focusing on other types of review questions, for example, about the needs and/or preferences of particular population groups or the causes of particular social and/or health problems.

Our aim is to provide broad guidance on ways in which the process of narrative synthesis can be made more systematic and transparent and on how bias introduced by the evidence itself (as a result of methodological shortcomings in the included studies) and/or by decisions made by reviewers (for example, through the process of inclusion and exclusion) can be minimised. The guidance does not provide a set of definitive prescriptive rules on the conduct of narrative synthesis. In our experience the most appropriate approach and the selection of specific tools and techniques for data management and manipulation depends on the nature of the particular review being conducted.

In this chapter we describe a generic framework that identifies four elements of the narrative synthesis process and various tools and techniques that can be used to manage data, manipulate and synthesise findings from multiple studies and present the results of the synthesis. In the following two chapters we describe in detail the practical application of the guidance and particular tools and techniques to the synthesis of two bodies of research-based evidence: one concerned with the effects of an intervention the other concerned with factors influencing the implementation of an intervention.

3.1 A general framework for narrative synthesis

For the purpose of this guidance we have identified four main elements to a narrative synthesis process:

- Developing a theory of how the intervention works, why and for whom
- Developing a preliminary synthesis of findings of included studies
- Exploring relationships in the data
- Assessing the robustness of the synthesis

Figure 2 describes the purpose of each of these four elements of a synthesis in relation to a systematic review focusing on (1) the effects and (2) the factors impacting on the implementation of an intervention/programme.

We are not suggesting that narrative synthesis should proceed in a linear fashion with these elements being undertaken sequentially. In practice, reviewers will move in an iterative manner among the activities we have suggested make up these four elements. We have separated them out and presented them sequentially simply to provide a structure to the guidance. In the following sections we focus on these elements in turn in order to explain the aims of each in more detail. We then provide brief descriptions of tools and/or techniques that may be utilised in the conduct of a narrative synthesis before moving on in the subsequent chapters to demonstrate the practical application of the narrative synthesis framework and the specific tools and techniques.

Main elements of synthesis	Effectiveness Reviews	Implementation Reviews
1. Developing a theoretical model of how the interventions work, why and for whom	*Purpose:* • To inform decisions about the review question and what types of studies to review • To contribute to the interpretation of the review's findings • To assess how widely applicable those findings may be	*Purpose:* • To inform decisions about the review question and what types of studies to review • To contribute to the interpretation of the review's findings • To assess how widely applicable those findings may be
2.Developing a preliminary synthesis	*Purpose:* • To organise findings from included studies to describe patterns across the studies in terms of: ○ The direction of effects[1] ○ The size of effects	*Purpose:* • To organise findings from included studies in order to: ○ Identify and list the facilitators and barriers to implementation reported ○ Explore the relationship between reported facilitators and barriers
3. Exploring relationships in the data	*Purpose:* • To consider the factors that might explain any differences in direction and size of effect across the included studies	*Purpose:* • To consider the factors that might explain any differences in the facilitators and/or barriers to successful implementation across included studies • To understand how and why interventions have an effect
4. Assessing the robustness of the synthesis product	*Purpose:* • To provide an assessment of the strength of the evidence for: ○ Drawing conclusions about the likely size and direction of effect ○ Generalising conclusions on effect size to different population groups and/or contexts	*Purpose:* • To provide an assessment of the strength of the evidence for drawing conclusions about the facilitators and/or barriers to implementation identified in the synthesis. Generalising the product of the synthesis to different population groups and/or contexts

Figure 2. The main elements in a narrative synthesis

Element 1: The role of theory in evidence synthesis

Although not all reviewers may choose to do this, it can be useful to develop a model of what Weiss refers to as an intervention's "theory of change" to inform a systematic review. The "theory of change" describes "*the chain of causal assumption that link programme resources, activities, intermediate outcomes and ultimate goals*".[7] It is concerned with how the intervention works, why, and for whom. Reviewers would normally develop their theory of change at an early stage of a review before the synthesis proper begins. If done early enough an understanding of the theory behind the intervention can inform decisions about the review question and the types of studies to include. In terms of the narrative synthesis, a "theory of change" can contribute to the interpretation of the review's findings and will be valuable in assessing how widely applicable those findings may be. Information on programme theory may come from explicit statements in study reports on the goals of the intervention (who it is intended to affect, in what way and how) and from other reviews. The theory can be presented in narrative form or as a diagram like the one reproduced below in Figure 3.

Theory building and theory testing is a neglected aspect of systematic reviews. Shadish (1996) has pointed out that meta-analysis for example has focused too much on descriptive causation (simply describing the size of an effect) and too little on the development of explanatory theories.[8] Yet systematic reviews - whether of qualitative or quantitative research - are likely to be much more powerful than single studies for these purposes. In turn systematic reviews can contribute to developing and testing the limits of theories, by examining how contextual or temporal variables moderate outcomes. Theories themselves can also be the subject of systematic reviews.[9-13]

[1] The notion of 'effects' should not be taken for granted. In some reviews the synthesis process will involve the reviewers in a process intended to help to understand what the effects of a particular interventions or programme are. This is particularly the case when the effects are presented in narrative form rather than in numerical form or derived from structured questionnaires/indicators.

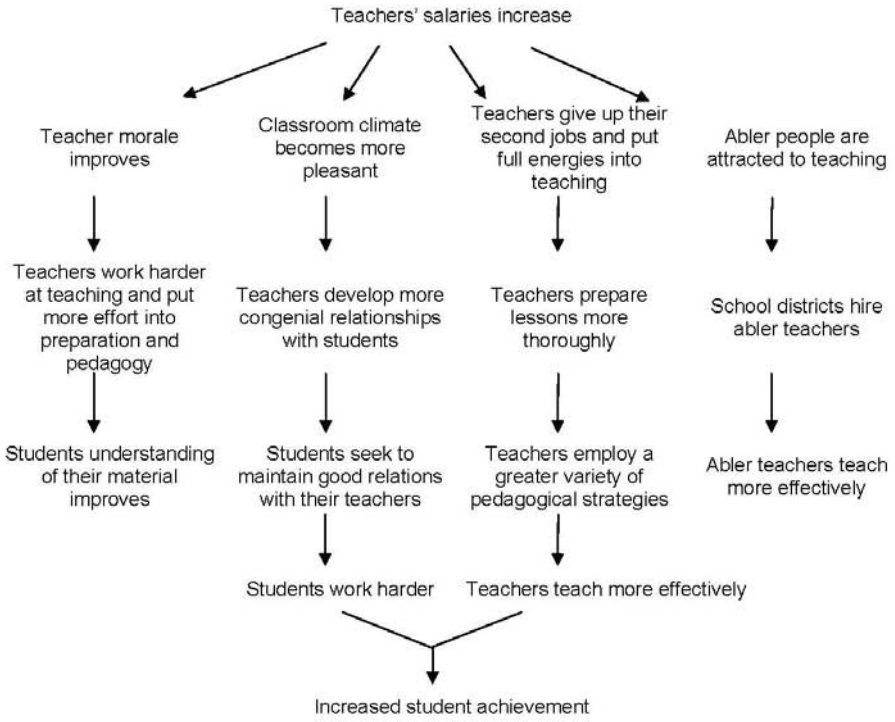


Figure 3. Example of a Programme Theory model: mechanisms by which higher teachers' pay may be linked to increased student achievement (from Weiss, 1998)

Element 2: Developing a preliminary synthesis

Whatever the focus of the review, the purpose of the preliminary synthesis is to develop an initial description of the results of included studies. It is important to remember that the product of this initial process will only be ***preliminary,*** rather than an end in itself. It will always be necessary to interrogate the preliminary synthesis to identify factors that have influenced the results reported in included studies i.e. to begin to construct an explanation of how and why a particular intervention had the effects reported; of how and why particular factors/processes impinged on implementation, and to test the robustness of the results of the synthesis. This is the purpose of other elements of the synthesis process described below.

During the preliminary synthesis, reviewers focusing on the effects of an intervention will need to organise the results of the included studies so they are able to describe patterns across them in terms of both the direction and size of the effects reported. In relation to a review on implementation, the studies need to be organised so that patterns in the factors/processes that are reported as impacting in some way on the implementation of an intervention can be identified across the studies. Assuming that study quality appraisal has been carried out at the same time as data extraction these details will be available during the whole of the synthesis process although quality was not examined in our demonstration synthesis reported later until near the end of the synthesis.

Element 3: Exploring relationships within and between studies

As patterns across study results begin to emerge from preliminary attempts at a synthesis reviewers should begin to subject these to rigorous interrogation in order to:

- Identify any factors that might explain differences in direction and size of effect across the included studies or in the type of facilitators and/or barriers to successful implementation
- To understand how and why interventions have or do not have an effect or why particular barriers and/or enablers to implementation operate

At this point in the synthesis the reviewers move beyond identifying, listing, tabulating and/or counting results to exploring relationships within and across the included studies. The relationships of interest are of two broad types:

- Those between characteristics of individual studies and their reported findings
- Those between the findings of different studies

Some of the studies included in a review may have reported information about relationships between study characteristics and reported findings, in which case the job of reviewers is to compare and contrast the ways in which the relationships have been identified and analysed across the studies. In other cases little attention may have been paid to these relationships. The practical work involves using data previously extracted from primary studies to look at the relationships between study results and key aspects of the primary studies, and comparing and contrasting these relationships across the studies. This element of a narrative synthesis can be very time consuming but it is critical to the quality of the process as a whole.

Exploring the influence of heterogeneity is important at this stage of the synthesis process. We have already noted that a primary reason for choosing a narrative approach to synthesis in a systematic review about the effects of an intervention is because there is considerable heterogeneity in the included studies in terms of methods, participants, interventions and via other unknown sources. There are also likely to be differences between studies in terms of their findings – whether quantitative or qualitative. This too may be due to known differences between the studies, including methodological differences, and differences in the baseline characteristics of populations being studied.[13] Narrative methods have long been recognised as useful for investigating heterogeneity across primary studies and developing an understanding of which aspects of an intervention may be responsible for its success[14] or investigating the possibility that study variation is attributable to theoretical variables.[15]

Many social or behavioural interventions are complex because of the characteristics of the interventions, study population/s, outcomes, or other methodological issues relating to the conduct of the primary studies.[16] Further complexity is introduced because any effects of the interventions may be modified by context, and the intervention itself may vary when it is being implemented.[17-19] Because of these variations, reviewers of complex interventions may expect considerable heterogeneity across studies and need to consider this when synthesising results.

"Social" heterogeneity may incorporate not only socio-demographic and individual differences, but also historical, cultural, spatial and other differences that may affect both the delivery and impact of the interventions being reviewed. Some of the main sources of variability that reviewers need to consider when 'testing' the robustness of the patterns emerging from the included studies are outlined below (adapted from guidance produced by the Cochrane Health Promotion and Public Health Field).[20]

Variability in outcomes

In systematic reviews of clinical interventions variation in outcomes is termed clinical heterogeneity. Variation also exists in social research, however, given the longer causal chains for many social interventions (including public health interventions), proximal/immediate, intermediate, and distal/long term outcomes may be reported. Whilst the synthesis would ideally seek to address all these outcomes in practice it is often not feasible to do this.

Variability in study designs

Methodological diversity is common in systematic reviews of social interventions. Where the main potential sources of variation are known, heterogeneity between effects can be explored by means of subgroup analysis, based for example on theories about how the intervention works, and for which

groups. For many social and public health interventions, theories about mechanisms and interactions may be under-developed and the exploration and interpretation of heterogeneity complex. It may therefore be difficult to anticipate the main sources of heterogeneity *a priori*.

Variability in study populations, interventions and settings
The content of complex social interventions may vary between specific settings or populations. Some of the variability may be intentional as interventions are tailored to local needs (including characteristics which may influence the outcomes of interest such as race, gender, and socio-economic position).

As noted earlier an understanding of the interventions 'theory of change' will be particularly valuable when exploring the influence of heterogeneity especially when interpreting differences between subgroups of studies (post-hoc sub group analyses). The findings of individual studies will vary with study characteristics such as intervention type, quality and extent of implementation, and the study setting, and may vary between different subgroups of participants. Developing plausible explanations for these differences (some of which will be due to chance) is difficult but sub-group findings that are supported by an *a priori* rationale (that is, which have been described in the programme theory) are more plausible than those which are not.

The extent to which reviewers are able to consider the impact of context in systematic reviews evaluating the effects of interventions or factors impacting on implementation will depend on the availability of relevant information in the included studies. Typically, reviews focusing on effects do not consider the context in which an intervention is implemented in great depth. Given that implementation studies are focusing specifically on how dimensions of context (alongside other factors) impinge on implementation, the data available in these studies should be much richer. However, research has suggested that there may be a particular problem with inadequate reporting of research methods in these studies.[21] The dimensions of context which might be relevant to exploring differences in the reported results of included studies will depend on the nature of the intervention with which the review is concerned.

Other factors to be considered in this exploration of factors mediating the impact of an intervention, or explanations of how or why it has a particular impact, may not be able to be extracted from studies as 'data'. These include information about the general approach taken by the researchers both in terms of theory and methods.

Element 4: Assessing the robustness of the synthesis
The notion of robustness in relation to evidence synthesis is complex. Most straightforwardly robustness can be used to refer to the methodological quality of the primary studies included in the review and/or the trustworthiness of the product of the synthesis process. Obviously, these are related. The trustworthiness of a synthesis will depend on both the quality and the quantity of the evidence base it is built on. If primary studies of poor methodological quality are included in the review in an uncritical manner then this will affect the trustworthiness of the synthesis.

The trustworthiness of the synthesis will also depend on the methods used in the synthesis. This will depend on the measures taken to minimize bias, ensuring, for example, that studies judged to be of equal technical quality are given equal weight or if not providing a sound justification for not doing so. Another less straightforward aspect of robustness that can impact on the trustworthiness of the synthesis is the extent to which reviewers have enough information to judge that individual studies meet the criteria for inclusion. This can be a significant problem with reviews of complex interventions. Authors of primary studies often fail to provide adequate information on the intervention they are focusing on and there can be inconsistency between studies in the definition of what constitutes a particular intervention. It is particularly important that reviewers give detailed information about the interventions they plan to include and exclude from a review: for example, stating that 'psychological interventions are eligible' is unlikely to be adequate.

Towards the end of the synthesis process, therefore, the analysis of relationships within and between studies described above should lead into an overall assessment of the strength of the evidence available for drawing conclusions on the basis of a narrative synthesis. This should include systematic attention to all three elements of robustness discussed above.

It is particularly important that the results of any appraisal of the methodological quality of included studies be considered in a systematic manner. Whilst there are well-established methods for assessing the quality of intervention studies, this is not the case in relation to studies of implementation processes, qualitative research or mixed methods research in general so there are no approaches to quality assessment that can be recommended in these situations. Additionally, the results of the appraisal process may or may not have been used to exclude some studies on methodological grounds. Whatever approach to quality appraisal is adopted, (probably at an earlier stage of the review process) this information should inform the assessment of the strength or weight of the evidence available to support conclusions drawn on the basis of the synthesis process.[6]

3.2 Tools and techniques for narrative synthesis

In this section we provide brief descriptions of the tools and techniques we have identified which can be used in the process of narrative synthesis. We have divided these into those which appear to be most appropriate for use in each of the three analytical elements of the synthesis.

At the beginning of each sub-section below the main tools and techniques are listed in a table. As we have noted, decisions about which of these are appropriately used in a specific synthesis will be determined by the nature of the evidence being synthesised as will be illustrated in the practical applications of the guidance.

Before describing the tools and techniques a general comment about the visual representation of data from included studies is warranted. Many of the specific tools and techniques described involve visual representation and this can be invaluable at all stages of a synthesis. However, it is important to recognise that visual representation of data is not sufficient in itself as a synthesis. As Evans[22] has argued, for example, tabulation and other visual representations of data tend to reduce studies to their key characteristics neglecting aspects that could be important in understanding the patterns revealed. He draws a distinction between 'descriptive synthesis' and 'interpretive synthesis' and is critical of the heavy reliance placed by some reviewers on synthesis by tabulation. For commentators such as Evans, the relationship between the visual representation of data (the descriptive synthesis) and the narrative elaboration of the patterns identified (the interpretative synthesis) is critical to the quality of a narrative synthesis.

Element 1: Tools and techniques for developing a theory of change

We have not identified specific tools or techniques for use in the development of a theory of change although some of those described for use at other points in the synthesis process may also inform theory development and elaboration - as highlighted in the practical application of the guidance in chapters four and five.

Element 2: Tools and techniques for developing a preliminary synthesis

1. Textual descriptions of studies
2. Groupings and clusters
3. Tabulation
4. Transforming data into a common rubric
5. Vote counting as a descriptive tool
6. Translating data; thematic analysis
7. Translating data: content analysis

Textual descriptions

A simple starting point in a preliminary synthesis is to produce a descriptive paragraph on each included study - it may also be useful for recording purposes to do this for all excluded studies as well. In many reviews this will have been completed at an early stage in the review process and it can be done for any type of study. It is important that these narrative descriptions are produced in a systematic way, including the same information for all studies if possible and in the same order. Some reviewers have suggested that studies considered more important in terms of what they offer the review may be discussed at greater length, while briefer discussion may be afforded to less central or informative studies.[23] In theory this is a way of giving more weight to higher quality or larger studies within a narrative synthesis. However, it is difficult to determine how much "weight" in terms of description/discussion should be allotted to individual studies and how this should vary with methodological quality, for example. Additionally, if textual descriptions are produced at an early

stage of the review process it will not be possible to give more weight to one study over another and hence a fuller description because methodological quality and other aspects of relevance will not yet have been assessed. Whilst textual descriptions are a useful way for reviewers to become familiar with the included studies and to begin to compare and contrast findings across studies, it can be very difficult to discern patterns across studies from these textual descriptions, particularly when there are a large number of studies.

Groupings and clusters

There can be considerable variation in the number of studies included in systematic reviews. Some Cochrane reviews, for example, conduct the synthesis on a very small number of studies, often because of very tightly defined inclusion/exclusion criteria and/or to a paucity of research addressing the question of interest. Other reviews include large numbers of studies in the pool to be synthesised. In most cases the number of studies included will be determined by the size and quality of the existing literature. Whilst including findings from large numbers of studies can be labour intensive, the analytical process involved in statistical meta-analyses can readily manage large numbers. This is not the case with narrative synthesis. Usually therefore, a process of narrative synthesis will involve organising the included studies into smaller groups to make the process more manageable. Although the reviewers may start to group the included studies at an early stage of the review, it may be necessary to refine these initial groups as the synthesis develops.

Organising the included studies into groups can also be a useful way of aiding the process of description and analysis and looking for patterns within and across these groups. It is important to use the review question(s) to inform decisions about how to group the included studies. Studies can be grouped according to one or a combination of the following: the type of intervention being studied; the setting or context for the intervention (school or community based interventions for example); the group at whom it is being directed (different age groups, for example); the study design; and/or the nature of the results being reported (different outcome measures for example, or different types of factors impacting on implementation).

Tabulation

Tabulation is a common approach used in all types of systematic review to represent both quantitative and/or qualitative data visually - indeed many of the examples of approaches to description included in this guidance are presented in tabular form. Tabulation can be useful at any stage of the preliminary synthesis process according to the preference of reviewers. It can be particularly useful in helping to develop an initial description of the included studies and to begin to identify patterns across studies. They are typically used to provide details of study design, results of study quality assessment, outcome measures and other results. These data may be presented in different columns in the same table or in different tables. Used thoughtfully, tabulation can be a valuable tool in the preliminary synthesis of results across studies and can provide important building blocks for future elements of the synthesis process.[22, 24-26]

Some authors stress the need to take care with the layout of tables, arguing that the way in which data are tabulated may affect readers' impression of the relationships between studies. For example, 'placing a results column adjacent to any of the characteristics or quality columns could invite speculation about correlation and association'.[24] These notes of caution point to the importance of the reviewers' attempting some narrative interpretation of tabulated data.[22, 24]

Transforming data: Constructing a common rubric across quantitative studies

The results of studies included in a review may take different numerical and/or statistical forms.[2] In these situations reviewers need to transform results into a common numerical/statistical rubric if possible. When extracting data from quantitative studies, it is standard practice to extract the raw or summary data from included studies wherever possible, so a common statistic can be calculated for each study, e.g. converting dichotomous data into odds ratios or relative risks and continuous data (if from different measurement scales) into standardised mean differences (SMD). In a review of effectiveness which incorporates a statistical meta-analysis these results would be pooled to provide a single estimate of effect. In a narrative synthesis study results will not be pooled statistically, so the process cannot provide a new single estimate of effect. However, transforming study results into a

[2] The distinction being made here, between numerical and statistical, relates to the possibility that figures provided as percentages, for example, would not accurately be described as statistics.

common rubric will allow reviewers to develop a meaningful summary of study results and a more robust assessment of the range of effects that would be anticipated from a particular intervention.

Vote-counting as a descriptive tool

Although some commentators[27] have argued strongly against 'vote counting' calculating the frequency of different types of results across included studies can be a useful way of producing an initial description of patterns across the included studies.[28] Indeed, it could be argued to be an intrinsic element of the preliminary stages of any narrative synthesis. In the case of reviews evaluating the effects of an intervention, a simple approach to vote-counting might involve the tabulation of statistically significant and non-significant findings. Some reviewers have developed more complex approaches to vote counting, both in terms of the categories used and by assigning different weights or scores to different categories.

The interpretation of the results of any vote counting exercise is a complex task. According to some methodologists writing about vote counting, the category with the most studies "wins".[29] Similarly in the context of reviews of effects, some commentators argue that the statistical significance category 'containing the largest number of studies represents the direction of the true relationship'.[30] However, it has also been argued that, this approach to synthesis "tends to give equal weight to studies with different sample sizes and effect sizes at varying significance levels, resulting in misleading conclusions".[31] There are examples where vote counting has been compared with other methods of synthesis and major differences in findings have been reported.[32-34] So, whilst vote counting can be a useful step in a preliminary synthesis the interpretation of the results must be approached with caution and these should be subjected to further exploration of relationships between data/findings within and across the included studies.

Translating data: thematic and content analysis

Where results are presented in the form of themes or concepts, as is the case in qualitative research or some surveys, studies focusing on similar topics may have conceptual overlaps, even if these are not apparent from the way the results are reported. Alternatively, apparently similar concepts in different studies may actually be referring to different phenomena. In this context a process of 'translation' of primary themes or concepts reported across studies can be used to explore similarities and/or differences between different studies.[35] Where studies involve both qualitative and quantitative data reviewers may decide to construct a common rubric for the synthesis – this could involve transforming qualitative findings into quantitative form or vice versa. Both thematic analysis and content analysis can help in this process of 'translation' or 'interpretation' as it is sometimes referred to.

Thematic analysis

Thematic analysis, a common technique used in the analysis of qualitative data in primary research, can be used to identify systematically the main, recurrent and/or most important (based on the review question) themes and/or concepts across multiple studies. Although usually used with qualitative data some people have argued that it could be used with studies involving quantitative data or data from mixed method studies. For example the variable labels included in survey research may be extracted as 'themes' in the same way as conceptual themes are extracted from qualitative research reports.[36] Thematic analysis provides a means of organising and summarising the findings from large, diverse bodies of research. The analysis would typically, but not invariably, be developed in an inductive manner; i.e. without a complete set of a priori themes to guide data extraction and analysis from the outset. Thematic analysis tends to work with, and reflect directly, the main ideas and conclusions across studies, rather than developing new knowledge although this is possible.

There are problems with thematic analysis from the perspective of a systematic review. The process can, for example, be associated with a lack of transparency – it can be difficult to understand how and at what stage themes were identified. The results of the synthesis might look very different if an entirely a priori, theoretically-driven approach had been used as against an inductive approach. In this context it is important that reviewers give as much detail as possible about how a thematic analysis was conducted.

Content analysis

Content analysis was developed as an analytical approach for primary research, but it is readily applied to the synthesis of findings from multiple studies. Content analysis has been defined as 'a

systematic, replicable technique for compressing many words of text into fewer content categories based on explicit rules of coding.'[37] Unlike thematic analysis, it is essentially a quantitative method, since all the data are eventually converted into frequencies, though qualitative skills and knowledge of underlying theory may be needed to identify and characterise the categories into which findings are to be fitted.

Element 3: Tools and techniques for exploring relationships

1.	Graphs, frequency distributions, funnel plots, forest plots and L'Abbe plots
2.	Moderator variables and sub-group analyses
3.	Idea webbing and conceptual mapping
4.	Translation : reciprocal and refutational
5.	Qualitative case descriptions
6.	Investigator/methodological triangulation
7.	Conceptual triangulation

Graphs, frequency distributions, funnel plots, forest plots and L'Abbe plots.
There are several visual or graphical tools that can help reviewers explore relationships within and between studies, although these are typically only useful in the context of quantitative data. These include:

- presenting results in graphical form
- plotting findings (e.g. effect size or factors impacting on implementation) against study quality
- plotting confidence intervals; and/or plotting outcome measures

Frequency distributions, funnel plots, forest plots, and L'Abbé plots are other possibilities. These tools do not provide any overall interpretative synthesis of the data presented in the plot. There may be good reasons for reviewers not to provide an overall interpretative synthesis of the data presented graphically but it is normally good practice to do so and if not done then it is important that reviewers 'explain' their reasons for not presenting an overall narrative synthesis of these types of representations of data.

Moderator variables and subgroup analyses
There is a growing consensus that when evaluating the impacts of interventions the important questions are "what works, for whom, and in what circumstances". One approach to answering these questions when findings are quantitative is by means of analysing moderator variables – variables which can be expected to moderate the main effects being examined by the review. This can be done at the study level, by examining characteristics that vary between studies (such as study quality, study design or study setting) or by analysing characteristics of the sample (such as groups of outcomes, or participants), based on some underlying theory as to the effects of those variables on outcomes. An analysis of moderator variables can be guided by questions such as:

- What are the moderators that the authors of the primary studies identify?
- What are the contributing factors that appear to recur across the studies even if they have not been explicitly identified by authors as moderators?
- How much difference do the likely moderators appear to make to the study results?
- What possible relationships are there among the moderators?

One approach currently used to explore moderators is to examine the effects of interventions across different social groups. Systematic reviewers have argued for some years for the importance of exploring moderator effects in systematic reviews.[8, 38, 39] Methodological groups working within the Cochrane Collaboration have also contributed extensive empirical and other work on these issues. For example the Cochrane Methods Group in Subgroup Analysis has demonstrated some of the methodological and epistemological pitfalls. A new Joint Cochrane Campbell Methods Group has also been formed focusing on equity issues in systematic reviews and exploring the effects of socio-economic moderators will be an important focus for this group. Explorations of effects in subgroups can also play an important role in testing and developing theory in systematic reviews. They can be an important tool for assessing the strength of relationships, for testing the limits of theoretical concepts and explanations, and can contribute to the development of new theories.[10 13]

Developing conceptual models

There are a number of approaches to exploring relationships within and across the studies included in a systematic review that can be broadly described as conceptual models. The basic idea underpinning these approaches is (i) to group findings that reviewers decide are empirically and/or conceptually similar and (ii) to identify (again on the basis of empirical evidence and/or conceptual/theoretical arguments) relationships between these groupings. The approaches often involve visual methods to help to construct groupings and relationships and to represent the final product of this process. Three specific approaches were identified in the methodological literature review conducted to support the production of this guidance: idea webbing, conceptual mapping and conceptual triangulation. Although we describe them separately below they are very similar as we discuss in the demonstration syntheses reported in chapter 4 and 5. It is perhaps worth noting that these tools can also be used to develop review questions and to begin to identify moderator variables to be explored in more detail before the synthesis begins but we do not discuss these uses in this guidance.

Ideas webbing

Ideas webbing suggested by Clinkenbeard,[29] as a method for conceptualising and exploring connections among the findings reported by the studies included in a review. This approach uses spider diagrams to develop a visual picture of possible relationships across study results.

Concept mapping

Mulrow, Langhorne & Grimshaw[40] describe a similar process which we refer to as concept mapping. Their approach involves linking multiple pieces of evidence extracted from across individual studies included in a review to construct a model highlighting key concepts or issues relevant to the review question and representing the relationships between these. This approach uses diagrams and flow charts to visually represent the relationships being explored. The notion of *conceptual* triangulation described by Foster appears to be very similar in that it is concerned to explore relationships between data drawn from within and between studies.[41] Foster argues that this approach alleviates 'concerns about combining numbers and text because both qualitative and quantitative results can be portrayed conceptually". The approach relies heavily on tables to facilitate the analysis and produces a number of possible *models* through which the phenomenon of interest may be better understood on the basis of the diverse sources of evidence synthesised.

Translation as an approach to exploring relationships

Translation as a process for synthesis is typically associated with the work of Noblit & Hare on meta-ethnography.[35] It is a way of using qualitative research techniques to synthesise findings from multiple studies. The term 'meta' in this context refers to the *translation* of studies into one another. Although developed for use with qualitative research, the approach could be used with a mixture of qualitative and quantitative evidence. Translation focuses on seeking a common rubric for salient categories of meaning, rather than the literal translation of words or phrases. Noblit and Hare identify two different types of 'translation':

1. *Reciprocal translation* (accounts are directly comparable)
2. *Refutational translation* (the accounts are oppositional)

In practice there are few examples of refutational translation in the literature. Having translated the studies into one another, they suggest that reviewers should develop a 'line of argument' drawing inferences from the results of the translation. The line of argument is developed by examining similarities and differences between cases to integrate them in a new interpretation that 'fits' all the studies. Meta-ethnography is a specialist approach to synthesis (akin to statistical meta-analysis with quantitative studies) and not therefore an approach to be utilised in full in the context of a narrative synthesis. However, the translational process may be of value as a way of exploring relationships across studies. The inductive nature of the process means it is emergent, the initial question or area of interest may be adapted or redirected, and there are numerous judgement calls along the way. Of course the same can be argued for other types of synthesis.

Qualitative case descriptions

As Light and Pillemer note, formal statistical procedures may be able to detect subtle differences in effectiveness but they do not necessarily explain them.[14] These authors argue that 'qualitative case descriptions' are particularly valuable in helping with the interpretation of statistical findings. However,

they give relatively little practical advice about how one would go about doing this type of case description. In general terms qualitative case description would seem to include any process in which descriptive data from studies included in a systematic review are used to try to explain differences in statistical findings, such as why one intervention outperforms another (ostensibly similar) intervention or why some studies are statistical outliers. As an example of this process they suggest that in a review of the effectiveness of educational programmes the reviewers might use a range of information from the included studies to seek to answer questions such as:

- What are the characteristics of successful implementations?
- How were the teachers trained?
- How were parents involved?
- What were the details of the educational programme?

This kind of descriptive information may or may not be reported in the original study reports. The textual descriptions of studies described earlier would be a potential resource for this type of work.

Investigator triangulation and methodological triangulation
Approaches to triangulation focus on the methodological and theoretical approaches adopted by the researchers undertaking the primary studies included in a systematic review. Consideration of how these differ across the included studies may be helpful in exploring the nature and impact of moderators in quantitative research or broader relationships in qualitative research. Some authors argue that by working with a number of different triangulation approaches reviewers can develop a better understanding of how the various factors involved in the intervention and its evaluation may have impacted on the results reported in included studies.[42]

Investigator triangulation was developed by Begley to explore the extent to which heterogeneity in study results may be attributable to the diverse approaches taken by different researchers.[42] The approach involves analysing the data in relation to the context in which they were produced, notably the disciplinary perspectives and expertise of the researchers producing the data.[42] Begley is focusing on primary research but this approach could be valuable for evidence synthesis too. It works from the understanding that each disciplinary approach may have produced different kinds of findings. Considering what kinds of evidence and what kinds of outcomes emerge from studies conducted by researchers from particular disciplinary and epistemological positions is potentially an illuminating way to think about possible sources of heterogeneity. This approach will be easier if the review is being undertaken by a multidisciplinary research team "allowing data to be subjected to a range of disciplinary gazes".[43]

Methodological triangulation was developed by Maggs-Rapport and offers a broadly similar approach.[44] Both of these approaches serve as a reminder that the evidence being synthesised in a systematic review does not offer a series of discrete 'answers' to a specific question. Rather, each 'piece' of evidence offers a partial picture of the phenomenon of interest. The product of the systematic review, particularly in the case of narrative synthesis, may not be a 'meta-answer' to the review question, but a theoretical insight and/or a new model that informs understanding about the mechanisms underlying the results reported.

Element 4: Tools and techniques for assessing robustness of the synthesis

1.	Weight of Evidence – e.g. the EPPI approach
2.	Best Evidence Synthesis
3.	Use of validity assessment – e.g. the CDC approach
4.	Reflecting critically on the synthesis process
5.	Checking the synthesis with authors of primary studies

Weight of Evidence – the EPPI approach
The Weight of Evidence approach developed by staff of the EPPI-Centre is used in many EPPI-Centre reviews.[45] In the EPPI approach relevance criteria are set for a particular review and studies are then assessed for relevance using these. Those that are judged to be relevant are then assessed for methodological quality.

Best Evidence Synthesis (BES)
BES deals with the robustness in terms of the methodological quality of included studies though the application of inclusion criteria. This is based on an approach described by the educational researcher Robert Slavin.[46, 47] In BES, only studies that meet minimal standards of methodological adequacy and relevance to the review are included, and information is extracted in a common standard format from each study, with a systematic approach to the assessment of study quality and study relevance. This approach is not prescriptive about the study designs which can to be included in a review – this can vary, depending on the review question. BES aims to identify and synthesise sources of evidence no matter how diverse. It has been suggested however that BES is simply an example of good systematic review practice albeit with some problems. Suri, for example, suggests that in extracting data from the primary studies BES tends towards calculating the median effect size, rather than calculating a weighted mean effect size, as is standard meta-analytic practice.[31]

Although BES accounts cover the whole review process the approach focuses in particular on the selection of studies into a systematic review rather than focusing on the synthesis, thus emphasising that decisions about study quality should be taken early in the review process to ensure that the review is based on robust evidence. The decision about "strength of evidence" is therefore made early in the review process, and its practical application can be seen in the inclusion and exclusion criteria. For this reason the demonstrations of the application of the narrative synthesis guidance reported in the next two chapters were not able to utilise the approach to check the robustness of the synthesis findings.

Use of validity assessment – Centre for Disease Control (CDC) approach
Other approaches to assessing the strength of evidence included in evidence synthesis have been developed. For example, specific rules may be used to define explicitly what is meant by "weak", "moderate" or "good" evidence. There are numerous examples of this form of synthesis but few from the social sciences. One recent example from healthcare comes from the CDC Community Guide to Preventive Services. [48] In this approach, the reasons for determining that the evidence is insufficient are: A. Insufficient designs or executions, B. Too few studies, C. Inconsistent, D. Effect size too small, E. Expert opinion not used. The categories are not mutually exclusive. While the criteria can be debated, the grounds on which the decision about strength of evidence is made are at least explicit. Many other healthcare evidence grading systems use a similar approach.

Reflecting critically on the synthesis process
Busse et al[49] recommend that in reporting the results of a systematic review a summary discussion section should be provided including the following:

- Methodology of the synthesis used (especially focusing on its limitations and their influence on the results)
- Evidence used (quality, validity, generalisability) – with emphasis on the possible sources of bias from the sources of evidence used and their potential influence on results of the synthesis
- Assumptions made
- Discrepancies and uncertainties identified (the way that any discrepancies in findings between included evidence were dealt with in the synthesis should be discussed and wherever the evidence is weak or non-existent, areas where future research is needed can be highlighted)
- Expected changes in technology or evidence (e.g. identified ongoing studies)
- Aspects that may have an influence on the implementation of the technology and its effectiveness in real settings
- Such a summary would enable the analysis of robustness to temper the synthesis of evidence as well as indicating how generalisable the synthesis might be.

Checking the synthesis with authors of primary studies
In the context of their meta-ethnography of qualitative research Britten et al suggest consulting the authors of included primary studies in order to test the validity of the interpretations developed during the synthesis and the extent to which they are supported by the primary data.[50] This is most likely to be useful where the number of primary studies is small but the authors of the primary studies may have useful insights into the possible accuracy and generalisability of the synthesis.

3.3 Conclusion

In this chapter we have provided an overview of the four main elements of the narrative synthesis process that we have identified and briefly described various tools and techniques that can be used at different points in the synthesis process. In the next two chapters we describe in detail the practical application of the guidance, including the use of particular tools and techniques, to the synthesis of two bodies of research evidence. Chapter four focuses on a narrative synthesis of the findings of the 11 RCTs included in the Cochrane systematic review of interventions for promoting smoke alarm ownership and function.[51] The original Cochrane review involved a meta-analysis which means we are able to compare the results/conclusions of the two approaches to synthesis. Chapter five focuses on the narrative synthesis of studies of the implementation of domestic smoke alarm promotion interventions. This is linked to an earlier pilot review and some comparisons with the outcomes of this earlier work are made.[21, 52]

CHAPTER 4: APPLYING THE GUIDANCE 1: A NARRATIVE SYNTHESIS OF STUDIES OF THE EFFECTIVENESS OF INTERVENTIONS FOR PROMOTING SMOKE ALARM OWNERSHIP AND FUNCTION

4.1 Introduction

The aims of this chapter are to:

- Illustrate in practical terms the decision making processes involved in the application of the guidance to a specific narrative synthesis
- Identify factors that should inform choices about the use of particular tools and techniques in the context of a specific synthesis
- Provide examples of how particular tools and techniques can be used in the synthesis of evidence on effectiveness
- Demonstrate the type of outcomes achieved by a narrative synthesis
- Compare the outcomes of a narrative synthesis of the effect of an intervention with those produced by meta-analysis.

The review selected for comparison was a Cochrane review investigating the effects of interventions for promoting smoke alarm ownership and function.[51] This review was selected because it was methodologically sound, had incorporated a meta-analysis, had analysed a 'manageable' number of studies (11 RCTs) and because it complemented the systematic review of the implementation of smoke alarm promotion interventions that was also being resynthesised in a concurrent 'testing' of the guidance.

- Developing a theory of how the intervention works, why and for whom
- Developing a preliminary synthesis
- Exploring relationships within and between studies
- Assessing the robustness of the synthesis

Within each of these sections, the guidance presents a number of related tools and techniques that can be used to complete the various stages of the synthesis. To apply this guidance to the narrative synthesis, each of the sections was read through in sequential order, and for each element the tools and/or techniques that appeared to be useful and relevant to the synthesis at hand were selected. The reasons for selecting or rejecting a tool or technique are given within each section. Where possible, the tools and techniques employed were used to derive conclusions about the effects of interventions for promoting smoke alarm ownership and function. Where tools or techniques proved to be less useful, this is discussed. A flow chart summarising the synthesis process is presented in figure 4.

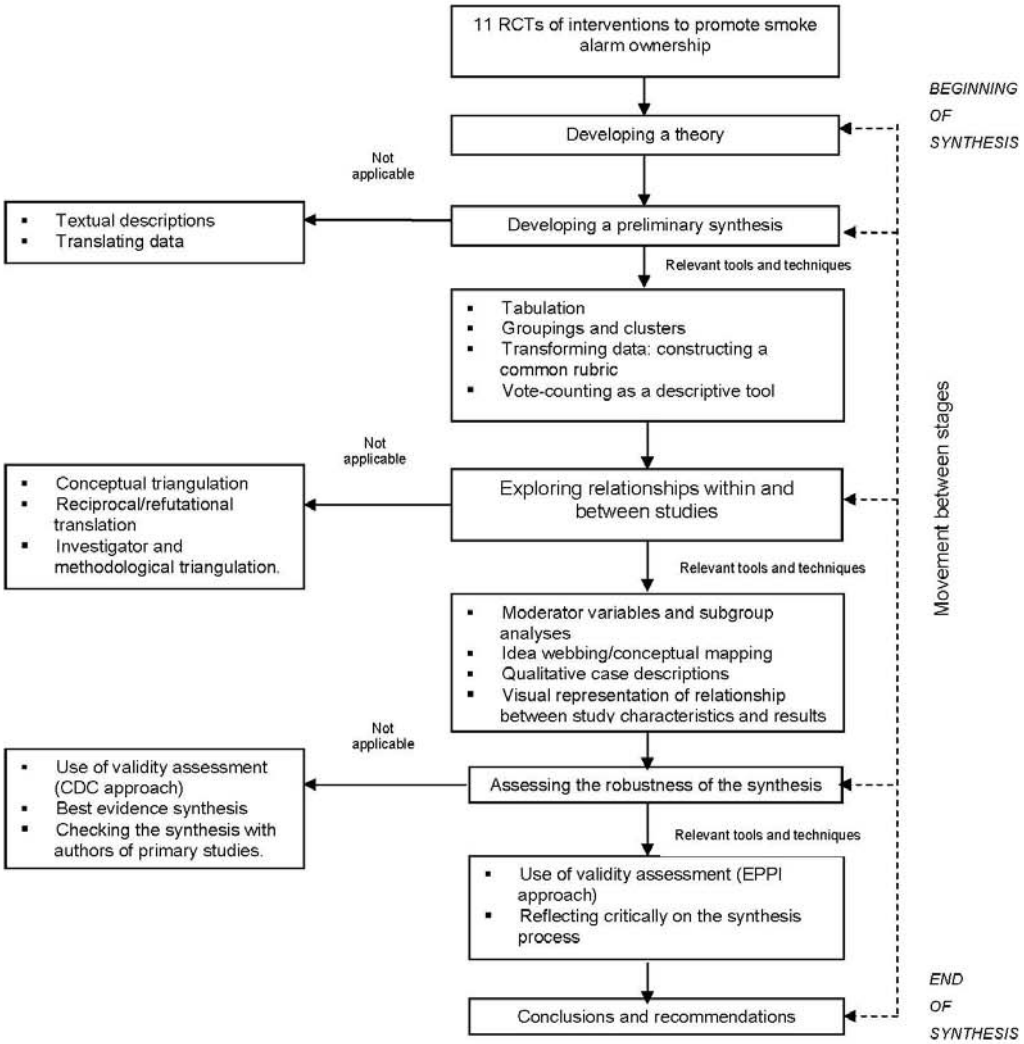


Figure 4: Synthesis process

4.2 Developing a theory

The majority of studies aimed to increase smoke alarm ownership and function through the use of educational interventions with or without the addition of free or discounted smoke alarms for participants. The primary studies did not clearly describe the theoretical basis of the evaluated interventions, but the implicit theory underlying most educational interventions was that education can increase knowledge of potential fire/burns risks, change risk perceptions and lead to behaviour change (i.e. acquisition of smoke alarms). The use of discounted or free smoke alarms as an intervention to increase ownership and function (usually in lower income families) suggests that authors consider cost to be a barrier to smoke alarm acquisition.

4.3 Developing a preliminary synthesis

It is stated in the guidance that "how a reviewer approaches the preliminary synthesis... will depend in part on whether the evidence to be synthesised is quantitative, qualitative or both". In the case of this example, the data to be synthesised were anticipated to be predominantly quantitative and, more specifically, derived entirely from randomised controlled trials. With this in mind each of the tools and techniques presented in the 'preliminary synthesis' section of chapter three were evaluated as to whether they would be relevant for the synthesis at hand (see table 1 below).

Table 1: Selection of tools and techniques in developing a preliminary synthesis

Name of tool/technique	Thoughts/ideas/comments in relation to current synthesis	Should this tool/technique be applied here?
Textual descriptions	Need to determine which aspects of each study will be drawn from the reports. These might be the same as the table headings	Possibly, but not necessarily as a first step
Groupings and clusters	If possible, organise studies by intervention type, context, target population, study design, outcomes. Maybe have 'primary clusters' e.g. (intervention type, population) and have 'secondary clusters' (e.g. study design, context) within these	Yes
Transforming data: constructing a common rubric	Odds ratios or relative risks for dichotomous data, weighted or standardised mean difference for continuous data	Yes
Translating data	Inappropriate given predominantly quantitative data and the effectiveness focus of this review	No
Tabulation	Describe study characteristics and results. Will quality be assessed here? How? Predefined categories or just voicing methodological concerns that occur when reading the studies? Present these in text, tables, or both? Perhaps use the text descriptions to highlight any important aspects about *individual* studies that might not be apparent from the tables (issues across studies are more likely to fit into the next section on 'exploring relationships')	Yes
Vote-counting as a descriptive tool	Would be possible here if all data had been converted to odds ratios/relative risks/mean differences	Yes

Consequently, five of the six tools/techniques described in the guidance were applied to the synthesis and were carried out in the order described below.

Tabulating the data

It was decided that extracting data from the primary studies in tabular form might be the most natural starting point for the synthesis. This was done by using the same format as the Cochrane review's 'characteristics of included studies' table (participants, interventions, outcomes, notes) and adding further information, including country of origin, duration and provider of the intervention, number of participants in each group, context in which intervention was delivered, and results (see Table 4).

Study validity/quality is not addressed in detail in this section of the guidance. However, the Cochrane review did report some aspects of study validity (e.g. concealment of allocation) in the data extraction tables. It seemed sensible at this stage of the narrative synthesis (where the papers were being read in detail and some broad judgements about their content are starting to be made) to consider study quality. Consequently, a column including data on methods/quality was included in the table and structured comments were included regarding individual papers, based on Jadad et al's scale for evaluating RCTs.[53]

It became apparent that there were some discrepancies between the outcomes extracted for tabulation in the narrative synthesis and those in the Cochrane review. However, upon contacting the Cochrane review's authors, all these discrepancies could be explained by the inclusion of unpublished data or statistical adjustments for clustering. In these cases, to ensure comparability, data from the Cochrane review were used in the narrative synthesis.

At this stage, it became clear that the majority of studies were concerned with child safety, and that most included some measure of smoke alarm ownership/function as a main outcome. Only two studies[54, 55] looked at injuries as an outcome, but neither of these presented separate data on fire/smoke/burn related injuries.

Textual descriptions

It was not entirely clear what these might add to the preliminary synthesis over and above the information presented in the data extraction tables. Immediately after having constructed the data extraction tables, this seemed like an unnecessary duplication of effort, though it was considered that 'textual descriptions' might actually be useful for describing the interventions in more depth than can be usefully given in the tables. Consequently, the use of this technique was delayed until a later stage of the synthesis process.

Groupings and clusters

The presence of natural groups or clusters of studies was investigated, primarily to determine whether studies could be clustered according to the characteristics in the data-extraction tables (such as intervention, participants, setting, outcomes etc). The most obvious difference between studies in terms of the populations included is that all the studies deal with children and/or their families, with the exception of the Ploeg study that includes only participants aged 65+ years.[56] This study was therefore excluded from later comparisons. Secondly, studies could be clearly be grouped according to which of the four smoke alarm/ownership outcomes (specified a priori in the Cochrane review) they reported.

Developing a common rubric

As mentioned previously, data were only available for the four smoke alarm ownership/function outcomes. As these data were dichotomous, odds ratios and relative risks were calculated. Absolute risks differences and percentage smoke alarm ownership in the control group were also calculated for each smoke alarm ownership outcome and tabulated (an example for the 'final smoke alarm ownership' outcome is shown in table 5).

These tables showed that the effects of most interventions were generally fairly small for most smoke alarm ownership and function outcomes (absolute differences ranged from 0% to 12.4%). However, they generally favoured intervention over control (only two of the 10 studies that measured final smoke alarm ownership were negative for this outcome and one of the four studies reported a very small negative finding (absolute difference –0.1%) for 'smoke alarms acquired'.

Smoke alarm ownership in the control groups of each study was generally quite high, with one clear exception (Kelly et al),[54] 11%). As might be expected, there was greater range of odds ratios than corresponding relative risks for each outcome, as odds ratios are frequently more extreme (i.e. further from 1) than relative risks.

This approach proved a useful first step in comparing the effects observed across the included studies.

Table 2: Characteristics of included studies

Reference	Intervention	Participants	Setting/context	Outcomes	Results	Methods/quality	Other notes
Barone (1988) USA	*Content:* I: Usual safety education, plus slides and handouts on burn prevention, motor vehicle safety education and video; bath water thermometer; hot water gauge. (n=41) C: Usual safety education (n= 38) *Duration:* 4 x 2h weekly meetings. *Delivered by:* Unclear	Couples or individuals attending "Parenting the toddler" classes	Classes conducted at suburban hospital, family homes	Home inspection 6 months after class 1) Final smoke alarm ownership 2) Final functioning smoke alarms	1) Final smoke alarm ownership I = 32/34 C = 26/29 2) Final functioning smoke alarms: I = 39/41 C = 34/38 I = 32/34 C = 26/29 No significant difference between groups	Allocation by coin toss within paired classes Outcome assessment not blinded Withdrawals: 27% of parents attending randomised classes did not enrol in trial	
Clamp (1998) UK	*Content:* I: Safety advice, leaflets, discount safety devices for low income families (n=83 families) C: Routine child health surveillance and routine consultations without intervention (n=82 families) *Duration:* Unclear *Delivered by:* Health visitors/practice nurses	Families of children <5 yrs on GP list	Delivered during child health surveillance consultations, opportunistically during other consultations, or the family was asked to make an appointment specifically for the intervention	Telephone/ mail survey 6 weeks after visit: 1) Smoke alarms acquired 2) Functioning smoke alarms acquired 3) Final smoke alarm ownership 4) Final functioning smoke alarms	1) Smoke alarms acquired: I = 8/83 C = 0/82 2) Functioning smoke alarms acquired I = 7/83 C = 4/82 3) Final smoke alarm ownership: I: 82/83 C: 71/82 4) Final functioning smoke alarms: I: 80/83, C: 71/82	Allocation by random numbers table numbered 1-165, the first 83 numbers on the list were allocated to the intervention group. Allocation was done by a researcher blinded to the number given to each family at the time of allocation Outcome assessment not blinded Withdrawals: None	

Reference	Intervention	Participants	Setting/context	Outcomes	Results	Methods/quality	Other notes
Davis (1987) USA	*Content:* I: Fire safety lessons with workbook, demonstrations, teacher training, materials, take home materials for parents (n=439) C: Usual lessons (n=418) *Duration:* 6 x 1-hour lessons *Delivered by:* Teacher	Children in grade 4-6 classes	School	In school survey, immediately after class: 1) Final smoke alarm ownership	Final smoke alarm ownership: I = 221/314 C = 195/299 I = 309/439 C = 272/418	Method of random allocation unclear Outcome assessment not blinded Withdrawals: I = 1% C = 0%	The study uses repeated hypothesis testing
Jenkins (1996) Canada	*Content:* I: Discharge teaching book about burn care and prevention; routine discharge teaching (n=62 families) C: Routine discharge teaching (n=61 families) *Duration:* One session *Delivered by:* Physical therapist, occupational therapist or nurse	Families of children <17 years in burn unit	Delivered at discharge from burn unit	Interview in clinic at first follow-up visit (time since intervention unclear): Final smoke alarm ownership	Final smoke alarm ownership: I = 45/62 C = 46/61	Allocation by random numbers table read by independent person Outcome assessment blinded Withdrawals: 13% overall (unclear for each group)	48% of children in the study were of Native American Indian origin. Families were less likely to have safety devices, and less likely to speak English as a first language

Reference	Intervention	Participants	Setting/context	Outcomes	Results	Methods/quality	Other notes
Kelly (1987) USA	*Content:* I: Developmentally oriented child safety education, hazard assessment and handout at 6, 9 and 12-month well child visits. (n=55 families) C: Usual 6, 9 and 12-month well child visits (n=54 families) *Duration:* Each visit approx 15 mins *Delivered by:* I = Principal investigator C = primary caretaker (paediatric resident, fellow, faculty member, or nurse practitioner	Families of children aged 6 months seen for well child care	Family home	1) Final smoke alarm ownership (from home inspection, 1 month after 12-month visit) 2) Accidents and/or hospitalisations (from hospital record review)	1) Final smoke alarm ownership: I = 8/55 C = 6/54 No significant difference between groups 2) ER/primary care visits for accidents: I = 15/55 C = 11/54 Accidents requiring treatment: I = 3/55 C = 4/54 Hospitalisations for accidents: I = 1/55 C = 1/54	Method of random allocation unclear Outcome assessment blinded. Withdrawals: I = 35% C = 37%	
Kendrick (1999) UK	*Content:* I: Age specific advice, cheap safety equipment for low income families, home safety checks, first aid training. Checklists, information sheets and literature provided throughout (18 centres randomised, n=1124) C: Usual care (no further description) (18 centres randomised, n=1028) *Duration:* Unclear. *Delivered by:* Health visitors and practice nurses	Children aged 3-12 months	Community	a) Record review of injuries b) Postal survey of safety practices at 25 month follow-up: 1) Smoke alarms acquired 2) Functioning smoke alarms acquired 3) Final smoke alarm ownership 4) Final functioning smoke alarms	1) Smoke alarms acquired: I = 15/274 C = 11/277 2) Functioning smoke alarms acquired: I = 20/274 C = 14/277 3) Final smoke alarm ownership: I = 254/274 C = 248/277 4) Final functioning smoke alarms: I = 243/274 C = 241/277	Allocation by random numbers table by investigator blind to the identity of the practices Outcomes assessment blinded Withdrawals: I = 67% C = 64%	Not all participants received all aspects of the intervention

Reference	Intervention	Participants	Setting/context	Outcomes	Results	Methods/quality	Other notes
King (2001) Canada	*Content:* I: Home safety inspection and tailored education, safety device coupons; reinforcement (by telephone) at 4 and 8 months, plus a letter from the local project director (n=482 families) C: Home safety inspection and general safety pamphlet only (n=469 families) *Duration:* Unclear *Delivered by:* "Home visitor"	Families of children aged <8 years hospitalised for injuries	Family home	Home inspection at 1 year follow-up: 1) Smoke alarms acquired 2) Functioning smoke alarms acquired 3) Final smoke alarm ownership 4) Final functioning smoke alarms	1) Smoke alarms acquired: I = 14/476 C = 14/464 2) Functioning smoke alarms acquired: I = 44/440 C = 36/435 3) Final smoke alarm ownership: I = 460/479 C = 454/465 1.45 (0.94, 2.22), p=0.05. 4) Final functioning smoke alarms: I = 412/459 C = 401/447 1.01 (0.79, 1.30)	Allocation by opening sealed, serially numbered, opaque envelopes Outcome assessment blinded. Withdrawals: I = 20% C = 18%	Though generally not given feedback after home safety inspection, control group families were informed if non-functioning smoke alarms were discovered

Reference	Intervention	Participants	Setting/context	Outcomes	Results	Methods/quality	Other notes
Mathews (1988) USA	*Content:* I: Home safety inspection, video, handouts, modelling re: safety and managing dangerous child behaviour; hot water thermometers; choke tube. (n=12 families) C: Home visit with video, handouts, modelling on language simulation (n=12 families) *Duration:* Home visits 1.5 – 2 hours, intervention 45-60 mins *Delivered by:* Psychologist	Mothers of toddlers (12-14 months at first contact) from clinics, day care centres	Family home	Home inspection 2 weeks after home visit: 1) Smoke alarms acquired 2) Functioning smoke alarms acquired 3) Final smoke alarm ownership 4) Final functioning smoke alarms	1) Smoke alarms acquired: I = 0/12 C = 0/12 2) Functioning smoke alarms acquired: I = 0/12 C = 0/12 3) Final smoke alarm ownership: Pre-test: I = 10/12 C = 9/12 4) Final functioning smoke alarm ownership: I = 6/12 C = 6/12 There were no significant differences between groups or trials on these outcomes	First eight participants allocated in odd-even manner, remainder using open random numbers table Blinding unclear Withdrawals: 8% in total	

Reference	Intervention	Participants	Setting/context	Outcomes	Results	Methods/quality	Other notes
Ploeg (1994) Canada	*Content:* I: Safety behaviour promotion A safety checklist developed from the injury prevention literature, used with clients to discuss personal, home and community safety and to address strategies to improve safety. (n=148) C: Influenza immunisation promotion (n=211) *Duration:* One visit Duration unclear *Delivered by:* Public health nurses	English speaking public health clients aged 65 or over Mean age 77.2 years, 67% female	Delivered during a visit to the client's home	Telephone survey after 2-3 months: Smoke alarms acquired	Smoke alarms acquired: I = 3/146 C = 1/197	Allocation by random numbers table read by independent person Outcome assessment blinded Withdrawals: I = 1% C = 7%	
Thomas (1984) USA	*Content:* I: Well-baby classes with standard safety information plus burn prevention education lecture, pamphlet, handouts and discount coupon for smoke alarm purchase (9 classes: n=29) C: Well-baby classes with standard safety information (6 classes: n=26) *Duration:* I/C: 1 x 90min session *Delivered by:* Paediatric nurse practitioners	Volunteer parents of infants enrolled with a single HMO No further information provided	Hospital? (conference room)	Home inspection 4-6 weeks after class: Final smoke alarm ownership	Final smoke alarm ownership: I = 27/28 C = 21/25	Randomised using coin toss. Blinding unclear No withdrawals mentioned	Smoke alarm ownership was very high in both groups (actual numbers not given for C group)

Reference	Intervention	Participants	Setting/context	Outcomes	Results	Methods/quality	Other notes
Williams (1988) USA	*Content:* I: Usual safety education plus 1 hour lecture, handouts on burn prevention; motor vehicle safety education and video (n=40). C: Usual safety education plus 1 hour lecture, handouts on infant stimulation and feeding (n=35) *Duration:* Unclear *Delivered by:* "Trainer"	New mothers identified while attending prenatal classes	Unclear	Home inspection 4-70 weeks after delivery: 1) Final smoke alarm ownership	Outcome data not available. The authors state that there was no difference between I and C groups, with both groups showing usage rates for smoke alarms of over 77%	Allocation by random numbers table by independent statistician Outcome assessment not blinded Withdrawals: 55% of women attending randomised classes did not enrol in trial	

I = Intervention group
C = Control group

Table 3: Final smoke alarm ownership (common rubric and vote count)

Reference	Absolute difference (%)	Relative risk (95% CI)	Odds ratio (95% CI)	Vote count RR	Vote count OR	% smoke alarm ownership in control group
Barone (1988)	4.5	1.05 (0.90, 1.22)	1.85 (0.29, 11.89)			90
Clamp (1998)	12.2	1.14 (1.04, 1.25)	12.7 (1.6, 100.85)	✓	✓	87
Davis (1987)	5.2	1.08 (0.97, 1.20)	1.27 (0.9, 1.78)			65
Jenkins (1996)	-2.8	0.96 (0.78, 1.19)	0.86 (0.39, 1.93)			75
Kelly (1987)	3.4	1.31 (0.49, 3.52)	1.36 (0.44, 4.23)			11
Kendrick (1999)	3.2	1.04 (0.98, 1.09)	1.49 (0.82, 2.7)			90
King (2001)	-1.6	0.98 (0.96, 1.01)	0.59 (0.28, 1.25)			98
Mathews (1988)	8.3	1.11 (0.74, 1.68)	1.67 (0.22, 12.35)			75
Thomas (1984)	12.4	1.15 (0.95, 1.38)	5.14 (0.53, 49.5)			84
Williams (1988)	No stats	No stats	No stats	No stats	No stats	>77

Key to table colour coding

- Significantly favours intervention
- Trend towards intervention
- No difference
- Trend towards control
- Significantly favours control

Vote counting as a descriptive tool

Tables showing two approaches to vote counting were developed: (i) only using ticks where the effect of the intervention was positive and statistically significant; (ii) using colours (superimposed on the rows of the table) to grade both the direction and statistical significance of each outcome (see table 5 for an example showing the 'final smoke alarm ownership' outcome).

In terms of the vote-count there were no differences between the relative risks and odds ratios calculated previously. The study by Williams[57] reported that there was no statistically significant difference between the experimental and control groups but did not provide data to calculate the measures in this table. For the subsequent steps, the relative risk and the more "informative" (colour coded) vote count were both used.

The vote-count supported the observations previously made by looking across the absolute risk values. Where several studies report the same outcome, the majority of these studies show a tendency to favour the intervention over control, though the relative risk is usually small. Only one study reported any statistically significant differences between intervention and control groups (Clamp reported statistically significant positive effects of intervention on final smoke alarm ownership and final functioning smoke alarms).[58]

In this case, the colour-coded descriptive vote-count allows the reader to see the outcome data as either a simple vote-count or as a statistical value, depending upon the 'focus' they adopt when examining the outcome table.

4.4 Exploring relationships within and between studies

Tools/techniques described in this section of the guidance are described in the table below.

Table 4: Selection of tools and techniques for exploring relationships between studies

Name of tool/technique	Thoughts/ideas/comments in relation to current synthesis	Should this tool/technique be applied here?
Moderator variables and subgroup analyses	Most likely sources of potential moderator variables are likely to be variations in intervention, population or possibly setting	Yes
Idea webbing/conceptual mapping	This may help structure the investigation of moderator variables	Yes
Conceptual triangulation	This approach would be more appropriate to a synthesis of implementation studies, in which more qualitative information is likely to be available and there is greater scope for model development	No
Reciprocal/refutational translation	Insufficient qualitative evidence in this review	No
Qualitative case descriptions	This appears to be essentially the same as the 'textual descriptions' described earlier. However, here the approach is presented in the context of investigating differences between, rather than simply describing, the studies. Might be worthwhile to revisit the studies and extract detailed data from them, with an eye to any potential moderator variables	Yes
Visual representation of relationship between study characteristics and results	This is possible given the quantitative data available for each study	Yes
Investigator and methodological triangulation	More applicable to qualitative studies. As all studies here were RCTs, there should not be any systematic difference in results between authors from different disciplines (if there was, bias would be a very serious concern). Data on the disciplinary perspective/expertise of investigators was not available for all studies	No

The four main tools and techniques for exploring relationships within and between studies were conducted in the order described below.

Moderator variables and subgroup analyses

It would be useful to know any variables that might moderate the main effects being examined by the review. Two further types of table were drawn up to help investigate whether there were any clear moderators of effect. The first table was constructed to show the various components that make up the intervention for each study and the overlap between the different interventions in terms of these components (table 5).

The table indicates that there is little overlap between the studies in terms of the specific components employed within the interventions they evaluate. Seven of the ten studies concerned with children and/or their families used handouts and four used 'burn education', money-off coupons or discounted devices and home safety inspections. However, this lack of overlap is possibly due to the fact that studies were, on the whole, very poorly described. Even when sufficient information was reported allow extraction, there was still variation in the terms and definitions used by different authors, making direct comparisons even more difficult.

The second set of tables is an adaptation of the outcomes/vote count table, with further information taken from the data extraction table and the intervention components table described above (table 6 gives an example for the 'final smoke alarm ownership' outcome). Intervention, population and setting columns were included to identify potential subgroups/moderators. These are described as briefly as possible (1-5 words) to allow visual comparison across the table. The description of the intervention is broken into 3 separate cells to facilitate such visual comparisons for the complex interventions.

Looking at the outcome of 'final smoke alarm ownership' (for which the majority of studies provide data), four studies stand out from the majority of positive but statistically non-significant findings: Williams (no difference),[57] Clamp (significantly positive),[58] Jenkins and King (both non-significantly negative).[59, 60] Williams reports that "there were no differences between experimental and control groups",[57] though whether this means there was truly no difference between the groups or that any observed differences were not statistically significant is unclear. Either way, it is difficult to determine why the studied intervention had little or no effect based on this one study alone. The intervention studied by Clamp included safety advice, discounted safety devices and handouts and resulted in a significant increase in final smoke alarm ownership and function.[58] However, these particular intervention components were common to other studies that differed from Clamp's study in terms of both magnitude and statistical significance of effect. The two negative studies on the ownership outcome (Jenkins and King) evaluate two different interventional approaches.[59, 60] However, these studies do share a common characteristic that is not present in the 'positive' studies: the intervention was delivered to the families of children that had been previously hospitalised for an injury.

Qualitative case reports/textual descriptions

The two 'tools' 'textual descriptions' and 'qualitative case descriptions' would seem to be very similar. It was decided that writing a short summary of each study at this stage of the synthesis (i.e. having already organised, described and examined them) would provide an opportunity to check the previous stages for accuracy, and allow the reviewer to draw out in detail any aspects of individual studies that may not have seemed relevant at the start of the synthesis, but have become of interest during the subsequent stages of describing and exploring the study data. These summaries were structured such that they provided details of the setting, participants, intervention, comparison, and outcomes, along with any other factors of interest (for an example of one such description see Box 1).

Table 5: Table showing various components of the evaluated interventions

	Burn education	Slides	Handouts	Safety advice	Discount devices /coupons	First aid training	Home safety inspection	Tailored education	Reinforcement	Video	Modelling	Free thermometer /choke tube	School fire safety lessons	Child safety education
Barone (1988)	✓	✓	✓											
Clamp (1998)			✓	✓	✓									
Davis (1987)			✓										✓	
Jenkins (1996)	✓													
Kelly (1987)			✓				✓							✓
Kendrick (1999)				✓	✓	✓	✓							
King (2001)					✓		✓	✓	✓					
Mathews (1988)			✓				✓			✓	✓	✓		
Thomas (1984)	✓		✓		✓									
Williams (1988)	✓		✓	✓						✓				

*Above studies relate to children/families. Ploeg (included only ptps aged >65yrs)

Table 6: Final smoke alarm ownership (potential moderator variables)

Reference	Intervention			Population	Setting	Absolute difference (%)	% smoke alarm ownership in control group
Barone (1988)	Burn education	Slides	Handouts	Parents of toddlers	Hospital, family home	4.5	90
Clamp (1998)	Safety advice	Discount devices	Handouts	Parents of children <5yrs	Family home, other	12.2	87
Davis (1987)	Fire safety lessons		Take home material for parents.	Children	School	5.2	65
Jenkins (1996)	Discharge teaching book on burn care/prevention			Children <17yrs	Hospital burn unit	-2.8	75
Kelly (1987)	Child safety education		Home safety inspection Handouts	Families of babies <6mths	Family home	3.4	11
Kendrick (1999)	Safety advice First aid training	Discount devices	Home safety inspection	Families of babies 3-12mths	Community	3.2	90
King (2001)	Tailored education Reinforcement	Discount coupons	Home safety inspection	Families of hospitalized children <8yrs	Family home	-1.6	98
Mathews (1988)	Video Modeling re safety	Free thermometers and choke tube.	Home safety inspection Handouts	Mothers of toddlers (12-18mths)	Family home	8.3	75
Thomas (1984)	Well-baby classes plus burn prevention education lecture.	Discount smoke alarm coupon	Handouts Pamphlet	Parents of infants	Hospital(?)	12.4	84
Williams (1988)	Burn prevention lecture		Handouts	Pregnant women (last trimester)	Unclear	No stats	>77

Key to colour coding:

- Significantly favours intervention
- Trend towards intervention
- No difference
- Trend towards control
- Significantly favours control

A number of questions arose from the process of writing these summaries:

- Does the immediate on-site availability of smoke alarms in the intervention setting increase uptake?
- Are lower income families more likely than higher income families to respond to interventions incorporating discounted smoke alarms?
- Does having experienced a child injury prior to intervention increase uptake of the recommendations given in the intervention?
- Do interventions that focus on burn injuries/fire prevention have different effects to interventions that relate to safety more generally?
- Does advice being age-specific alter outcomes? Would advice regarding fire safety always be the same, independent of child age?
- Does attrition have an effect?
- Is length of follow-up an important factor?
- Is sample size important? Studies may be powered to detect differences on other outcomes.
- Several studies attribute any lack of effect to the fact that an active effort is required to install smoke alarms. Is there a relationship between intervention effectiveness and amount of active effort required?

Box 1 – Example of textual descriptions/qualitative case descriptions of included studies

Barone (1988)

Setting: US suburban hospital.

Participants: Individuals or couples attending a continuing-education series on "Parenting the Toddler". Couples were predominantly of middle- and upper-middle class socioeconomic status and generally well educated.

Intervention: Parenting information, with specific information and materials on burn prevention and child restraints. Included a slide presentation on falls, strangulations, drownings, poisonings, and fire hazards, plus additional slides on the hazard of hot tap water, use of smoke detectors and the advantages of child car seats. 4 weekly sessions, each of 2 hours duration. 41 participants.

Comparison: Parenting information, with general child safety information. Included a slide presentation on falls, strangulations, drownings, poisonings, and fire hazards. 38 participants.

Outcomes: A researcher inspecting participants' homes looked for and tested any smoke alarms, 6 months after the classes.

Other:

- The protocol for this intervention is very similar to that described by Williams (both are from the same University in the same year).
- The author suggests that the very high rate of smoke alarm ownership might be due to previous health promotion efforts.
- The author also suggests that it would have been possible for participants in the control group to be 'warned' in advance what the researchers were looking for and testing during home inspections by other participants whose homes had already been inspected.

This suggests that the production of summaries can be a helpful prelude to identifying and assessing impact of moderator variables, building on data extraction and developing conceptual models.

Developing conceptual models/idea webbing/concept mapping

These three tools/techniques are also very similar although the implementation narrative synthesis illustrates differences between them. The aim of using these techniques in this example was to make transparent the logic behind the subgroup analyses/investigation of moderator variables (see figure 5). In working through this process it became apparent that it also incorporates aspects of grouping and clustering. The resulting figure is also in part a way to link the previously described processes and the resulting issues/ideas together in order to structure the synthesis. It represents the product of

a process whereby variables or patterns are identified in one of the previously described tables or documents and then re-examined from the viewpoint of the remaining tables/documents. For example, the characteristic most fully explored in the figure is that of the included study population, as described in the table of potential moderator variables and in the textual descriptions. Studies of children/families were grouped by age of the included children according to the moderator tables. Within these groups, further participant variables such as socioeconomic status were identified from the textual descriptions.

The 'outcomes' and 'quality' nodes are connected to one another via 'loss to follow-up'. The withdrawal rates vary substantially across this group of studies, from 0% to 67%. Where high dropout rates are discussed in these studies, it is attributed non-attendance over time or unavailability of participants at final follow-up.

Though identified as potential moderators, no clear or consistent effect on smoke alarm ownership could be seen across studies for intervention variables such as the use of home inspections or free/discounted devices, or for fire/burn-specific education alone versus general safety information that incorporates fire/burn material.

Initially idea webbing was very useful approach to guiding the synthesis. However, the ad hoc use of the approach led to a natural impulse to seek out any association, no matter how spurious. Given this it may be better to use these types of approaches early on in the synthesis (or even protocol development) process to identify *a priori* the characteristics to be investigated and to structure the synthesis before seeing the data itself.

Alternatively, it might be useful to employ this approach at both points in the review process (protocol development and exploring relationships in the final synthesis), placing more "weight" on investigations from the *a priori* idea web (i.e. using it to help develop conclusions about effects and moderators), and using the idea web constructed after interrogating the data purely for suggesting areas in which further research might be worthwhile.

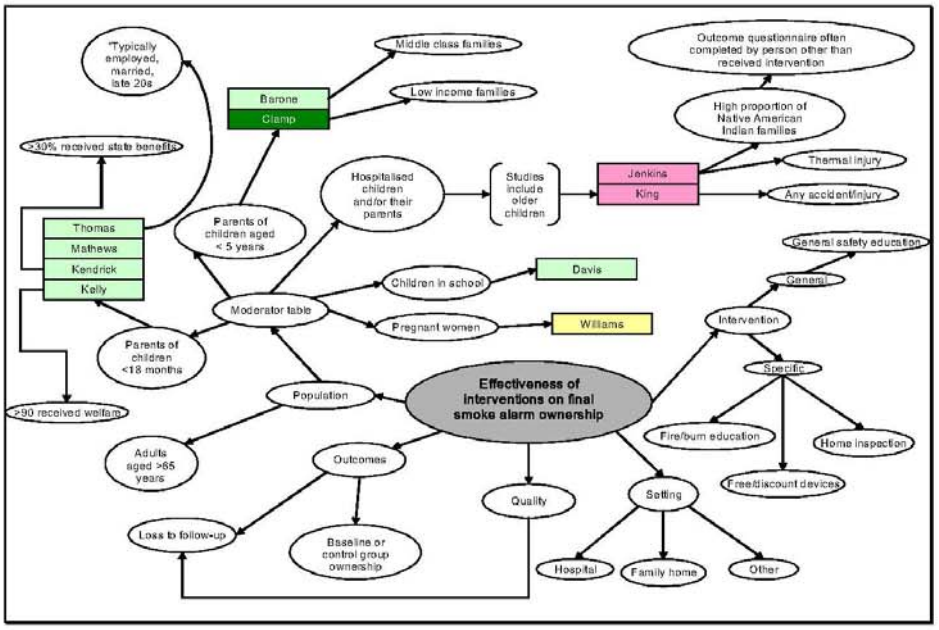


Figure 5: Conceptual mapping/idea webbing

Visual representation of relationships between study characteristics and results

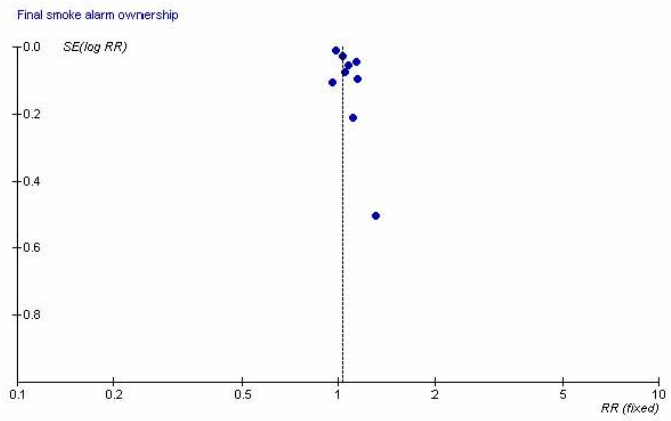
Funnel plots to examine the relationship between study sample size/variance and effect size were constructed by plotting relative risk against standard error (see figure 6). Due to the small number of studies reporting data on the outcomes of interest, these proved to be largely uninformative. The plot for 'final smoke alarm ownership' shows that the study with the lowest precision is that with the most strongly positive effect, but this alone does not provide strong evidence for publication bias.

These proved unhelpful but may be more useful in larger reviews where enough quantitative data are reported to allow a visual display. However this may not be the case for many systematic reviews of social interventions.

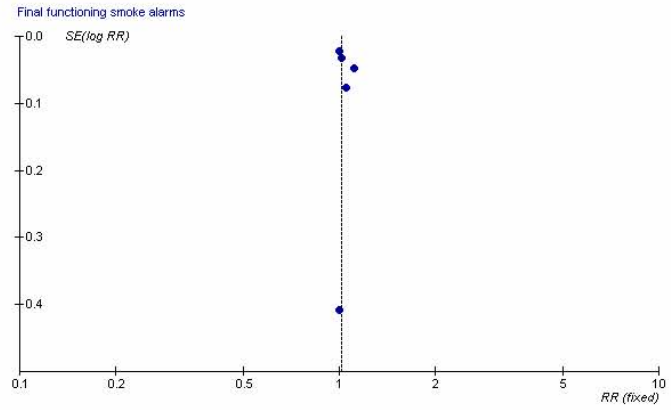
Forest plots showing the point estimates and 95% confidence intervals for each study for each of the main outcomes (but without a pooled estimate) were also drawn, as suggested in the guidance (figure 7). These provide a clear visual representation of the relative risks and associated 95% confidence intervals previously presented as in table 5.

Figure 6: Funnel plots showing standard error versus relative risk for each outcome

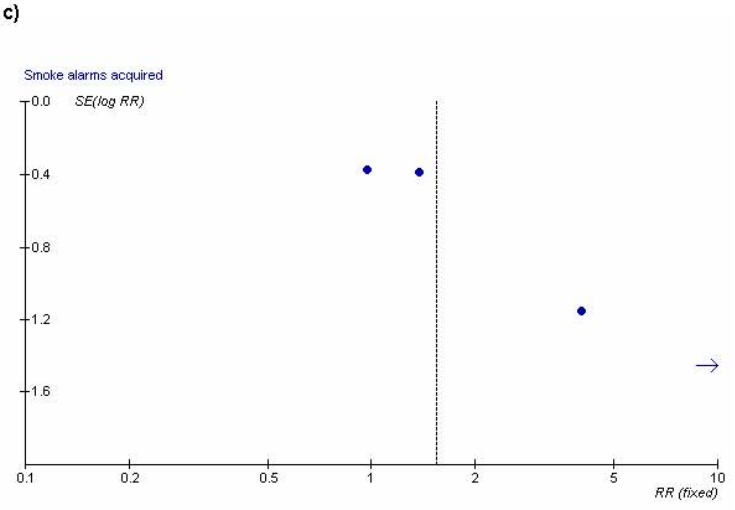
a)



b)



c)


Smoke alarms acquired
0.0
SE(log RR)
0.4
0.8
1.2
1.6
0.1
0.2
0.5
1
2
5
10
RR (fixed)

d)

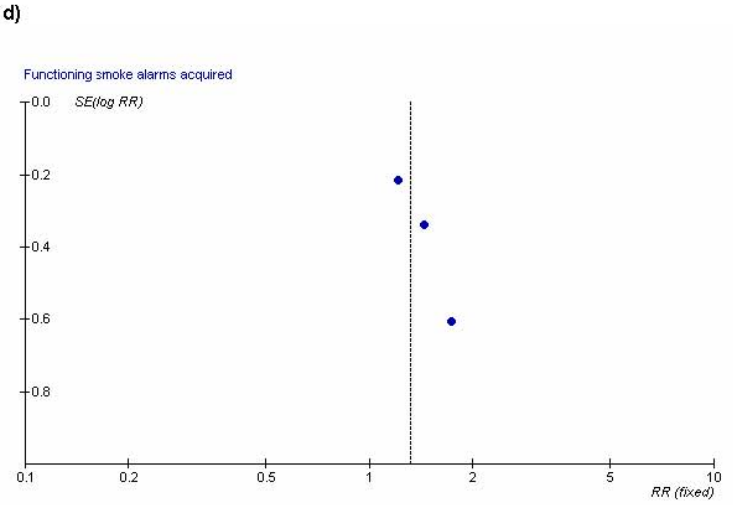
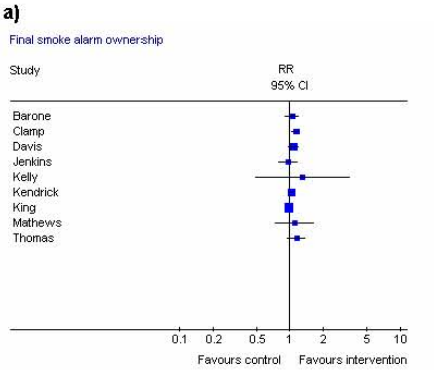
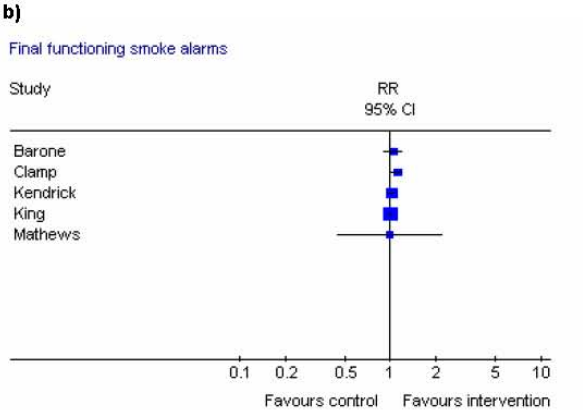

Functioning smoke alarms acquired
0.0
SE(log RR)
0.2
0.4
0.6
0.8
0.1
0.2
0.5
1
2
5
10
RR (fixed)

Figure 7: Forest plots (without pooled data) for each outcome

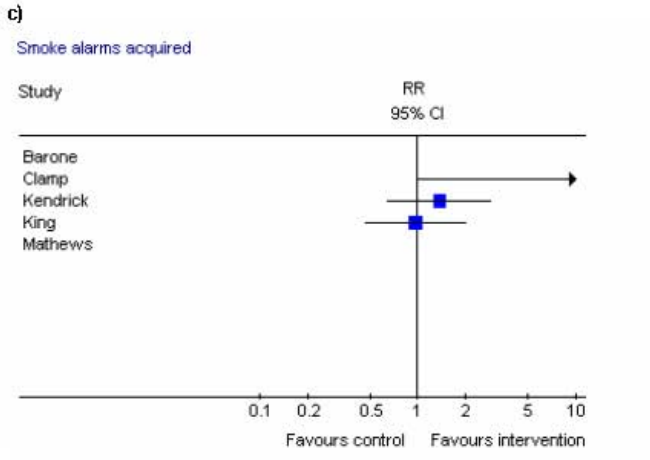
a)



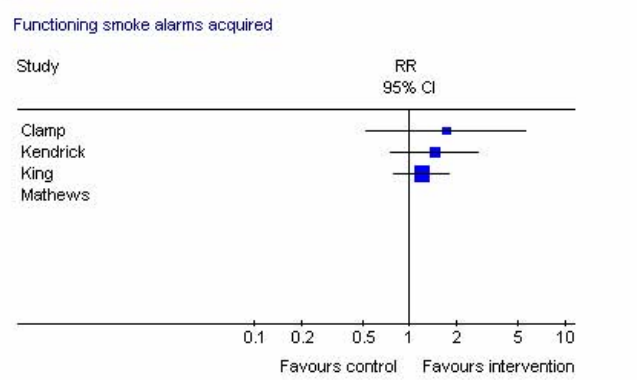
b)



c)


Smoke alarms acquired
Study
RR
95% CI
Barone
Clamp
Kendrick
King
Mathews
0.1
0.2
0.5
1
2
5
10
Favours control
Favours intervention

d)


Functioning smoke alarms acquired
Study
RR
95% CI
Clamp
Kendrick
King
Mathews
0.1
0.2
0.5
1
2
5
10
Favours control
Favours intervention

4.5 Assessing the robustness of the synthesis

Tools and techniques described in this section are presented in the table below.

Table 7. Selection of tools and techniques to assess the robustness of the synthesis

Name of tool/technique	Thoughts/ideas/comments in relation to current synthesis	Should this tool/technique be applied here?
Best evidence synthesis	Not really appropriate since this technique is primarily concerned with the selection of studies, and all studies in this synthesis are RCTs	No.
Use of validity assessment (EPPI-centre approach, CDC approach)	EPPI approach may be possible, using internal validity data presented in the summary tables. CDC approach needs further clarification before it could be applied (e.g. what is a "sufficient" effect size?)	Yes (EPPI). No (CDC).
Checking the synthesis with authors of primary studies.	Not possible given the time available for this synthesis.	No.
Reflecting critically on the synthesis process	Although partly done throughout this process, it might be useful to have a dedicated section discussing issues that arose from the synthesis.	Yes.

Strength of evidence (EPPI approach)

The guidance states that "four criteria are used to appraise each study: (1) the study's methodological soundness, (2) the appropriateness of the study design to answering the review question, (3) the study relevance, and (4) an assessment of the overall weight of evidence which the study provides. The first three criteria contribute to the assessment of (4) study "weight". These are described elsewhere by EPPI review authors as (1) Trustworthiness, (2) Appropriateness, (3) Relevance, and (4) Overall weight. An attempt was made to tabulate these characteristics for the studies included here (table 8), with criterion (1) based upon the validity evaluations in the first data extraction table (as these are derived from the Jadad scale, scores of 3-5 are considered 'high' quality. In this example, a score of 2 was described as 'medium' and a score of 0 or 1 as 'low').

Of the ten studies of children or their families, three received an overall 'high' weight, five were classified as 'medium' and two were given an overall weight of 'low'. These 'overall weights' corresponded exactly to the 'trustworthiness' scores that relate to internal validity. This is because there was little to distinguish between the studies in terms of appropriateness (all were RCTs – a design appropriate to this kind of evaluative research) and relevance (studies were selected for relevance early in the review by the application of inclusion criteria). The only study that was not considered 'highly' relevant in its focus was by Davis, as this was delivered to exclusively to schoolchildren, whereas other studies involved parents in the intervention.

It is possible that these 'overall weights' overemphasise the differences between the included studies. All of the studies scored 1, 2 or 3 on the Jadad scale and were consequently labelled 'low', 'medium' and 'high' respectively. All of the studies described themselves as RCTs, and (partly because of the nature of the intervention) none were double-blind or used an indistinguishable control intervention. Therefore, overall study weighting was dictated solely by whether the studies included descriptions of allocation concealment and/or withdrawals.

Three studies received an overall weighting of "high" (Clamp, Kendrick and King).[55, 58, 60] However, these were conducted in different settings and, for final smoke alarm ownership, reported differing results from one another. Consequently, this quality assessment approach does not greatly impact on the current synthesis, though could prove more useful in syntheses where there is greater variation in the quality of the studies being synthesised.

Table 8: Weighting of studies by quality, according to four criteria

Study	A Trustworthiness	B Appropriateness	C Relevance	D Overall weight
Barone (1988)	Medium	High	High	Medium
Clamp (1998)	High	High	High	High
Davis (1987)	Medium	High	Medium	Medium
Jenkins (1996)	Medium	High	High	Medium
Kelly (1987)	Medium	High	High	Medium
Kendrick (1999)	High	High	High	High
King (2001)	High	High	High	High
Mathews (1988)	Low	High	High	Low
Thomas (1984)	Low	High	High	Low
Williams (1988)	Medium	High	High	Medium

4.6 Reflecting critically on the synthesis process

Methodology of the synthesis used

There were some limitations to the approach taken in this synthesis, relating to the potential for bias. For example, the selection and arrangement of intervention components included in the moderator table was to some extent subjective. Similarly, the themes emerging from the textual descriptions that seemed most important were chosen at least partly subjectively. This is another argument for 'down-weighting' conclusions based on moderators only identified through extensive examination of the primary studies.

In the case of this particular synthesis, only RCTs were included. Subsequently, there was less methodological heterogeneity than in many narrative syntheses. This precluded the use of several techniques (although it is unlikely that any synthesis would be able to make use of *all* the tools and techniques described in the guidance). As all the studies were RCTs, the techniques that were appropriate were often variations on those used when undertaking a meta-analysis. It also meant that the variation in quality between studies was relatively small and difficult to incorporate usefully into the synthesis.

4.7 Conclusions

Interventions that provide safety information directly to families of young children appear in general to have a small beneficial effect on smoke alarm ownership and function. No conclusions can be made about the effect of such interventions in terms of fire-related injury or burn prevention, as these outcomes were not reported separately. It is unclear from the synthesis of RCTs presented here how specific fire-related safety education compares with general safety advice. Neither is there a clear relationship between the incorporation of home inspections or discount devices/coupons and the effect of interventions on smoke alarm ownership/function.

However, examination of the studies indicated several implications for the conduct of research in this area:

Implications for research on smoke alarms

- Future RCTs of similar interventions should measure relevant fire-related injury and burn outcomes after an appropriately long follow-up, preferably from hospital record review or similar method that reduces the potential for bias and attrition inherent in the questionnaire methods employed in several of the currently published RCTs.
- Any future studies should provide full and detailed descriptions of the intervention being evaluated and each of its components.
- Theory should be incorporated into the design and evaluation of any such intervention. Those designing evaluations of this type of intervention should consider the causal pathways between providing the intervention and the outcomes, and the barriers to its

adoption, and ensure that data is provided on each of the steps (or events) in the pathway.

- Randomised studies should take into account confounding due to concurrent community-wide initiatives and legislation to increase fire injury awareness and smoke alarm ownership.
- The rates of smoke alarm ownership at baseline might be investigated as a potential variable that influences intervention effectiveness within the target population.
- The only studies with negative findings in this synthesis were those in which participants were children, or the families of children that had been hospitalised for an injury. Whether this was a chance finding or indicative of a true lack of effect for these interventions in families of previously injured children may be of interest.

Comparison of narrative synthesis and meta-analysis

The Cochrane review,[51] based on the meta-analysis of the same group of RCTs, reached very similar broad conclusions to the narrative synthesis. The Cochrane authors reported that fire-related injury outcomes were not available and the main meta-analyses of RCTs showed that "smoke alarm ownership at follow-up appeared somewhat more likely in the intervention group (OR = 1.26; 95% C.I., 0.87 to 1.82). Similarly modest positive, statistically non significant effects on functioning smoke alarms, and on new acquisitions of smoke alarms and functioning smoke alarms, were found". They summarised that there were "only modest potential benefits from education to promote smoke alarms".

As in the narrative synthesis, the apparent lack of effect of intervention in the two trials involving families of injured children was noted in the Cochrane review. The Cochrane authors state that "exclusion of these trials from the meta-analyses results in a stronger, statistically significant intervention effect on alarm ownership (OR = 1.43; 95% C.I., 1.07 to 1.90) and other alarm outcomes". In the narrative synthesis, this was considered an area of potential interest for future research. The Cochrane authors suggest in the discussion section of their review, "Having an injured child may lead to safety behaviour changes so large that they obscure any safety education effects", but they do not mention this as one of their implications for future research.

In addition, the Cochrane review concluded that smoke alarms delivered as part of child health surveillance may be more effective. The effects on final smoke alarm ownership were statistically significant (OR = 1.96; 95% C.I., 1.03 to 3.72), with strong, non-significant effects on the other ownership and function outcomes. The authors state that these subgroup analyses were based on few trials and were heavily influenced by a single trial.(Kendrick)[55]

The relationship between offering discount devices/coupons and the effect of interventions on smoke alarm ownership/function was not obvious in the narrative synthesis. The results of a subgroup meta-analysis suggested that offering discounted alarms had a modestly stronger effect on smoke alarm ownership (OR = 1.83; 95% C.I., 0.63 to 5.28) than did education alone, but the trial results were significantly heterogeneous (p=0.015). Another subgroup meta-analysis indicated that the removal of the one study in which a research assistant delivered the intervention[60] resulted in a stronger positive effect of intervention on three of the reported outcomes.

The Cochrane authors concluded that the quality of the available evidence is limited, with sensitivity analyses showing that pooled trials with blinded outcomes assessment indicated little apparent effect on ownership or function, whereas unblended studies indicated strong effects.

Recommendations derived from the meta-analyses were similar to those in the narrative synthesis: "Further trials to evaluate the effect of smoke alarm promotion as part of child health surveillance in primary care... should assess their impact on fire-related injuries, using adequate allocation concealment and blinded outcomes assessment". These recommendations did not stretch to improvements in outcome measurement, description of interventions, use of theory in designing interventions, or adjusting for potential confounding from concurrent fire safety initiatives/policies, as they did in the narrative synthesis.

On the whole, the findings of the narrative synthesis and the meta-analyses were very similar. However, the differences mentioned above appear to be attributable to two main factors: the

possibility of undertaking sensitivity and subgroup pooled analyses in the meta-analyses and the close scrutiny of studies undertaken in the narrative synthesis. Consequently, conclusions about the impact of moderators on effect appeared to be 'firmer' when derived from the meta-analysis than from narrative synthesis, whereas implications for future research appeared to be more extensive and detailed when derived from the narrative synthesis. However, the Cochrane review authors mention caveats in relation to some of the 'additional' findings derived from subgroup analyses (e.g. that the apparent increase in effect attributable to offering discounted alarms was based on a meta-analysis of highly heterogeneous studies). Meta-analysis allowed the authors of the Cochrane review to observe the impact of specific aspects of study validity (allocation of concealment and blinded outcome measurement) on results. In the narrative synthesis, validity was considered more broadly and showed no obvious correlation with study results. Although the differences in the conclusions of the two syntheses were relatively minor, it is unclear whether it would be possible to eradicate them altogether, considering that by definition narrative synthesis precludes statistical pooling.

The process of comparing these two syntheses has highlighted the important contribution that the guidance makes to increasing the transparency of the narrative synthesis approach. As a reader, it was possible to check the conclusions derived from the narrative synthesis by examining the synthesis itself and the associated tables and figures, much as it is possible to examine and interpret data presented in a series of forest plots.

CHAPTER 5: *APPLYING THE GUIDANCE 2:* A NARRATIVE SYNTHESIS OF STUDIES INFORMING THE IMPLEMENTATION OF DOMESTIC SMOKE ALARM PROMOTION INTERVENTIONS.

5.1 Introduction

This chapter, like the previous one, provides a practical example of a narrative synthesis. In this case, however, the focus is on the synthesis of evidence on factors influencing the ***implementation of interventions*** rather than effectiveness. The specific aims of the chapter are to:

- Illustrate in practical terms the decision making processes involved in the application of the guidance to a specific narrative synthesis
- Identify factors that should inform choices about the use of particular tools and techniques in the context of a specific synthesis
- Provide examples of how particular tools and techniques can be used in the synthesis of evidence on the implementation of specific interventions
- Demonstrate the type of outcomes achieved by a narrative synthesis
- Comment on the way in which the present synthesis compares with the earlier work on which we drew

The evidence on implementation synthesised in this chapter is drawn from an earlier, exploratory review of evidence on the implementation of interventions aiming to promote the use and functioning of domestic smoke alarms and broader community based injury prevention interventions.[61] The focus of the narrative synthesis reported here has been restricted to domestic smoke alarm interventions. A new search was conducted to update the studies identified for the original review and a purposive sample of seven papers judged by the review team to provide relatively rich data on factors influencing the implementation of the interventions was selected for inclusion in the narrative synthesis. Details of the seven papers, and the interventions reported on, are shown below in Table 9.

Table 9. The seven papers

Paper/report authors	Location of trial/initiative	Type of trial/ initiative	Type of intervention
1. Campbell DeLong Resources Inc. (2003)[62]	Oregon, USA	Non RCT	Fire safety awareness campaign (aimed primarily at landlords)
2. Camit, M. (1998)[63]	New South Wales, Australia	Non RCT	Provision of discounted smoke alarms, plus written fire safety information and alarm demonstration
3. Camit, M. (2002)[64]	New South Wales, Australia	Non RCT	Provision of discounted smoke alarms, plus written fire safety information and alarm demonstration
4. DiGuiseppi, C., Slater, S., Roberts, I., Adams, L., Sculpher, M., Wade, A. and McCarthy, M. (1999)[65]	London, UK	RCT	Provision and installation of free smoke alarms, plus written fire safety information
5. McConnell, C.F., Dwyer, W.O. and Leeming, F.C. (1996)[66]	Memphis, USA	Non RCT	Fire safety education (including use of video, written and other material)

6.	Roberts, H., Curtis, K., Liabo, K., Rowland, D., DiGuiseppi, C. and Roberts, I. (2004)[67]	London, UK	RCT	Provision and installation of smoke alarms, plus written fire safety information
7.	Young, M., Camit, M. and Mihajlovic, M. (1999)[68]	New South Wales, Australia	Non RCT	Provision of discounted smoke alarms, plus written fire safety information and alarm demonstration

Before describing the NS work undertaken, we first consider the nature of the research evidence likely to be included in reviews of factors shaping the successful implementation of particular interventions.

5.2 A note on extracting data on implementation.

There are at least four issues that reviewers have to deal with when extracting data on implementation from reports of interventions. These are:

- Locating the data in the text
- Establishing the nature and type of implementation data
- Ascertaining their provenance and reliability
- Extracting the data in preparation for analysis

In relation to the first, it should be noted that reports of interventions tend not to contain a great deal of detail on process issues - implementation is rarely the focus of reports of interventions. This may be because interventions are often reported in journal papers, and there is not the space to describe the factors affecting implementation. If there are no additional documents available describing the implementation of an initiative some of these data may be found interspersed throughout the text, but are more likely to be located predominantly in the discussion section of papers, where authors usually attempt to provide an explanation for the effectiveness or otherwise of an intervention.

Once data have been identified the reviewers need to establish the nature and type of implementation data found in the text. Most data of this kind consist of simple narrative observations made by authors of the factors affecting the campaign or initiative. It is relatively rare for primary data (direct quotations from users, for instance) to be found to support these observations.

Ascertaining the provenance and reliability of data on implementation can be the most difficult and frustrating part of the process of extracting data on implementation. It is rare to find details of where these data (either in the form of narrative author observations or quotations) come from, or what they are based on. Are authors drawing on fieldworkers' observations? Or unreported focus group findings? Are these just 'hunches'? Where data are 'thicker' (perhaps where qualitative data have been presented) additional maybe unused and/or unreported data on implementation may be available.

Finally, how should those attempting a synthesis extract the data in preparation for analysis? Where authors have made simple observations about implementation, these can be taken whole from the text (and the presentation/reporting of this will depend on the number and complexity of such observations). Quotations or other qualitative data can also be taken directly from the text and the data entered into tables with appropriate headings (for example: author/paper; location of implementation data in the report; type and nature of data).

5.3 The narrative synthesis

The guidance is structured around four elements in the synthesis process:

- Developing a theory of how the intervention works, why and for whom
- Developing a preliminary synthesis
- Exploring relationships within and between studies
- Assessing the robustness of the synthesis

In the NS reported here no *prima facie* attempt was made to develop a theoretical basis for the work so this element of the guidance was not applied. The NS was carried out separately by two different members of the study team, each using the draft guidance to test its usefulness and clarity. One reviewer worked with all seven papers, and the other worked with only three papers.

The guidance presents a number of related tools and techniques that can be used to complete the various elements of the synthesis. To apply the guidance to the narrative synthesis of implementation evidence, each of the sections was read through in sequential order, and for each element the tools and/or techniques that appeared to be useful and relevant to the synthesis at hand were selected. The results of this exercise are shown in Table 10. The tools and techniques were either used in the worked examples (Yes), not used and probably not relevant (No), or not used but potentially relevant (Potentially yes). The reasons for selecting or rejecting a tool or technique are discussed in more detail within each section. Where tools or techniques proved to be less useful, this is also discussed.

Table 10. Tools and techniques for the implementation syntheses

Tools and techniques	Useful for NS of implementation data?
Preliminary Synthesis	
Textual descriptions	Yes
Tabulation	Yes
Groupings and clusterings	Yes
Constructing a common rubric	No
Thematic analysis	Yes
Content analysis	No
Vote counting	No
Exploring relationships	
Variability in outcomes	Potentially yes
Variability in study design	Potentially yes
Variability in study population	Potentially yes
Moderator variables and subgroup analyses	No
Idea webbing and subgroup analyses	Yes
Conceptual triangulation	Potentially yes
Translation	Yes
Case descriptions	Potentially yes
Visual representation	No
Investigator triangulation and methodological triangulation	Potentially yes
Assessing robustness	
Weight of evidence	Potentially yes
Best evidence synthesis	Potentially yes
Checking with authors	Potentially yes
Critical reflection	Yes

In the following sections, we illustrate the practical application of the various tools and techniques discussed in earlier chapters providing a worked example of a narrative synthesis of data on the

factors affecting the implementation of interventions aiming to increase the uptake of domestic smoke alarms. A flow chart summarising the synthesis process as a whole is presented in figure 8 below. Some of the worked examples have been limited by the relatively small size of the evidence base as several of the tools and techniques are only relevant when synthesising a large body of literature.

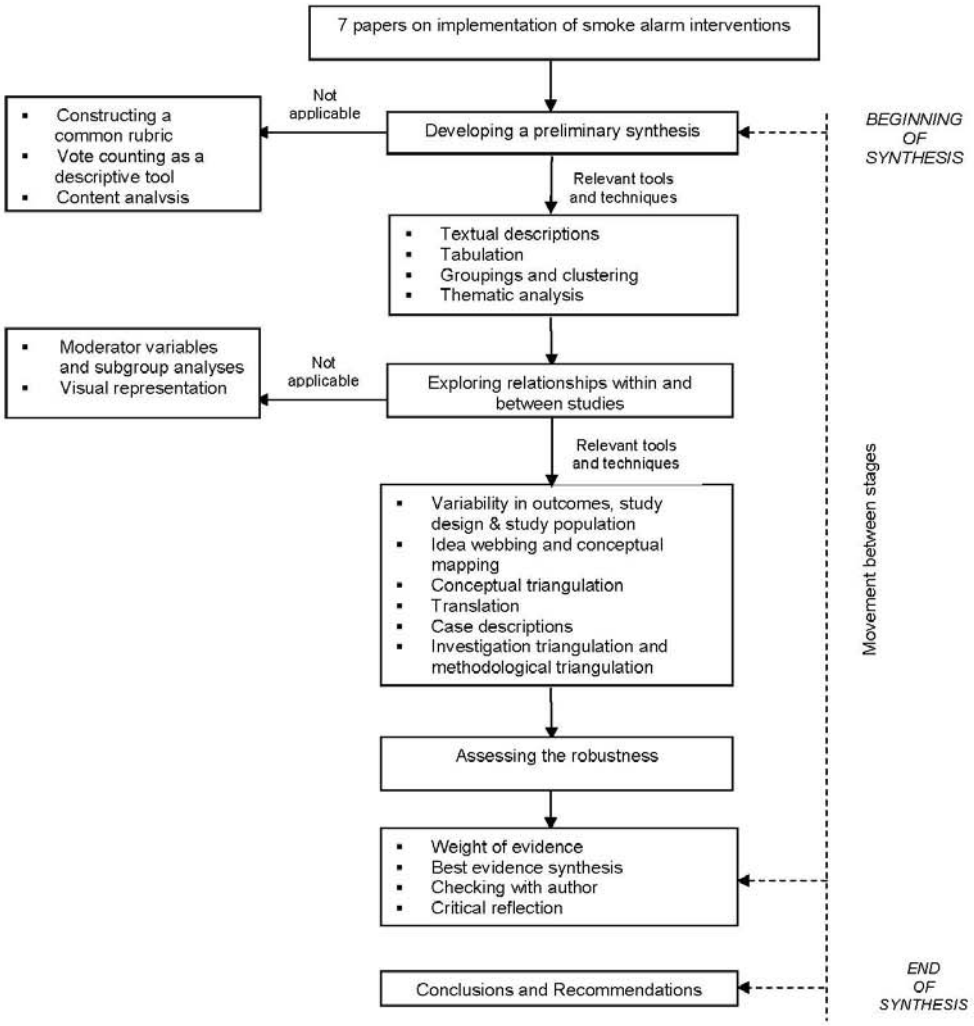


Figure 8: Synthesis process

Developing a preliminary synthesis

Textual description

This was used by both reviewers at an early stage, and was found useful as a way of summarising the papers and beginning to extract information in a systematic way. Textual descriptions offer the potential to include more details than, for example, tabulations.

Examples of textual description:

Example 1

> In McConnell et al,[66] the **target population** was new heads of households in public housing residences of the Memphis Housing Authority (MHA), USA and they were predominantly female Afro Americans living with children. The MHA policy is to ensure that a functioning smoke detector is in every unit when rented, but a spot check of 325 units in 1992 found that less than 8% had a working smoke detector. The 35 minute **intervention** (delivered during mandatory orientation sessions for new MHA heads of household) consisted of the following components: a pre test; videotape accompanied by lecturettes delivered by one of 36 uniformed fire fighters, one MHA supervisor or one civilian educator; behavioural contract, post test, and fire-safety reminder card. The **outcomes** were fire incidence data (after possibly 15 months, timescale not clear); residents' evaluations of the programme; changes in their fire safety knowledge; and their commitments to fire safety behaviours. The method of **evaluation** was an uncontrolled comparison between trained and untrained residents, using contemporary and historical comparison groups. The evaluation **data** were all quantitative. The **results** showed a lower incidence of fires in trained residents compared with untrained residents (1 fire for every 4312 renter months in trained residents compared with 1 fire for every 780 renter months in untrained residents; a relative risk of 5.5). Comparing trained residents with untrained residents over the 9 year baseline period gave a relative risk of 4.8. Comparisons between newer and older residents from the MHA records suggested that newer residents were more likely to experience fires, thus countering the suggestion that the results can be explained by the fact that the trained residents were also new residents. No data were provided on the proportion of working smoke detectors post intervention.

Example 2

> Young et al (1999),[68] Camit (2002)[64] and Camit (1998)[63] report on the effectiveness and implementation of a smoke alarm promotion campaign in NSW Australia oriented to the needs of Arabic, Chinese and Vietnamese communities. Qualitative data were collected in focus groups and interviews. Survey data were also collected. Their main observations in relation to implementation are that among the target community there was a lack of awareness of the need for smoke alarms. Living in rented property where the landlord was thought to be unsympathetic to the need for a smoke alarm also created barriers to the installation of smoke alarms.

Tabulation

Both reviewers felt that tabulation and textual descriptions were very similar, possibly using the same headings but laid out differently. In a table, however, it was easier to compare data across different studies.

Table 11. Example of tabulation

Author & year	Location & setting	Target population	Method	Main findings
Roberts et al (2004)[67]	London, UK Urban	All residents on estate (n=40000 households)	Focus groups and interviews	Problems with smoke alarms (sensitivity, false alarming) identified as major barriers to implementation
Camit (2002)[64] Young et al (1999[66]) Camit (1998)[63]	NSW, Australia Mixed	Chinese, Vietnamese, Arabic-speaking (not given)	Focus groups	Implementation successful using multi-faceted, language appropriate approach
Campbell De Long Resources (2003)[62]	Oregon, USA Mixed	All residents, but focus on Latino-speaking (sample population varied according to element of intervention)	Interviews	Successful implementation heavily dependent on landlords' attitudes

Groupings and clusters

This technique is more useful when there are larger numbers of papers. The type of groups identified is likely to depend on the reviewers categorisations, but may also depend on the document type (is it describing a trial or a campaign, for example). The result is similar to tabulation, except that the relationships (or differences) between groups of studies can be made more explicit when they are clustered in this way.

Table 12. Example of grouping

Grouping according to:		
Location	***Focus of report***	***Population***
UK (DiGuiseppi et al 1999; Roberts et al 2004)	Broad, general factors affecting programme (DiGuiseppi et al 1999; Camit 1998; Camit 2002)	Ethnically mixed (Camit, 1998; Camit, 2002; Young et al, 1999)
USA: (Campbell De Long Inc 2003; McConnell et al, 1996)	Individual-factors affecting programme (Young et al, 1999; McConnell et al 1996; Roberts et al, 2004; Campbell de Long Inc 2003)	Ethnically mixed and low income (Campbell de Long Inc, 2003; McConnell et al 1996; DiGuiseppi et al 1999; Roberts et al 2004
Australia: Camit 1998; Camit 2002; Young et al 1999)		

Transforming the data: constructing a common rubric

This technique was not found useful by the reviewers, as there was no unit of measurement that could form the basis of a common rubric. The data from the intervention studies in this review (installation rates of purchased smoke alarms, fire incidence data, and percentages of working smoke alarms) did not lend themselves to transformation in this way, though it may be feasible for other reviews where the outcomes are more homogeneous.

Translating data: thematic analysis

Different reviewers, and the same reviewer at different times, identified different themes. These themes were identified on an inductive basis by reading and re-reading the papers. They fell into two categories: those aspects of the interventions that seemed to act as barriers or facilitators, in the view of the author or the reviewer; and themes identified by the authors from their qualitative data.

Example of thematic analysis

Example 1

- The smoke alarm
- The individual and
- The community

Table 13: Example 2

	McConnell[66]	Young[68]	Roberts[67]
Facilitators	**Aspects of intervention**: Mandatory orientation sessions Formal written commitment Involvement of fire service Tailored intervention; involvement of residents in its design Reminder card Penalties for arson and disabling smoke detectors explained Testing of knowledge Smoke detectors already installed Commitment of landlord (MHA)		Early warning in the event of fire and increased sense of security Feeling at high risk for fire Fitting of the smoke alarms by installers
Barriers		**Lack of awareness** Lack of awareness of the need for smoke alarms Unsympathetic landlords Tendency to overestimate installation costs Underestimation of fire risk Lack of awareness of danger of smoke Perception that fire is a hazard only for wooden homes Not relating potential benefits to their own circumstances **Rented accommodation** Landlords holding residents liable for installation damage Frequent moving Having to leave alarms behind when moving Landlords withholding permission Difficulty of applying to estate agents **Installation** Lack of time	Concerns about strangers entering their home Suspicion of an intervention provided free Feeling oneself "too old to be worth spending money on" Alarms as a source of stress Problems with maintenance Alarm sensitivity Alarms as nuisance Alarms as a threat to immediate well-being

Table14: Example 3

- Barriers/levers to the acquisition of smoke alarms
- Barriers/levers to the installation of smoke alarms
- Barriers/levers to the continued use of smoke alarms

1) Barriers/levers to acquisition of smoke alarms		
General		
	Barriers	*Levers*
	Problems accessing communities/gatekeepers	Gaining trust of key community 'players' and leaders
	Suspicion of 'authority' or local government	Emphasising separation from distrusted authority/alliance with trusted partners
Specific to smoke alarm campaigns		
	Lack of awareness of benefits of smoke alarms	Running well-coordinated, culturally appropriate awareness campaign
	Perceived cost of smoke alarms	Giveaway or availability of reduced price alarms
	Perception that household is not at risk of fire (due to type of house or characteristics of household members)	Awareness campaign
2) Barriers/levers to installation of smoke alarms		
General		
	Barriers	*Levers*
	Anxiety about damage to property	Landlord approval/permission for installation, or landlord example of installation
Specific to smoke alarm campaigns		
	Inability/unwillingness to install alarm	Installation of alarm by project worker
3) Barriers/levers to continued use of smoke alarms		
	Barriers	*Levers*
	False alarms	Education about triggers for false alarms/re-installation of alarm
	Problems with maintenance	Project workers offer to maintain alarms/education about maintenance

The differences over time or between reviewers do not suggest that the synthesis is flawed but rather draws attention to different ways of interpreting the same data. Both reviewers identified a typology including facilitators and barriers as did some of the study authors. Whether the data are seen ecologically or in stages, the idea of barriers and facilitators are common to both. In a final synthesis, specific factors that act as barriers/levers, the notion of stages (temporality) and the organisation of these factors within domains at different levels (ecological perspective) could be brought together.

Translating the data: Content analysis

This technique was not found useful by either reviewer as the data did not lend themselves to conversion into frequencies.

Vote counting as a descriptive tool

Similarly, this technique was not found useful by either reviewer. The description of vote counting in the guidance focuses primarily on effectiveness reviews. Although in theory it could be used to 'count' up facilitators and barriers it would probably only be appropriate with a larger number of studies included in the review and reviewers would need to be aware of the disadvantages of using this technique inappropriately as discussed earlier in this guidance.

Exploring relationships within and between studies

Exploring the influence of heterogeneity[69]

Only one reviewer attempted to use this technique, which focused attention on the characteristics of the different studies and their potential relationships to the findings. The other reviewer did not find this technique useful.

Example of variability in outcomes

> The quantitative outcomes of the three studies relate to three different stages of the trajectory of a domestic fire: installation of a smoke alarm; continued working of a smoke alarm; incidence of domestic fires. In the Young paper, installation rates varied from 65% to 35% after 10-15 days (with some suggestion that these rates rose after 2 months).[68] Given that the installation rate in the Roberts paper[67] was 100% for all respondents, but the proportion of working alarms at 15 months was only 51%, this suggests that the proportion of working alarms in the Young study would have been even lower. It seems unlikely that these proportions of working alarms would have produced the relative risk results of the McConnell paper (assuming of course that their relative risk figures are robust). Given that the McConnell paper[66] did not measure the proportion of working alarms in their trained and untrained groups, this is speculative; it may be the case that in such environments, low proportions of working alarms can deliver measurable improvements in the incidence of domestic fires. This is supported by McConnell's data on pledges made about fire safety behaviours: only 24% of the pledges made by trained residents were commitments to keep smoke detectors in working order. However one of the five points on the reminder card in the McConnell paper was "Keep your smoke detector working and check it often". I would have to look beyond these papers to try and assess the comparability of the outcomes of these papers. My own working assumption is that the intervention evaluated in the McConnell study was more successful, perhaps because it showed an improvement in the target (fire incidence), as well as proxy (residents' fire safety knowledge), outcomes.

Example of variability in interventions

> The interventions were different: an educational intervention for residents with already installed smoke alarms (McConnell); the sale of smoke alarms (Young); and the free installation of smoke alarms (Roberts). However one aspect of the interventions is potentially explanatory. The intervention in the McConnell study was drawn up following focus groups and individual residents in which they were asked about the best approaches to be used; the Roberts paper concluded that "It may well be that a more appropriate design would have started with qualitative work, improving take up and maintenance instructions". The Young study was not drawn up in consultation with the local community. Thus the involvement of the local population and the development of a

tailored intervention may well have influenced the success of the McConnell study. The 5.5 minute videotape in the McConnell study "depicted actual MHA residents, apartments and structures" and was described by MHA residents who previewed it as "extremely powerful and likely to make a lasting impression". The McConnell intervention was underpinned by a model of community psychology, the goal of which is "to optimise the well being of communities and individuals with innovative and alternative interventions designed in collaboration with affected community members and with other related disciplines inside and outside of psychology". Other components of the McConnell intervention seem likely to have influenced its success: the mandatory orientation sessions which meant that all new residents received the intervention whether they wanted to or not; the formal written commitments obtained at the end of the educational session; the involvement of the fire service who delivered some of the interventions themselves; and the reminder card.

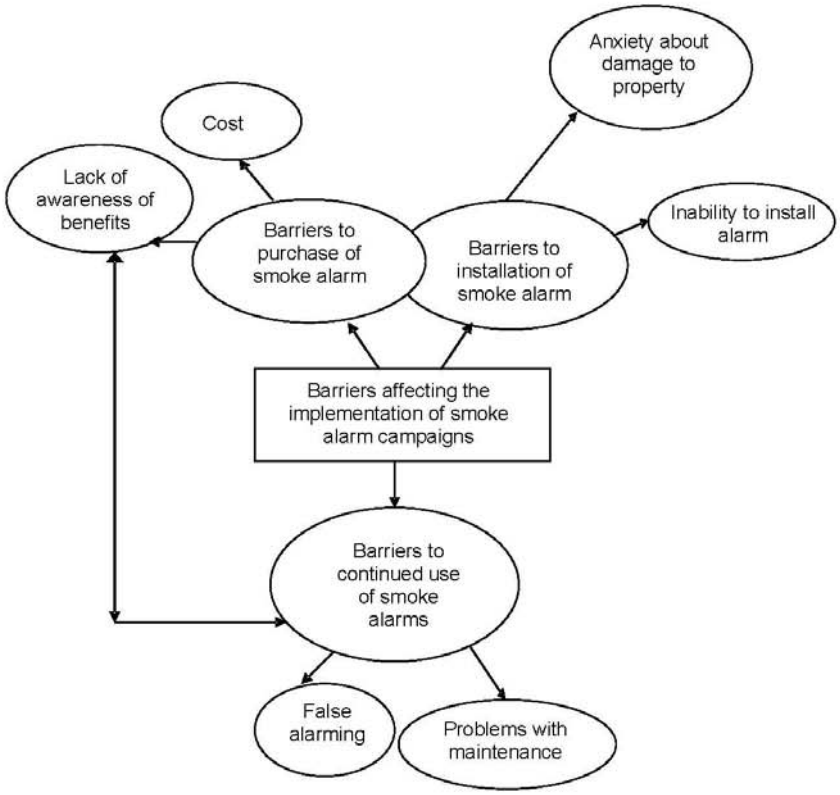
Moderator variables and subgroup analyses
These were not found useful by either reviewer, as they are primarily helpful in effectiveness syntheses.

Idea webbing and conceptual mapping
We found it impossible to distinguish between these two techniques and so have combined them. The use of figures/diagrams was found to be helpful for exploring issues on implementation. One reviewer does not usually use these kinds of visual methods and only did so to test the guidance; but then felt that it had helped her to think about the relationships between the themes rather more than she would have done otherwise.

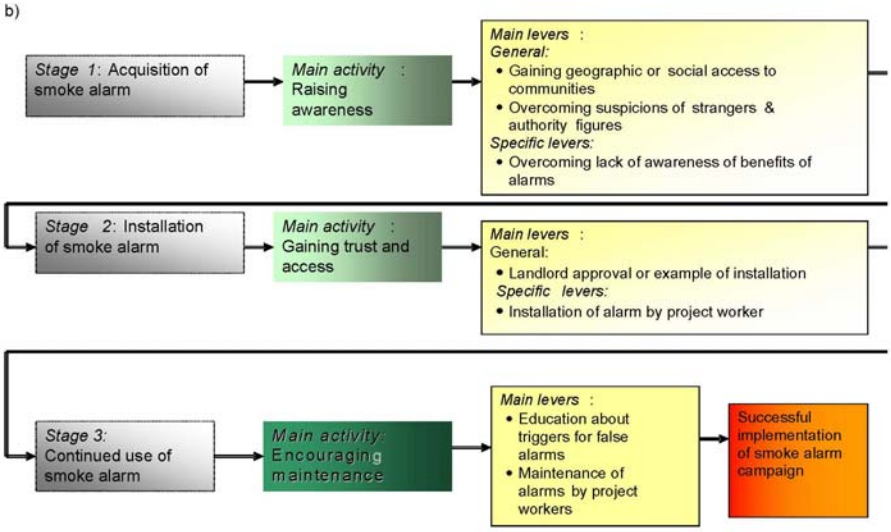
Much depends on how complex and detailed the figures are, and how easy they are to 'read'. There may be too much information in them (lots of bubbles or balloons making it hard to see what the message is). The structure of the figures tended to replicate themes identified by the thematic analysis. In a sense these figures represented an early product of the synthesis. It was clear that at this stage they were specific to smoke alarms and could not necessarily be applied to other types of intervention.

Figure 9. Examples of idea webbing

a)

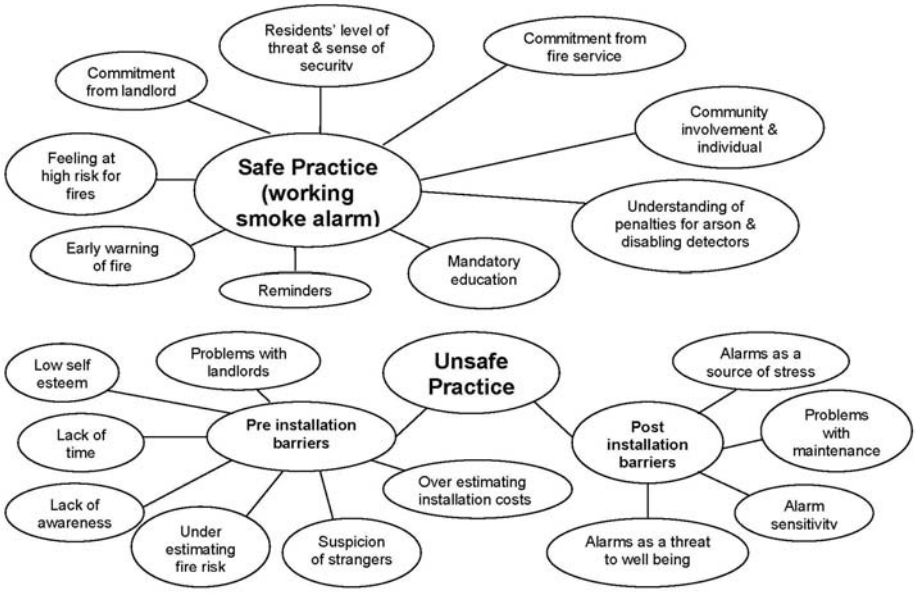


b)


Stage 1: Acquisition of smoke alarm
Main activity : Raising awareness
Main levers :
General:
• Gaining geographic or social access to communities
• Overcoming suspicions of strangers & authority figures
Specific levers:
• Overcoming lack of awareness of benefits of alarms
Stage 2: Installation of smoke alarm
Main activity : Gaining trust and access
Main levers :
General:
• Landlord approval or example of installation
Specific levers:
• Installation of alarm by project worker
Stage 3: Continued use of smoke alarm
Main activity: Encouraging maintenance
Main levers :
• Education about triggers for false alarms
• Maintenance of alarms by project workers
Successful implementation of smoke alarm campaign

62

c)


Residents' level of threat & sense of securitv
Commitment from fire service
Commitment from landlord
Community involvement & individual
Feeling at high risk for fires
Safe Practice (working smoke alarm)
Understanding of penalties for arson & disabling detectors
Early warning of fire
Mandatory education
Reminders
Low self esteem
Problems with landlords
Unsafe Practice
Alarms as a source of stress
Pre installation barriers
Post installation barriers
Problems with maintenance
Lack of time
Over estimating installation costs
Lack of awareness
Alarm sensitivitv
Under estimating fire risk
Suspicion of strangers
Alarms as a threat to well being

63

Conceptual triangulation
Neither reviewer found this technique useful for this exercise but felt that it would be useful with a larger number of studies.

Translation as an approach to exploring relationships
One reviewer with no previous experience of meta-ethnography attempted unsuccessfully to use it. The other reviewer, who does have previous experience of meta-ethnography, felt that translation was the most useful technique at this stage.

Example of reciprocal translation

> There are three concepts that seem to offer themselves for translation across studies. The **first** is the one of landlord commitment/lack of commitment. The difficulties with landlords discussed in the Young paper seem to be the exact opposite of the commitment demonstrated by the MHA in the McConnell paper, but not explicitly commented on by the authors. The **second** concept is risk perception: feeling oneself at high risk (Roberts), or underestimating the risk of fire (Young). The McConnell intervention presumably increased residents' estimates of their own risk but no information is provided about this. The **third** and less robust concept is residents' level of trust (or something like that). In the McConnell paper, residents were involved in the development of the intervention and presumably this generated a certain amount of trust (even though the education was mandatory). In the Roberts paper, the fact that some residents were uncomfortable with strangers coming into their homes to fit smoke alarms, and suspicious of anything offered for free, suggests a lack of trust. These three concepts may just represent the potential for translation between the three studies. However they may also start to characterise the elements necessary for a successful intervention: landlord commitment, risk perception, and residents' level of trust.

This kind of translation could eventually lead to an explanation or theory, which might form one output of the synthesis and which might itself inform future interventions.

Qualitative case descriptions
Neither reviewer found this useful, partly because there was little practical advice about how to use it and partly because it appeared similar to the earlier technique of textual description. It could be useful in building up some kind of composite picture of successful interventions and providing the kind of detail that could be useful for those wanting to design interventions themselves.

Visual representation of relationships between study characteristics and results
These techniques seemed to be appropriate for statistical data rather than the kinds of data in the implementation studies.

Investigator triangulation and methodological triangulation
Neither reviewer used these techniques due to lack of sufficient data in the primary studies, but both felt that they could be useful if there were more data available.

Assessing the robustness of the synthesis
Comparison with earlier review
The results achieved using the guidance were compared to the earlier review referred to at the beginning of this chapter.[61] This involved a simple thematic analysis of the data, with no attempt at synthesis using other tools or techniques. The conclusions identified features present in papers containing a 'thicker' description of implementation processes. The authors concluded that where interventions had been successfully implemented, the programmes were likely to have the following features:

- A relatively detailed description of the intervention, its strengths and weaknesses and its suitability for the targeted population:
    - The type of intervention used, and its appropriateness for the target population, affects the outcome of trials and other initiatives.
- Some consideration of the context within which the trials take place.

    - This is frequently limited to a discussion of the problems encountered by those implementing the programme rather than those receiving it and rarely substantiated by reference to data collected during the trial (or, indeed, any other data). From a methodological perspective, this constrains the use of these insights.

- Some recognition of the discrepancy between the design and orientation of an intervention and its implementation in an everyday setting.
- Some exploration of the reasons for anomalous results and findings.
- Some description of the factors that affect implementation.
    - Includes: the importance of understanding the people and the community receiving the intervention; the need to consider the role of community leaders and other key local figures to programme success; recognition that the characteristics of the community affect programme success.

These are useful though rather general insights into factors affecting implementation. In fact, this is less a synthesis than a list of insights about implementation drawn from the papers/reports.

Weight of evidence
This was not used as no assessment was made of the quality of individual studies. This technique could be useful if there were more primary studies available for review.

Best evidence synthesis
Similarly, this technique was not used for the worked example but could be useful if there were more primary studies.

Checking with the authors of primary studies
This is potentially useful if time allows, but depends on the accessibility and generosity of authors in providing further information. (Those we identified in the original study were helpful.) Given the general paucity of information relating to implementation of interventions in the demonstration studies, it would have been very useful to gain more details about the interventions and the contexts in which they were implemented.

Critical reflection
The synthesis of the implementation studies was based on a small number of studies, better suited for an exercise like this one where we are considering the processes of narrative synthesis than for providing a definitive synthesis product or "answer" (though it should be remembered that at present, many, perhaps most reviews are based on similarly small numbers of studies with similarly scant data on implementation).

By comparing the work of two reviewers we can examine inter-researcher differences. They used the different tools and techniques in parallel ways, not necessarily identifying precisely the same themes or concepts. Indeed, the reviewer working with all seven studies carried out two thematic analyses at different times, and identified different themes in her later reflections. The product of the synthesis will reflect the experience of the reviewers as well as their familiarity with tools and techniques. The reviewer with previous experience of meta-ethnography for example, found it easy to use the technique of translation to compare the concepts of different studies. No attempt was made in the demonstration project to achieve consensus between the reviewers, as we were interested in documenting their separate attempts and describing differences.

In comparison with the earlier exploratory review, the guidance enabled a more systematic overview of the different papers and a more nuanced appreciation of the evidence. As a justification for narrative synthesis is based in part on its claim to address the potential for bias, this demonstration review has shown that the use of specific tools and techniques can provide transparency of process. A more robust product is likely to be achieved if at least two reviewers work independently and then compare their findings to produce a mutually agreed (or a transparently divergent) final version.

CHAPTER 6: THE NEXT STEPS

Narrative approaches to synthesis are widespread in systematic reviews yet as we have noted these approaches do not rest on an authoritative body of knowledge. The guidance presented here has been developed on the basis of an extensive review of methodological literature and it has been applied to two contrasting bodies of evidence – one focusing on the effects of interventions to promote the use of domestic smoke alarms and the other focusing on evidence to inform the implementation of such interventions. In undertaking these demonstration syntheses detailed notes were kept of all major decisions taken and the reasoning behind them. This approach of prospectively documenting the synthesis process was a helpful aid to transparency and recall. We would recommend this to all reviewers adopting a narrative approach.

We do not claim to have produced the definitive guide to narrative synthesis – there is much work still to be done to develop and refine this approach to evidence synthesis. However, we do believe that the guidance offers both a general framework and specific tools and techniques that can help to increase the transparency and trustworthiness of systematic reviews involving narrative synthesis. We would also stress that whilst the guidance describes a range of tools and techniques that if used appropriately will improve the process of narrative synthesis these will not remove the need for reviewers to combine sound methodology with creative interpretative work.

We hope that people will find the guidance useful and that they will let us have feedback so that we can revise the guidance in light of this. The guidance is to be made available on the project website (address to be added) and comments can be sent by email to j.popay@lancaster.ac.uk

REFERENCES

1. NHS Centre for Reviews and Dissemination. *Undertaking systematic reviews of research on effectiveness: CRD's guidance for carrying out or commissioning reviews. CRD report 4.* 2nd ed. York: NHS Centre for Reviews and Dissemination, 2001.
2. Dixon-Woods M, Agarwal S, Young B, Jones D, Sutton A. *Integrative approaches to qualitative and quantitative evidence.* London: NHS Health Development Agency, 2004. Available from: http://www.hda.nhs.uk/documents/integrative_approaches.pdf
3. Human Resources at Massachusetts Institute of Technology. *Performance review forms [monograph online].* Human Resources at Massachusetts Institute of Technology; [cited Jan 2006]. Available from: http://web.mit.edu/hr/performance/forms.html.
4. Eisen GM, Weinberg DS. Narrative review: screening for colorectal cancer in patients with a first-degree relative with colonic neoplasia. *Ann Intern Med* 2005;143:190-98.
5. Higgins JPT, Green S. *Cochrane handbook for systematic reviews of interventions 4.2.5 [updated May 2005].* Available from: http://www.cochrane.org/resources/handbook/index.htm (accessed 31st May 2005)
6. EPPI Centre. *EPPI-Centre Review Group Manual: working document version 1.1 [monograph online].* London: EPPI Centre, 2001. Available from: http://eppi.ioe.ac.uk/EPPIWebContent/downloads/RG_manual_version_1_1.pdf
7. Weiss CH. *Evaluation: methods for studying programs and policies.* New Jersey: Prentice-Hall, 1998.
8. Shadish W. Meta-analysis and the exploration of causal mediating processes: a primer of examples, methods, and issues. *Psychological Methods* 1996;1:47-65.
9. Pawson R. *Does Megan's Law work? A theory-driven systematic review [monograph online].* London: ESRC Centre for Evidence Based Policy and Practice, 2002. Available from: http://www.evidencenetwork.org/Documents/wp8.pdf
10. Miller N, Pollock VE. Meta-analytic synthesis for theory development. In: Cooper H, Hedges LV, editors. *The handbook of research synthesis.* New York: Russell Sage Foundation; 1994. p. 457-83.
11. Powell S, Tod J. A systematic review of how theories explain learning behaviour in school contexts. In: *Research Evidence in Education Library.* London: EPPI-Centre, Social Science Research Unit, Institute of Education; 2004. Available from: http://eppi.ioe.ac.uk/EPPIWeb/home.aspx?&page=/reel/reviews.htm
12. Pawson R, Greenhalgh T, Harvey G, Walshe K. *Realist synthesis: an introduction.* RMP Methods Paper 2/2004. Manchester: ESRC Research Methods Programme; 2004. Available from: http://www.ccsr.ac.uk/methods/publications/documents/RMPmethods2.pdf
13. Petticrew M, Roberts H. *Systematic reviews in the social sciences: a practical guide.* Oxford: Blackwell, 2006.
14. Light RJ, Pillemer DB. *Summing up: the science of reviewing research.* Cambridge, MA, USA: Harvard University Press, 1984.
15. Cook TD, Leviton LC. Reviewing the literature: a comparison of traditional methods versus meta-analysis. *J Pers* 1980;48:449-72.
16. Grimshaw J, Freemantle N, Wallace S, Russell IT, Hurwitz B, Watt I, et al. Developing and implementing clinical guidelines. *Qual Health Care* 1995;4:55-64.
17. Rychetnik L, Frommer M, Hawe P, Shiell A. Criteria for evaluating evidence on public health interventions. *J Epidemiol Community Health* 2002;56.
18. Hawe P, Shiell A, Riley T. Complex interventions: how "out of control" can a randomised controlled trial be? *BMJ* 2004;2004:1561-63.
19. Campbell M, Fitzpatrick R, Haines A, Kinmonth A, Sandercock P, Spiegelhalter D, et al. Framework for design and evaluation of complex interventions to improve health. *BMJ* 2000;321:694-6.
20. Jackson N, Cochrane Health Promotion and Public Health Field, Victorian Health Promotion Foundation. *Handbook: systematic reviews of health promotion and public health interventions.* Victoria, Australia: Deakin University, 2004.
21. Roen K, Arai L, Roberts H, Popay J. Extending systematic reviews to include evidence on implementation: methodological work on a review of community-based initiatives to prevent injuries [in press]. *Soc Sci Med* 2006.
22. Evans D. Systematic reviews of interpretive research: interpretive data synthesis of processed data. *Aust J Adv Nurs* 2002;20:22-6.

23. Long TJ, Convey JJ, Chwalek AR. *Completing dissertations in the behavioral sciences and education: a systematic guide for graduate students*. San Francisco: Jossey-Bass, 1985.
24. Burls A, Cummins C, Fry-Smith A, Gold L, Hyde C, Jordan R, et al. *West Midlands Development and Evaluation Service handbook [monograph online]*. Birmingham: West Midlands Development and Evaluation Service, 2000. Available from: http://www.publichealth.bham.ac.uk/wmhtac/pdf/wmhandbook.pdf
25. Mays N, Roberts E, Popay J. Synthesising research evidence. In: Fulop N, Allen P, Clarke A, Black N, editors. *Studying the organisation and delivery of health services: research methods*. London: Routledge; 2001. p. 188-220.
26. Mulrow C, Cook D. *Systematic reviews: synthesis of best evidence for health care decisions*. Philadelphia: American College of Physicians, 1998.
27. Hedges LV, Olkin I. *Statistical methods for meta-analysis*. London: Academic Press, 1985.
28. Cwikel J, Behar L, Rabson-Hare J. A comparison of a vote-count and meta-analysis review of intervention research with adult cancer patients. *Res Soc Work Pract* 2000;10:139-58.
29. Clinkenbeard PR. Beyond summary: constructing a review of the literature. In: Buchanan NK, Feldhusen JF, editors. *Conducting research and evaluation in gifted education: a handbook of methods and applications*. New York, NY: Teachers College Press; 1991. p. 33-50.
30. Curlette WL, Cannella KS. Going beyond the narrative summarization of research findings: the meta-analysis approach. *Res Nurs Health* 1985;8:293-301.
31. Suri H. A critique of contemporary methods of research synthesis. *Post-Script* 2000;1:49-55.
32. Hunt M. Is it true that "throwing money at schools" does no good? In: Hunt M, editor. *How science takes stock: the story of meta-analysis*. New York: Russell Sage Foundation; 1997. p. 54-57.
33. Hanushek EA. Throwing money at schools. *Journal of Policy Analysis and Management* 1981;1:19-41.
34. Hedges LV, Laine RD, Greenwald R. Does money matter? A meta-analysis of studies of the effects of differential school inputs on student outcomes. *Educational Researcher* 1994;23:5-14.
35. Noblit GW, Hare RD. *Meta-ethnography: synthesizing qualitative studies*. London: Sage, 1988.
36. Mays N, Pope C, Popay J. Systematically reviewing qualitative and quantitative evidence to inform management and policy-making in the health field. *J Health Serv Res Policy* 2005;10 Suppl 1:6-20.
37. Stemler S. An overview of content analysis. *Practical Assessment, Research and Evaluation [serial online]* 2001 [cited 2006 Jan 5];7. Available from: http://PAREonline.net/getvn.asp?v=7&n=17
38. Cooper H, Hedges LV, editors. *The handbook of research synthesis*. New York: Russell Sage Foundation, 1994.
39. Cooper HM. *Synthesizing research: a guide for literature reviews*. 3rd ed. London: Sage, 1998.
40. Mulrow C, Langhorne P, Grimshaw J. Integrating heterogeneous pieces of evidence in systematic reviews. In: Mulrow C, Cook D, editors. *Systematic reviews: synthesis of best evidence for health care decisions*. Philadelphia: American College of Physicians; 1998. p. 103-12.
41. Foster RL. Addressing epistemologic and practical issues in multimethod research: a procedure for conceptual triangulation. *ANS Adv Nurs Sci* 1997;20:1-12.
42. Begley CM. Using triangulation in nursing research. *J Adv Nurs* 1996;24:122-8.
43. Barbour RS, Barbour M. Evaluating and synthesizing qualitative research: the need to develop a distinctive approach. *J Eval Clin Pract* 2003;9:179-86.
44. Maggs-Rapport F. Combining methodological approaches in research: ethnography and interpretive phenomenology. *J Adv Nurs* 2000;31:219-25.
45. Gough D, Oakley Elbourne D. Methodologies for the systematic synthesis of experimental and non-experimental designs [poster]. In: *1st Campbell Collaboration Methods Group Conference*; 2002 Sep 17-19; Baltimore, Maryland, USA. 2002.
46. Slavin RE. Best-evidence synthesis: an alternative to meta-analytic and traditional reviews. In: Shadish WRJ, Reichardt CS, editors. *Evaluation studies: review annual*. Thousand Oaks, CA: Sage Publications, Inc.; 1987. p. 667-73.
47. Slavin RE. Best evidence synthesis: an intelligent alternative to meta-analysis. *J Clin Epidemiol* 1995;48:9-18.
48. Centers for Disease Control and Prevention. *The guide to community preventive services: what works to promote health? Task Force on Community Preventive Services [monograph online]*. Atlanta, GA: Centers for Disease Control and Prevention, 2005. Available from: http://www.thecommunityguide.org/

49. Busse R, Orvain J, Velasco M, Perleth M, Drummond M, Gurtner F, et al. Best practice in undertaking and reporting health technology assessments. Working group 4 report. *Int J Technol Assess Health Care* 2002;18:361-422.
50. Britten N, Campbell R, Pope C, Donovan J, Morgan M, Pill R. Using meta ethnography to synthesise qualitative research: a worked example. *J Health Serv Res Policy* 2002;7:209-15.
51. DiGuiseppi C, Higgins JPT. *Interventions for promoting smoke alarm ownership and function.* In: The Cochrane Database of Systematic Reviews: Reviews 2001 Issue 2 John Wiley & Sons, Ltd Chichester, UK DOI: 10.1002/14651858.
52. Arai L, Roen K, Roberts H, Popay J. It might work in Oklahoma but will it work in Oakhampton? Context and implementation in the effectiveness literature on domestic smoke detectors. *Inj Prev* 2005;11:148-51.
53. Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJ, Gavaghan DJ, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Control Clin Trials* 1996;17:1-12.
54. Kelly B, Sein C, McCarthy PL. Safety education in a pediatric primary care setting. *Pediatrics* 1987;79:818-24.
55. Kendrick D, Marsh P, Fielding K, Miller P. Preventing injuries in children: cluster randomised controlled trial in primary care. *BMJ* 1999;318:980-3.
56. Ploeg J, Black MF, Hutchison BG, Walter SD, Scott EAF, Chambers IW. Personal, home and community safety promotion with community-dwelling elderly persons: response to a public health nurse intervention. *Can J Public Health* 1994;85:188-91.
57. Williams GE. *An analysis of prenatal education classes: an early start to injury prevention [dissertation].* Kansas City, KS: University of Kansas, 1988.
58. Clamp M, Kendrick D. A randomised controlled trial of general practitioner safety advice for families with children under 5 years. *BMJ* 1998;316:1576-9.
59. Jenkins HM, Blank V, Miller K, Turner J, Stanwick RS. A randomized single-blind evaluation of a discharge teaching book for pediatric patients with burns. *J Burn Care Rehabil* 1996;17:49-61.
60. King WJ, Klassen TP, LeBlanc J, Bernard-Bonnin AC, Robitaille Y, Pham B, et al. The effectiveness of a home visit to prevent childhood injury. *Pediatrics* 2001;108:382-8.
61. Popay J, Arai L, Roberts H, Roen K. *Exploring methodological and practical issues in the systematic review of factors affecting the implementation of child injury prevention initiatives.* London: Health Development Agency, 2003.
62. Campbell DeLong Resources Inc. *Barriers to maintaining working smoke alarms: service area demographic study and interview research.* Portland, Oregan: Campbell DeLong Resources Inc., 2003.
63. Camit M. *Smoke alarms wake you up if there is a fire. Report on the Arabic Chinese and Vietnamese communities smoke alarm purchase and installation program 1998.* NSW Australia: NSW Multicultural Health Communication Service, 1998.
64. Camit M. Smoke alarms wake you up if there's a fire: a smoke alarm campaign targeting Arabic, Chinese, and Vietnamese communities in New South Wales. *Social Marketing Quarterly* 2002;8:52-54.
65. DiGuiseppi C, Slater S, Roberts I, Adams L, Sculpher M, Wade A, et al. The "Let's Get Alarmed!" initiative: a smoke alarm giveaway programme. *Inj Prev* 1999;5:177-82.
66. McConnell CF, Dwyer WO, Leeming FC. A behavioral approach to reducing fires in public housing. *J Community Psychol* 1996;24:201-12.
67. Roberts H, Curtis C, Liabo K, Rowland D, DiGuiseppi C, Roberts I. Putting public health evidence into practice: increasing the prevalence of working smoke alarms in disadvantaged inner city housing. *J Epidemiol Community Health* 2004;58:280-85.
68. Young M, Camit M, Mihajlovic M. A smoke alarm campaign in Arabic, Chinese and Vietnamese communities. *NSW Public Health Bulletin* 1999;10:133-35.
69. Thompson SG. Systematic review: why sources of heterogeneity in meta-analysis should be investigated. *BMJ* 1994;309:1351-5.

APPENDIX 1: METHODS USED IN THE PRODUCTION OF THE GUIDANCE

Literature Search

It was suggested by the funders that rather than reviewing the literature in this area systematically, the applicants could use their existing knowledge of this question to identify the relevant methodological literature relating to narrative synthesis of research based evidence. The team compiled an initial list of 52 works, but subsequently reviewed the decision to search in this way and initiated a limited literature search of key databases and the websites of relevant organisations. It was expected that this search might retrieve literature beyond that already identified, and that it would reduce bias introduced by a reliance on references known to the research team.

Furthermore, it was expected that relevant literature was more likely to be found in textbooks, reports and guidelines than in journal articles (it should be noted that many electronic bibliographic databases only index journals). A complication, which made devising a search strategy problematic, was the lack of a definitive terminology for 'narrative synthesis'. Searching for 'systematic review' or 'meta-analysis' would have been too sensitive and produced unmanageable results. An attempt was made to find definitions and search terms in the database indexes, textbooks and journal articles already identified, and through citation searches, but little additional information was found. There were many possible terms, none of which was consistently used.

Because of the difficulties encountered retrieving useful references from searching electronic databases, an internet search was also undertaken. It was expected that there would be more success looking at potentially relevant sites associated with conducting systematic reviews/meta-analyses, centres interested in evidence based health and social care research (and research methodology), health technology assessment organisations, and similar sites. Further searches of conference proceedings, reference lists, bibliographies and other resources identified while searching the Internet were undertaken. The public catalogues of the British Library and the Library of Congress were searched for general systematic review/meta-analysis books. Finally, a search using general search engines (Copernic and Google) and information gateway sites (OMNI and SOSIG) was undertaken.

Terminology

A number of books and journal articles were searched for definitions or descriptions of 'narrative synthesis'. Some citation searching was also undertaken. Terms found included 'realistic evaluation', 'realist synthesis', 'collective interpretation', 'interpretative synthesis', and 'explanatory synthesis', as well as methods of synthesis such as 'meta-ethnography' and 'triangulation'. The books and articles searched included the following:

- Busse R, Orvain J, Velasco M, *et al*. Best practice in undertaking and reporting health technology assessments. Working group 4 report. *Int J Technol Assess Health Care* 2002;**18**:361-422.
- Cooper H, Hedges LV. *The handbook of research synthesis*. New York: Rusell Sage Foundation, 1994.
- Egger M, Davey Smith G, Altman DG. *Systematic reviews in health care: meta analysis in context*. 2nd ed. London: BMJ Publishing, 2001.
- Forbes A, Griffiths P. Methodological strategies for the identification and synthesis of 'evidence' to support decision-making in relation to complex healthcare systems and practices. *Nurs Inq* 2002;**9**:141-55.
- Glasziou P, Irwig L, Bain C, *et al*. *Systematic reviews in health care: a practical guide*. Cambridge: Cambridge University Press, 2001.

- Khan KS, Ter Riet G, Glanville J, *et al. Undertaking systematic reviews of research on effectiveness: CRD's guidance for carrying out or commissioning reviews: CRD report 4.* 2nd ed. York: NHS Centre for Reviews and Dissemination, 2001.

- Mays N, Roberts E, Popay J. Synthesising research evidence. In: Fulop N, Allen P, Clarke A, Black N, editors. *Studying the organisation and delivery of health services: research methods.* London: Routledge, 2001.

- Williamson JW, Weir CR, Turner CW, *et al. Healthcare informatics and information synthesis; developing and applying clinical knowledge to improve outcomes.* Thousand Oaks, CA, USA: Sage Publications, 2002.

Internet searches

Websites on evidence-based policy and practice were browsed for publications, guidelines, ongoing research and other information of potential interest. Searches of organisation websites where search engines were available were undertaken using single terms and phrases such as 'narrative', 'synthesis' 'systematic review' or 'meta-analysis' or combinations of these. Anything of potential interest was added to an EndNote Library (bibliographic management software).

The sites searched, with dates and results were as follows:

Centre for Reviews and Dissemination (CRD).
University of York. 8th September 2003.
http://www.york.ac.uk/inst/crd/
Found: CRD Report 4 Undertaking systematic reviews of research on effectiveness: CRD's guidance for carrying out or commissioning reviews (2nd edition 2001) and useful links to other sites.

EPPI-Centre (Evidence for Policy and Practice Information Co-ordinating Centre).
SSRU ,University of London. 8th September 2003.
http://eppi.ioe.ac.uk/EPPIWeb/home.aspx
Found: A number of reviews about young people including methodological advances for using and synthesising qualitative data were found, the Centre's Review group manual was downloaded and some further useful links were noted.

Health Evidence Bulletins.
University of Wales, Cardiff. 8th September 2003.
http://hebw.uwcm.ac.uk/
Found: Project methodology was downloaded.

Scottish Intercollegiate Guidelines Network (SIGN).
Edinburgh. 8th September 2003.
http://www.sign.ac.uk/index.html
Found: SIGN 50. Guideline developer's handbook.

Agency for Healthcare Research and Quality (AHRQ).
U.S. Department of Health and Human Services. 8th September 2003.
http://www.ahcpr.gov/
Found: Reviews, assessments and guidelines were identified.

Aggressive Research Intelligence Facility (ARIF).
University of Birmingham. 8th September 2003.
http://www.bham.ac.uk/arif/
Found: Nothing of relevance was identified.

Bandolier.
Oxford. 8th September 2003.
http://www.jr2.ox.ac.uk/bandolier/index.html
Found: Nothing of relevance.

Centre for Evidence-Based Medicine.
Oxford. 8th September 2003.
http://www.cebm.net/

Centre for Evidence-Based Child Health.
London. 8th September 2003.
http://www.ich.ucl.ac.uk/

Centre for Evidence-Based Mental Health.
Oxford. 8th September 2003.
http://www.cebmh.org

Centre for Evidence-Based Dentistry.
Oxford. 8th September 2003.
http://www.ihs.ox.ac.uk/cebd/index.htm

Centre for Evidence-Based Pharmacotherapy.
Birmingham. 8th September 2003.
http://www.aston.ac.uk/lhs/teaching/pharmacy/cebp/

Centre for Evidence-Based Nursing.
York. 8th September 2003.
http://www.york.ac.uk/healthsciences/centres/evidence/cebn.htm
Found: Nothing of relevance on any of these evidence-based centre sites was identified.

Guidelines International Network. (G-I-N).
8th September 2003.
http://www.g-i-n.net/index.cfm?fuseaction=home
Found: Nothing of relevance was identified, though this is a new organisation and website with little content as yet.

Health Development Agency.
London. 8th September 2003.
http://www.hda-online.org.uk/
Found: In the Evidence Based Medicine section of the site papers about good quality evidence, presentations by Hammersley and Marks and process and policy paper including a section on 'synthesising the evidence'.

Health Information Research Unit (HIRU).
McMaster University. 8th September 2003.
http://hiru.mcmaster.ca/default.htm
Found: Nothing of relevance was identified.

National electronic Library for Health (NeLH).
8th September 2003.
http://www.nelh.nhs.uk/
Found: Nothing of relevance was identified.

National Institute of Clinical Excellence (NICE).
London. 8th September 2003.
http://www.nice.org.uk/cat.asp
Found: Nothing of relevance was identified. The guidelines for undertaking technology assessments were not available online.

International Network of Agencies for Health Technology Assessment (INAHTA).
8th September 2003.
http://www.inahta.org/
Found: Nothing of relevance was found on the INAHTA site, or on any of the member sites. Any guides or guidelines about conducting health technology assessment did not describe how to undertake narrative synthesis. However searching on a number of the sites was problematic because we could only search in English.

INAHTA member sites searched:

Australia (ASERNIP, MSAC)
Austria (ITA)
Canada (AETMIS, AHFMR, CCOHTA)
Chile (ETESA)
Cuba (INHEM)
Denmark (DACEHTA, DSI)
Finland (FinOHTA)
France (ANAES, CEDIT)
Germany (DIMDI)
Netherlands (CVZ, GR, TNO, ZonMW)
New Zealand (NZHTA)
Norway (SMM)
Spain (AETS, AETSA, CAHTA, OSTEBA)
Sweden (CMT, SBU)
Switzerland (SNHTA, TA-SWISS)

School of Health and Related Research (ScHARR) – Netting the Evidence.
University of Sheffield. 8th September 2003.
http://www.nettingtheevidence.org.uk/
Found: Nothing of relevance was identified.

King's Fund.
London. 8th September 2003.
http://www.kingsfund.org.uk/
Found: Nothing of relevance was identified.

OMNI (Organising Medical Networked Information).
8th September 2003.
http://omni.ac.uk/
Internet gateway to evaluated, quality Internet resources in health and medicine. Searched for 'narrative synthesis', 'synthesis' 'systematic review' and 'meta-analysis'. Nothing of relevance was identified.

TRIP Database (Turning Research into Practice).
8th September 2003.
http://www.update-software.com/trip/about.htm
Searched for 'narrative synthesis', 'synthesis' 'systematic review' and 'meta-analysis'. Nothing of relevance was identified.

Evidence Network.
The Economic and Social Research Council (ESRC) site for Evidence Based Policy and Practice Research in the UK. 11th September 2003.
http://www.evidencenetwork.org/home.asp
Found: A number of potentially relevant publications were found.
The websites of the seven research teams involved in the Network were also searched individually with a couple of additional publications of interest being identified. (What works for Children, Centre for Neighbourhood Research, Centre for Evidence-Based Public Health Policy, Centre for Economic Evaluation, Research Unit for Research Utilisation, Centre for Evidence in Ethnicity, Health and Diversity, and the Centre for Comparative European Policy Evaluation).

Health Care Practice Research and Development Unit (HCPRDU).
University of Salford. 11th September 2003.
http://www.fhsc.salford.ac.uk/hcprdu/
Found: A section on systematic reviewing was downloaded, and reference to 2 reviews unavailable online, including narrative syntheses of the evidence was noted.

Social Care Institute for Excellence (SCIE).
London. 11th September 2003.
http://www.scie.org.uk/
Found: Nothing of relevance was identified.

electronic Library for Social Care (eLSC).
11th September 2003.
http://www.elsc.org.uk/
Found: The site houses the CareData bibliographic database, which was searched and identified a number of useful references. (See database searches below).

Centre for Evidence-Based Social Services.
University of Exeter. 11th September 2003.
http://www.ex.ac.uk/cebss/
Found: Nothing of relevance was identified.

The Qual-Quan Evidence Synthesis Group.
University of Leicester. 11th September 2003.
http://www.prw.le.ac.uk/research/qualquan/index.htm
Found: Reference to a number of ongoing projects was found and completed publications about the synthesis of qualitative and quantitative data in systematic reviews were downloaded.

The Campbell Collaboration.
11th September 2003.
http://www.campbellcollaboration.org/
Found: Guidelines for the preparation of reviews, presentations given at the 2003 Campbell Colloquium, newsletters of the Campbell Methods Group, and protocols and references via the Cochrane Qualitative Research Methods Group and Campbell Process Implementation Methods Group were downloaded. A methods protocol was found on the Reviews database (C2-RIPE), but nothing specifically about narrative synthesis was found on the trials register (C2-SPECTR).

The Cochrane Collaboration.
11th September 2003.
http://www.cochrane.de/beta/index0.htm
Found: Cochrane Reviewers Handbook version 4.2 was identified. We looked at all 50 Cochrane Review Group websites for any additional guidance to the Reviewers Handbook.
See also searches of the Cochrane Library Methodology Register and *Conference Proceedings* below.

SOSIG (The Social Science Information Gateway).
11th September 2003.
http://www.sosig.ac.uk/
Internet gateway to evaluated, quality Internet resources in the social sciences, business and law. Searched for 'narrative synthesis', 'synthesis' 'systematic review' and 'meta-analysis'. Nothing of relevance was identified.

British Library Public Catalogue.
11th September 2003.
http://blpc.bl.uk/
Searched in title and abstract for 'narrative synthesis', 'systematic review' and 'meta-analysis'. The bibliographic details of any that looked of potentially relevant were retained.

Library of Congress Online Catalog.
11th September 2003.
http://www.loc.gov/
Searched for 'narrative synthesis', 'systematic review' and 'meta-analysis'. Almost all potentially relevant titles had already been found in the British Library Catalogue search. One reference of potential interest was retained.

Copernic (meta-search engine).
12th September 2003.
http://www.copernic.com
Searched for 'narrative synthesis', 'systematic review' and 'meta-analysis'. Browsed through hits, but found nothing of relevance that had not already been identified.

Google (general search engine).
12th September 2003.
http://www.google.com
Searched for 'narrative synthesis', 'systematic review' and 'meta-analysis'. Browsed through the first 100 hits, but found nothing of relevance that had not already been identified.

Any references of potential interest were saved and details added to an EndNote Library.

Additional Internet searches

Additional Internet sites were searched. These were identified from previous searches or were suggested by members of the project team.

If anything of relevance was found the details were added to the Internet Endnote library.

Additional websites were searched on 22nd September 2003 and included:
Joanna Briggs Institute (http://www.joannabriggs.edu.au/about/home.php), US Preventive Services Task Force (http://www.ahcpr.gov/clinic/uspstfix.htm), Canadian Task Force on Preventive Health Care (http://www.ctfphc.org/), AHRQ Evidence Based Practice Centers (http://www.ahcpr.gov/clinic/epc/), Alberta Heritage Foundation for Medical Research (http://www.ahfmr.ab.ca/).

Database searches

Although it was recognised that database searching would be difficult, a number of electronic databases were searched. The strategies used were fairly limited in the use of search terms because of the lack of definitive terms for 'narrative synthesis'. Ideally a search for 'synthesis' alone would have retrieved further relevant records, but would also have resulted in an unmanageable number of irrelevant references. Even combining 'synthesis' with terms used to identify systematic reviews/meta-analyses would have produced a large number of references (again almost entirely irrelevant). This is a problem experienced when searching for any reports of methodological research, but particularly in this case where there is no definitive terminology.

The databases, strategies, dates and results of the database searches were as follows:

MEDLINE: Ovid Gateway. Internet. 1966-2003/August week 4. 9th September 2003.
The MEDLINE search covered the date range 1966 to August 2003. 347 records were identified.
1. review.ab.
2. review.pt.
3. meta-analysis.ab.
4. meta-analysis.pt.
5. meta-analysis.ti.
6. 1 or 2 or 3 or 4 or 5
7. letter.pt.
8. editorial.pt.
9. comment.pt.
10. 7 or 8 or 9
11. 6 not 10
12. (narrativ$ adj3 (synth$ or summar$ or description$ or analy$ or finding$ or form or forms))
13. realistic evaluation$
14. collective interpret$
15. meta ethnograp$
16. meta stud$

17. grounded theory
18. realist synth
19. interp$ synth
20. meta synth$
21. (meta matrix) or (meta matrices)
22. mini synth$
23. explanatory synth$
24. triangulation
25. theory led
26. bayesian adj2 hierarch$
27. or/12-26
28. 11 and 27

Sociological Abstracts: WebSPIRS. Internet. 1963-2003/6. 9th September 2003.
The Sociological Abstracts search covered the date range 1963 to June 2003. 92 records were identified.

#1 review in ti,ab,de
#2 meta analy*
#3 #1 or #2
#4 narrative near3 (synth* or summar* or analy* or description* or finding* or form or forms)
#5 realistic evaluation*
#6 collective interp*
#7 meta ethnograp*
#8 meta stud*
#9 grounded theory
#10 realist synth*
#11 interp* synth*
#12 meta synth*
#13 mini synth*
#14 explanatory synth*
#15 triangulation
#16 (meta matrix) or (meta matrices)
#17 theory led
#18 bayesian near3 hierarch*
#19 #4 or #5 or #6 or #7 or #8 or #9 or #10 or #11 or #12 or #13 or #14 or #15 or #16 or #17 or #18
#20 #3 and #19

Social Science Citation Index (SSCI): Web of Science. Internet. 1981-2003/8. 9th September 2003.
The SSCI search covered the date range 1981 to August 2003. 195 records were identified.
TS=metaanalysis
TS=meta analysis
TS=systematic SAME TS=review*
TS=systematic SAME TS=overview*
TS=literature SAME TS=review*
#1 or #2 or #3 or #4 or #5
TS=narrative SAME TS=synth*
TS=narrative SAME (TS=summar* or TS=description*)
TS=narrative SAME (TS=finding* or TS=review*)
TS=narrative SAME (TS=form or TS=forms)
TS=meta SAME (TS=ethnography OR TS=synthesis OR TS=study)
(TS=realistic evaluation) or (TS=collective interp*)
TS=synthesis SAME (TS=interp* OR TS=explanatory)
TS=synthesis SAME (TS=mini OR TS=realist)
TS=grounded theory
(TS=meta matrix) or (TS=theory led)
TS=bayesian SAME TS=hierarch*

TS=triangulation
#7 or #11 or #8 or #9 or #10 or #11 or #12 or #13 or #14 or #15 or #16 or #17 or #18
#6 and #19

PsycINFO: BIDS. Internet. 1872-2003/9. 9th September 2003.
The PsycINFO search covered the date range 1872 to September 2003. This search identified 352 records.
#1 META-ANALYSIS in PT
#2 LITERATURE-REVIEW-RESEARCH-REVIEW in PT
#3 metaanaly* in ti,de
#4 meta-analy* in ti,de
#5 (review* or overview*) in ti
#6 (review literature) in ti
#7 synthes* near3 ((literature* or research or studies or data) in ti)
#8 ((review* or overview*) in ti) near10 ((systematic* or methodologic* or quantitativ* or research* or literature or studies or trial* or effective*) in ti)
#9 #1 or #2 or #3 or #4 or #5 or #6 or #7 or #8
#10 narrative near3 (synth* or summar* or analy* or description* or finding* or form or forms)
#11 realistic evaluation*
#12 collective interpret*
#13 meta ethnograp*
#14 meta stud*
#15 grounded theory
#16 realist synth*
#17 interp* synth*
#18 meta synth*
#19 mini synth*
#20 explanatory synth*
#21 triangulation
#22 (meta matrix) or (meta matrices)
#23 theory led
#24 bayesian near3 hierarch*
#25 #10 or #11 or #12 or #13 or #14 or #15 or #16 or #17 or #18 or #19 or #20 or #21 or #22 or #23 or #24
#26 #9 and #25

Cochrane Library Methodology Register: Internet. 2003:Issue 3. 9th September 2003.
The Cochrane Library Methodology Register search identified 8 records.
#1 (narrativ* next synth*)
#2 (narrativ* next summar*)
#3 (narrativ* next description*)
#4 (narrativ* next finding*)
#5 (narrativ* next form*)
#6 (realistic next evaluation*)
#7 (collective next interp*)
#8 (meta next ethnograp*)
#9 (grounded next theory)
#10 (realist next synth*)
#11 (interp* next synt*)
#12 (meta next synth*)
#13 (mini next synth*)
#14 (explanatory next synth*)
#15 (#1 or #2 or #3 or #4 or #5 or #6 or #7 or #8 or #9 or #10 or #11 or #12 or #13 or #14)

CareData. Internet. 9th September 2003.
http://www.elsc.org.uk/caredata/caredata.htm
CareData produced 23 unique records. The search interface available for CareData does not allow for sophisticated search strategies. Separate phrase searching was undertaken, firstly in the abstract field and then in the keyword field. The phrases searched in the abstract field were 'narrative

synthesis', 'synthesis' 'systematic review' and 'meta analysis'. The keyword field had an index and the terms 'literature review' and 'research methods' were combined. The results of the 5 separate searches were pooled, and the duplicate references removed.

DARE (Database of Abstracts of Reviews of Effects): Internal CRD administration database. CAIRS T System. 1994-2003/8. 9th September 2003.

The internal CRD administration version of DARE was searched for methodology papers identified as part of the DARE production process, and CRD records which are not available on the public DARE database. This search identified 32 records.

s narrativ$(w3)(synth$ or summar$ or description$ or analy$ or finding$ or form or forms or review$)
s realistic(w)evaluation$
s collective(w)interpret$
s meta(w)ethnograp$
s meta(w)stud$
s grounded(w)theory
s realist(w)synth$
s interp$(w)synth$
s meta(w)synth$
s mini(w)synth$
s explanatory(w)synth$
s theory(w)led
s (meta(w)matrix) or (meta(w)matrices)
s triangulation
s Bayesian(w3)hierarch$
s s1 or s2 or s3 or s4 or s5 or s6 or s7 or s8 or s9 or s10 or s11 or s12 or s13 or s14 or s15
s m/st1
s s16 and s17

Applied Social Sciences Index and Abstracts (ASSIA): Cambridge Scientific Abstracts (CSA). Internet. 1987-2003. 9th September 2003.
The ASSIA search covered the date range 1987 to date. The search identified 55 records
((synthesis) OR (narrative)) AND ((systematic review) OR (meta analysis) OR KW=(systematic reviews) OR (meta analysis))

Educational Resources Information Center (ERIC): Dialog. Internet. 1966-2003/6. 9th September 2003.
The ERIC search covered the date range 1966 to June 2003. The search identified 176 records.
EXPLANATORY SYNTH? OR MINI SYNTH? OR META SYNTH? OR INTERP? SYNTH? OR REALIST SYNTH? OR GROUNDED THEORY OR META STUD? OR META ETHNOGRAP? OR COLLECTIVE INTERPRET? OR REALISTIC EVALUATION? OR NARRATIV? WITH(3) (SYNTH? OR SUMMAR? OR DESCRIPTION? OR ANALY? OR FINDING? OR FORM OR FORMS OR REVIEW?) AND META ANALYSIS OR 1 term(s): ERIC Subject Headings=("META ANALYSIS") OR REVIEW

All references were downloaded into an EndNote Library and deduplicated.

Additional database searches

The search strategies were rerun to include further search terms on 21st October 2003. The new search terms were derived from references identified by the original searches or by reviewer hand searches. The following search terms were added to each strategy:
'Cross design', 'evaluation synthesis', 'descriptive synthesis' and 'best evidence synthesis'.

The updated Medline search strategy as an example is presented below. The other database strategies were amended in similar ways.

MEDLINE: Ovid.
1. review.ab.
2. review.pt.
3. meta-analysis.ab.
4. meta-analysis.pt.
5. meta-analysis.ti.
6. 1 or 2 or 3 or 4 or 5
7. letter.pt.
8. editorial.pt.
9. comment.pt.
10. 7 or 8 or 9
11. 6 not 10
12. (narrativ$ adj3 (synth$ or summar$ or description$ or analy$ or finding$ or form or forms))
13. realistic evaluation$
14. collective interpret$
15. meta ethnograp$
16. meta stud$
17. grounded theory
18. realist synth
19. interp$ synth
20. meta synth$
21. (meta matrix) or (meta matrices)
22. mini synth$
23. explanatory synth$
24. triangulation
25. theory led
26. bayesian adj2 hierarch$
27. cross design$
28. evaluation synth$
29. (best evidence adj3 synth$)

30. descrip$ synth$.mp
31. or/12-30
32. 11 and 31

The results of the searches were added to an Endnote library and then deduplicated against the original search results. A further 192 references were identified from all the databases using these additional terms.

Conference Proceedings
A number of conference proceedings were handsearched, and any potentially relevant abstracts were added to the EndNote Library. The following conference proceedings were searched:

- *19th Annual Meeting of the International Society of Technology Assessment in Health Care*; 2003 Jun 22-25; Canmore, Canada.
- *18th Annual Meeting of the International Society of Technology Assessment in Health Care*; 2002 Jun 9-12; Berlin, Germany.
- *17th Annual Meeting of the International Society of Technology Assessment in Health Care*; 2001 Jun 3-6; Philadelphia, USA.
- *16th Annual Meeting of the International Society of Technology Assessment in Health Care*; 2000 Jun 18-21; The Hague, The Netherlands.
- *15th Annual Meeting of the International Society of Technology Assessment in Health Care*; 1999 Jun 20-23; Edinburgh, Scotland.
- *14th Annual Meeting of the International Society of Technology Assessment in Health Care*; 1998 Jun 7-10; Ottawa, Canada.
- *13th Annual Meeting of the International Society of Technology Assessment in Health Care*; 1997 May 25-28; Barcelona, Spain.
- *12th Annual Meeting of the International Society of Technology Assessment in Health Care*; 1996 Jun 23-26; San Francisco, USA.
- *11th Annual Meeting of the International Society of Technology Assessment in Health Care*; 1995 Jun; Stockholm, Sweden.
- *10th Annual Meeting of the International Society of Technology Assessment in Health Care*; 1994; Baltimore, USA.
- *4th Symposium on Systematic Reviews: Pushing the Boundaries*; 2002 Jul 2-4; Oxford, UK.
- *3rd Symposium on Systematic Reviews*; 2000 Jul 3-5 Oxford, UK.
- *2nd Symposium on Systematic Reviews: Beyond the Basics*; 1999 Jan 5-7; Oxford, UK.
- *1st Symposium on Systematic Reviews: Beyond the Basics*; 1998 Jan 8-9; Oxford, UK.
- *11th Cochrane Colloquium: Evidence, health care and culture*; 2003 Oct 26-31; Barcelona, Spain.
- *10th Cochrane Colloquium*; 2002 31 Jul-3 Aug; Stavanger, Norway.
- *9th Cochrane Colloquium: The evidence dissemination process: how to make it more efficient*; 2001 Oct 9-13; Lyon, France.

- *8th Cochrane Colloquium: Evidence for action: challenges for the Cochrane Collaboration in the 21st century*; 2000 Oct 25-29; Cape Town, South Africa.
- *7th Cochrane Colloquium: The best evidence for healthcare: the role of the Cochrane Collaboration*; 1999 Oct 5-9; Rome, Italy.
- *6th Cochrane Colloquium: Systematic reviews: evidence for action*; 1998 Oct 22-26; Baltimore, USA.
- *5th Cochrane Colloquium: Using the evidence*; 1997 Oct 8-12; Amsterdam, Holland.
- *4th Cochrane Colloquium*; 1996; Adelaide, Australia.
- *3rd Cochrane Colloquium*; 1995 Oct 4-8; Oslo, Norway.
- *2nd Cochrane Colloquium*; 1994; Hamilton, Canada

Methods texts selection process

A total of 1,307 articles were retrieved from the literature searches. Two reviewers independently selected articles from the titles and abstracts available from the searches. Articles were included if they offered guidance on the conduct of reviews or combining data from different studies. Where reviewers disagreed, the full article was included for further investigation. This resulted in a total of 260 full publications being ordered for further assessment. One reviewer then selected all published articles that reported any tool or technique meeting the following criteria:

1) Was concerned with the synthesis of primary research
2) Was not a strictly statistical technique (e.g. meta-analysis)
3) Could conceivably be applied or adapted to the context of a systematic review of the literature.

A total of 69 studies were selected on the basis of these criteria, and were used to inform our guidance.

The majority of included articles were initially identified from the internet searches (36%) and database searches (33%) (see figure 1.1). Thirteen (19%) of the initial texts identified by the project team were included in the final 66 selected articles. A further eight articles (12%) were identified by handsearching/scanning of reference lists.

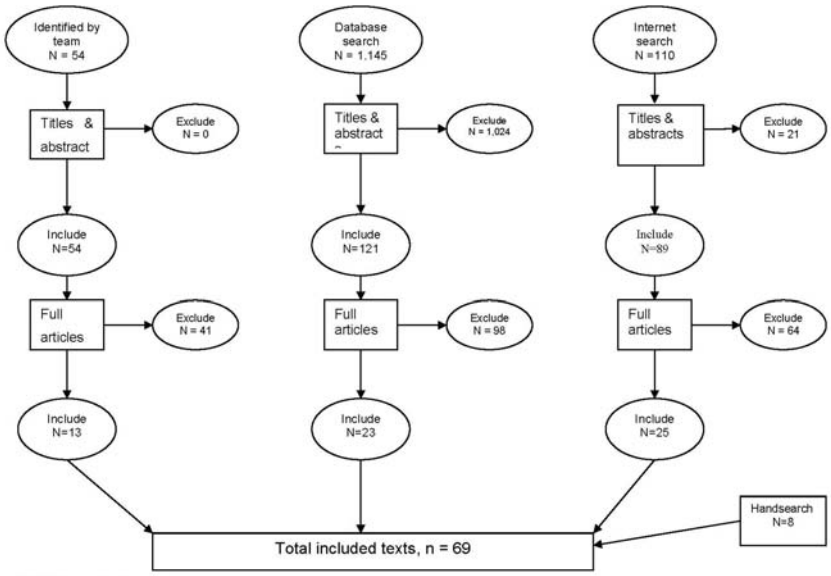


Figure 1.1: Methods text selection process.

APPENDIX 2: BIBLIOGRAPHY OF METHODOLOGICAL TEXTS USED IN THE PRODUCTION OF THE GUIDANCE

Extracted methods texts

1. Bangert Drowns RL, Wells Parker E, Chevillard I. Assessing the methodological quality of research in narrative reviews and meta-analyses. In: Bryant KJ, Windle M, editors. *The science of prevention: methodological advances from alcohol and substance abuse research*. Washington, DC, US: American Psychological Association, 1997. p. 405-429.

2. Barbour RS, Barbour M. Evaluating and synthesizing qualitative research: the need to develop a distinctive approach. *J Eval Clin Pract* 2003;9:179-186.

3. Beck CT. Mothering multiples: a meta-synthesis of qualitative research. *MCN, American Journal of Maternal Child Nursing*. 2002;27:214-21.

4. Becker BJ. Examining theoretical models through research synthesis: the benefits of model driven meta-analysis. *Eval. Health Prof.* 2001;24:190-217.

5. Begley CM. Using triangulation in nursing research. *J Adv Nurs* 1996;24:122-8.

6. Bennett JA. A case for theory triangulation. *Nursing Science Quarterly*. 1997;10:97-102; discussion 103-6.

7. Brannen J. *Mixing methods: qualitative and quantitative research*. Aldershot: Avebury, 1992.

8. Britten N, Campbell R, Pope C, Donovan J, Morgan M, Pill R. Using meta ethnography to synthesise qualitative research: a worked example. *J Health Serv Res Policy* 2002;7:209-15.

9. Burls A, Cummins C, Fry-Smith A, Gold L, Hyde C, Jordan R, et al. *West Midlands Development and Evaluation Service handbook [monograph online]*. Birmingham: West Midlands Development and Evaluation Service, 2000. Available from: http://www.publichealth.bham.ac.uk/wmhtac/pdf/wmhandbook.pdf

10. Busse R, Orvain J, Velasco M, Perleth M, Drummond M, Gurtner F, et al. Best practice in undertaking and reporting health technology assessments. Working group 4 report. *Int J Technol Assess Health Care* 2002;18:361-422.

11. Campbell R, Pound P, Pope C, Britten N, Pill R, Morgan M, et al. Evaluating meta-ethnography: a synthesis of qualitative research on lay experiences of diabetes and diabetes care. *Soc. Sci. Med*. 2003;56:671-84.

12. Clinkenbeard PR. Beyond summary: constructing a review of the literature. In: Buchanan NK, Feldhusen JF, editors. *Conducting research and evaluation in gifted education: a handbook of methods and applications*. New York, NY: Teachers College Press, 1991. p. 33-50.

13. The Cochrane Collaboration. *Cochrane reviewers' handbook (Version 4.2.0: March, 2003) [monograph online]*, 2003. Available from: http://www.cochrane.dk/cochrane/handbook/hbook.htm

14. Cook TD, Leviton LC. Reviewing the literature: a comparison of traditional methods versus meta-analysis. *J Pers* 1980;48:449-72.

15. Cooper H, Hedges LV, editors. *The handbook of research synthesis*. New York: Russell Sage Foundation, 1994.

16. Curlette WL, Cannella KS. Going beyond the narrative summarization of research findings: the meta-analysis approach. *Research in Nursing & Health*. 1985;8:293-301.

17. Dixon-Woods M, Fitzpatrick R, Roberts KA. Including qualitative research in systematic reviews: opportunities and problems. *J Eval Clin Pract* 2001;7:125-133.

18. Eaves YD. A synthesis technique for grounded theory data analysis. *J Adv Nurs* 2001;35:654-63.

19. EPPI Centre. *EPPI-Centre Review Group Manual: working document version 1.1 [monograph online]*. London: EPPI Centre, 2001. Available from: http://eppi.ioe.ac.uk/EPPIWebContent/downloads/RG_manual_version_1_1.pdf

20. Evans D. Systematic reviews of interpretive research: interpretive data synthesis of processed data. *Australian Journal of Advanced Nursing* 2002;20:22-6.

21. Flinspach SL. Interpretive synthesis: a methodology for reviewing qualitative case-study research. *Dissertation Abstracts International, A: The Humanities and Social Sciences* 2001;61:4971-A.

22. Forbes A, Griffiths P. Methodological strategies for the identification and synthesis of 'evidence' to support decision-making in relation to complex healthcare systems and practices. *Nurs Inq* 2002;9:141-55.

23. Foster RL. Addressing epistemologic and practical issues in multimethod research: a procedure for conceptual triangulation. *Advances in Nursing Science* 1997;20:1-12.

24. Fulop N, Allen P, Clarke A, Black N. *Studying the organisation and delivery of health services: research methods*. London: Routledge, 2001.

25. Jensen LA, Allen MN. Meta-synthesis of qualitative findings. *Qualitative Health Research* 1996;6:553-60.

26. Jones K. *Mission drift in qualitative research, or moving toward a systematic review of qualitative studies, moving back to a more systematic narrative review [monograph online]*. Leicester: Centre for Evidence in Ethnicity, Health & Diversity, De Montfort University, 2003. Available from: http://users.wbs.warwick.ac.uk/cms_attachment_handler.cfm?f=d250896e-82a2-48e6-b356-c310d3be6861&t=mission_drift.pdf

27. Light RJ, Pillemer DB. *Summing up: the science of reviewing research*. Cambridge, MA, USA: Harvard University Press, 1984.

28. Long TJ, Convey JJ, Chwalek AR. *Completing dissertations in the behavioral sciences and education: a systematic guide for graduate students*. San Francisco: Jossey-Bass, 1985.

29. Maggs-Rapport F. Combining methodological approaches in research: ethnography and interpretive phenomenology. *J Adv Nurs* 2000;31:219-25.

30. Mulrow C, Cook D. *Systematic reviews: synthesis of best evidence for health care decisions*. Philadelphia: American College of Physicians, 1998.

31. Khan KS, Ter Riet G, Glanville J, Sowdon AJ, Kleijnen J. *Undertaking systematic reviews of research on effectiveness: CRD's guidance for carrying out or commissioning reviews: CRD report 4*. 2nd ed. York: NHS Centre for Reviews and Dissemination, 2001.

32. Noblit GW, Hare RD. Meta-ethnography: issues in the synthesis and replication of qualitative research. In: *Annual Meeting of the American Educational Research Association*; 1983 Apr 11-15; Montreal, Quebec. 1983. p. 23.

33. Noblit GW, Hare RD. *Meta-ethnography: synthesizing qualitative studies*. London: Sage, 1988.

34. Nurius PS, Yeaton WH. Research synthesis reviews: an illustrated critique of "hidden" judgments, choices, and compromises. *Clin Psychol Rev* 1987;7:695-714.

35. Pawson R. Evidence-based policy: the promise of realist synthesis. *Evaluation* 2002;8:340-358.

36. Popay J, Rogers A, Williams G. Rationale and standards for the systematic review of qualitative literature in health services research. *Qualitative Health Research* 1998;8:341-51.

37. Ragin CC. *The comparative method: moving beyond qualitative and quantitative strategies*. Berkeley, CA: University of California Press, 1987.

38. Rees R, Harden A, Shepherd J, Brunton G, Oliver S, Oakley A. *Young people and physical activity: a systematic review of research on barriers and facilitators [monograph online]*. London: EPPI Centre, 2001. Available from: http://eppi.ioe.ac.uk/EPPIWebContent/hp/reports/physical_activity01/physical_activity.pdf

39. Risjord MW, Dunbar SB, Moloney MF. A new foundation for methodological triangulation. *J Nurs Scholarsh* 2002;34:269-75.

40. Shepherd J, Harden A, Rees R, Brunton G, Garcia J, Oliver S, et al. *Young people and healthy eating: a systematic review of research on barriers and facilitators [monograph online]*. London: EPPI Centre, 2001. Available from: http://eppi.ioe.ac.uk/EPPIWebContent/hp/reports/health_eating01/healthy_eating_yp.pdf

41. Slavin RE. Best-evidence synthesis: an alternative to meta-analytic and traditional reviews. In: Shadish WRJ, Reichardt CS, editors. *Evaluation studies: review annual*. Thousand Oaks, CA: Sage Publications, Inc., 1987. p. 667-673.

42. Slavin RE. Best evidence synthesis: an intelligent alternative to meta-analysis. *J Clin Epidemiol* 1995;48:9-18.

43. Suri H, Clarke D. Revisiting methods of literature synthesis. In: *Annual Meeting of the American Educational Research Association*; 1999 Apr 19-23; Montreal, Canada. 1999. p. 2-15.

44. Suri H. The process of synthesising qualitative research: a case study. In: *Annual Conference of the Association for Qualitative Research*; 1999 Jul 6-10; Melbourne. 1999. Available from: http://www.latrobe.edu.au/aqr/offer/papers/HSuri.htm

45. Suri H. A methodologically inclusive model for research synthesis [monograph online]. In: *Australian Association for Research in Education (AARE) and the New-Zealand Association for Research in Education (NZARE) Joint-Conference*; 1999 Dec; Melbourne. 1999. Available from: http://www.aare.edu.au/99pap/sur99673.htm

46. Suri H. A critique of contemporary methods of research synthesis. *Post-Script* 2000;1:49-55. Available from: http://www.edfac.unimelb.edu.au/student/insight/postscriptfiles/vol1/suri.pdf

47. Sutton A, Abrams KR, Jones DR, Sheldon TA, Song F. Systematic reviews of trials and other studies. *Health Technol Assess* 1998;2.

48. Thomas J, Harden A, Oakley A, Oliver S, Sutcliffe K, Rees R, et al. Integrating qualitative research with trials in systematic reviews. *BMJ* 2004;328:1010-1012.

49. Williamson JW, Weir CR, Turner CW, Lincoln MJ, Cofrin KMW. *Healthcare informatics and information synthesis; developing and applying clinical knowledge to improve outcomes*. Thousand Oaks, CA, USA: Sage Publications, 2002.

50. Yin RK, Heald KA. Using the case survey method to analyze policy studies. *Administrative Science Quarterly* 1975;20:371-81.

Methods texts covering material similar or identical to previously extracted texts

1. ACC/AHA Task Force on Practice Guidelines. Section II: tools and methods for creating guidelines. Step four: synthesize and interpret the evidence [monograph online]. In: *Manual for ACC/AHA Guideline Writing Committees.Methodologies and Policies from the ACC/AHA Task Force on Practice Guidelines*. Bethesda, MD: American College of Cardiology, American Heart Association, 2002. Available from: http://www.acc.org/clinical/manual/manual_IIstep4.htm

2. Booth A. Cochrane or cock-eyed? How should we conduct systematic reviews of qualitative research? [monograph online]. In: *Qualitative Evidence-Based Practice Conference*; 2001 May 14-16; Coventry University. 2001. Available from: http://www.leeds.ac.uk/educol/documents/00001724.htm

3. Bryman A. *Social research methods*. Oxford: Oxford University Press, 2001.

4. Canadian Task Force on Preventive Health Care. *Canadian Task Force Methodology [monograph online]*. London, Ontario, Canada: Canadian Task Force on Preventive Health Care, 1998. Available from: http://www.ctfphc.org/Methodology.html

5. Smith MC, Stullenbarger E. A prototype for integrative review and meta-analysis of nursing research. *J Adv Nurs* 1991;16:1272-83.

6. Gall MD, Borg WR, Gall JP. *Educational research. an introduction*. 7th ed. Boston, MA: Allyn & Bacon, 2003.

7. Gough D, Elbourne D. Systematic research synthesis to inform policy, practice and democratic debate. *Social Policy and Society* 2002;1:225-236.

8. Green BN, Johnson CD, Adams A. Writing narrative literature reviews for peer-reviewed journals: secrets of the trade. *Journal of Sports Chiropractic & Rehabilitation* 2001;15:5-19.

9. Letzel H. 'Best evidence synthesis:an intelligent alternative to meta-analysis': discussion. a case of "either or" or "as well" (comment). *J Clin Epidemiol* 1995;48:19-21.

10. Pawson R. *Assessing the quality of evidence in evidence-based policy: why, how and when? Draft [monograph online]*. London: ESRC, 2003. Available from: http://www.ccsr.ac.uk/methods/publications/Pawson.pdf

11. Pawson R. Evidence-based policy: in search of a method. *Evaluation* 2002;8:157-181.

12. Pearson A, Wiechula R, Long L. QARI: a systematic approach to the appraisal, extraction and synthesis of the findings of qualitative research [abstract]. In: *11th Cochrane Colloquium: Evidence, health care and culture*; 2003 Oct 26-31; Barcelona, Spain. 2003.

13. Pearson A, Wiechula R, Long L. The role of text in enhancing Cochrane reviews of effectiveness [abstract]. In: *11th Cochrane Colloquium: Evidence, health care and culture*; 2003 Oct 26-31; Barcelona, Spain. 2003.

14. Swann C, Falce C, Morgan A, Kelly M. *HDA Evidence Base: process and quality standards manual for evidence briefings. Draft [monograph online]*. London: Health Development Agency Public Health Evidence Steering Group, 2002. Available from: http://www.hda.nhs.uk/evidence/ebmanual.pdf

15. Thomas J, Harden A, Sutcliffe K, Oakley A, Rees R, Oliver S, et al. New methods for integrating qualitative and quantitative research in a systematic review: an example from public health [abstract]. In: *11th Cochrane Colloquium: Evidence, health care and culture*; 2003 Oct 26-31; Barcelona, Spain. 2003.

Methods texts not yet received/extracted

1. Cwikel J, Behar L, Rabson-Hare J. A comparison of a vote-count and meta-analysis review of intervention research with adult cancer patients. *Res Soc Work Pract* 2000;10:139-158.

2. Edwards A, Russell I, Stott N. Signal versus noise in the evidence base for medicine: an alternative to hierarchies of evidence? *Fam Pract* 1998;15:319-322.

3. Tashakkori A, Teddlie C. *Mixed Methodologogy: combining qualitative and quantitative approaches*. Thousand Oaks, CA: Sage Publications, 1998.

4. Wachter KW, Straf ML, editors. *The future of meta-analysis*. New York, NY, US: Russell Sage Foundation, 1990.

ANEXO 7: HANDBOOK OF COCHCHRANE (CHAPTER 13)

Table of Contents

4 Guide to the contents of a Cochrane review and protocol

Editors: Nynke Smidt, Jonathan Deeks and Theresa Moore.

[This chapter includes sections from: Higgins JPT, Green S (editors). Cochrane Handbook for Systematic Reviews of Interventions Version 5.0.0 (updated February 2008). The Cochrane Collaboration, 2008. Available from http://www.cochrane.org/resources/handbook.]

4.1 Format of a review of diagnostic test accuracy

All systematic reviews of diagnostic test accuracy published in the *Cochrane Database of Systematic Reviews* have the same format. There are several reasons for this. It helps readers to find the results of research quickly and to assess the validity, applicability and implications of those results. It guides authors to report their work explicitly and concisely, and minimises the effort required to do this. The format is also suited to electronic publication and updating, and facilitates the production of statistical summary figures and tables, which are informative and readable when viewed on a computer monitor or printed.

The Review Manager software (RevMan5) is designed to help authors construct reviews in the appropriate format and to prepare files required to transfer reviews electronically. Standard headings and tables embedded in RevMan5 guide review authors when preparing their report and are used by the publisher to structure the publication and link readers to the sections which are of particular interest to them. The content that should follow each heading is described in this chapter.

Each protocol consists of:

- Title;
- Protocol information – details of authors, contact person, important dates, 'What's new' and 'History';
- Main text of the protocol – consisting of the background, objectives, methods, acknowledgements and declaration of interests;
- Tables – relevant to the background or methods;
- Studies and references – other references relevant for the background or methods;
- Sources of support – split into internal and external sources of support;
- Appendices – relevant to the background or methods.

Each review consists of:

- Title;
- Review information – details of authors, contact person, important dates, 'What's new' and 'History';
- Structured abstract;
- Plain language summary;

- Text of the review – consisting of the background, objectives, methods, results, discussion, authors' conclusions, acknowledgements, contribution of authors, declaration of interests, differences between protocol and review, and published notes;
- Tables – showing characteristics and methodological quality of the included studies, a summary of results table, a log of the studies that were excluded and reasons why, and additional tables relevant to the review;
- Studies and references – presenting the references of the included studies, excluded studies and other references;
- Data and analyses – presenting the data (2x2 table) by test or by study;
- Figures – summarizing the methodological quality of the included studies, forest plots of sensitivity and specificity, figures of summary receiver operating characteristic (ROC) plots, and any additional imported figures;
- Sources of support – split into internal and external sources of support;
- Feedback;
- Appendices.

4.2 Protocol and review information

This section of a review includes information about the title of the review, details of the contact person and co-authors, and key dates relating to the production and publication of the review and protocol.

4.2.1 Title of the review

The essence of the objective should be captured in the review's title. Typically this involves stating the diagnostic technology together with the key characteristics of the people to whom it is applied and the purpose for which it is used. The key components of the title are therefore:

- the patients (how they present, where they present to, what tests have been done before);
- the target condition (disease, disease stage, or sub-type of a disease eligible for a specific treatment);
- the test or tests being evaluated.

The test that is being evaluated is known as the index test. A review may evaluate and compare the diagnostic accuracy of several index tests, and may elect one as a comparator test with which the diagnostic accuracy of the other index tests is compared, particularly if this test is currently standard diagnostic practice. The target condition is the condition of interest that the index and comparator test(s) are attempting to detect. The clinical reference standard is usually the test or tests representing the best available method of detecting the target condition. Reference standards, which give results with very little error, are known as 'gold standards'. Please note the difference between the reference standard and the comparator test: the reference standard is the best test available to detect the target condition (and may not routinely be used in clinical practice) while the comparator test is a routinely used test, the diagnostic accuracy of which we wish to compare with other index tests to decide which is the best for detecting the target condition.

The review title in RevMan5 is structured to ensure that the correct information is reported to reflect the objective of the review for example, "What is the diagnostic accuracy of 'index

test' for diagnosing 'target condition' in 'patient description'?" Four title options are possible (Table 4.2.a). These vary in the number of tests being evaluated (options 1 and 2 are for two tests, options 3 and 4 work for either a single test or several tests), and whether the patient description is required (options 1 and 3). Options 2 and 4, which do not include the patient description, should only be used where the target condition clearly implies a particular patient group.

Table 4.2.a: Structure for titles of Cochrane systematic reviews of diagnostic test accuracy

Option 1	Format	<Index test 1> *versus* <index test 2†> *for* <target condition(s)> *in* <patient description>
	Example	<Chlamydia antibody titre testing> *versus* <hysterosalpingography> *for diagnosing* <tubal pathology> *in* <subfertile women>
Option 2	Format	<Index test 1> *versus* <index test 2†> *for* <target condition(s)>
	Example	<MRI> *versus* <ultrasound> *for diagnosing* <ischaemic stroke>
Option 3	Format	<Index test(s)> *for* <target condition(s)> *in* <patient description>
	Example	<Physical examination> *for the diagnosis of* <lumbar radiculopathy due to disc herniation> *in* <patients with low back pain>
Option 4	Format	<Index test(s)> *for* <target condition(s)>
	Example	<Physical tests> *for* <detection of shoulder impingements>

† if the comparison is with current diagnostic practice, the second index test will be the comparator test

4.2.2 Authors

Authorship of all scientific papers (including Cochrane protocols and reviews) establishes accountability, responsibility and credit (Rennie 1997, Flanagin 1998, Rennie 1998). When deciding who should go in the by-line for Cochrane reviews, it is important to distinguish individuals who have made a substantial contribution to the review (and who should be listed) and those who have made other contributions, which should be noted in the acknowledgements section. Ideally, the order of authors should relate to their relative contributions to the review. The person who contributed most should be listed first. Authorship should be based on substantial contributions to all of the following three steps, based on the requirements for submitting papers to biomedical journals, developed by the International Committee of Medical Journal Editors (ICMJE 2006):

- Conception and design of study, or analysis and interpretation of data;
- Drafting the review or revising it critically for important intellectual content;
- Final approval of the version to be published;

Affiliations of authors will be published within the completed protocol or review, so authors should ensure that these fields are completed and up to date in 'Archie', the web interface of

The Cochrane Collaboration's Information Management System (IMS). The fields that must be completed are the First name(s) and Last name(s) of the author, Organisation and Country. If a co-author does not have a publishable address, but should still appear in the by-line for the citation, then the Organisation and Country should be those of the Review Group (for example Smith J. c/o Cochrane Pregnancy and Childbirth Group, UK). Group authorship is possible but requires an entry to be created in 'Archie' in the name of the group (see Chapter 2).

4.2.3 Contact Person

Contact details (i.e. name, address, e-mail, telephone and fax number) for the person to whom correspondence about the review should be addressed, and who has agreed to take responsibility for maintaining and developing the review, are automatically taken from 'Archie', the web interface of The Cochrane Collaboration's Information Management System (IMS). This person usually takes responsibility for developing and organising the review, communicates with the editorial base, ensures that the review is prepared within agreed timescales, submits it to the editorial base, communicates feedback to co-authors and ensures that the updates are prepared.

The contact author does not need to be the first author, and the choice of contact author will not affect the citation for the review. If the contact author no longer wishes to be responsible for a published review and another member of the review team does not wish to take responsibility for it, then the Review Group Co-ordinator (RGC) should be listed as the contact author, and the former contact author listed as a co-author. The RGC need not be listed as a co-author.

4.2.4 Dates

There are several dates associated with a Cochrane review. Some of these are automatically generated by RevMan, and some need to be entered by the review author. These dates are important both to inform readers of the review and to facilitate management of review publication. It is essential that authors apply these definitions when entering dates into relevant fields during an update or amendment to a review.

For considerations to make when updating a review please see Chapter 11 'Updating and maintaining reviews'.

4.2.4.1 Assessed as Up-to-date

This date is entered by review authors for full reviews only (not protocols). On publication, this date is reproduced in a prominent place in the review to inform readers of how recently the review has been assessed as up to date. The criteria for assessing a review as up to date are listed in Box 4.2.a.

A review might be considered to be up to date even if it has received only minimal edits for many years, for example if a recent search for studies identifies no new evidence since the review was published. All reviews submitted for publication must include a date on which the review was last assessed as being up to date. The date should be entered by the authors, and will often coincide with the date on which the authors submit the review for consideration to be published in the *Cochrane Database of Systematic Reviews*. It may be appropriate to amend the date on approval of the review for publication.

Box 4.2. a: Guidance for declaring a review as being up to date

The date a review is assessed as being up to date must be chosen so that the review (new, updated or amended) meets the following key criterion:

1. The evidence is up to date on the performance of the test(s)

The list of included studies should include all available evidence, and should result from a most recent search typically being within six months of the date on which the review is assessed as being up to date;

In addition, it is highly desirable, but not mandatory, that

2. The methods of the review are up to date

All mandatory methods for Cochrane reviews (as described in the current version of the *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy*) should be incorporated;

3. Factual statements are correct

Factual statements, for example, in the Background and Discussion, should not be unreasonably out-dated.

4.2.4.2 Date of search

This date is entered by review authors for full reviews only (i.e. not protocols). 'Search' here refers to the searches of all the databases searched for the review. If different databases were searched on different dates, the most recent date of the search for each database should be given within the text of the review and the earliest of the dates should be put in this field. For example, if the most recent searches of the following databases were on the following dates (MEDLINE 5 June 2007, EMBASE 12 June 2007 and CENTRAL 28 June 2007) the 'Date of search' would be 5 June 2007. For further details on how to document your searches please see Chapter 7.

4.2.4.3 Next stage expected

This is entered by review authors as:

- For protocols: the date on which the full review is expected (e.g. in 1 year);
- For full reviews: the date on which the next update is expected (e.g. in 2 years).

4.2.4.4 Protocol first published, Review first published, Last citation issue

These dates are automatically recorded in RevMan and generated at the time of publication of the protocol or full review.

4.2.5 What's new and History

This should describe the changes to the protocol or review since it was last published in the *Cochrane Database of Systematic Reviews*. At each update of a review, substantive or not, the 'What's new' field should contain the calendar date of the change, the event and a description of what was changed. The following options for the events in the protocol are: feedback incorporated, amendment, declare new citation version (major change or minor

change). In the full review examples of events to note include: Update (new search for studies), feedback incorporated, amendment, declare new citation version (conclusions changed or not changed) and declare review as no longer being updated. In the description, the authors should give a brief summary of how much new information has been added to the review (for example, number of studies, participants or extra analyses) and any important changes to the conclusions, results or methods of the review.

Changes which are made between versions prior to the penultimate version should be listed as events under 'History'. The 'What's New' section should only indicate changes made between the previous version and the current version.

4.3 Abstract

All full reviews must include an abstract of not more than 400 words. It should be kept as brief as possible without sacrificing important content. Abstracts to Cochrane reviews are published in MEDLINE and the Science Citation Index, and are made freely accessible on the Internet, so will often be read as stand-alone documents. They should, therefore, summarize the key methods and results of the review and not contain any material that is not in the review. The content must be consistent with the text, data and conclusions of the review and not include references to any information outside the review. Links to other parts of the review (such as references, studies, additional tables and additional figures) may not be inserted in the abstract. An example is included in Box 4.3.a.

Abstracts should be made as readable as possible without compromising scientific integrity. They should primarily be targeted at healthcare decision makers (clinicians, consumers and policy makers) rather than just researchers. Terminology should be reasonably comprehensible to a general rather than a specialist healthcare audience. Abbreviations should be avoided, except where they are widely understood (for example, HIV). Where essential, other abbreviations should be spelt out (with the abbreviations in brackets) on first use.

It is important that Cochrane reviews of diagnostic test accuracy are identifiable in MEDLINE and other electronic bibliographic databases from their titles, abstracts and keywords. Indexing of diagnostic terms is currently poor, and the guidelines for the titles of Cochrane reviews rule out the use of phrases such as 'systematic review' or 'diagnostic test accuracy'. Their titles may therefore be difficult to distinguish from reviews of interventions. "Screening for type 2 diabetes mellitus", for example, may refer to an intervention review investigating the effectiveness of screening for type 2 diabetes mellitus in reducing morbidity and mortality associated with diabetes, but it could also describe a diagnostic test accuracy review investigating the accuracy of screening tests for type 2 diabetes mellitus. Diagnostic test accuracy reviews will be clearly identifiable by 'flags' within *The Cochrane Library*, but these will be lost when the abstracts are transferred to bibliographic databases.

To overcome this problem it is recommended that the phrase 'diagnostic accuracy' be included somewhere in the abstract – the 'objectives' section may provide a natural home for this phrase. This is a key phrase that is highly likely to be used in searching by researchers who want to identify reviews of diagnostic tests. Combining this phrase with the phrase for the *Cochrane Database of Systematic Reviews* as a journal title (for example, 'Cochrane Database of Syst Rev' in PubMed) will retrieve Cochrane reviews of diagnostic test accuracy. Ultimately, the introduction of specific database indexing terms into MEDLINE (such as Publication Types 'diagnostic test accuracy review' and 'diagnostic test accuracy study') should improve the efficiency of searching.

The content under each heading in the abstract should be as follows:

Background
One or two sentences explaining the context, the purpose and rationale for the review.

Objectives
A precise statement of the primary objective of the review, ideally in a single sentence. Where possible the style should be of the form 'To determine the diagnostic accuracy of [Index test] for diagnosing [target condition] in [patient description]'. The exact phrase 'diagnostic accuracy' is important as this can then be used as a search term to retrieve Cochrane systematic reviews of diagnostic test accuracy from MEDLINE and the Science Citation Index, and other publications of these reviews in other electronic bibliographic databases.

Search methods
List the sources and the dates of the search for each source, using the active form 'We searched....' For example 'We searched MEDLINE from January 1950 to December 2006'. Search terms should not be listed here, and if a large number of databases have been searched only key databases and the number of databases should be stated. The date range of the search for each database should be given. For most databases such as MEDLINE, it should be in the form 'MEDLINE (January 1950 to December 2006)'. Searching of bibliographies for relevant citations can be covered in a generic phrase 'reference lists of articles'. If there were any constraints based on language or publication status, these should be listed. If individuals or organisations were contacted to locate studies this should be noted and it is preferable to use, 'We contacted pharmaceutical companies' rather than a listing of all the pharmaceutical companies contacted. If journals were specifically handsearched for the review, this should be noted.

Selection criteria
Briefly list the main criteria used to select studies for inclusion in the review. Please include; type of study design (for example consecutive patient series and /or case-control studies); index and /or comparator tests; target condition; eligible reference standards, and key characteristics of the study population.

Data collection and analysis
Describe how data extraction and quality assessment of studies were done, whether by more than one person, and whether multiple assessments were done independently. The description should be restricted to how data were extracted and assessed, and not include details of what data were extracted. It should be stated whether a meta-analysis was done and if so the statistical methods used should be named, and the summary statistics estimated stated. If comparisons are made between index and/or comparator tests it should be stated whether comparisons were direct (made within studies) or indirect (made between studies).

Results
This section should begin with the total number of studies and participants included in the analysis, and brief details pertinent to the interpretation of the results, such as a summary of study quality, diversity in study characteristics and heterogeneity in study results. It should address the primary objective and be restricted to the main results. Wherever possible, accuracy should be expressed using summary statistics most likely to help someone make a decision about whether or not to use a particular test. Any summary statistics in the abstract

should be the same as those presented in the review and in the summary of results table. Summary statistics should be stated together with confidence intervals.

Authors' conclusions

The primary purpose of the review should be to present information, rather than to offer advice. The authors' conclusions should be succinct, address the review question and draw directly from the findings of the review so that they directly and obviously reflect the main results. Assumptions should not be made about practice circumstances, values, preferences and tradeoffs, and the giving of advice or recommendations should generally be avoided. Any important limitations of data and analyses should be noted. Important conclusions about the implications for research should also be included.

Box 4.3.a: Example of an abstract for a Cochrane Systematic Review of Diagnostic Test Accuracy: Magnetic resonance imaging for provisional diagnosis of multiple sclerosis in patients with suspected disease.

Background

Detection of clinically silent brain lesions on magnetic resonance imaging (MRI) in patients with a single episode of neurological dysfunction may allow earlier diagnosis of multiple sclerosis (MS) if MRI lesions accurately predict the occurrence of a second episode required for a definitive MS diagnosis.

Objective

To determine the diagnostic accuracy of MRI for the provisional diagnosis of MS in patients with suspected disease.

Search methods

We searched MEDLINE (January 1950 to November 2004) and 11 other electronic databases (from inception to November 2004). We conducted citation searches, and screened reference lists of included studies.

Selection criteria

Diagnostic accuracy studies that compared MRI (or diagnostic criteria incorporating MRI) with standard clinical diagnosis of MS.

Data collection and analysis

Two authors independently screened titles and abstracts for relevance. Screening for inclusion, data extraction, and quality assessment were carried out by one author and checked by a second. Studies were assessed for methodological quality using QUADAS. HSROC meta-analytical methods were used to estimate summary ROC curves and diagnostic odds ratios, and to investigate the impact of study design and length of follow up on diagnostic accuracy.

Results

29 studies (18 cohort studies, 11 other designs) recruiting 2714 patients were included. Test accuracy depended on both study design and length of follow-up, hence we focused on the results of the two higher quality studies with cohort designs and clinical follow-up of 10 years or more. The presence of many lesions on MRI (>8 or > 10), could not accurately rule in MS (positive likelihood ratios of 3.6; 95% CI 0.9 to 14.4 and 1.9; 95% CI 1.0 to 3.7). The absence of lesions had limited ability to rule out a diagnosis of MS (negative likelihood ratios of 0.1; 95% CI 0.04 to 0.3 and 0.5; 95% CI 0.4 to 0.6).

Conclusions

Use of MRI to confirm MS after a single attack of neurological dysfunction has poor accuracy, and may lead to both misdiagnosis and inappropriate treatment.

4.4 Writing a plain language summary

The format and structure of plain language summaries for systematic reviews of diagnostic test accuracy are currently the subject of discussion with the Cochrane Steering Group. This section will be updated once these discussions are completed.

4.5 Main text

The text of the review should be as succinct and readable as possible. Although there is no formal word limit on Cochrane reviews, review authors should consider 10,000 words an absolute maximum unless there is a special reason to write a longer review (for example, where a large number of index tests are evaluated). The majority of reviews should be substantially shorter than this. A review should be written so that someone who is not an expert in the area can understand it, in light of the following policy statement, reported in Cochrane News 1999; 15: 14:

> "The target audience for Cochrane reviews is people making decisions about health-care. This includes health-care professionals, consumers and policy makers with a basic understanding of the underlying disease or problem.
>
> It is a part of the mission and a basic principle of The Cochrane Collaboration to promote the accessibility of systematic reviews of the effects of healthcare interventions to anyone wanting to make a decision about healthcare. However, this does not mean that Cochrane reviews must be understandable to anyone, regardless of their background. This is not possible, any more than it would be possible for Cochrane reviews to be written in a single language that is understandable to everyone in the world. It is important to translate the content, or elements of the content, of reviews into different languages and formats targeted at different audiences including healthcare professionals, consumers and policy makers in a variety of circumstances.
>
> Cochrane reviews should be written so that they are easy to read and understand by someone with a basic sense of the topic who may not necessarily be an expert in the area. Some explanation of terms and concepts is likely to be helpful, and perhaps even essential. However, too much explanation can detract from the readability of a review. Simplicity and clarity are also vital to readability. The readability of Cochrane reviews should be comparable to that of a well-written article in a general medical journal."

The text of a Cochrane review contains a number of fixed headings and subheadings that are available in the RevMan5 document structure. Additional optional subheadings are also included in the structure, but authors should not be limited by these and should add their own subheadings where appropriate. Authors are encouraged to use the optional subheadings included in the structure where possible, but not if they make individual sections needlessly short.
Text cannot be placed immediately under the following Level 1 headings: 'Abstract', 'Methods', 'Criteria for considering studies for this review', 'Results', or 'Authors conclusions'. Text for these sections starts below the first subsequent subheading.

Background [fixed, level 1 heading]

Well-formulated review questions do not appear out of thin air. They occur in the context of an existing body of knowledge. This context should be explained in the background section of

the review. The background helps set the rationale for the review, and should explain why the questions being asked are important and why closely related questions are not being covered. The background section of the review should inform the readers of the review about why the review is being done, particularly where there are existing systematic reviews published outside The Cochrane Collaboration. This can generally be done within 1 to 1.5 printed pages.

The background section contains 4 optional subheadings which can be used to help structure the text:

Target condition being diagnosed **[optional, level 2 heading]**

A description of the target condition of interest (frequency, severity, prognosis and possible treatments). If there are Cochrane reviews of interventions for the target condition they should be cross-referenced here.

Index test(s) **[optional, level 2 heading]**

A description of the index tests that are being evaluated in this review, whether they are currently used in clinical practice, and the roles being considered for the tests (for example, as replacement for a comparator test because of lower cost and invasiveness).

Alternative test(s) **[optional, level 2 heading]**

A description of the possible diagnostic tests and strategies that could (and are) used in clinical practice, irrespective of whether they are evaluated in this review. This will help place the review in context of all available diagnostic options.

Rationale **[optional, level 2 heading]**

The background helps set the rationale for the review, and should explain why the questions being asked are important. It might also mention why this review was undertaken and how it might relate to a wider review of a general problem.

Objectives [fixed, level 1 heading]

The review question should be clearly stated. The primary objective should be described and include the index test(s), the target condition verified by the reference standard ('gold standard') and the study population. Where possible the style should be 'To determine the diagnostic accuracy of [index test] for detecting [target condition] in [patient description]'. It should also be clear from the objective whether the authors want to investigate the accuracy of an index test as triage for the current normal test regime, as a replacement for it, or in combination with it (see Chapter 5). The term 'comparator' is specifically used to refer to what is thought to be the most commonly used current normal test regime. Where the aim of the review is to make comparisons between tests these should be clearly stated.

Secondary objectives **[optional, level 2 heading]**

Complex reviews that investigate multiple questions may categorise their objectives as 'Primary Objectives' and 'Secondary Objectives'. For example, the primary objectives may be to compare test accuracy between two tests; the secondary objectives may estimate test accuracy for each test at pre-specified thresholds. However, secondary objectives related to investigating heterogeneity between study results should not be listed under this subheading but under the next subheading.

Investigation of sources of heterogeneity **[optional, level 2 heading]**

Heterogeneity investigations explore factors which may affect diagnostic accuracy, and are essential because they may provide a framework by which the expected heterogeneity may be explained a priori and to provide a more clinically useful review. For example, the test accuracy of troponin may vary depending upon when the test is done after the start of symptoms. Providing a primary objective only, for example a comparison of troponin against Creatine Kinase-MB for patients with chest pain to detect myocardial infarction, is not sufficient for clinical decision making when patients present at different time points after the onset of chest pain, and the test performance of troponin may vary accordingly. Alternative subgroups can be aspects of study design and execution, age groups, setting, differences in the presence or degree of clinical characteristics or differences in the operation of either the index test or the reference test.

Methods [fixed, level 1 heading]

The Methods section in a protocol should be written in the future tense. Cochrane reviews are updated as new evidence accumulates, therefore methods outlined in the protocol should generally anticipate a sufficiently large number of studies to address the review's objectives (even if it is known this is not the case).

The Methods section in a full review should be written in the past tense, and should describe what was done to obtain the results and conclusions of the current version of the review. Often a review is unable to implement all of the methods outlined in the protocol, usually because there is insufficient evidence. In such circumstances, it is recommended that the methods that were not implemented still be outlined in the review, so that it serves as a protocol for future updates of the review. Some Cochrane review groups (CRGs) have policies on this issue, and these should be available from the Review Group Co-ordinator (RGC). Examples include adding an additional subsection at the end of 'Methods of the review', or including the methods for future updates in an additional table.

In the methods section, the authors should clearly describe the selection criteria for considering studies for the review (Chapter 5 and Chapter 6), the methods used to identify relevant studies (Chapter 7), the process used for selection of studies and collecting data (Chapter 8) and how the methodological quality of the included studies is assessed (Chapter 9). A statistician may best write the section for describing the statistical analysis and data synthesis (Chapter 10). In addition, information about how to investigate sources of heterogeneity and any pre-planned sensitivity analyses should be described clearly here (Chapter 10).

Criteria for considering studies for this review [fixed, level 2 heading]

The eligibility criteria required for studies to be included in the review must be clearly stated.

Types of studies [fixed, level 3 heading]

Eligible study designs should be stated here, along with any thresholds for inclusion based on the conduct or quality of the studies. For example, 'All consecutive series of patients and case-control studies' or 'All study designs". Exclusion of particular types of studies (for example, case-control studies, retrospective studies) should briefly be justified. Restrictions based on the use of particular reference standards should not be listed here.

Participants [fixed, level 3 heading]

Specify the participants for whom the test would be applicable, including any restrictions on diagnoses, age groups and settings. Planned subgroup analyses related to participant characteristics should not be listed here.

Index tests [fixed, level 3 heading]

Specify the test or tests under evaluation.

Comparator tests [optional, level 3 heading]

This subheading is optional. The comparator test is the testing regime used in practice, which the index test may be seeking to replace. Not all reviews will have a comparator test, hence it is an optional subheading. If a comparator is stated in the review question it should be stated here. There is potential to confuse the comparator with the reference standard. However the reference standard used in the context of research to establish test accuracy will not usually be the same as the test or test combination used in day-to-day practice.

Target conditions [fixed, level 3 heading]

The target condition is a particular disease or disease stage that the index test is intended to identify. Tests may occasionally be used to differentiate between several target conditions – if this is the case they should all be listed here.

Reference standards [fixed, level 3 heading]

Describe the clinical reference standards ('gold standard') that are considered appropriate to establish the presence or absence of the target condition in the tested population. If particular reference standards are commonly used but considered inadequate they should be stated here as exclusion criteria.

Search methods for identification of studies [fixed, level 2 heading]

Electronic searches [fixed, level 3 heading]

The methods used to identify studies should be summarized. Further details of the content of these sections are discussed in Chapter 7. The bibliographic databases searched, the dates and periods searched and any constraints, such as language, should be stated. The full search strategies for each database should be listed in an appendix to the review.

Searching other resources [optional, level 3 heading]

List grey literature sources, such as reports and conference proceedings. If journals are specifically handsearched for the review, this should also be noted. List people (for example, researchers, experts) and/or organisations who were contacted. List any other sources, which may include, for example, reference lists, the World Wide Web or personal collections of articles.

This text may be organised under the following four subheadings:

- Grey literature,
- Handsearching,
- References lists and
- Correspondence.

These subheadings are not included in the RevMan5 structure, so if required an author will need to create them. They can be used either in place of 'Searching other resources' or as subheadings to it.

Data collection and analysis [fixed, level 2 heading]

Selection of studies [fixed, level 3 heading]

The method used to apply the selection criteria should be described stating whether the criteria were applied independently by more than one author, and how any disagreements were resolved (see Chapter 8).

Data extraction and management [fixed, level 3 heading]

State the method used to extract or obtain data from published reports or from primary authors of the included studies (for example, using a data extraction/data collection form). Any reanalysis of individual patient data should be described. Whether data are extracted independently by more than one author should be stated, along with how any disagreements are resolved. If relevant, methods for processing data in preparation for analysis should be described (see Chapter 8).

Assessment of methodological quality [fixed, level 3 heading]

Assessment of methodological quality involves describing both the tool, and the method by which it was applied. The tool(s) used (i.e. QUADAS) should be described or referenced, with an indication of how quality assessments were incorporated into the interpretation of the results (see Chapter 9). Operational definitions of items within the quality assessment tool should be stated (possibly using an additional table). For example, the QUADAS scale requires authors to explicitly state what they consider to be 'appropriate spectrum' and an 'appropriate reference standard' before applying the tool. The method used to assess methodological quality should be described, stating whether the tool was applied independently by more than one author and how any disagreements were resolved.

Statistical analysis and data synthesis [fixed, level 3 heading]

In this section, the descriptive and inferential statistical methods should be described. Descriptive methods consist of tabulation, graphical displays of estimates of diagnostic accuracy (for example sensitivities and specificities), and plotting the study results in ROC space (see Chapter 10). Inferential statistical methods are used for the estimation of summary ROC curves and average operating points, testing of differences between tests, and investigations of heterogeneity. Details should be given of the statistical method and model used, what parameters were estimated, whether random effects were estimated, and the software used.

Investigations of heterogeneity [optional, level 3 heading]

Indicate how the sources of heterogeneity listed in the objectives were investigated (see Chapter 10).

Sensitivity analyses [optional, level 3 heading]

Pre-planned sensitivity analyses should be stated here. These could include restricting analyses to a particular subgroup of patients, or to studies without a particular methodological shortcoming for example verification bias or review bias (see Chapter 10).

Assessment of reporting bias [optional, level 3 heading]

If any tests or investigations were undertaken to detect reporting biases the methods used should be explained here (see Chapter 10).

Results [fixed, level 1 heading]

Results of the search [fixed, level 2 heading]

The results section of the full review should start with a summary of the results of the search, reporting the number of citations identified by the electronic searches, the number for which full reports were retrieved, the number of citations that were finally included in the review, and the number of unique studies that they report. Brief details should be given of the occurrence of duplicate reports citations of the same studies. Similar details should be given for the results of searching other sources where possible (for example handsearching, correspondence and reference lists).

An overview of the selection process should be given. This can be summarized using a flow diagram, and it is recommended that authors consider including one as an additional Figure (see Chapter 7).

The number of included studies should be clearly stated, together with the numbers of participants and the number who have the target condition. If a review evaluates more than one test these numbers should be given per test. Where pairwise comparisons are based on direct (within study) comparisons, the number of studies and patients available for each comparison should be stated.

Individual details of the studies are tabulated in the 'Characteristics of included studies' table. A succinct summary of the key characteristics of the design, participants, index tests and other methodological issues presented in this table should be given. In some circumstances an additional table may be useful to give a tabular summary of the different index tests that were encountered, or other aspects of the study. If such a table is produced a link to it must be created at an appropriate place in this section.

The number of excluded studies should be mentioned here and refer to the information contained in the 'Characteristics of Excluded Studies' table, providing a succinct summary of why studies were excluded from the review.

Methodological quality of included studies [fixed, level 2 heading]

A figure summarising overall methodological quality, and a table of the methodological quality ratings for individual studies can be produced automatically in RevMan5 (see Figures below), and should be linked to the text in this section. The text should summarize the general quality of the included studies, its variability across studies and any important flaws in individual studies which will threaten the validity of the results. The criteria that were used to assess the methodological quality bias should be described in the 'Methods' and not here. For reviews that evaluate several tests in separate studies, reporting of methodological quality by test should be considered if there are important differences.

Findings [fixed, level 2 heading]

The main findings on the diagnostic accuracy of the index tests studied in the review should be presented, together with the results of the planned comparisons between the index tests, or between the index tests and the comparator tests. The section should directly address the objectives of the review rather than list the findings of the included studies in turn. The focus should be on reporting the pattern of results across all the included studies. Numerical summaries and results of statistical analyses may best be summarized in additional tables and a link to the table placed in the text where they are discussed. Links to relevant forest plots and summary ROC plots may also be placed in the text.

Subheadings are encouraged if they make reading easier (for example, for each index test if a review addresses more than one). Any exploration of heterogeneity, sensitivity analyses and investigations of possible biases that were undertaken should be reported.

Inferences should be avoided in the results section.

Discussion [fixed, level 1 heading]

A structured discussion can aid the systematic consideration of the implications of the review (Docherty 1999).

Summary of main results [fixed, level 2 heading]

The authors should summarize the main findings of the diagnostic accuracy of the index test(s). This description should mirror the summary of results table (it may be easiest to create this table before writing this section). The authors should summarize the relevance of the findings of investigations of heterogeneity. There should also be comment on the homogeneity and methodological quality of the evidence, and the completeness of the evidence in terms of whether it has been possible to address all the objectives of the review and the degree to which uncertainties remain. Consideration should be given to the amount of data available to address the primary hypotheses.

Strengths and weaknesses of the review [fixed, level 2 heading]

The author should state the strengths and limitations of the review with regard to preventing bias. These may be factors within, or outside, the control of the review authors. The discussion might include whether all relevant studies were identified, whether all relevant data could be obtained, or whether the review methods used (for example, searching, study selection, data extraction, and analysis) could have introduced bias. The degree to which robust conclusions can be drawn if the study results are notably heterogeneous should be discussed.

Comments on how the included studies fit into the context of other published reviews might be included, stating clearly whether the other evidence was systematically reviewed.

Applicability of findings to clinical practice and policy [fixed, level 2 heading]

To assess the applicability of their findings authors consider whether the studies that were identified were sufficient to address all of the objectives of the review. They should consider the relevance of the patients, tests and settings that were included in the review to the review objectives, and identify the limits of the situations to which the evidence from the included studies definitely applies. For example, studies from secondary care may have limited relevance to primary care, or the evaluations located may have only evaluated an outmoded version of the test technology. Comments on how the results of the review fit into the context of current practice should also be included here, although authors should bear in mind that current practice might vary internationally.

Authors' conclusions [fixed, level 1 heading]

The primary purpose of the review should be to present information, rather than to offer advice. Conclusions of the authors are divided into two sections:

Implications for practice [fixed, level 2 heading]

The implications for practice should be as practical and unambiguous as possible. They should not go beyond the evidence that was reviewed and be justifiable by the data presented in the review.

Implications for research [fixed, level 2 heading]

This section of Cochrane reviews is used increasingly by people making decisions about future research, and authors should try to write something that will be useful for this purpose. As with the 'Implications for Practice', the content should be based on the available evidence and should avoid the use of information that was not included or discussed within the review.

In preparing this section, authors should consider the different aspects of research, perhaps using types of studies, index tests, study population and target condition verified by the reference standard as a framework. Implications for *how* research might be done and reported should be distinguished from *what* future research should be done. It is important that this section is as clear and explicit as possible. General statements that contain little or no specific information, such as "Future research should be better conducted" or "More research is needed" are of little use to people making decisions, and should be avoided.

Acknowledgements [fixed, level 1 heading]

This section should be used to acknowledge any individuals or organisations who have contributed but who are not listed among the authors. This would include any previous authors of the Cochrane review and might include the contributions of the editorial team of the CRG. Permission should be obtained from persons acknowledged.

Contribution of authors [fixed, level 1 heading]

The names and contribution of the present co-authors should be described in this section. One author, usually the contact author, should be identified as the guarantor of the review. All authors should discuss and agree on their respective descriptions of contribution before the review is submitted for publication in the *Cochrane Database of Systematic Reviews*. When the review is updated, this section should be checked and revised as necessary to ensure that it is accurate and up to date.

The following potential contributions have been adapted from Yank et al (Yank 1999). This is a suggested scheme and the section should describe what people did, rather than attempt to identify within which of these categories someone's contribution falls. Ideally, the contributors should describe their contribution in their own words:

- Conceiving the review;
- Designing the review;
- Co-ordinating the review;
- Data collection for the review
- Designing search strategies;
- Undertaking searches;
- Screening search results;
- Organising retrieval of papers;
- Screening retrieved papers against inclusion criteria;
- Appraising quality of papers;

- Extracting data from papers;
- Writing to authors of papers for additional information;
- Providing additional data about papers;
- Obtaining and screening data on unpublished studies;
- Data management for the review;
- Entering data into RevMan5;
- Analysis of data;
- Interpretation of data;
- Providing a methodological perspective;
- Providing a clinical perspective;
- Providing a policy perspective;
- Providing a consumer perspective;
- Writing the review;
- Providing general advice on the review;
- Securing funding for the review;
- Performing previous work that was the foundation of the current study.

Declarations of interest [fixed, level 1 heading]

Authors should report any conflict of interest that might be perceived by others as being capable of influencing their judgements, including personal, political, academic and other possible conflicts, as well as financial conflicts. Authors must state if they have been involved in a study included in the review. Details of The Cochrane Collaboration's policy on conflicts of interest appear in Chapter 2 of the *Cochrane Handbook for Systematic Reviews of Interventions* (Higgins 2006a).

Financial conflicts of interest cause the most concern, and should be avoided, but must be reported if there are any. Any secondary interest (such as personal conflicts) that might unduly influence judgements made in a review (concerning, for example, the inclusion or exclusion of studies, assessments of the validity of included studies or the interpretation of results) should be reported.

If there are no conflicts of interest, this should be stated explicitly, for example, by writing 'None known'.

Differences between protocol and review [fixed, level 1 heading]

Authors should report and clearly explain why the differences between the methods reported in the protocol and the full review exist.

Published notes [fixed, level 1 heading]

These will be published in the *Cochrane Database of Systematic Reviews*. They may include:

- Editorial notes and comments from the CRG and the Diagnostic Test Accuracy Editorial Board, for example where issues highlighted by editors or referees are believed worthy of publication alongside the review;

- A summary of previous changes to the review. Changes since the previous published version must be stated under 'What's new'.

The published notes must be completed for all withdrawn publications to give the reason for withdrawal. Only the cover sheet and published notes are published for withdrawn protocols and reviews.

4.6 Tables

4.6.1 Characteristics of included studies

This is a standard table with eight entries for each study including the following fixed items: (1) Study ID, (2) Clinical features and settings, (3) Participants, (4) Study design, (5) Target condition and reference standard(s), (6) Index and comparator test(s), (7) Follow up and (8) Notes. Footnotes should be used for explanations of any abbreviations used (these will be published in the *Cochrane Database of Systematic Reviews*).

Table 4.6.a: Guide to the contents of the 'Characteristics of included studies table'.

Heading	Content
Study ID	First author, year of publication
Clinical features and settings	Report the presenting clinical signs and symptoms, and previous test results, clinical setting, and referral routes.
Participants	Report the sample size, socio-demographic items (age, gender, ethnic group etc), co-morbidities, geographic region.
Study design	Was the sample selected as a single group, or as separate groups with and without the target condition? Were participants consecutively enrolled in the study? How was comparability of the groups ensured? Were participants identified retrospectively or prospectively? If studies evaluated more than one test, how were tests allocated to individuals, or did each individual receive all tests?
Target condition and reference standard(s)	State the target condition, the definition and description of reference standard(s), and the criteria used to define an individual to have the target condition. Give details of test operators (who performed the reference standard). If applicable, report the timing, manufacturer, technical characteristics, etc. of the reference standard. Report the prevalence of the target condition in the sample.
Index and comparator tests	For all index and comparator tests, report the test(s) definition and description, criteria for positive test results, and details of test operators. If applicable, report the timing, manufacturer and technical characteristics.
Follow up	How much of the sample was lost, or has missing or uninterpretable test

	results? Were adverse events noted that could be caused by the test?
Notes	Other information that the author considers relevant to report, for example sources of funding, abbreviations, and other issues not covered elsewhere.

4.6.2 Assessment of methodological quality table

This table is used to record the methodological quality of each study. The table contains eleven predefined quality assessment items from the QUADAS tool (see Table 4.6.b). These items are strongly recommended, but can be deleted in exceptional circumstances when not relevant. In addition it is possible to enter custom quality items by creating 'user defined' fields. Details of how to edit the table can be found in RevMan5 Help. Full details of how to assess quality and how to decide which quality items to assess can be found in Chapter 9.

For each item in the quality assessment the author must enter both a description of how the study addressed the issue, and a classification as to whether this was adequate ('yes'), inadequate ('no'), or whether not enough detail was presented for a judgement to be made ('unclear'). For both the standard and the user defined items details of the criteria for classifying studies as 'yes', 'no' and 'unclear' must be stated in the review protocol, probably using an additional table (see Section 4.6.1.7).

It is possible to attribute quality assessments either to a study, or to each test within each study. In the former, all tests within a study are classified the same way, in the latter, tests within the same study can have different quality assessments. In RevMan5 this is described as 'by test' or 'by study'. It is possible to specify a mixture of 'by test' and 'by study' quality assessment items within a review. For details of how to define the quality assessment table please refer to RevMan5 Help.

Table 4.6.b: Guide to the contents of the Assessment of methodological quality table

Representative spectrum?	Was the spectrum of patients representative of the patients who will receive the test in practice?
Acceptable reference standard?	Is the reference standard likely to correctly classify the target condition?
Acceptable delay between tests?	Is the time period between reference standard and index test short enough to be reasonably sure that the target condition did not change between the two tests?
Partial verification avoided?	Did the whole sample or a random selection of the sample, receive verification using a reference standard of diagnosis?
Differential verification avoided?	Did patients receive the same reference standard regardless of the index test result?
Incorporation avoided?	Was the reference standard independent of the index test? (i.e. the index test did not form part of the reference standard).
Reference standard results blinded?	Were the reference standard results interpreted without knowledge of the results of the index test?
Index test results blinded?	Were the index test results interpreted without knowledge of the results of the reference standard?

Relevant clinical information?	Were the same clinical data available when test results were interpreted as would be available when the test is used in practice?
Uninterpretable results reported?	Were uninterpretable/ intermediate test results reported?
Withdrawals explained?	Were withdrawals from the study explained?

4.6.3 Characteristics of excluded studies

Studies for which reports were retrieved based on the abstract and title, but finally excluded based on the content of the report, should be listed and the key reason for exclusion should be given. This should be kept brief, and a single reason for exclusion is usually sufficient.

4.6.4 Characteristics of studies awaiting classification

Details of studies for which insufficient information is available to include them or exclude them from the review should be reported in this table, using the same structure as the included studies table. For example, translations may be awaited, or confirmation of details required from authors, which mean some studies cannot be included when the review is first published. Studies which are discovered just before publication could also be listed here, but all attempts should be made to include all available literature in the review.

4.6.5 Characteristics of ongoing studies

Studies that may be eligible but are still ongoing are reported here. Please record principal investigator and/or contact person, location and country, web site address (if provided), start date, target condition, reference standard and tests which are being evaluated.

4.6.6 Summary of results tables

In order to make Cochrane systematic reviews of diagnostic test accuracy informative and as accessible as possible for readers, they must contain a summary of results table (see Chapter 11) as well as a summary in the text. It is essential however that the two are consistent with each other.

4.6.7 Additional tables

Additional tables may be used for information that cannot be conveniently placed in the text or in fixed tables. Examples include:

- information to support the background for example details about the index test, comparator test or reference standard;
- explicit details of the methodological quality criteria assessed;
- numerical statistical summaries.

4.7 Studies and references

Authors should check all references for accuracy (Dickersin 1986, Eichorn 1987).

4.7.1 References to studies

Studies are organized under four fixed headings:

Included studies
Studies that specifically meet the inclusion criteria and are included in the review should be listed here.

Excluded studies
Studies for which reports were retrieved but were found not to meet the inclusion criteria should be listed here.

Studies awaiting assessment
Relevant studies that have been identified, but cannot be assessed for inclusion until additional data or information are obtained, should be listed here. These need not be cited in the text of the review.

Ongoing studies
Studies that are ongoing but meet (or appear to meet) the inclusion criteria should be listed here.

Each of these headings can include multiple studies (or no studies). A study is identified by a 'Study ID'. A year can be associated with each study (usually the year of completion, or the publication year of the primary reference to that study). In addition, each study should be assigned a category of 'Data source' from among the following.

- Published data only;
- Published and unpublished data;
- Unpublished data only;
- Unpublished data sought but not used.

Each study can have multiple references. Each reference has its own 'Reference ID'. A single reference for each study should be awarded the status of 'Primary reference', usually the reference that most completely reports the study.

4.7.2 Other references

References other than those to studies are divided into three categories:

Additional references
Other references cited in the text should be listed here, including those cited in the background and methods sections. If a report of a study is cited in the text for some reason other than referring to the study (for example, because of some background or methodological information in the report), it should be listed here as well as under the relevant study.

Other published versions of this review
References to other published versions of the review in a journal, textbook or the *Cochrane Database of Systematic Reviews* should be listed here.

Classification pending references

RevMan5 also includes an optional subheading 'Classification pending references' category to facilitate **organisation** of references while preparing a review. Any references remaining in this category when the review is submitted are not published.

4.8 Data and analyses

4.8.1 Study results

Results of the studies included in a review can be organised either by entering all data from a single study together even when it includes several tests (data tables by study), or by entering all data for a single test from several studies together (data tables by test). Each test requires entry of the number of true positives, false positives, false negatives and true negatives. Where results are available for a test at several thresholds these can either be added as separate results with a covariate (see below) specifying the threshold, or several tests defined each with a specific threshold. When the review is published all of the data entered will be displayed in paired sensitivity-specificity forest plots, even if these are not included in the review. This flexible approach to organise and enter data differs from that for Cochrane reviews of interventions where the data structure (comparisons, outcomes, subgroups) defines the analyses which will be undertaken. In reviews of test accuracy all data are entered before the structure of the analyses is specified.

4.8.2 Covariates

For each test and study, authors can define one or more covariates for example setting, age, severity of disease, or specific features of the test such as threshold. These covariates can be used for identifying subgroups that may have different test accuracy results.

4.8.3 Analyses

In RevMan5, authors can select analyses from a list of five options:

- a simple analysis of a single test;
- a simple analysis of a single test that investigates heterogeneity (subgroups);
- an analysis comparing several tests;
- an analysis comparing several tests that investigates heterogeneity (subgroups);
- an analysis of two tests restricted to paired data.

For each analysis paired forest plots of sensitivity and specificity can be created and summary ROC plots produced. In addition, plots of the summary ROC curve, based on the Littenberg and Moses (Littenberg 1993) linear regression model can be presented. However, summary ROC curves, average operating points including 95% confidence intervals and 95% prediction regions can also be produced in RevMan5 using estimates from more valid statistical models: the hierarchical summary ROC (HSROC) model and the bivariate model.

4.9 Figures

RevMan5 can create four types of figures for automatic inclusion in reviews:

- a methodological quality graph (a bar chart showing overall compliance with the methodological criteria);
- a methodological quality summary (an indication of whether individual studies meet the methodological criteria);

- paired forest plots of sensitivity and specificity (which can be ordered and grouped by covariates);
- summary ROC plots (with variations for including lines, operating points, confidence and prediction regions, and indicating different tests and subgroups using colours).

Authors can also import their own additional figures. All figures need to be created within Figures, and captions specified. To be published with the review the Figures need to be linked to from the text.

Review authors can also include additional figures created by other software. Additional figures should never be used for content that can be included in other ways in RevMan5, for example as standard RevMan5 graphs, tables or created as additional tables. Guidance on technical aspects of preparing additional figures, including appropriate file size, guidance on labelling and captions is available in RevMan5 help files. The images authors upload as additional figures will not be edited or otherwise improved by others, but will be published 'as is'. It is therefore important that images are fully fit for publication. Important results from all additional figures should be summarized in the Results section of the review text. Wherever numerical results taken from a figure are reported in the text of the review their meaning and derivation should be clear, and a reference to the relevant figure should be provided.

If permission to publish a copyrighted figure is granted, the final phrase of the figure caption must be: "Copyright © [Year] [Name of copyright holder, or other required wording]: reproduced with permission." Warning! Large images take up lots of disk space. A single large image can easily take up ten times the total space used for the text and tables of the review. This leads to very large export files. Scanned images can be especially space-consuming because the resolution may be much higher than needed. Always use images with a good balance between resolution and detail, and include as few images as possible.

4.10 Sources of support to the review

Authors should give details of grants that supported the review and other forms of support, such as support from their university or institution in the form of a salary. Sources of support are divided into 'internal' (provided by the institutions at which the review was produced) and 'external' (provided by other institutions or funding agencies).

4.11 Feedback

This section is used to present feedback and comments from people who have read the published protocol or review. There are three subheadings: **Summary, Reply** and **Contributors**. The summary should be prepared by the feedback editor for the CRG in consultation, if necessary, with the person submitting the comment. A reply to this should then be prepared by the author(s) of the review. Details of the people who contributed to this process should be given. Further information on the comments and feedback and the updating of reviews is given in Chapter 10 of the *Cochrane Handbook for Systematic Reviews of Interventions* (Higgins 2006b).

4.12 Appendices

Appendices provide a place to publish additional material that might be of interest to readers of the review. Appendices should be considered as supplementary information, as they may not appear in some reduced formats of the published review. Examples of the types of material that authors may wish to present in Appendices include:

- Details of the search strategies used to identify studies for the review in each of the databases searched (the recommended place to put these);
- Detailed statistical methods;
- Data extraction forms;
- Correspondence with authors of studies.

4.13 Contributions to this chapter

Editors: Nynke Smidt, Jonathan Deeks, Theresa Moore.

Contributors: This chapter was based on:
Higgins JPT, Green S (editors). Chapter 4: Guide to the contents of a Cochrane protocol and review. In: Higgins JPT, Green S (editors). Cochrane Handbook for Systematic Reviews of Interventions Version 5.0.0 (updated February 2008). The Cochrane Collaboration, 2008. Available from http://www.cochrane.org/resources/handbook.

And

Schünemann HJ, Oxman AD, Higgins JPT, Vist GE, Glasziou P, Guyatt GH. Chapter 11: Presenting results and 'Summary of findings tables'. In: Higgins JPT, Green S (editors), Cochrane Handbook for Systematic Reviews of Interventions Version 5.0.0 (updated February 2008). The Cochrane Collaboration, 2008. Available from http://www.cochrane.org/resources/handbook.

Acknowledgements: We thank Anne Eisinga, Chris Hyde and Carol Lefevbre for helpful comments, Penny Whiting for providing an example abstract.

4.14 References

Dickersin 1986
Dickersin K, Hewitt P. Look before you quote. British Medical Journal 1986;293:1000-2.

Docherty 1999
Docherty M, Smith R. The case for structuring the discussion of scientific papers. BMJ 1999;318:1224-5.

Eichorn 1987
Eichorn P, Yankauer A. Do authors check their references? A survey of accuracy of references in three public health journals. American Journal of Public Health 1987;77:1011-12.

Flanagin 1998
Flanagin A, Carey LA, Fontanarosa PB, Philips SG, Pace BP, Lundberg GD, Rennie D. Prevalence of articles with honorary articles and ghost authors in peer-reviewed medical journals. JAMA 1998; 280:222-4.

Higgins 2006a
Higgins JPT, Green S, editors. Conflict of interest and commercial sponsorship. *Cochrane Handbook for Systematic Reviews of Interventions* 4.2.6 [updated September 2006]; Section 2.6. In: *The Cochrane Library*, Issue 4, 2006. Chichester, UK: John Wiley & Sons, Ltd.

Higgins 2006b
Higgins JPT, Green S, editors. Responding to criticisms. *Cochrane Handbook for Systematic Reviews of Interventions* 4.2.6 [updated September 2006]; Section 10.6. In: The *Cochrane Library*, Issue 4, 2006. Chichester, UK: John Wiley & Sons, Ltd.

ICMJE 2006
International Committee of Medical Journal Editors. Uniform requirements for manuscripts submitted to biomedical journals. Writing and editing for biomedical publication [Updated February 2006]. http://www.icmje.org (accessed 12 October 2007).

Littenberg 1993
Littenberg B, Moses LE. Estimating diagnostic accuracy from multiple conflicting reports: a new analytic method. Medical Decision Making 1993; 13: 313-21.

Rennie 1997
Rennie D, Emanuel L, Yank V. When authorship fails: a proposal to make contributors accountable. JAMA 1997;278:579-85.

Rennie 1998
Rennie D, Yank V. If authors become contributors, everyone would gain, especially the reader. American Journal of Public Health 1998;88:828-30.

Yank 1999
Yank V, Rennie D. Disclosure of researcher contributions: a study of original research articles in The Lancet. Annals of Internal Medicine 1999; 130: 661-70.

ANEXO 8: PRISMA

The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration

Alessandro Liberati,[1 2] Douglas G Altman,[3] Jennifer Tetzlaff,[4] Cynthia Mulrow,[5] Peter C Gøtzsche,[6] John P A Ioannidis,[7] Mike Clarke,[8 9] P J Devereaux,[10] Jos Kleijnen,[11 12] David Moher[4 13]

[1]Università di Modena e Reggio Emilia, Modena, Italy
[2]Centro Cochrane Italiano, Istituto Ricerche Farmacologiche Mario Negri, Milan, Italy
[3]Centre for Statistics in Medicine, University of Oxford, Oxford
[4]Ottawa Methods Centre, Ottawa Hospital Research Institute, Ottawa, Ontario, Canada
[5]Annals of Internal Medicine, Philadelphia, Pennsylvania, USA
[6]Nordic Cochrane Centre, Copenhagen, Denmark
[7]Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece
[8]UK Cochrane Centre, Oxford
[9]School of Nursing and Midwifery, Trinity College, Dublin, Republic of Ireland
[10]Departments of Medicine, Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada
[11]Kleijnen Systematic Reviews, York
[12]School for Public Health and Primary Care (CAPHRI), University of Maastricht, Maastricht, Netherlands
[13]Department of Epidemiology and Community Medicine, Faculty of Medicine, Ottawa, Ontario, Canada

Correspondence to: alesslib@mailbase.it

Accepted: 5 June 2009

Cite this as: *BMJ* 2009;339:b2700
doi: 10.1136/bmj.b2700

Introduction

Systematic reviews and meta-analyses are essential tools for summarising evidence accurately and reliably. They help clinicians keep up to date; provide evidence for policy makers to judge risks, benefits, and harms of healthcare behaviours and interventions; gather together and summarise related research for patients and their carers; provide a starting point for clinical practice guideline developers; provide summaries of previous research for funders wishing to support new research;[1] and help editors judge the merits of publishing reports of new studies.[2] Recent data suggest that at least 2500 new systematic reviews reported in English are indexed in Medline annually.[3]

Unfortunately, there is considerable evidence that key information is often poorly reported in systematic reviews, thus diminishing their potential usefulness.[3-6] As is true for all research, systematic reviews should be reported fully and transparently to allow readers to assess the strengths and weaknesses of the investigation.[7] That rationale led to the development of the QUOROM (quality of reporting of meta-analysis) statement; those detailed reporting recommendations were published in 1999.[8] In this paper we describe the updating of that guidance. Our aim is to ensure clear presentation of what was planned, done, and found in a systematic review.

Terminology used to describe systematic reviews and meta-analyses has evolved over time and varies across different groups of researchers and authors (see box 1 at end of document). In this document we adopt the definitions used by the Cochrane Collaboration.[9] A systematic review attempts to collate all empirical evidence that fits pre-specified eligibility criteria to answer a specific research question. It uses explicit, systematic methods that are selected to minimise bias, thus providing reliable findings from which conclusions can be drawn and decisions made. Meta-analysis is the use of statistical methods to summarise and combine the results of independent studies. Many systematic reviews contain meta-analyses, but not all.

The QUOROM statement and its evolution into PRISMA

The QUOROM statement, developed in 1996 and published in 1999,[8] was conceived as a reporting guidance for authors reporting a meta-analysis of randomised trials. Since then, much has happened. First, knowledge about the conduct and reporting of systematic reviews has expanded considerably. For example, the Cochrane Library's Methodology Register (which includes reports of studies relevant to the methods for systematic reviews) now contains more than 11 000 entries (March 2009). Second, there have been many conceptual advances, such as "outcome-level" assessments of the risk of bias,[10 11] that apply to systematic reviews. Third, authors have increasingly used systematic reviews to summarise evidence other than that provided by randomised trials.

However, despite advances, the quality of the conduct and reporting of systematic reviews remains well short of ideal.[3-6] All of these issues prompted the need for an update and expansion of the QUOROM statement. Of note, recognising that the updated statement now addresses the above conceptual and methodological issues and may also have broader applicability than the original QUOROM statement, we changed the name of the reporting guidance to PRISMA (preferred reporting items for systematic reviews and meta-analyses).

Development of PRISMA

The PRISMA statement was developed by a group of 29 review authors, methodologists, clinicians, medical editors, and consumers.[12] They attended a three day meeting in 2005 and participated in extensive post-meeting electronic correspondence. A consensus process that was informed by evidence, whenever possible, was used to develop a 27-item checklist (table 1) and a four-phase flow diagram (fig 1) (also available as extra items on bmj.com for researchers to download and re-use). Items deemed essential for transparent reporting of a systematic review were included in the checklist. The flow diagram originally proposed by QUOROM was also modified to show numbers of identified records, excluded articles, and included studies. After 11 revisions the group approved the checklist, flow diagram, and this explanatory paper.

The PRISMA statement itself provides further details regarding its background and development.[12] This

accompanying explanation and elaboration document explains the meaning and rationale for each checklist item. A few PRISMA Group participants volunteered to help draft specific items for this document, and four of these (DGA, AL, DM, and JT) met on several occasions to further refine the document, which was circulated and ultimately approved by the larger PRISMA Group.

Scope of PRISMA

PRISMA focuses on ways in which authors can ensure the transparent and complete reporting of systematic reviews and meta-analyses. It does not address directly or in a detailed manner the conduct of systematic reviews, for which other guides are available.[13-16]

We developed the PRISMA statement and this explanatory document to help authors report a wide array of systematic reviews to assess the benefits and harms of a healthcare intervention. We consider most of the checklist items relevant when reporting systematic reviews of non-randomised studies assessing the benefits and harms of interventions. However, we recognise that authors who address questions relating to aetiology, diagnosis, or prognosis, for example, and who review epidemiological or diagnostic accuracy studies may need to modify or incorporate additional items for their systematic reviews.

How to use this paper

We modeled this explanation and elaboration document after those prepared for other reporting guidelines.[17-19] To maximise the benefit of this document, we encourage people to read it in conjunction with the PRISMA statement.[11]

We present each checklist item and follow it with a published exemplar of good reporting for that item. (We edited some examples by removing citations or web addresses, or by spelling out abbreviations.) We then explain the pertinent issue, the rationale for including the item, and relevant evidence from the literature, whenever possible. No systematic search was carried out to identify exemplars and evidence. We also include seven boxes at the end of the document that provide a more comprehensive explanation of certain thematic aspects of the methodology and conduct of systematic reviews.

Although we focus on a minimal list of items to consider when reporting a systematic review, we indicate places where additional information is desirable to improve transparency of the review process. We present the items numerically from 1 to 27; however, authors need not address items in this particular order in their reports. Rather, what is important is that the information for each item is given somewhere within the report.

The PRISMA checklist

Title and abstract

Item 1: Title

Identify the report as a systematic review, meta-analysis, or both.

Examples "Recurrence rates of video-assisted thoracoscopic versus open surgery in the prevention of recurrent pneumothoraces: a systematic review of randomised and non-randomised trials"[20]

"Mortality in randomised trials of antioxidant supplements for primary and secondary prevention: systematic review and meta-analysis"[21]

Explanation Authors should identify their report as a systematic review or meta-analysis. Terms such as "review" or "overview" do not describe for readers whether the review was systematic or whether a meta-analysis was performed. A recent survey found that 50% of 300 authors did not mention the terms "systematic review" or "meta-analysis" in the title or abstract of their systematic review.[3] Although sensitive search strategies have been developed to identify systematic reviews,[22] inclusion of the terms systematic review or meta-analysis in the title may improve indexing and identification.

We advise authors to use informative titles that make key information easily accessible to readers. Ideally, a title reflecting the PICOS approach (participants, interventions, comparators, outcomes, and study design) (see item 11 and box 2) may help readers as it provides key information about the scope of the review. Specifying the design(s) of the studies included, as shown in the examples, may also help some readers and those searching databases.

Some journals recommend "indicative titles" that indicate the topic matter of the review, while others require declarative titles that give the review's main conclusion. Busy practitioners may prefer to see the conclusion of the review in the title, but declarative titles can oversimplify or exaggerate findings. Thus, many journals and methodologists prefer indicative titles as used in the examples above.

Item 2: Structured summary

Provide a structured summary including, as applicable, background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; funding for the systematic review; and systematic review registration number.

Example "*Context*: The role and dose of oral vitamin D supplementation in nonvertebral fracture prevention have not been well established.

Objective: To estimate the effectiveness of vitamin D supplementation in preventing hip and nonvertebral fractures in older persons.

Data Sources: A systematic review of English and non-English articles using MEDLINE and the Cochrane Controlled Trials Register (1960-2005), and EMBASE (1991-2005). Additional studies were identified by contacting clinical experts and searching bibliographies and abstracts presented at the American Society for Bone and Mineral Research (1995-2004). Search terms included randomised controlled trial (RCT), controlled clinical trial, random allocation, double-blind method, cholecalciferol, ergocalciferol, 25-hydroxyvitamin D, fractures, humans, elderly, falls, and bone density.

Study Selection: Only double-blind RCTs of oral vitamin D supplementation (cholecalciferol, ergocalciferol) with or without calcium supplementation vs calcium supplementation or placebo in older persons (>60 years) that examined hip or nonvertebral fractures were included.

Data Extraction: Independent extraction of articles by 2 authors using predefined data fields, including study quality indicators.

Data Synthesis: All pooled analyses were based on random-effects models. Five RCTs for hip fracture (n=9294) and 7 RCTs for nonvertebral fracture risk (n=9820) met our inclusion criteria. All trials used cholecalciferol. Heterogeneity among studies for both hip and nonvertebral fracture prevention was observed, which disappeared after pooling RCTs with low-dose (400 IU/d) and higher-dose vitamin D (700-800 IU/d), separately. A vitamin D dose of 700 to 800 IU/d reduced the relative risk (RR) of hip fracture by 26% (3 RCTs with 5572 persons; pooled RR, 0.74; 95% confidence interval [CI], 0.61-0.88) and any nonvertebral fracture by 23% (5 RCTs with 6098 persons; pooled RR, 0.77; 95% CI, 0.68-0.87) vs calcium or placebo. No significant benefit was observed for RCTs with 400 IU/d vitamin D (2 RCTs with 3722 persons; pooled RR for hip fracture, 1.15; 95% CI, 0.88-1.50; and pooled RR for any nonvertebral fracture, 1.03; 95% CI, 0.86-1.24).

Conclusions: Oral vitamin D supplementation between 700 to 800 IU/d appears to reduce the risk of hip and any nonvertebral fractures in ambulatory or institutionalised elderly persons. An oral vitamin D dose of 400 IU/d is not sufficient for fracture prevention."[29]

Explanation Abstracts provide key information that enables readers to understand the scope, processes, and findings of a review and to decide whether to read the full report. The abstract may be all that is readily available to a reader, for example, in a bibliographic database. The abstract should present a balanced and realistic assessment of the review's findings that mirrors, albeit briefly, the main text of the report.

We agree with others that the quality of reporting in abstracts presented at conferences and in journal publications needs improvement.[24 25] While we do not uniformly favour a specific format over another, we generally recommend structured abstracts. Structured abstracts provide readers with a series of headings pertaining to the purpose, conduct, findings, and conclusions of the systematic review being reported.[26 27] They give readers more complete information and facilitate finding information more easily than unstructured abstracts.[28-32]

A highly structured abstract of a systematic review could include the following headings: *Context* (or *Background*); *Objective* (or *Purpose*); *Data sources*; *Study selection* (or *Eligibility criteria*); *Study appraisal* and *Synthesis methods* (or *Data extraction* and *Data synthesis*); *Results*; *Limitations*; and *Conclusions* (or *Implications*). Alternatively, a simpler structure could cover but collapse some of the above headings (such as label *Study selection* and *Study appraisal* as *Review methods*) or omit some headings such as *Background* and *Limitations*.

In the highly structured abstract mentioned above, authors use the *Background* heading to set the context for readers and explain the importance of the review question. Under the *Objectives* heading, they ideally use elements of PICOS (see box 2) to state the primary objective of the review. Under a *Data sources* heading, they summarise sources that were searched, any language or publication type restrictions, and the start and end dates of searches. *Study selection* statements then ideally describe who selected studies using what inclusion criteria. *Data extraction methods* statements describe appraisal methods during data abstraction and the methods used to integrate or summarise the data. The *Data synthesis* section is where the main results of the review are reported. If the review includes meta-analyses, authors should provide numerical results with confidence intervals for the most important outcomes. Ideally, they should specify the amount of evidence in these analyses (numbers of studies and numbers of participants). Under a *Limitations* heading, authors might describe the most important weaknesses of included studies as well as limitations of the review process. Then authors should provide clear and balanced *Conclusions* that are closely linked to the objective and findings of the review. Additionally, it would be helpful if authors included some information about funding for the review. Finally, although protocol registration for systematic reviews is still not common practice, if authors have registered their review or received a registration number, we recommend providing the registration information at the end of the abstract.

Taking all the above considerations into account, the intrinsic tension between the goal of completeness of the abstract and its keeping into the space limit often set by journal editors is recognised as a major challenge.

Introduction

Item 3: Rationale

Describe the rationale for the review in the context of what is already known.

Example "Reversing the trend of increasing weight for height in children has proven difficult. It is widely accepted that increasing energy expenditure and reducing energy intake form the theoretical basis for management. Therefore, interventions aiming to increase physical activity and improve diet are the foundation of efforts to prevent and treat childhood obesity. Such lifestyle interventions have been supported by recent systematic reviews, as well as by the Canadian Paediatric Society, the Royal College of Paediatrics and Child Health, and the American Academy of Pediatrics. However, these interventions are fraught with poor adherence. Thus, school-based interventions are theoretically appealing because adherence with interventions can be improved. Consequently, many local governments have enacted or are considering policies that mandate increased physical activity in schools, although the effect of such interventions on body composition has not been assessed."[33]

Explanation Readers need to understand the rationale behind the study and what the systematic review may add to what is already known. Authors should tell readers whether their report is a new systematic review or an update of an existing one. If the review is an update, authors should state reasons for the update, including what has been added to the evidence

base since the previous version of the review.

An ideal background or introduction that sets context for readers might include the following. First, authors might define the importance of the review question from different perspectives (such as public health, individual patient, or health policy). Second, authors might briefly mention the current state of knowledge and its limitations. As in the above example, information about the effects of several different interventions may be available that helps readers understand why potential relative benefits or harms of particular interventions need review. Third, authors might whet readers' appetites by clearly stating what the review aims to add. They also could discuss the extent to which the limitations of the existing evidence base may be overcome by the review.

Item 4: Objectives

Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).

Example "To examine whether topical or intraluminal antibiotics reduce catheter-related bloodstream infection, we reviewed randomised, controlled trials that assessed the efficacy of these antibiotics for primary prophylaxis against catheter-related bloodstream infection and mortality compared with no antibiotic therapy in adults undergoing hemodialysis."[34]

Explanation The questions being addressed, and the rationale for them, are one of the most critical parts of a systematic review. They should be stated precisely and explicitly so that readers can understand quickly the review's scope and the potential applicability of the review to their interests.[35] Framing questions so that they include the following five "PICOS" components may improve the explicitness of review questions: (1) the patient population or disease being addressed (P), (2) the interventions or exposure of interest (I), (3) the comparators (C), (4) the main outcome or endpoint of interest (O), and (5) the study designs chosen (S). For more detail regarding PICOS, see box 2.

Good review questions may be narrowly focused or broad, depending on the overall objectives of the review. Sometimes broad questions might increase the applicability of the results and facilitate detection of bias, exploratory analyses, and sensitivity analyses.[35 36] Whether narrowly focused or broad, precisely stated review objectives are critical as they help define other components of the review process such as the eligibility criteria (item 6) and the search for relevant literature (items 7 and 8).

Methods

Item 5: Protocol and registration

Indicate if a review protocol exists, if and where it can be accessed (such as a web address), and, if available, provide registration information including the registration number.

Example "Methods of the analysis and inclusion criteria were specified in advance and documented in a protocol."[37]

Explanation A protocol is important because it pre-specifies the objectives and methods of the systematic review. For instance, a protocol specifies outcomes of primary interest, how reviewers will extract information about those outcomes, and methods that reviewers might use to quantitatively summarise the outcome data (see item 13). Having a protocol can help restrict the likelihood of biased post hoc decisions in review methods, such as selective outcome reporting. Several sources provide guidance about elements to include in the protocol for a systematic review.[16 38 39] For meta-analyses of individual patient-level data, we advise authors to describe whether a protocol was explicitly designed and whether, when, and how participating collaborators endorsed it.[40 41]

Authors may modify protocols during the research, and readers should not automatically consider such modifications inappropriate. For example, legitimate modifications may extend the period of searches to include older or newer studies, broaden eligibility criteria that proved too narrow, or add analyses if the primary analyses suggest that additional ones are warranted. Authors should, however, describe the modifications and explain their rationale.

Although worthwhile protocol amendments are common, one must consider the effects that protocol modifications may have on the results of a systematic review, especially if the primary outcome is changed. Bias from selective outcome reporting in randomised trials has been well documented.[42 43] An examination of 47 Cochrane reviews revealed indirect evidence for possible selective reporting bias for systematic reviews. Almost all (n=43) contained a major change, such as the addition or deletion of outcomes, between the protocol and the full publication.[44] Whether (or to what extent) the changes reflected bias, however, was not clear. For example, it has been rather common not to describe outcomes that were not presented in any of the included studies.

Registration of a systematic review, typically with a protocol and registration number, is not yet common, but some opportunities exist.[45 46] Registration may possibly reduce the risk of multiple reviews addressing the same question,[45-48] reduce publication bias, and provide greater transparency when updating systematic reviews. Of note, a survey of systematic reviews indexed in Medline in November 2004 found that reports of protocol use had increased to about 46%[3] from 8% noted in previous surveys.[49] The improvement was due mostly to Cochrane reviews, which, by requirement, have a published protocol.[3]

Item 6: Eligibility criteria

Specify study characteristics (such as PICOS, length of follow-up) and report characteristics (such as years considered, language, publication status) used as criteria for eligibility, giving rationale.

Examples Types of studies: "Randomised clinical trials studying the administration of hepatitis B vaccine to CRF [chronic renal failure] patients, with or without dialysis. No language, publication date, or publication status restrictions were imposed…"

Types of participants: "Participants of any age with

CRF or receiving dialysis (haemodialysis or peritoneal dialysis) were considered. CRF was defined as serum creatinine greater than 200 µmol/L for a period of more than six months or individuals receiving dialysis (haemodialysis or peritoneal dialysis)…Renal transplant patients were excluded from this review as these individuals are immunosuppressed and are receiving immunosuppressant agents to prevent rejection of their transplanted organs, and they have essentially normal renal function..."

Types of intervention: "Trials comparing the beneficial and harmful effects of hepatitis B vaccines with adjuvant or cytokine co-interventions [and] trials comparing the beneficial and harmful effects of immunoglobulin prophylaxis. This review was limited to studies looking at active immunisation. Hepatitis B vaccines (plasma or recombinant (yeast) derived) of all types, dose, and regimens versus placebo, control vaccine, or no vaccine…"

Types of outcome measures: "Primary outcome measures: Seroconversion, ie, proportion of patients with adequate anti-HBs response (>10 IU/L or Sample Ratio Units). Hepatitis B infections (as measured by hepatitis B core antigen (HBcAg) positivity or persistent HBsAg positivity), both acute and chronic. Acute (primary) HBV [hepatitis B virus] infections were defined as seroconversion to HBsAg positivity or development of IgM anti-HBc. Chronic HBV infections were defined as the persistence of HBsAg for more than six months or HBsAg positivity and liver biopsy compatible with a diagnosis or chronic hepatitis B. Secondary outcome measures: Adverse events of hepatitis B vaccinations… [and]…mortality."[50]

Explanation Knowledge of the eligibility criteria is essential in appraising the validity, applicability, and comprehensiveness of a review. Thus, authors should unambiguously specify eligibility criteria used in the review. Carefully defined eligibility criteria inform various steps of the review methodology. They influence the development of the search strategy and serve to ensure that studies are selected in a systematic and unbiased manner.

A study may be described in multiple reports, and one report may describe multiple studies. Therefore, we separate eligibility criteria into the following two components: study characteristics and report characteristics. Both need to be reported. Study eligibility criteria are likely to include the populations, interventions, comparators, outcomes, and study designs of interest (PICOS, see box 2), as well as other study-specific elements, such as specifying a minimum length of follow-up. Authors should state whether studies will be excluded because they do not include (or report) specific outcomes to help readers ascertain whether the systematic review may be biased as a consequence of selective reporting.[42 43]

Report eligibility criteria are likely to include language of publication, publication status (such as inclusion of unpublished material and abstracts), and year of publication. Inclusion or not of non-English language literature,[51-55] unpublished data, or older data can influence the effect estimates in meta-analyses.[56-59] Caution may need to be exercised in including all identified studies due to potential differences in the risk of bias such as, for example, selective reporting in abstracts.[60-62]

Item 7: Information sources

Describe all information sources in the search (such as databases with dates of coverage, contact with study authors to identify additional studies) and date last searched.

Example "Studies were identified by searching electronic databases, scanning reference lists of articles and consultation with experts in the field and drug companies…No limits were applied for language and foreign papers were translated. This search was applied to Medline (1966 - Present), CancerLit (1975 - Present), and adapted for Embase (1980 - Present), Science Citation Index Expanded (1981 - Present) and Pre-Medline electronic databases. Cochrane and DARE (Database of Abstracts of Reviews of Effectiveness) databases were reviewed…The last search was run on 19 June 2001. In addition, we handsearched contents pages of Journal of Clinical Oncology 2001, European Journal of Cancer 2001 and Bone 2001, together with abstracts printed in these journals 1999 - 2001. A limited update literature search was performed from 19 June 2001 to 31 December 2003."[63]

Explanation The National Library of Medicine's Medline database is one of the most comprehensive sources of healthcare information in the world. Like any database, however, its coverage is not complete and varies according to the field. Retrieval from any single database, even by an experienced searcher, may be imperfect, which is why detailed reporting is important within the systematic review.

At a minimum, for each database searched, authors should report the database, platform, or provider (such as Ovid, Dialog, PubMed) and the start and end dates for the search of each database. This information lets readers assess the currency of the review, which is important because the publication time-lag outdates the results of some reviews.[64] This information should also make updating more efficient.[65] Authors should also report who developed and conducted the search.[66]

In addition to searching databases, authors should report the use of supplementary approaches to identify studies, such as hand searching of journals, checking reference lists, searching trials registries or regulatory agency websites,[67] contacting manufacturers, or contacting authors. Authors should also report if they attempted to acquire any missing information (such as on study methods or results) from investigators or sponsors; it is useful to describe briefly who was contacted and what unpublished information was obtained.

Item 8: Search

Present the full electronic search strategy for at least one major database, including any limits used, such that it could be repeated.

Examples In text: "We used the following search terms to search all trials registers and databases: immunoglobulin*; IVIG; sepsis; septic shock; septicaemia; and septicemia…"[68]

In appendix: "Search strategy: MEDLINE (OVID)
01. immunoglobulins/
02. immunoglobulin$.tw.
03. ivig.tw.
04. 1 or 2 or 3
05. sepsis/
06. sepsis.tw.
07. septic shock/
08. septic shock.tw.
09. septicemia/
10. septicaemia.tw.
11. septicemia.tw.
12. 5 or 6 or 7 or 8 or 9 or 10 or 11
13. 4 and 12
14. randomised controlled trials/
15. randomised-controlled-trial.pt.
16. controlled-clinical-trial.pt.
17. random allocation/
18. double-blind method/
19. single-blind method/
20. 14 or 15 or 16 or 17 or 18 or 19
21. exp clinical trials/
22. clinical-trial.pt.
23. (clin$ adj trial$).ti,ab.
24. ((singl$ or doubl$ or trebl$ or tripl$) adj (blind$)).ti,ab.
25. placebos/
26. placebo$.ti,ab.
27. random$.ti,ab.
28. 21 or 22 or 23 or 24 or 25 or 26 or 27
29. research design/
30. comparative study/
31. exp evaluation studies/
32. follow-up studies/
33. prospective studies/
34. (control$ or prospective$ or volunteer$).ti,ab.
35. 30 or 31 or 32 or 33 or 34
36. 20 or 28 or 29 or 35
37. 13 and 36"[68]

Explanation The search strategy is an essential part of the report of any systematic review. Searches may be complicated and iterative, particularly when reviewers search unfamiliar databases or their review is addressing a broad or new topic. Perusing the search strategy allows interested readers to assess the comprehensiveness and completeness of the search, and to replicate it. Thus, we advise authors to report their full electronic search strategy for at least one major database. As an alternative to presenting search strategies for all databases, authors could indicate how the search took into account other databases searched, as index terms vary across databases. If different searches are used for different parts of a wider question (such as questions relating to benefits and questions relating to harms), we recommend authors provide at least one example of a strategy for each part of the objective.[69] We also encourage authors to state whether search strategies were peer reviewed as part of the systematic review process.[70]

We realise that journal restrictions vary and that having the search strategy in the text of the report is not always feasible. We strongly encourage all journals, however, to find ways—such as a "web extra," appendix, or electronic link to an archive—to make search strategies accessible to readers. We also advise all authors to archive their searches so that (1) others may access and review them (such as replicate them or understand why their review of a similar topic did not identify the same reports), and (2) future updates of their review are facilitated.

Several sources provide guidance on developing search strategies.[71-73] Most searches have constraints, such as relating to limited time or financial resources, inaccessible or inadequately indexed reports and databases, unavailability of experts with particular language or database searching skills, or review questions for which pertinent evidence is not easy to find. Authors should be straightforward in describing their search constraints. Apart from the keywords used to identify or exclude records, they should report any additional limitations relevant to the search, such as language and date restrictions (see also eligibility criteria, item 6).[51]

Item 9: Study selection

State the process for selecting studies (that is, for screening, for determining eligibility, for inclusion in the systematic review, and, if applicable, for inclusion in the meta-analysis).

Example "Eligibility assessment…[was] performed independently in an unblinded standardized manner by 2 reviewers…Disagreements between reviewers were resolved by consensus."[74]

Explanation There is no standard process for selecting studies to include in a systematic review. Authors usually start with a large number of identified records from their search and sequentially exclude records according to eligibility criteria. We advise authors to report how they screened the retrieved records (typically a title and abstract), how often it was necessary to review the full text publication, and if any types of record (such as letters to the editor) were excluded. We also advise using the PRISMA flow diagram to summarise study selection processes (see item 17 and box 3).

Efforts to enhance objectivity and avoid mistakes in study selection are important. Thus authors should report whether each stage was carried out by one or several people, who these people were, and, whenever multiple independent investigators performed the selection, what the process was for resolving disagreements. The use of at least two investigators may reduce the possibility of rejecting relevant reports.[75] The benefit may be greatest for topics where selection or rejection of an article requires difficult judgments.[76] For these topics, authors should ideally tell readers the level of inter-rater agreement, how commonly arbitration about selection was required, and what efforts were made to resolve disagreements (such as by contact with the authors of the original studies).

Item 10: Data collection process

Describe the method of data extraction from reports (such as piloted forms, independently by two reviewers) and any processes for obtaining and confirming data from investigators.

Example "We developed a data extraction sheet (based on the Cochrane Consumers and Communication Review Group's data extraction template), pilot-tested it on ten randomly-selected included studies, and refined it accordingly. One review author extracted the following data from included studies and the second author checked the extracted data…Disagreements were resolved by discussion between the two review authors; if no agreement could be reached, it was planned a third author would decide. We contacted five authors for further information. All responded and one provided numerical data that had only been presented graphically in the published paper."[77]

Explanation Reviewers extract information from each included study so that they can critique, present, and summarise evidence in a systematic review. They might also contact authors of included studies for information that has not been, or is unclearly, reported. In meta-analysis of individual patient data, this phase involves collection and scrutiny of detailed raw databases. The authors should describe these methods, including any steps taken to reduce bias and mistakes during data collection and data extraction.[78] (See box 3)

Some systematic reviewers use a data extraction form that could be reported as an appendix or "Web extra" to their report. These forms could show the reader what information reviewers sought (see item 11) and how they extracted it. Authors could tell readers if the form was piloted. Regardless, we advise authors to tell readers who extracted what data, whether any extractions were completed in duplicate, and, if so, whether duplicate abstraction was done independently and how disagreements were resolved.

Published reports of the included studies may not provide all the information required for the review. Reviewers should describe any actions they took to seek additional information from the original researchers (see item 7). The description might include how they attempted to contact researchers, what they asked for, and their success in obtaining the necessary information. Authors should also tell readers when individual patient data were sought from the original researchers.[41] (see item 11) and indicate the studies for which such data were used in the analyses. The reviewers ideally should also state whether they confirmed the accuracy of the information included in their review with the original researchers, for example, by sending them a copy of the draft review.[79]

Some studies are published more than once. Duplicate publications may be difficult to ascertain, and their inclusion may introduce bias.[80 81] We advise authors to describe any steps they used to avoid double counting and piece together data from multiple reports of the same study (such as juxtaposing author names, treatment comparisons, sample sizes, or outcomes). We also advise authors to indicate whether all reports on a study were considered, as inconsistencies may reveal important limitations. For example, a review of multiple publications of drug trials showed that reported study characteristics may differ from report to report, including the description of the design, number of patients analysed, chosen significance level, and outcomes.[82] Authors ideally should present any algorithm that they used to select data from overlapping reports and any efforts they used to solve logical inconsistencies across reports.

Item 11: Data items

List and define all variables for which data were sought (such as PICOS, funding sources) and any assumptions and simplifications made.

Examples "Information was extracted from each included trial on: (1) characteristics of trial participants (including age, stage and severity of disease, and method of diagnosis), and the trial's inclusion and exclusion criteria; (2) type of intervention (including type, dose, duration and frequency of the NSAID [non-steroidal anti-inflammatory drug]; versus placebo or versus the type, dose, duration and frequency of another NSAID; or versus another pain management drug; or versus no treatment); (3) type of outcome measure (including the level of pain reduction, improvement in quality of life score (using a validated scale), effect on daily activities, absence from work or school, length of follow up, unintended effects of treatment, number of women requiring more invasive treatment)."[83]

Explanation It is important for readers to know what information review authors sought, even if some of this information was not available.[84] If the review is limited to reporting only those variables that were obtained, rather than those that were deemed important but could not be obtained, bias might be introduced and the reader might be misled. It is therefore helpful if authors can refer readers to the protocol (see item 5) and archive their extraction forms (see item 10), including definitions of variables. The published systematic review should include a description of the processes used with, if relevant, specification of how readers can get access to additional materials.

We encourage authors to report whether some variables were added after the review started. Such variables might include those found in the studies that the reviewers identified (such as important outcome measures that the reviewers initially overlooked). Authors should describe the reasons for adding any variables to those already pre-specified in the protocol so that readers can understand the review process.

We advise authors to report any assumptions they made about missing or unclear information and to explain those processes. For example, in studies of women aged 50 or older it is reasonable to assume that none were pregnant, even if this is not reported. Likewise, review authors might make assumptions about the route of administration of drugs assessed. However, special care should be taken in making assumptions about qualitative information. For example, the upper age limit for "children" can vary from 15 years to 21 years, "intense" physiotherapy might mean very different things to different researchers at different times and for different patients, and the volume of blood associated with "heavy" blood loss might vary widely depending on the setting.

Item 12: Risk of bias in individual studies

Describe methods used for assessing risk of bias in individual studies (including specification of whether this was done at the study or outcome level, or both), and how this information is to be used in any data synthesis.

Example "To ascertain the validity of eligible randomized trials, pairs of reviewers working independently and with adequate reliability determined the adequacy of randomization and concealment of allocation, blinding of patients, health care providers, data collectors, and outcome assessors; and extent of loss to follow-up (i.e. proportion of patients in whom the investigators were not able to ascertain outcomes)."[85]

"To explore variability in study results (heterogeneity) we specified the following hypotheses before conducting the analysis. We hypothesised that effect size may differ according to the methodological quality of the studies."[86]

Explanation The likelihood that the treatment effect reported in a systematic review approximates the truth depends on the validity of the included studies, as certain methodological characteristics may be associated with effect sizes.[87 88] For example, trials without reported adequate allocation concealment exaggerate treatment effects on average compared with those with adequate concealment.[88] Therefore, it is important for authors to describe any methods that they used to gauge the risk of bias in the included studies and how that information was used.[89] Additionally, authors should provide a rationale if no assessment of risk of bias was undertaken. The most popular term to describe the issues relevant to this item is "quality," but for the reasons that are elaborated in box 4 we prefer to name this item as "assessment of risk of bias."

Many methods exist to assess the overall risk of bias in included studies, including scales, checklists, and individual components.[90 91] As discussed in box 4, scales that numerically summarise multiple components into a single number are misleading and unhelpful.[92 93] Rather, authors should specify the methodological components that they assessed. Common markers of validity for randomised trials include the following: appropriate generation of random allocation sequence;[94] concealment of the allocation sequence;[95] blinding of participants, health care providers, data collectors, and outcome adjudicators;[95-98] proportion of patients lost to follow-up;[99 100] stopping of trials early for benefit;[101] and whether the analysis followed the intention-to-treat principle.[100 102] The ultimate decision regarding which methodological features to evaluate requires consideration of the strength of the empiric data, theoretical rationale, and the unique circumstances of the included studies.

Authors should report how they assessed risk of bias; whether it was in a blind manner; and if assessments were completed by more than one person, and if so, whether they were completed independently.[103 104] Similarly, we encourage authors to report any calibration exercises among review team members that were done. Finally, authors need to report how their assessments of risk of bias are used subsequently in the data synthesis (see item 16). Despite the often difficult task of assessing the risk of bias in included studies, authors are sometimes silent on what they did with the resultant assessments.[89] If authors exclude studies from the review or any subsequent analyses on the basis of the risk of bias, they should tell readers which studies they excluded and explain the reasons for those exclusions (see item 6). Authors should also describe any planned sensitivity or subgroup analyses related to bias assessments (see item 16).

Item 13: Summary measures

State the principal summary measures (such as risk ratio, difference in means).

Examples "Relative risk of mortality reduction was the primary measure of treatment effect."[105]

"The meta-analyses were performed by computing relative risks (RRs) using random-effects model. Quantitative analyses were performed on an intention-to-treat basis and were confined to data derived from the period of follow-up. RR and 95% confidence intervals for each side effect (and all side effects) were calculated."[106]

"The primary outcome measure was the mean difference in $\log_{10}$ HIV-1 viral load comparing zinc supplementation to placebo..."[107]

Explanation When planning a systematic review, it is generally desirable that authors pre-specify the outcomes of primary interest (see item 5) as well as the intended summary effect measure for each outcome. The chosen summary effect measure may differ from that used in some of the included studies. If possible the choice of effect measures should be explained, though it is not always easy to judge in advance which measure is the most appropriate.

For binary outcomes, the most common summary measures are the risk ratio, odds ratio, and risk difference.[108] Relative effects are more consistent across studies than absolute effects,[109 110] although absolute differences are important when interpreting findings (see item 24).

For continuous outcomes, the natural effect measure is the difference in means.[108] Its use is appropriate when outcome measurements in all studies are made on the same scale. The standardised difference in means is used when the studies do not yield directly comparable data. Usually this occurs when all studies assess the same outcome but measure it in a variety of ways (such as different scales to measure depression).

For time-to-event outcomes, the hazard ratio is the most common summary measure. Reviewers need the log hazard ratio and its standard error for a study to be included in a meta-analysis.[111] This information may not be given for all studies, but methods are available for estimating the desired quantities from other reported information.[111] Risk ratio and odds ratio (in relation to events occurring by a fixed time) are not equivalent to the hazard ratio, and median survival times are not a reliable basis for meta-analysis.[112] If authors have used these measures they should describe their methods in the report.

Item 14: Planned methods of analysis

Describe the methods of handling data and combining results of studies, if done, including measures of consistency (such as I^2) for each meta-analysis.

Examples "We tested for heterogeneity with the

Breslow-Day test, and used the method proposed by Higgins et al. to measure inconsistency (the percentage of total variation across studies due to heterogeneity) of effects across lipid-lowering interventions. The advantages of this measure of inconsistency (termed I^2) are that it does not inherently depend on the number of studies and is accompanied by an uncertainty interval."[113]

"In very few instances, estimates of baseline mean or mean QOL [Quality of life] responses were obtained without corresponding estimates of variance (standard deviation [SD] or standard error). In these instances, an SD was imputed from the mean of the known SDs. In a number of cases, the response data available were the mean and variance in a pre study condition and after therapy. The within-patient variance in these cases could not be calculated directly and was approximated by assuming independence."[114]

Explanation The data extracted from the studies in the review may need some transformation (processing) before they are suitable for analysis or for presentation in an evidence table. Although such data handling may facilitate meta-analyses, it is sometimes needed even when meta-analyses are not done. For example, in trials with more than two intervention groups it may be necessary to combine results for two or more groups (such as receiving similar but non-identical interventions), or it may be desirable to include only a subset of the data to match the review's inclusion criteria. When several different scales (such as for depression) are used across studies, the sign of some scores may need to be reversed to ensure that all scales are aligned (such as so low values represent good health on all scales). Standard deviations may have to be reconstructed from other statistics such as P values and *t* statistics,[115 116] or occasionally they may be imputed from the standard deviations observed in other studies.[117] Time-to-event data also usually need careful conversions to a consistent format.[111] Authors should report details of any such data processing.

Statistical combination of data from two or more separate studies in a meta-analysis may be neither necessary nor desirable (see box 5 and item 21). Regardless of the decision to combine individual study results, authors should report how they planned to evaluate between-study variability (heterogeneity or inconsistency) (box 6). The consistency of results across trials may influence the decision of whether to combine trial results in a meta-analysis.

When meta-analysis is done, authors should specify the effect measure (such as relative risk or mean difference) (see item 13), the statistical method (such as inverse variance), and whether a fixed-effects or random-effects approach, or some other method (such as Bayesian) was used (see box 6). If possible, authors should explain the reasons for those choices.

Item 15: Risk of bias across studies

Specify any assessment of risk of bias that may affect the cumulative evidence (such as publication bias, selective reporting within studies).

Examples "For each trial we plotted the effect by the inverse of its standard error. The symmetry of such 'funnel plots' was assessed both visually, and formally with Egger's test, to see if the effect decreased with increasing sample size."[118]

"We assessed the possibility of publication bias by evaluating a funnel plot of the trial mean differences for asymmetry, which can result from the non publication of small trials with negative results…Because graphical evaluation can be subjective, we also conducted an adjusted rank correlation test and a regression asymmetry test as formal statistical tests for publication bias...We acknowledge that other factors, such as differences in trial quality or true study heterogeneity, could produce asymmetry in funnel plots."[119]

Explanation Reviewers should explore the possibility that the available data are biased. They may examine results from the available studies for clues that suggest there may be missing studies (publication bias) or missing data from the included studies (selective reporting bias) (see box 7). Authors should report in detail any methods used to investigate possible bias across studies.

It is difficult to assess whether within-study selective reporting is present in a systematic review. If a protocol of an individual study is available, the outcomes in the protocol and the published report can be compared. Even in the absence of a protocol, outcomes listed in the methods section of the published report can be compared with those for which results are presented.[120] In only half of 196 trial reports describing comparisons of two drugs in arthritis were all the effect variables in the methods and results sections the same.[82] In other cases, knowledge of the clinical area may suggest that it is likely that the outcome was measured even if it was not reported. For example, in a particular disease, if one of two linked outcomes is reported but the other is not, then one should question whether the latter has been selectively omitted.[121 122]

Only 36% (76 of 212) of therapeutic systematic reviews published in November 2004 reported that study publication bias was considered, and only a quarter of those intended to carry out a formal assessment for that bias.[3] Of 60 meta-analyses in 24 articles published in 2005 in which formal assessments were reported, most were based on fewer than 10 studies; most displayed statistically significant heterogeneity; and many reviewers misinterpreted the results of the tests employed.[123] A review of trials of antidepressants found that meta-analysis of only the published trials gave effect estimates 32% larger on average than when all trials sent to the drug agency were analysed.[67]

Item 16: Additional analyses

Describe methods of additional analyses (such as sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.

Example "Sensitivity analyses were pre-specified. The treatment effects were examined according to quality components (concealed treatment allocation, blinding of patients and caregivers, blinded outcome assessment), time to initiation of statins, and the type of statin. One post-hoc sensitivity analysis was conducted including unpublished data from a trial using cerivastatin."[124]

Explanation Authors may perform additional

analyses to help understand whether the results of their review are robust, all of which should be reported. Such analyses include sensitivity analysis, subgroup analysis, and meta-regression.[125]

Sensitivity analyses are used to explore the degree to which the main findings of a systematic review are affected by changes in its methods or in the data used from individual studies (such as study inclusion criteria, results of risk of bias assessment). Subgroup analyses address whether the summary effects vary in relation to specific (usually clinical) characteristics of the included studies or their participants. Meta-regression extends the idea of subgroup analysis to the examination of the quantitative influence of study characteristics on the effect size.[126] Meta-regression also allows authors to examine the contribution of different variables to the heterogeneity in study findings. Readers of systematic reviews should be aware that meta-regression has many limitations, including a danger of over-interpretation of findings.[127 128]

Even with limited data, many additional analyses can be undertaken. The choice of which analysis to undertake will depend on the aims of the review. None of these analyses, however, is exempt from producing potentially misleading results. It is important to inform readers whether these analyses were performed, their rationale, and which were pre-specified.

Results

Item 17: Study selection

Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.

Examples In text: "A total of 10 studies involving 13 trials were identified for inclusion in the review. The search of Medline, PsycInfo and Cinahl databases provided a total of 584 citations. After adjusting for duplicates 509 remained. Of these, 479 studies were discarded because after reviewing the abstracts it appeared that these papers clearly did not meet the criteria. Three additional studies…were discarded because full text of the study was not available or the paper could not be feasibly translated into English. The full text of the remaining 27 citations was examined in more detail. It appeared that 22 studies did not meet the inclusion criteria as described. Five studies…met the inclusion criteria and were included in the systematic review. An additional five studies...that met the criteria for inclusion were identified by checking the references of located, relevant papers and searching for studies that have cited these papers. No unpublished relevant studies were obtained."[129]

See flow diagram in fig 2.

Explanation Authors should report, ideally with a flow diagram, the total number of records identified from electronic bibliographic sources (including specialised database or registry searches), hand searches of various sources, reference lists, citation indices, and experts. It is useful if authors delineate for readers the number of selected articles that were identified from the different sources so that they can see, for example, whether most articles were identified through electronic bibliographic sources or from references or experts. Literature identified primarily from references or experts may be prone to citation or publication bias.[131 132]

The flow diagram and text should describe clearly the process of report selection throughout the review. Authors should report unique records identified in searches, records excluded after preliminary screening (such as screening of titles and abstracts), reports retrieved for detailed evaluation, potentially eligible reports that were not retrievable, retrieved reports that did not meet inclusion criteria and the primary reasons for exclusion, and the studies included in the review. Indeed, the most appropriate layout may vary for different reviews.

Authors should also note the presence of duplicate or supplementary reports so that readers understand the number of individual studies compared with the number of reports that were included in the review. Authors should be consistent in their use of terms, such as whether they are reporting on counts of citations, records, publications, or studies. We believe that reporting the number of studies is the most important.

A flow diagram can be very useful; it should depict all the studies included based on fulfilling the eligibility criteria, and whether data have been combined for statistical analysis. A recent review of 87 systematic reviews found that about half included a QUOROM flow diagram.[133] The authors of this research recommended some important ways that reviewers can improve the use of a flow diagram when describing the flow of information throughout the review process, including a separate flow diagram for each important outcome reported.[133]

Item 18: Study characteristics

For each study, present characteristics for which data were extracted (such as study size, PICOS, follow-up period) and provide the citation.

Examples In text: "*Characteristics of included studies*

Methods

All four studies finally selected for the review were randomised controlled trials published in English. The duration of the intervention was 24 months for the RIO-North America and 12 months for the RIO-Diabetes, RIO-Lipids and RIO-Europe study. Although the last two described a period of 24 months during which they were conducted, only the first 12-months results are provided. All trials had a run-in, as a single blind period before the randomisation.

Participants

The included studies involved 6625 participants. The main inclusion criteria entailed adults (18 years or older), with a body mass index greater than 27 kg/m^2 and less than 5 kg variation in body weight within the three months before study entry.

Intervention

All trials were multicentric. The RIO-North America was conducted in the USA and Canada, RIO-Europe

in Europe and the USA, RIO-Diabetes in the USA and 10 other different countries not specified, and RIO-Lipids in eight unspecified different countries.

The intervention received was placebo, 5 mg of rimonabant or 20 mg of rimonabant once daily in addition to a mild hypocaloric diet (600 kcal/day deficit).

Outcomes

Primary

In all studies the primary outcome assessed was weight change from baseline after one year of treatment and the RIO-North America study also evaluated the prevention of weight regain between the first and second year. All studies evaluated adverse effects, including those of any kind and serious events. Quality of life was measured in only one study, but the results were not described (RIO-Europe).

Secondary and additional outcomes

These included prevalence of metabolic syndrome after one year and change in cardiometabolic risk factors such as blood pressure, lipid profile, etc.

No study included mortality and costs as outcome.

The timing of outcome measures was variable and could include monthly investigations, evaluations every three months or a single final evaluation after one year."[134]

In table: See table 2.

Explanation For readers to gauge the validity and applicability of a systematic review's results, they need to know something about the included studies. Such information includes PICOS (box 2) and specific information relevant to the review question. For example, if the review is examining the long term effects of antidepressants for moderate depressive disorder, authors should report the follow-up periods of the included studies. For each included study, authors should provide a citation for the source of their information regardless of whether or not the study is published. This information makes it easier for interested readers to retrieve the relevant publications or documents.

Reporting study-level data also allows the comparison of the main characteristics of the studies included in the review. Authors should present enough detail to allow readers to make their own judgments about the relevance of included studies. Such information also makes it possible for readers to conduct their own subgroup analyses and interpret subgroups, based on study characteristics.

Authors should avoid, whenever possible, assuming information when it is missing from a study report (such as sample size, method of randomisation). Reviewers may contact the original investigators to try to obtain missing information or confirm the data extracted for the systematic review. If this information is not obtained, this should be noted in the report. If information is imputed, the reader should be told how this was done and for which items. Presenting study-level data makes it possible to clearly identify unpublished information obtained from the original researchers and make it available for the public record.

Typically, study-level characteristics are presented as a table as in the example (table 2). Such presentation ensures that all pertinent items are addressed and that missing or unclear information is clearly indicated. Although paper based journals do not generally allow for the quantity of information available in electronic journals or Cochrane reviews, this should not be accepted as an excuse for omission of important aspects of the methods or results of included studies, since these can, if necessary, be shown on a website.

Following the presentation and description of each included study, as discussed above, reviewers usually provide a narrative summary of the studies. Such a summary provides readers with an overview of the included studies. It may, for example, address the languages of the published papers, years of publication, and geographic origins of the included studies.

The PICOS framework is often helpful in reporting the narrative summary indicating, for example, the clinical characteristics and disease severity of the participants and the main features of the intervention and of the comparison group. For non-pharmacological interventions, it may be helpful to specify for each study the key elements of the intervention received by each group. Full details of the interventions in included studies were reported in only three of 25 systematic reviews relevant to general practice.[84]

Item 19: Risk of bias within studies

Present data on risk of bias of each study and, if available, any outcome-level assessment (see item 12).

Example See table 3.

Explanation We recommend that reviewers assess the risk of bias in the included studies using a standard approach with defined criteria (see item 12). They should report the results of any such assessments.[89]

Reporting only summary data (such as "two of eight trials adequately concealed allocation") is inadequate because it fails to inform readers which studies had the particular methodological shortcoming. A more informative approach is to explicitly report the methodological features evaluated for each study. The Cochrane Collaboration's new tool for assessing the risk of bias also requests that authors substantiate these assessments with any relevant text from the original studies.[11] It is often easiest to provide these data in a tabular format, as in the example. However, a narrative summary describing the tabular data can also be helpful for readers.

Item 20: Results of individual studies

For all outcomes considered (benefits and harms), present, for each study, simple summary data for each intervention group and effect estimates and confidence intervals, ideally with a forest plot.

Examples See table 4 and fig 3.

Explanation Publication of summary data from individual studies allows the analyses to be reproduced and other analyses and graphical displays to be investigated. Others may wish to assess the impact of excluding particular studies or consider subgroup analyses not reported by the review authors. Displaying the results of each treatment group in included studies also enables inspection of individual study features. For example,

if only odds ratios are provided, readers cannot assess the variation in event rates across the studies, making the odds ratio impossible to interpret.[138] Additionally, because data extraction errors in meta-analyses are common and can be large,[139] the presentation of the results from individual studies makes it easier to identify errors. For continuous outcomes, readers may wish to examine the consistency of standard deviations across studies, for example, to be reassured that standard deviation and standard error have not been confused.[138]

For each study, the summary data for each intervention group are generally given for binary outcomes as frequencies with and without the event (or as proportions such as 12/45). It is not sufficient to report event rates per intervention group as percentages. The required summary data for continuous outcomes are the mean, standard deviation, and sample size for each group. In reviews that examine time-to-event data, the authors should report the log hazard ratio and its standard error (or confidence interval) for each included study. Sometimes, essential data are missing from the reports of the included studies and cannot be calculated from other data but may need to be imputed by the reviewers. For example, the standard deviation may be imputed using the typical standard deviations in the other trials[116 117] (see item 14). Whenever relevant, authors should indicate which results were not reported directly and had to be estimated from other information (see item 13). In addition, the inclusion of unpublished data should be noted.

For all included studies it is important to present the estimated effect with a confidence interval. This information may be incorporated in a table showing study characteristics or may be shown in a forest plot.[140] The key elements of the forest plot are the effect estimates and confidence intervals for each study shown graphically, but it is preferable also to include, for each study, the numerical group-specific summary data, the effect size and confidence interval, and the percentage weight (see second example, fig 3). For discussion of the results of meta-analysis, see item 21.

In principle, all the above information should be provided for every outcome considered in the review, including both benefits and harms. When there are too many outcomes for full information to be included, results for the most important outcomes should be included in the main report with other information provided as a web appendix. The choice of the information to present should be justified in light of what was originally stated in the protocol. Authors should explicitly mention if the planned main outcomes cannot be presented due to lack of information. There is some evidence that information on harms is only rarely reported in systematic reviews, even when it is available in the original studies.[141] Selective omission of harms results biases a systematic review and decreases its ability to contribute to informed decision making.

Item 21: Syntheses of results

Present the main results of the review. If meta-analyses are done, include for each, confidence intervals and measures of consistency.

Examples "Mortality data were available for all six trials, randomizing 311 patients and reporting data for 305 patients. There were no deaths reported in the three respiratory syncytial virus/severe bronchiolitis trials; thus our estimate is based on three trials randomizing 232 patients, 64 of whom died. In the pooled analysis, surfactant was associated with significantly lower mortality (relative risk =0.7, 95% confidence interval =0.4–0.97, P=0.04). There was no evidence of heterogeneity (I^2=0%)."[142]

"Because the study designs, participants, interventions, and reported outcome measures varied markedly, we focused on describing the studies, their results, their applicability, and their limitations and on qualitative synthesis rather than meta-analysis."[143]

"We detected significant heterogeneity within this comparison (I^2=46.6%, χ^2=13.11, df=7, P=0.07). Retrospective exploration of the heterogeneity identified one trial that seemed to differ from the others. It included only small ulcers (wound area less than 5 cm^2). Exclusion of this trial removed the statistical heterogeneity and did not affect the finding of no evidence of a difference in healing rate between hydrocolloids and simple low adherent dressings (relative risk=0.98, [95% confidence interval] 0.85 to 1.12, I^2=0%)."[144]

Explanation Results of systematic reviews should be presented in an orderly manner. Initial narrative descriptions of the evidence covered in the review (see item 18) may tell readers important things about the study populations and the design and conduct of studies. These descriptions can facilitate the examination of patterns across studies. They may also provide important information about applicability of evidence, suggest the likely effects of any major biases, and allow consideration, in a systematic manner, of multiple explanations for possible differences of findings across studies.

If authors have conducted one or more meta-analyses, they should present the results as an estimated effect across studies with a confidence interval. It is often simplest to show each meta-analysis summary with the actual results of included studies in a forest plot (see item 20).[140] It should always be clear which of the included studies contributed to each meta-analysis. Authors should also provide, for each meta-analysis, a measure of the consistency of the results from the included studies such as I^2 (heterogeneity, see box 6); a confidence interval may also be given for this measure.[145] If no meta-analysis was performed, the qualitative inferences should be presented as systematically as possible with an explanation of why meta-analysis was not done, as in the second example above.[143] Readers may find a forest plot, without a summary estimate, helpful in such cases.

Authors should in general report syntheses for all the outcome measures they set out to investigate (that is, those described in the protocol, see item 4) to allow readers to draw their own conclusions about the implications of the results. Readers should be made aware of any deviations from the planned analysis. Authors should tell readers if the planned meta-analysis was not

thought appropriate or possible for some of the outcomes and the reasons for that decision.

It may not always be sensible to give meta-analysis results and forest plots for each outcome. If the review addresses a broad question, there may be a very large number of outcomes. Also, some outcomes may have been reported in only one or two studies, in which case forest plots are of little value and may be seriously biased.

Of 300 systematic reviews indexed in Medline in 2004, a little more than half (54%) included meta-analyses, of which the majority (91%) reported assessing for inconsistency in results.

Item 22: Risk of bias across studies

Present results of any assessment of risk of bias across studies (see item 15).

Example "Strong evidence of heterogeneity (I^2=79%, P<0.001) was observed. To explore this heterogeneity, a funnel plot was drawn. The funnel plot [fig 4] shows evidence of considerable asymmetry."[146]

"Specifically, four sertraline trials involving 486 participants and one citalopram trial involving 274 participants were reported as having failed to achieve a statistically significant drug effect, without reporting mean HRSD [Hamilton Rating Scale for Depression] scores. We were unable to find data from these trials on pharmaceutical company Web sites or through our search of the published literature. These omissions represent 38% of patients in sertraline trials and 23% of patients in citalopram trials. Analyses with and without inclusion of these trials found no differences in the patterns of results; similarly, the revealed patterns do not interact with drug type. The purpose of using the data obtained from the FDA was to avoid publication bias, by including unpublished as well as published trials. Inclusion of only those sertraline and citalopram trials for which means were reported to the FDA would constitute a form of reporting bias similar to publication bias and would lead to overestimation of drug–placebo differences for these drug types. Therefore, we present analyses only on data for medications for which complete clinical trials' change was reported."[147]

Explanation Authors should present the results of any assessments of risk of bias across studies. If a funnel plot is reported, authors should specify the effect estimate and measure of precision used, presented typically on the x axis and y axis, respectively. Authors should describe if and how they have tested the statistical significance of any possible asymmetry (see item 15). Results of any investigations of selective reporting of outcomes within studies (as discussed in item 15) should also be reported. Also, we advise authors to tell readers if any pre-specified analyses for assessing risk of bias across studies were not completed and the reasons (such as too few included studies).

Item 23: Additional analyses

Give results of additional analyses, if done (such as sensitivity or subgroup analyses, meta-regression [see item 16]).

Example "...benefits of chondroitin were smaller in trials with adequate concealment of allocation compared with trials with unclear concealment (P for interaction =0.050), in trials with an intention-to-treat analysis compared with those that had excluded patients from the analysis (P for interaction =0.017), and in large compared with small trials (P for interaction =0.022)."[148]

"Subgroup analyses according to antibody status, antiviral medications, organ transplanted, treatment duration, use of antilymphocyte therapy, time to outcome assessment, study quality and other aspects of study design did not demonstrate any differences in treatment effects. Multivariate meta-regression showed no significant difference in CMV [cytomegalovirus] disease after allowing for potential confounding or effect-modification by prophylactic drug used, organ transplanted or recipient serostatus in CMV positive recipients and CMV negative recipients of CMV positive donors."[149]

Explanation Authors should report any subgroup or sensitivity analyses and whether they were pre-specified (see items 5 and 16). For analyses comparing subgroups of studies (such as separating studies of low and high dose aspirin), the authors should report any tests for interactions, as well as estimates and confidence intervals from meta-analyses within each subgroup. Similarly, meta-regression results (see item 16) should not be limited to P values but should include effect sizes and confidence intervals,[150] as the first example reported above does in a table. The amount of data included in each additional analysis should be specified if different from that considered in the main analyses. This information is especially relevant for sensitivity analyses that exclude some studies; for example, those with high risk of bias.

Importantly, all additional analyses conducted should be reported, not just those that were statistically significant. This information will help avoid selective outcome reporting bias within the review as has been demonstrated in reports of randomised controlled trials.[42 44 121 151 152] Results from exploratory subgroup or sensitivity analyses should be interpreted cautiously, bearing in mind the potential for multiple analyses to mislead.

Discussion

Item 24: Summary of evidence

Summarise the main findings, including the strength of evidence for each main outcome; consider their relevance to key groups (such as healthcare providers, users, and policy makers).

Example "Overall, the evidence is not sufficiently robust to determine the comparative effectiveness of angioplasty (with or without stenting) and medical treatment alone. Only 2 randomized trials with long-term outcomes and a third randomized trial that allowed substantial crossover of treatment after 3 months directly compared angioplasty and medical treatment…the randomized trials did not evaluate enough patients or did not follow patients for a sufficient duration to allow definitive conclusions to be made about clinical outcomes, such as mortality and cardiovascular or kidney failure events.

Some acceptable evidence from comparison of medical treatment and angioplasty suggested no difference in long-term kidney function but possibly better blood pressure control after angioplasty, an effect that may be

limited to patients with bilateral atherosclerotic renal artery stenosis. The evidence regarding other outcomes is weak. Because the reviewed studies did not explicitly address patients with rapid clinical deterioration who may need acute intervention, our conclusions do not apply to this important subset of patients."[143]

Explanation Authors should give a brief and balanced summary of the nature and findings of the review. Sometimes, outcomes for which little or no data were found should be noted due to potential relevance for policy decisions and future research. Applicability of the review's findings—to different patients, settings, or target audiences, for example—should be mentioned. Although there is no standard way to assess applicability simultaneously to different audiences, some systems do exist.[153] Sometimes, authors formally rate or assess the overall body of evidence addressed in the review and can present the strength of their summary recommendations tied to their assessments of the quality of evidence (such as the GRADE system).[10]

Authors need to keep in mind that statistical significance of the effects does not always suggest clinical or policy relevance. Likewise, a non-significant result does not demonstrate that a treatment is ineffective. Authors should ideally clarify trade-offs and how the values attached to the main outcomes would lead different people to make different decisions. In addition, adroit authors consider factors that are important in translating the evidence to different settings and that may modify the estimates of effects reported in the review.[153] Patients and healthcare providers may be primarily interested in which intervention is most likely to provide a benefit with acceptable harms, while policy makers and administrators may value data on organisational impact and resource utilisation.

Item 25: Limitations

Discuss limitations at study and outcome level (such as risk of bias), and at review level (such as incomplete retrieval of identified research, reporting bias).

Examples Outcome level: "The meta-analysis reported here combines data across studies in order to estimate treatment effects with more precision than is possible in a single study. The main limitation of this meta-analysis, as with any overview, is that the patient population, the antibiotic regimen and the outcome definitions are not the same across studies."[154]

Study and review level: "Our study has several limitations. The quality of the studies varied. Randomization was adequate in all trials; however, 7 of the articles did not explicitly state that analysis of data adhered to the intention-to-treat principle, which could lead to overestimation of treatment effect in these trials, and we could not assess the quality of 4 of the 5 trials reported as abstracts. Analyses did not identify an association between components of quality and re-bleeding risk, and the effect size in favour of combination therapy remained statistically significant when we excluded trials that were reported as abstracts.

Publication bias might account for some of the effect we observed. Smaller trials are, in general, analyzed with less methodological rigor than larger studies, and an asymmetrical funnel plot suggests that selective reporting may have led to an overestimation of effect sizes in small trials."[155]

Explanation A discussion of limitations should address the validity (that is, risk of bias) and reporting (informativeness) of the included studies, limitations of the review process, and generalisability (applicability) of the review. Readers may find it helpful if authors discuss whether studies were threatened by serious risks of bias, whether the estimates of the effect of the intervention are too imprecise, or if there were missing data for many participants or important outcomes.

Limitations of the review process might include limitations of the search (such as restricting to English-language publications), and any difficulties in the study selection, appraisal, and meta-analysis processes. For example, poor or incomplete reporting of study designs, patient populations, and interventions may hamper interpretation and synthesis of the included studies.[64] Applicability of the review may be affected if there are limited data for certain populations or subgroups where the intervention might perform differently or few studies assessing the most important outcomes of interest; or if there is a substantial amount of data relating to an outdated intervention or comparator or heavy reliance on imputation of missing values for summary estimates (item 14).

Item 26: Conclusions

Provide a general interpretation of the results in the context of other evidence, and implications for future research.

Example Implications for practice: "Between 1995 and 1997 five different meta-analyses of the effect of antibiotic prophylaxis on infection and mortality were published. All confirmed a significant reduction in infections, though the magnitude of the effect varied from one review to another. The estimated impact on overall mortality was less evident and has generated considerable controversy on the cost effectiveness of the treatment. Only one among the five available reviews, however, suggested that a weak association between respiratory tract infections and mortality exists and lack of sufficient statistical power may have accounted for the limited effect on mortality."

Implications for research: "A logical next step for future trials would thus be the comparison of this protocol against a regimen of a systemic antibiotic agent only to see whether the topical component can be dropped. We have already identified six such trials but the total number of patients so far enrolled (n=1056) is too small for us to be confident that the two treatments are really equally effective. If the hypothesis is therefore considered worth testing more and larger randomised controlled trials are warranted. Trials of this kind, however, would not resolve the relevant issue of treatment induced resistance. To produce a satisfactory answer to this, studies with a different design would be necessary. Though a detailed discussion goes beyond the scope of this paper, studies in which the intensive care unit rather than the individual patient is the unit of randomisation and in which the occurrence of antibiotic resistance is monitored over a long period of time should be undertaken."[156]

Explanation Systematic reviewers sometimes draw conclusions that are too optimistic[157] or do not consider the harms equally as carefully as the benefits, although some evidence suggests these problems are decreasing.[158] If conclusions cannot be drawn because there are too few reliable studies, or too much uncertainty, this should be stated. Such a finding can be as important as finding consistent effects from several large studies.

Authors should try to relate the results of the review to other evidence, as this helps readers to better interpret the results. For example, there may be other systematic reviews about the same general topic that have used different methods or have addressed related but slightly different questions.[159 160] Similarly, there may be additional information relevant to decision makers, such as the cost-effectiveness of the intervention (such as health technology assessment). Authors may discuss the results of their review in the context of existing evidence regarding other interventions.

We advise authors also to make explicit recommendations for future research. In a sample of 2535 Cochrane reviews, 82% included recommendations for research with specific interventions, 30% suggested the appropriate type of participants, and 52% suggested outcome measures for future research.[161] There is no corresponding assessment about systematic reviews published in medical journals, but we believe that such recommendations are much less common in those reviews.

Clinical research should not be planned without a thorough knowledge of similar, existing research.[162] There is evidence that this still does not occur as it should and that authors of primary studies do not consider a systematic review when they design their studies.[163] We believe systematic reviews have great potential for guiding future clinical research.

Funding

Item 27: Funding

Describe sources of funding or other support (such as supply of data) for the systematic review, and the role of funders for the systematic review.

Examples "The evidence synthesis upon which this article was based was funded by the Centers for Disease Control and Prevention for the Agency for Healthcare Research and Quality and the U.S. Prevention Services Task Force."[164]

"Role of funding source: The funders played no role in study design, collection, analysis, interpretation of data, writing of the report, or in the decision to submit the paper for publication. They accept no responsibility for the contents."[165]

Explanation Authors of systematic reviews, like those of any other research study, should disclose any funding they received to carry out the review, or state if the review was not funded. Lexchin and colleagues[166] observed that outcomes of reports of randomised trials and meta-analyses of clinical trials funded by the pharmaceutical industry are more likely to favor the sponsor's product compared with studies with other sources of funding. Similar results have been reported elsewhere.[167 168] Analogous data suggest that similar biases may affect the conclusions of systematic reviews.[169]

Given the potential role of systematic reviews in decision making, we believe authors should be transparent about the funding and the role of funders, if any. Sometimes the funders will provide services, such as those of a librarian to complete the searches for relevant literature or access to commercial databases not available to the reviewers. Any level of funding or services provided to the systematic review team should be reported. Authors should also report whether the funder had any role in the conduct or report of the review. Beyond funding issues, authors should report any real or perceived conflicts of interest related to their role or the role of the funder in the reporting of the systematic review.[170]

In a survey of 300 systematic reviews published in November 2004, funding sources were not reported in 41% of the reviews.[3] Only a minority of reviews (2%) reported being funded by for-profit sources, but the true proportion may be higher.[171]

Additional considerations for systematic reviews of non-randomised intervention studies or for other types of systematic reviews

The PRISMA statement and this document have focused on systematic reviews of reports of randomised trials. Other study designs, including non-randomised studies, quasi-experimental studies, and interrupted time series, are included in some systematic reviews that evaluate the effects of healthcare interventions.[172 173] The methods of these reviews may differ to varying degrees from the typical intervention review, for example regarding the literature search, data abstraction, assessment of risk of bias, and analysis methods. As such, their reporting demands might also differ from what we have described here. A useful principle is for systematic review authors to ensure that their methods are reported with adequate clarity and transparency to enable readers to critically judge the available evidence and replicate or update the research.

In some systematic reviews, the authors will seek the raw data from the original researchers to calculate the summary statistics. These systematic reviews are called individual patient (or participant) data reviews.[40 41] Individual patient data meta-analyses may also be conducted with prospective accumulation of data rather than retrospective accumulation of existing data. Here too, extra information about the methods will need to be reported.

Other types of systematic reviews exist. Realist reviews aim to determine how complex programmes work in specific contexts and settings.[174] Meta-narrative reviews aim to explain complex bodies of evidence through mapping and comparing different overarching storylines.[175] Network meta-analyses, also known as multiple treatments meta-analyses, can be used to analyse data from comparisons of many different treatments.[176 177] They use both direct and indirect comparisons and can be used to compare interventions that have not been directly compared.

We believe that the issues we have highlighted in this paper are relevant to ensure transparency and under-

standing of the processes adopted and the limitations of the information presented in systematic reviews of different types. We hope that PRISMA can be the basis for more detailed guidance on systematic reviews of other types of research, including diagnostic accuracy and epidemiological studies.

Discussion

We developed the PRISMA statement using an approach for developing reporting guidelines that has evolved over several years.[178] The overall aim of PRISMA is to help ensure the clarity and transparency of reporting of systematic reviews, and recent data indicate that this reporting guidance is much needed.[3] PRISMA is not intended to be a quality assessment tool and it should not be used as such.

This PRISMA explanation and elaboration document was developed to facilitate the understanding, uptake, and dissemination of the PRISMA statement and hopefully provide a pedagogical framework for those interested in conducting and reporting systematic reviews. It follows a format similar to that used in other explanatory documents.[17-19] Following the recommendations in the PRISMA checklist may increase the word count of a systematic review report. We believe, however, that the benefit of readers being able to critically appraise a clear, complete, and transparent systematic review report outweighs the possible slight increase in the length of the report.

While the aims of PRISMA are to reduce the risk of flawed reporting of systematic reviews and improve the clarity and transparency in how reviews are conducted, we have little data to state more definitively whether this "intervention" will achieve its intended goal. A previous effort to evaluate QUOROM was not successfully completed.[178] Publication of the QUOROM statement was delayed for two years while a research team attempted to evaluate its effectiveness by conducting a randomised controlled trial with the participation of eight major medical journals. Unfortunately that trial was not completed due to accrual problems (David Moher, personal communication). Other evaluation methods might be easier to conduct. At least one survey of 139 published systematic reviews in the critical care literature[179] suggests that their quality improved after the publication of QUOROM.

If the PRISMA statement is endorsed by and adhered to in journals, as other reporting guidelines have been,[17-19 180] there should be evidence of improved reporting of systematic reviews. For example, there have been several evaluations of whether the use of CONSORT improves reports of randomised controlled trials. A systematic review of these studies[181] indicates that use of CONSORT is associated with improved reporting of certain items, such as allocation concealment. We aim to evaluate the benefits (that is, improved reporting) and possible adverse effects (such as increased word length) of PRISMA and we encourage others to consider doing likewise.

Even though we did not carry out a systematic literature search to produce our checklist, and this is indeed a limitation of our effort, PRISMA was developed using an evidence based approach whenever possible. Checklist items were included if there was evidence that not reporting the item was associated with increased risk of bias, or where it was clear that information was necessary to appraise the reliability of a review. To keep PRISMA up to date and as evidence based as possible requires regular vigilance of the literature, which is growing rapidly. Currently the Cochrane Methodology Register has more than 11 000 records pertaining to the conduct and reporting of systematic reviews and other evaluations of health and social care. For some checklist items, such as reporting the abstract (item 2), we have used evidence from elsewhere in the belief that the issue applies equally well to reporting of systematic reviews. Yet for other items, evidence does not exist; for example, whether a training exercise improves the accuracy and reliability of data extraction. We hope PRISMA will act as a catalyst to help generate further evidence that can be considered when further revising the checklist in the future.

More than 10 years have passed between the development of the QUOROM statement and its update, the PRISMA statement. We aim to update PRISMA more frequently. We hope that the implementation of PRISMA will be better than it has been for QUOROM. There are at least two reasons to be optimistic. First, systematic reviews are increasingly used by healthcare providers to inform "best practice" patient care. Policy analysts and managers are using systematic reviews to inform healthcare decision making and to better target future research. Second, we anticipate benefits from the development of the EQUATOR Network, described below.

Developing any reporting guideline requires considerable effort, experience, and expertise. While reporting guidelines have been successful for some individual efforts,[17-19] there are likely others who want to develop reporting guidelines who possess little time, experience, or knowledge as to how to do so appropriately. The EQUATOR (enhancing the quality and transparency of health research) Network aims to help such individuals and groups by serving as a global resource for anybody interested in developing reporting guidelines, regardless of the focus.[7 180 182] The overall goal of EQUATOR is to improve the quality of reporting of all health science research through the development and translation of reporting guidelines. Beyond this aim, the network plans to develop a large web presence by developing and maintaining a resource centre of reporting tools, and other information for reporting research (www.equator-network.org/).

We encourage healthcare journals and editorial groups, such as the World Association of Medical Editors and the International Committee of Medical Journal Editors, to endorse PRISMA in much the same way as they have endorsed other reporting guidelines, such as CONSORT. We also encourage editors of healthcare journals to support PRISMA by updating their "instructions to authors" and including the PRISMA web address, and by raising awareness through specific editorial actions.

The following people contributed to this paper: Doug Altman, Centre

for Statistics in Medicine (Oxford, UK); Gerd Antes, University Hospital Freiburg (Freiburg, Germany); David Atkins, Health Services Research and Development Service, Veterans Health Administration (Washington DC, USA); Virginia Barbour, *PLoS Medicine* (Cambridge, UK); Nick Barrowman, Children's Hospital of Eastern Ontario (Ottawa, Canada); Jesse A Berlin, Johnson & Johnson Pharmaceutical Research and Development (Titusville NJ, USA); Jocalyn Clark, *PLoS Medicine* (at the time of writing, BMJ, London); Mike Clarke, UK Cochrane Centre (Oxford, UK) and School of Nursing and Midwifery, Trinity College (Dublin, Ireland); Deborah Cook, Departments of Medicine, Clinical Epidemiology and Biostatistics, McMaster University (Hamilton, Canada); Roberto D'Amico, Università di Modena e Reggio Emilia (Modena, Italy) and Centro Cochrane Italiano, Istituto Ricerche Farmacologiche Mario Negri (Milan, Italy); Jonathan J Deeks, University of Birmingham (Birmingham); P J Devereaux, Departments of Medicine, Clinical Epidemiology and Biostatistics, McMaster University (Hamilton, Canada); Kay Dickersin, Johns Hopkins Bloomberg School of Public Health (Baltimore MD, USA); Matthias Egger, Department of Social and Preventive Medicine, University of Bern (Bern, Switzerland); Edzard Ernst, Peninsula Medical School (Exeter, UK); Peter C Gøtzsche, Nordic Cochrane Centre (Copenhagen, Denmark); Jeremy Grimshaw, Ottawa Hospital Research Institute (Ottawa, Canada); Gordon Guyatt, Departments of Medicine, Clinical Epidemiology and Biostatistics, McMaster University; Julian Higgins, MRC Biostatistics Unit (Cambridge, UK); John P A Ioannidis, University of Ioannina Campus (Ioannina, Greece); Jos Kleijnen, Kleijnen Systematic Reviews (York, UK) and School for Public Health and Primary Care (CAPHRI), University of Maastricht (Maastricht, Netherlands); Tom Lang, Tom Lang Communications and Training (Davis CA, USA); Alessandro Liberati, Università di Modena e Reggio Emilia (Modena, Italy) and Centro Cochrane Italiano, Istituto Ricerche Farmacologiche Mario Negri (Milan, Italy); Nicola Magrini, NHS Centre for the Evaluation of the Effectiveness of Health Care—CeVEAS (Modena, Italy); David McNamee, Lancet (London, UK); David Moher, Ottawa Methods Centre, Ottawa Hospital Research Institute (Ottawa, Canada); Lorenzo Moja, Centro Cochrane Italiano, Istituto Ricerche Farmacologiche Mario Negri; Maryann Napoli, Center for Medical Consumers (New York, USA); Cynthia Mulrow, Annals of Internal Medicine (Philadelphia, Pennsylvania, US); Andy Oxman, Norwegian Health Services Research Centre (Oslo, Norway); Ba' Pham, Toronto Health Economics and Technology Assessment Collaborative (Toronto, Canada) (at the time of first meeting of the group, GlaxoSmithKline Canada, Mississauga, Canada); Drummond Rennie, University of California San Francisco (San Francisco CA, USA); Margaret Sampson, Children's Hospital of Eastern Ontario (Ottawa, Canada); Kenneth F Schulz, Family Health International (Durham NC, USA); Paul G Shekelle, Southern California Evidence Based Practice Center (Santa Monica CA, USA); Jennifer Tetzlaff, Ottawa Methods Centre, Ottawa Hospital Research Institute (Ottawa, Canada); David Tovey, Cochrane Library, Cochrane Collaboration (Oxford, UK) (at the time of first meeting of the group, BMJ, London); Peter Tugwell, Institute of Population Health, University of Ottawa (Ottawa, Canada).

Lorenzo Moja helped with the preparation and the several updates of the manuscript and assisted with the preparation of the reference list. AL is the guarantor of the manuscript.

Competing interests: None declared.

Provenance and peer review: Not commissioned; externally peer reviewed.

BOXES, TABLES AND REFERENCES FOLLOW

Table 1 | Checklist of items to include when reporting a systematic review or meta-analysis

Section/topic	Item No	Checklist item	Reported on page No
Title			
Title	1	Identify the report as a systematic review, meta-analysis, or both	
Abstract			
Structured summary	2	Provide a structured summary including, as applicable, background, objectives, data sources, study eligibility criteria, participants, interventions, study appraisal and synthesis methods, results, limitations, conclusions and implications of key findings, systematic review registration number	
Introduction			
Rationale	3	Describe the rationale for the review in the context of what is already known	
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS)	
Methods			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (such as web address), and, if available, provide registration information including registration number	
Eligibility criteria	6	Specify study characteristics (such as PICOS, length of follow-up) and report characteristics (such as years considered, language, publication status) used as criteria for eligibility, giving rationale	
Information sources	7	Describe all information sources (such as databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched	
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated	
Study selection	9	State the process for selecting studies (that is, screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis)	
Data collection process	10	Describe method of data extraction from reports (such as piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators	
Data items	11	List and define all variables for which data were sought (such as PICOS, funding sources) and any assumptions and simplifications made	
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis	
Summary measures	13	State the principal summary measures (such as risk ratio, difference in means).	
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (such as I^2) for each meta-analysis	
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (such as publication bias, selective reporting within studies)	
Additional analyses	16	Describe methods of additional analyses (such as sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified	
Results			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram	
Study characteristics	18	For each study, present characteristics for which data were extracted (such as study size, PICOS, follow-up period) and provide the citations	
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome-level assessment (see item 12).	
Results of individual studies	20	For all outcomes considered (benefits or harms), present for each study (a) simple summary data for each intervention group and (b) effect estimates and confidence intervals, ideally with a forest plot	
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency	
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see item 15)	
Additional analysis	23	Give results of additional analyses, if done (such as sensitivity or subgroup analyses, meta-regression [see item 16])	
Discussion			
Summary of evidence	24	Summarise the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (such as health care providers, users, and policy makers)	
Limitations	25	Discuss limitations at study and outcome level (such as risk of bias), and at review level (such as incomplete retrieval of identified research, reporting bias)	
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research	
Funding			
Funding	27	Describe sources of funding for the systematic review and other support (such as supply of data) and role of funders for the systematic review	

Table 2 | Example of summary of study characteristics: Summary of included studies evaluating the efficacy of antiemetic agents in acute gastroenteritis. Adapted from DeCamp et al[125]

Source	Setting	No of patients	Age range	Inclusion criteria	Antiemetic agent	Route	Follow-up
Freedman et al 2006	ED	214	6 months-10 years	GE with mild to moderate dehydration and vomiting in the preceding 4 hours	Ondansetron	PO	1-2 weeks
Reeves et al 2002	ED	107	1 month-22 years	GE and vomiting requiring IV rehydration	Ondansetron	IV	5-7 days
Roslund et al 2007	ED	106	1-10 years	GE with failed oral rehydration attempt in ED	Ondansetron	PO	1 week
Stork et al 2006	ED	137	6 months-12 years	GE, recurrent emesis, mild to moderate dehydration, and failed oral hydration	Ondansetron and dexamethasone	IV	1 and 2 days

ED = emergency department; GE = gastroenteritis; IV = intravenous; PO = by mouth.

Table 3 | Example of assessment of the risk of bias: Quality measures of the randomised controlled trials that failed to fulfil any one of six markers of validity. Adapted from Devereaux et al[96]

Trials	Concealment of randomisation	RCT stopped early	Patients blinded	Healthcare providers blinded	Data collectors blinded	Outcome assessors blinded
Liu	No	No	Yes	Yes	Yes	Yes
Stone	Yes	No	No	Yes	Yes	Yes
Polderman	Yes	Yes	No	No	No	Yes
Zaugg	Yes	No	No	No	Yes	Yes
Urban	Yes	Yes	No	No, except anaesthesiologists	Yes	Yes

RCT = randomised controlled trial.

Table 4 | Example of summary results: Heterotopic ossification in trials comparing radiotherapy to non-steroidal anti-inflammatory drugs after major hip procedures and fractures. Adapted from Pakos et al[136]

Author (year)	Radiotherapy		NSAID	
Kienapfel (1999)	12/49	24.5%	20/55	36.4%
Sell (1998)	2/77	2.6%	18/77	23.4%
Kolbl (1997)	39/188	20.7%	18/113	15.9%
Kolbl (1998)	22/46	47.8%	6/54	11.1%
Moore (1998)	9/33	27.3%	18/39	46.2%
Bremen-Kuhne (1997)	9/19	47.4%	11/31	35.5%
Knelles (1997)	5/101	5.0%	46/183	25.4%

NSAID = non-steroidal anti-inflammatory drug.

Box 1 |: Terminology

The terminology used to describe systematic reviews and meta-analyses has evolved over time and varies between fields. Different terms have been used by different groups, such as educators and psychologists. The conduct of a systematic review comprises several explicit and reproducible steps, such as identifying all likely relevant records, selecting eligible studies, assessing the risk of bias, extracting data, qualitative synthesis of the included studies, and possibly meta-analyses.

Initially this entire process was termed a meta-analysis and was so defined in the QUOROM statement.[8] More recently, especially in healthcare research, there has been a trend towards preferring the term systematic review. If quantitative synthesis is performed, this last stage alone is referred to as a meta-analysis. The Cochrane Collaboration uses this terminology,[9] under which a meta-analysis, if performed, is a component of a systematic review. Regardless of the question addressed and the complexities involved, it is always possible to complete a systematic review of existing data, but not always possible or desirable, to quantitatively synthesise results because of clinical, methodological, or statistical differences across the included studies. Conversely, with prospective accumulation of studies and datasets where the plan is eventually to combine them, the term "(prospective) meta-analysis" may make more sense than "systematic review."

For retrospective efforts, one possibility is to use the term systematic review for the whole process up to the point when one decides whether to perform a quantitative synthesis. If a quantitative synthesis is performed, some researchers refer to this as a meta-analysis. This definition is similar to that found in the current edition of the *Dictionary of Epidemiology*.[183]

While we recognise that the use of these terms is inconsistent and there is residual disagreement among the members of the panel working on PRISMA, we have adopted the definitions used by the Cochrane Collaboration.[9]

Systematic review A systematic review attempts to collate all empirical evidence that fits pre-specified eligibility criteria to answer a specific research question. It uses explicit, systematic methods that are selected with a view to minimising bias, thus providing reliable findings from which conclusions can be drawn and decisions made.[184 185] The key characteristics of a systematic review are (*a*) a clearly stated set of objectives with an explicit, reproducible methodology; (*b*) a systematic search that attempts to identify all studies that would meet the eligibility criteria; (*c*) an assessment of the validity of the findings of the included studies, such as through the assessment of risk of bias; and (*d*) systematic presentation and synthesis of the characteristics and findings of the included studies.

Meta-analysis Meta-analysis is the use of statistical techniques to integrate and summarise the results of included studies. Many systematic reviews contain meta-analyses, but not all. By combining information from all relevant studies, meta-analyses can provide more precise estimates of the effects of health care than those derived from the individual studies included within a review.

Box 2 |: Helping to develop the research question(s): the PICOS approach

Formulating relevant and precise questions that can be answered in a systematic review can be complex and time consuming. A structured approach for framing questions that uses five components may help facilitate the process. This approach is commonly known by the acronym "PICOS" where each letter refers to a component: the patient population or the disease being addressed (P), the interventions or exposure (I), the comparator group (C), the outcome or endpoint (O), and the study design chosen (S).[186] Issues relating to PICOS affect several PRISMA items (items 6, 8, 9, 10, 11, and 18).

- *P*—Providing information about the population requires a precise definition of a group of participants (often patients), such as men over the age of 65 years, their defining characteristics of interest (often disease), and possibly the setting of care considered, such as an acute care hospital.
- *I*—The interventions (exposures) under consideration in the systematic review need to be transparently reported. For example, if the reviewers answer a question regarding the association between a woman's prenatal exposure to folic acid and subsequent offspring's neural tube defects, reporting the dose, frequency, and duration of folic acid used in different studies is likely to be important for readers to interpret the review's results and conclusions. Other interventions (exposures) might include diagnostic, preventive, or therapeutic treatments; arrangements of specific processes of care; lifestyle changes; psychosocial or educational interventions; or risk factors.
- *C*—Clearly reporting the comparator (control) group intervention(s)—such as usual care, drug, or placebo—is essential for readers to fully understand the selection criteria of primary studies included in the systematic review, and might be a source of heterogeneity investigators have to deal with. Comparators are often poorly described. Clearly reporting what the intervention is compared with is important and may sometimes have implications for the inclusion of studies in a review—many reviews compare with "standard care," which is otherwise undefined; this should be properly addressed by authors.
- *O*—The outcomes of the intervention being assessed—such as mortality, morbidity, symptoms, or quality of life improvements—should be clearly specified as they are required to interpret the validity and generalisability of the systematic review's results.
- *S*—Finally, the type of study design(s) included in the review should be reported. Some reviews include only reports of randomised trials, whereas others have broader design criteria and include randomised trials and certain types of observational studies. Still other reviews, such as those specifically answering questions related to harms, may include a wide variety of designs ranging from cohort studies to case reports. Whatever study designs are included in the review, these should be reported.

Independently from how difficult it is to identify the components of the research question, the important point is that a structured approach is preferable, and this extends beyond systematic reviews of effectiveness. Ideally the PICOS criteria should be formulated a priori, in the systematic review's protocol, although some revisions might be required because of the iterative nature of the review process. Authors are encouraged to report their PICOS criteria and whether any modifications were made during the review process. A useful example in this realm is the appendix of the "systematic reviews of water fluoridation" undertaken by the Centre for Reviews and Dissemination.[187]

Box 3 |: Identification of study reports and data extraction

Comprehensive searches usually result in a large number of identified records, a much smaller number of studies included in the systematic review, and even fewer of these studies included in any meta-analyses. Reports of systematic reviews often provide little detail as to the methods used by the review team in this process. Readers are often left with what can be described as the "X-files" phenomenon, as it is unclear what occurs between the initial set of identified records and those finally included in the review.

Sometimes, review authors simply report the number of included studies; more often they report the initial number of identified records and the number of included studies. Rarely, although this is optimal for readers, do review authors report the number of identified records, the smaller number of potentially relevant studies, and the even smaller number of included studies, by outcome. Review authors also need to differentiate between the number of reports and studies. Often there will not be a 1:1 ratio of reports to studies and this information needs to be described in the systematic review report.

Ideally, the identification of study reports should be reported as text in combination with use of the PRISMA flow diagram. While we recommend use of the flow diagram, a small number of reviews might be particularly simple and can be sufficiently described with a few brief sentences of text. More generally, review authors will need to report the process used for each step: screening the identified records; examining the full text of potentially relevant studies (and reporting the number that could not be obtained); and applying eligibility criteria to select the included studies.

Such descriptions should also detail how potentially eligible records were promoted to the next stage of the review (such as full text screening) and to the final stage of this process, the included studies. Often review teams have three response options for excluding records or promoting them to the next stage of the winnowing process: "yes," "no," and "maybe."

Similarly, some detail should be reported on who participated and how such processes were completed. For example, a single person may screen the identified records while a second person independently examines a small sample of them. The entire winnowing process is one of "good bookkeeping" whereby interested readers should be able to work backwards from the included studies to come up with the same numbers of identified records.

There is often a paucity of information describing the data extraction processes in reports of systematic reviews. Authors may simply report that "relevant" data were extracted from each included study with little information about the processes used for data extraction. It may be useful for readers to know whether a systematic review's authors developed, a priori or not, a data extraction form, whether multiple forms were used, the number of questions, whether the form was pilot tested, and who completed the extraction. For example, it is important for readers to know whether one or more people extracted data, and if so, whether this was completed independently, whether "consensus" data were used in the analyses, and if the review team completed an informal training exercise or a more formal reliability exercise.

Box 4 |: Study quality and risk of bias

In this paper, and elsewhere,[11] we sought to use a new term for many readers, namely, risk of bias, for evaluating each included study in a systematic review. Previous papers[89 188] tended to use the term "quality." When carrying out a systematic review we believe it is important to distinguish between quality and risk of bias and to focus on evaluating and reporting the latter. Quality is often the best the authors have been able to do. For example, authors may report the results of surgical trials in which blinding of the outcome assessors was not part of the trial's conduct. Even though this may have been the best methodology the researchers were able to do, there are still theoretical grounds for believing that the study was susceptible to (risk of) bias.

Assessing the risk of bias should be part of the conduct and reporting of any systematic review. In all situations, we encourage systematic reviewers to think ahead carefully about what risks of bias (methodological and clinical) may have a bearing on the results of their systematic reviews.

For systematic reviewers, understanding the risk of bias on the results of studies is often difficult, because the report is only a surrogate of the actual conduct of the study. There is some suggestion[189 190] that the report may not be a reasonable facsimile of the study, although this view is not shared by all.[88 191] There are three main ways to assess risk of bias—individual components, checklists, and scales. There are a great many scales available,[192] although we caution against their use based on theoretical grounds[193] and emerging empirical evidence.[194] Checklists are less frequently used and potentially have the same problems as scales. We advocate using a component approach and one that is based on domains for which there is good empirical evidence and perhaps strong clinical grounds. The new Cochrane risk of bias tool[11] is one such component approach.

The Cochrane risk of bias tool consists of five items for which there is empirical evidence for their biasing influence on the estimates of an intervention's effectiveness in randomised trials (sequence generation, allocation concealment, blinding, incomplete outcome data, and selective outcome reporting) and a catch-all item called "other sources of bias".[11] There is also some consensus that these items can be applied for evaluation of studies across diverse clinical areas.[93] Other risk of bias items may be topic or even study specific—that is, they may stem from some peculiarity of the research topic or some special feature of the design of a specific study. These peculiarities need to be investigated on a case-by-case basis, based on clinical and methodological acumen, and there can be no general recipe. In all situations, systematic reviewers need to think ahead carefully about what aspects of study quality may have a bearing on the results.

Box 5 |: Whether to combine data

Deciding whether to combine data involves statistical, clinical, and methodological considerations. The statistical decisions are perhaps the most technical and evidence-based. These are more thoroughly discussed in box 6. The clinical and methodological decisions are generally based on discussions within the review team and may be more subjective.

Clinical considerations will be influenced by the question the review is attempting to address. Broad questions might provide more "license" to combine more disparate studies, such as whether "Ritalin is effective in increasing focused attention in people diagnosed with attention deficit hyperactivity disorder (ADHD)." Here authors might elect to combine reports of studies involving children and adults. If the clinical question is more focused, such as whether "Ritalin is effective in increasing classroom attention in previously undiagnosed ADHD children who have no comorbid conditions," it is likely that different decisions regarding synthesis of studies are taken by authors. In any case authors should describe their clinical decisions in the systematic review report.

Deciding whether to combine data also has a methodological component. Reviewers may decide not to combine studies of low risk of bias with those of high risk of bias (see items 12 and 19). For example, for subjective outcomes, systematic review authors may not wish to combine assessments that were completed under blind conditions with those that were not.

For any particular question there may not be a "right" or "wrong" choice concerning synthesis, as such decisions are likely complex. However, as the choice may be subjective, authors should be transparent as to their key decisions and describe them for readers.

Box 6 |: Meta-analysis and assessment of consistency (heterogeneity)

Meta-analysis: statistical combination of the results of multiple studies

If it is felt that studies should have their results combined statistically, other issues must be considered because there are many ways to conduct a meta-analysis. Different effect measures can be used for both binary and continuous outcomes (see item 13). Also, there are two commonly used statistical models for combining data in a meta-analysis.[195] The fixed-effect model assumes that there is a common treatment effect for all included studies;[196] it is assumed that the observed differences in results across studies reflect random variation.[196] The random-effects model assumes that there is no common treatment effect for all included studies but rather that the variation of the effects across studies follows a particular distribution.[197] In a random-effects model it is believed that the included studies represent a random sample from a larger population of studies addressing the question of interest.[198]

There is no consensus about whether to use fixed- or random-effects models, and both are in wide use. The following differences have influenced some researchers regarding their choice between them. The random-effects model gives more weight to the results of smaller trials than does the fixed-effect analysis, which may be undesirable as small trials may be inferior and most prone to publication bias. The fixed-effect model considers only within-study variability, whereas the random-effects model considers both within- and between-study variability. This is why a fixed-effect analysis tends to give narrower confidence intervals (that is, provides greater precision) than a random-effects analysis.[110 196 199] In the absence of any between-study heterogeneity, the fixed- and random-effects estimates will coincide.

In addition, there are different methods for performing both types of meta-analysis.[200] Common fixed-effect approaches are Mantel-Haenszel and inverse variance, whereas random-effects analyses usually use the DerSimonian and Laird approach, although other methods exist, including Bayesian meta-analysis.[201]

In the presence of demonstrable between-study heterogeneity (see below), some consider that the use of a fixed-effect analysis is counterintuitive because their main assumption is violated. Others argue that it is inappropriate to conduct any meta-analysis when there is unexplained variability across trial results. If the reviewers decide not to combine the data quantitatively, a danger is that eventually they may end up using quasi-quantitative rules of poor validity (such as vote counting of how many studies have nominally significant results) for interpreting the evidence. Statistical methods to combine data exist for almost any complex situation that may arise in a systematic review, but one has to be aware of their assumptions and limitations to avoid misapplying or misinterpreting these methods.

Assessment of consistency (heterogeneity)

We expect some variation (inconsistency) in the results of different studies due to chance alone. Variability in excess of that due to chance reflects true differences in the results of the trials, and is called "heterogeneity." The conventional statistical approach to evaluating heterogeneity is a ^{2} test (Cochran's Q), but it has low power when there are few studies and excessive power when there are many studies.[202] By contrast, the I^2 statistic quantifies the amount of variation in results across studies beyond that expected by chance and so is preferable to Q.[202 203] I^2 represents the percentage of the total variation in estimated effects across studies that is due to heterogeneity rather than to chance; some authors consider an I^2 value less than 25% as low.[202] However, I^2 also suffers from large uncertainty in the common situation where only a few studies are available,[204] and reporting the uncertainty in I^2 (such as 95% confidence interval) may be helpful.[145] When there are few studies, inferences about heterogeneity should be cautious.

When considerable heterogeneity is observed, it is advisable to consider possible reasons.[205] In particular, the heterogeneity may be due to differences between subgroups of studies (see item 16). Also, data extraction errors are a common cause of substantial heterogeneity in results with continuous outcomes.[139]

Box 7 |: Bias caused by selective publication of studies or results within studies

Systematic reviews aim to incorporate information from all relevant studies. The absence of information from some studies may pose a serious threat to the validity of a review. Data may be incomplete because some studies were not published, or because of incomplete or inadequate reporting within a published article. These problems are often summarised as "publication bias," although the bias arises from non-publication of full studies and selective publication of results in relation to their findings. Non-publication of research findings dependent on the actual results is an important risk of bias to a systematic review and meta-analysis.

Missing studies

Several empirical investigations have shown that the findings from clinical trials are more likely to be published if the results are statistically significant ($P<0.05$) than if they are not.[125 206 207] For example, of 500 oncology trials with more than 200 participants for which preliminary results were presented at a conference of the American Society of Clinical Oncology, 81% with $P<0.05$ were published in full within five years compared with only 68% of those with $P>0.05$.[208]

Also, among published studies, those with statistically significant results are published sooner than those with non-significant findings.[209] When some studies are missing for these reasons, the available results will be biased towards exaggerating the effect of an intervention.

Missing outcomes

In many systematic reviews only some of the eligible studies (often a minority) can be included in a meta-analysis for a specific outcome. For some studies, the outcome may not be measured or may be measured but not reported. The former will not lead to bias, but the latter could.

Evidence is accumulating that selective reporting bias is widespread and of considerable importance.[42 43] In addition, data for a given outcome may be analysed in multiple ways and the choice of presentation influenced by the results obtained. In a study of 102 randomised trials, comparison of published reports with trial protocols showed that a median of 38% efficacy and 50% safety outcomes per trial, respectively, were not available for meta-analysis. Statistically significant outcomes had higher odds of being fully reported in publications when compared with non-significant outcomes for both efficacy (pooled odds ratio 2.4 (95% confidence interval 1.4 to 4.0)) and safety (4.7 (1.8 to 12)) data. Several other studies have had similar findings.[210 211]

Detection of missing information

Missing studies may increasingly be identified from trials registries. Evidence of missing outcomes may come from comparison with the study protocol, if available, or by careful examination of published articles.[11] Study publication bias and selective outcome reporting are difficult to exclude or verify from the available results, especially when few studies are available.

If the available data are affected by either (or both) of the above biases, smaller studies would tend to show larger estimates of the effects of the intervention. Thus one possibility is to investigate the relation between effect size and sample size (or more specifically, precision of the effect estimate). Graphical methods, especially the funnel plot,[212] and analytic methods (such as Egger's test) are often used,[213-215] although their interpretation can be problematic.[216 217] Strictly speaking, such analyses investigate "small study bias"; there may be many reasons why smaller studies have systematically different effect sizes than larger studies, of which reporting bias is just one.[218] Several alternative tests for bias have also been proposed, beyond the ones testing small study bias,[215 219 220] but none can be considered a gold standard. Although evidence that smaller studies had larger estimated effects than large ones may suggest the possibility that the available evidence is biased, misinterpretation of such data is common.[123]

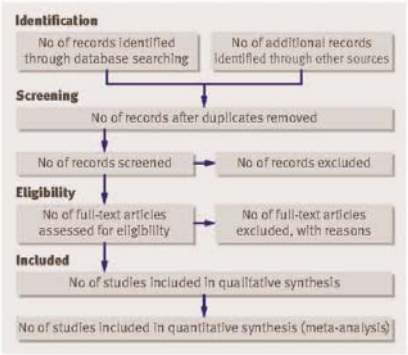


Fig 1 | Flow of information through the different phases of a systematic review.

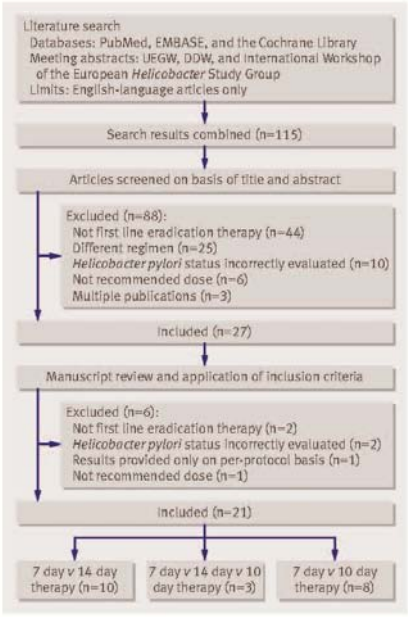


Fig 2 | Example flow diagram of study selection. DDW = Digestive Disease Week; UEGW = United European Gastroenterology Week. Adapted from Fuccio et al[130]

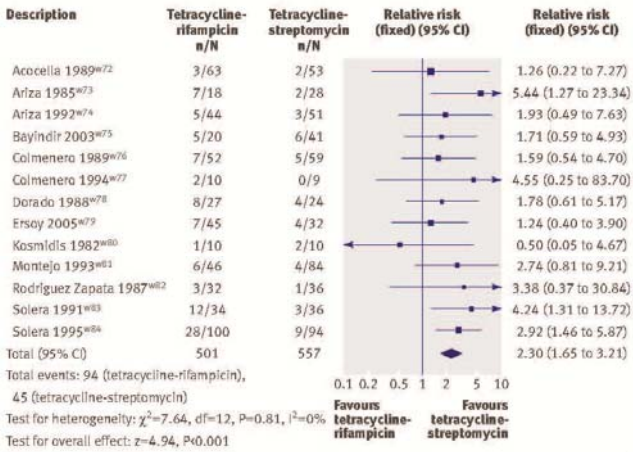


Fig 3 | Example of summary results: Overall failure (defined as failure of assigned regimen or relapse) with tetracycline-rifampicin versus tetracycline-streptomycin. Adapted from Skalsky et al[137]

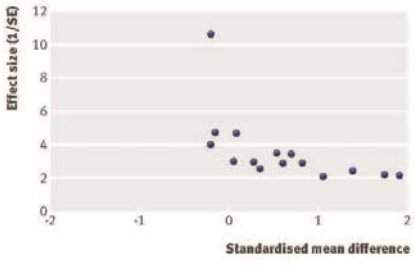


Fig 4 | Example of a funnel plot showing evidence of considerable asymmetry. SE = standard error. Adapted from Appleton et al[146]

1 Canadian Institutes of Health Research (2006) Randomized controlled trials registration/application checklist (12/2006). Available: http://www.cihr-irsc.gc.ca/e/documents/rct_reg_e.pdf. Accessed 26 May 2009.
2 Young C, Horton R. Putting clinical trials into context. *Lancet* 2005;366:107-108.
3 Moher D, Tetzlaff J, Tricco AC, Sampson M, Altman DG. Epidemiology and reporting characteristics of systematic reviews. *PLoS Med* 2007;4:e78. doi:10.1371/journal.pmed.0040078
4 Dixon E, Hameed M, Sutherland F, Cook DJ, Doig C. Evaluating meta-analyses in the general surgical literature: A critical appraisal. *Ann Surg* 2005;241:450-459.
5 Hemels ME, Vicente C, Sadri H, Masson MJ, Einarson TR. Quality assessment of meta-analyses of RCTs of pharmacotherapy in major depressive disorder. *Curr Med Res Opin* 2004;20:477-484.
6 Jin W, Yu R, LIW, Youping L, Ya L, et al. The reporting quality of meta-analyses improves: A random sampling study. *J Clin Epidemiol* 2008;61:770-775.
7 Moher D, Simera I, Schulz KF, Hoey J, Altman DG. Helping editors, peer reviewers and authors improve the clarity, completeness and transparency of reporting health research. *BMC Med* 2008;6:13.
8 Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, et al. Improving the quality of reports of meta-analyses of randomised controlled trials: The QUOROM statement. Quality of Reporting of Meta-analyses. *Lancet* 1999;354:1896-1900.
9 Green S, Higgins JPT, Alderson P, Clarke M, Mulrow CD, et al. Chapter 1: What is a systematic review? In: Higgins JPT, Green S, eds. Cochrane handbook for systematic reviews of interventions version 5.0.0 [updated February 2008]. The Cochrane Collaboration, 2008. Available: http://www.cochrane-handbook.org/. Accessed 26 May 2009.
10 Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, et al. GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 2008;336:924-926.
11 Higgins JPT, Altman DG. Chapter 8: Assessing risk of bias in included studies. In: Higgins JPT, Green S, eds. Cochrane handbook for systematic reviews of interventions version 5.0.0 [updated February 2008]. The Cochrane Collaboration, 2008. Available: http://www.cochrane-handbook.org/. Accessed 26 May 2009.
12 Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA Statement. *PLoS Med* 2008;6:e1000097. 10.1371/journal.pmed.1000097
13 Atkins D, Fink K, Slutsky J. Better information for better health care: The Evidence-based Practice Center program and the Agency for Healthcare Research and Quality. *Ann Intern Med* 2005;142:1035-1041.
14 Helfand M, Balshem H. Principles for developing guidance: AHRQ and the effective health-care program. *J Clin Epidemiol* 2009. In press.
15 Higgins JPT, Green S. Cochrane handbook for systematic reviews of interventions version 5.0.0 [updated February 2008]. The Cochrane Collaboration, 2008. Available: http://www.cochrane-handbook.org/. Accessed 26 May 2009.
16 Centre for Reviews and Dissemination. Systematic reviews: CRD's guidance for undertaking reviews in health care. York: University of York, 2009. Available: http://www.york.ac.uk/inst/crd/systematic_revfews_book.htm. Accessed 26 May 2009.
17 Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, et al. The revised CONSORT statement for reporting randomized trials: Explanation and elaboration. *Ann Intern Med* 2001;134:663-694.
18 Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, et al. The STARD statement for reporting studies of diagnostic accuracy: Explanation and elaboration. *Clin Chem* 2003;49:7-18.
19 Vandenbroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and elaboration. *PLoS Med* 2007;4:e297. doi:10.1371/journal.pmed.0040297
20 Barker A, Maratos EC, Edmonds L, Lim E. Recurrence rates of video-assisted thoracoscopic versus open surgery in the prevention of recurrent pneumothoraces: A systematic review of randomised and non-randomised trials. *Lancet* 2007;370:329-335.
21 Bjelakovic G, Nikolova D, Gluud LL, Simonetti RG, Gluud C. Mortality in randomized trials of antioxidant supplements for primary and secondary prevention: Systematic review and meta-analysis. *JAMA* 2007;297:842-857.
22 Montori VM, Wilczynski NL, Morgan D, Haynes RB. Optimal search strategies for retrieving systematic reviews from Medline: Analytical survey. *BMJ* 2005;330:68.
23 Bischoff-Ferrari HA, Willett WC, Wong JB, Giovannucci E, Dietrich T, et al. Fracture prevention with vitamin D supplementation: A meta-analysis of randomized controlled trials. *JAMA* 2005;293:2257-2264.
24 Hopewell S, Clarke M, Moher D, Wager E, Middleton P, et al. CONSORT for reporting randomised trials in journal and conference abstracts. *Lancet* 2008;371:281-283.
25 Hopewell S, Clarke M, Moher D, Wager E, Middleton P, et al. CONSORT for reporting randomized controlled trials in journal and conference abstracts: Explanation and elaboration. *PLoS Med* 2008;5:e20. doi:10.1371/journal.pmed.0050020
26 Haynes RB, Mulrow CD, Huth EJ, Altman DG, Gardner MJ. More informative abstracts revisited. *Ann Intern Med* 1990;113:69-76.
27 Mulrow CD, Thacker SB, Pugh JA. A proposal for more informative abstracts of review articles. *Ann Intern Med* 1988;108:613-615.
28 Froom P, Froom J. Deficiencies in structured medical abstracts. *J Clin Epidemiol* 1993;46:591-594.
29 Hartley J. Clarifying the abstracts of systematic literature reviews. *Bull Med Libr Assoc* 2000;88:332-337.
30 Hartley J, Sydes M, Blurton A. Obtaining information accurately and quickly: Are structured abstract more efficient? *J Infor Sci* 1996;22:349-356.
31 Pocock SJ, Hughes MD, Lee RJ. Statistical problems in the reporting of clinical trials. A survey of three medical journals. *N Engl J Med* 1987;317:426-432.
32 Taddio A, Pain T, Fassos FF, Boon H, Ilersich AL, et al. Quality of nonstructured and structured abstracts of original research articles in the British Medical Journal, the Canadian Medical Association Journal and the Journal of the American Medical Association. *CMAJ* 1994;150:1611-1615.
33 Harris KC, Kuramoto LK, Schulzer M, Retallack JE. Effect of school-based physical activity interventions on body mass index in children: A meta-analysis. *CMAJ* 2009;180:719-726.
34 James MT, Conley J, Tonelli M, Manns BJ, MacRae J, et al. Meta-analysis: Antibiotics for prophylaxis against hemodialysis catheter-related infections. *Ann Intern Med* 2008;148:596-605.
35 Counsell C. Formulating questions and locating primary studies for inclusion in systematic reviews. *Ann Intern Med* 1997;127:380-387.
36 Gotzsche PC. Why we need a broad perspective on meta-analysis. It may be crucially important for patients. *BMJ* 2000;321:585-586.
37 Grossman P, Niemann L, Schmidt S, Walach H. Mindfulness-based stress reduction and health benefits. A meta-analysis. *J Psychosom Res* 2004;57:35-43.
38 Brunton G, Green S, Higgins JPT, Kjeldstrøm M, Jackson N, et al. Chapter 2: Preparing a Cochrane review. In: Higgins JPT, Green S, eds. Cochrane handbook for systematic reviews of interventions version 5.0.0 [updated February 2008]. The Cochrane Collaboration, 2008. Available: http://www.cochrane-handbook.org/. Accessed 26 May 2009.
39 Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F. Systematic reviews of trials and other studies. *Health Technol Assess* 1998;2:1-276.
40 Ioannidis JP, Rosenberg PS, Goedert JJ, O'Brien TR. Commentary: meta-analysis of individual participants' data in genetic epidemiology. *Am J Epidemiol* 2002;156:204-210.
41 Stewart LA, Clarke MJ. Practical methodology of meta-analyses (overviews) using updated individual patient data. Cochrane Working Group. *Stat Med* 1995;14:2057-2079.
42 Chan AW, Hrobjartsson A, Haahr MT, Gøtzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: Comparison of protocols to published articles. *JAMA* 2004;291:2457-2465.
43 Dwan K, Altman DG, Arnaiz JA, Bloom J, Chan AW, et al. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLoS ONE* 2008;3:e3081. doi:10.1371/journal.pone.0003081
44 Silagy CA, Middleton P, Hopewell S. Publishing protocols of systematic reviews: Comparing what was done to what was planned. *JAMA* 2002;287:2831-2834.
45 Centre for Reviews and Dissemination. Research projects. York: University of York, 2009. Available: http://www.crd.york.ac.uk/crdweb. Accessed 26 May 2009.
46 The Joanna Briggs Institute. Protocols & work in progress, 2009. Available: http://www.joannabriggs.edu.au/pubs/systematic_reviews_prot.php. Accessed 26 May 2009.
47 Bagshaw SM, McAlister FA, Manns BJ, Ghali WA. Acetylcysteine in the prevention of contrast-induced nephropathy: A case study of the pitfalls in the evolution of evidence. *Arch Intern Med* 2006;166:161-166.
48 Biondi-Zoccai GG, Lotrionte M, Abbate A, Testa L, Remigi E, et al. Compliance with QUOROM and quality of reporting of overlapping meta-analyses on the role of acetylcysteine in the prevention of contrast associated nephropathy: Case study. *BMJ* 2006;332:202-209.
49 Sacks HS, Berrier J, Reitman D, Ancona-Berk VA, Chalmers TC. Meta-analyses of randomized controlled trials. *N Engl J Med* 1987;316:450-455.
50 Schroth RJ, Hitchon CA, Uhanova J, Noreddin A, Taback SP, et al. Hepatitis B vaccination for patients with chronic renal failure. *Cochrane Database Syst Rev* 2004;(3):CD003775, doi:10.1002/14651858.CD003775.pub2
51 Egger M, Zellweger-Zahner T, Schneider M, Junker C, Lengeler C, et al. Language bias in randomised controlled trials published in English and German. *Lancet* 1997;350:326-329.
52 Gregoire G, Derderian F, Le Lorier J. Selecting the language of the publications included in a meta-analysis: Is there a Tower of Babel bias? *J Clin Epidemiol* 1995;48:159-163.
53 Jüni P, Holenstein F, Sterne J, Bartlett C, Egger M. Direction and impact of language bias in meta-analyses of controlled trials: Empirical study. *Int J Epidemiol* 2002;31:115-123.
54 Moher D, Pham B, Klassen TP, Schulz KF, Berlin JA, et al. What contributions do languages other than English make on the results of meta-analyses? *J Clin Epidemiol* 2000;53:964-972.
55 Pan Z, Trikalinos TA, Kavvoura FK, Lau J, Ioannidis JP. Local literature bias in genetic epidemiology: an empirical evaluation of the Chinese literature. *PLoS Med* 2005;2:e334. doi:10.1371/journal.pmed.0020334

56 Hopewell S, McDonald S, Clarke M, Egger M. Grey literature in meta-analyses of randomized trials of health care interventions. *Cochrane Database Syst Rev* 2007;(2):MR000010, doi:10.1002/14651858.MR000010.pub3.
57 Melander H, Ahlqvist-Rastad J, Meijer G, Beermann B. Evidence b(l)ased medicine—Selective reporting from studies sponsored by pharmaceutical industry: review of studies in new drug applications. *BMJ* 2003;326:1171-1173.
58 Sutton AJ, Duval SJ, Tweedie RL, Abrams KR, Jones DR. Empirical assessment of effect of publication bias on meta-analyses. *BMJ* 2000;320:1574-1577.
59 Gotzsche PC. Believability of relative risks and odds ratios in abstracts: Cross sectional study. *BMJ* 2006;333:231-234.
60 Bhandari M, Devereaux PJ, Guyatt GH, Cook DJ, Swiontkowski MF, et al. An observational study of orthopaedic abstracts and subsequent full-text publications. *J Bone Joint Surg Am* 2002;84-A:615-621.
61 Rosmarakis ES, Soteriades ES, Vergidis PI, Kasiakou SK, Falagas ME. From conference abstract to full paper: Differences between data presented in conferences and journals. *Faseb J* 2005;19:673-680.
62 Toma M, McAlister FA, Bialy L, Adams D, Vandermeer B, et al. Transition from meeting abstract to full-length journal article for randomized controlled trials. *JAMA* 2006;295:1281-1287.
63 Saunders Y, Ross JR, Broadley KE, Edmonds PM, Patel S. Systematic review of bisphosphonates for hypercalcaemia of malignancy. *Palliat Med* 2004;18:418-431.
64 Shojania KG, Sampson M, Ansari MT, Ji J, Doucette S, et al. How quickly do systematic reviews go out of date? A survival analysis. *Ann Intern Med* 2007;147:224-233.
65 Bergerhoff K, Ebrahim S, Paletta G. Do we need to consider 'in process citations' for search strategies? Ottawa, Ontario, Canada: 12th Cochrane Colloqulum, 2-6 October 2004. Available: http://www.cochrane.org/colloquia/abstracts/ottawa/P-039.htm. Accessed 26 May 2009.
66 Zhang L, Sampson M, McGowan J. Reporting of the role of expert searcher in Cochrane reviews. *Evid Based Libr Info Pract* 2006;1:3-16.
67 Turner EH, Matthews AM, Linardatos E, Tell RA, Rosenthal R. Selective publication of antidepressant trials and its influence on apparent efficacy. *N Engl J Med* 2008;358:252-260.
68 Alejandria MM, Lansang MA, Dans LF, Mantaring JB. Intravenous immunoglobulin for treating sepsis and septic shock. *Cochrane Database Syst Rev* 2002;(1):CD001090, doi:10.1002/14651858.CD001090.
69 Golder S, McIntosh HM, Duffy S, Glanville J. Developing efficient search strategies to identify reports of adverse effects in MEDLINE and EMBASE. *Health Info Libr J* 2006;23:3-12.
70 Sampson M, McGowan J, Cogo E, Grimshaw J, Moher D, et al. An evidence-based practice guideline for the peer review of electronic search strategies. *J Clin Epidemiol* 2009; E-pub 2009 February 18.
71 Flores-Mir C, Major MP, Major PW. Search and selection methodology of systematic reviews in orthodontics (2000-2004). *Am J Orthod Dentofacial Orthop* 2006;130:214-217.
72 Major MP, Major PW, Flores-Mir C. An evaluation of search and selection methods used in dental systematic reviews published in English. *J Am Dent Assoc* 2006;137:1252-1257.
73 Major MP, Major PW, Flores-Mir C. Benchmarking of reported search and selection methods of systematic reviews by dental speciality. *Evid Based Dent* 2007;8:66-70.
74 Shah MR, Hasselblad V, Stevenson LW, Binanay C, O'Connor CM, et al. Impact of the pulmonary artery catheter in critically ill patients: Meta-analysis of randomized clinical trials. *JAMA* 2005;294:1664-1670.
75 Edwards P, Clarke M, DiGuiseppi C, Pratap S, Roberts I, et al. Identification of randomized controlled trials in systematic reviews: Accuracy and reliability of screening records. *Stat Med* 2002;21:1635-1640.
76 Cooper HM, Ribble RG. Influences on the outcome of literature searches for integrative research reviews. *Knowledge* 1989;10:179-201.
77 Mistiaen P, Poot E. Telephone follow-up, initiated by a hospital-based health professional, for postdischarge problems in patients discharged from hospital to home. *Cochrane Database Syst Rev* 2006(4):CD004510, doi:10.1002/14651858.CD004510.pub3.
78 Jones AP, Remmington T, Williamson PR, Ashby D, Smyth RL. High prevalence but low impact of data extraction and reporting errors were found in Cochrane systematic reviews. *J Clin Epidemiol* 2005;58:741-742.
79 Clarke M, Hopewell S, Juszczak E, Eisinga A, Kjeldstrom M. Compression stockings for preventing deep vein thrombosis in airline passengers. *Cochrane Database Syst Rev* 2006;(2):CD004002, doi:10.1002/14651858.CD004002.pub2.
80 Tramer MR, Reynolds DJ, Moore RA, McQuay HJ. Impact of covert duplicate publication on meta-analysis: A case study. *BMJ* 1997;315:635-640.
81 von Elm E, Poglia G, Walder B, Tramer MR. Different patterns of duplicate publication: An analysis of articles used in systematic reviews. *JAMA* 2004;291:974-980.
82 Gotzsche PC. Multiple publication of reports of drug trials. *Eur J Clin Pharmacol* 1989;36:429-432.
83 Allen C, Hopewell S, Prentice A. Non-steroidal anti-inflammatory drugs for pain in women with endometriosis. *Cochrane Database Syst Rev* 2005;(4):CD004753, doi:10.1002/14651858.CD004753.pub2.
84 Glasziou P, Meats E, Heneghan C, Shepperd S. What is missing from descriptions of treatment in trials and reviews? *BMJ* 2008;336:1472-1474.
85 Tracz MJ, Sideras K, Bolona ER, Haddad RM, Kennedy CC, et al. Testosterone use in men and its effects on bone health. A systematic review and meta-analysis of randomized placebo-controlled trials. *J Clin Endocrinol Metab* 2006;91:2011-2016.
86 Bucher HC, Hengstler P, Schindler C, Guyatt GH. Percutaneous transluminal coronary angioplasty versus medical treatment for non-acute coronary heart disease: Meta-analysis of randomised controlled trials. *BMJ* 2000;321:73-77.
87 Gluud LL. Bias in clinical intervention research. *Am J Epidemiol* 2006;163:493-501.
88 Plidal J, Hróbjartsson A, Jorgensen KJ, Hilden J, Altman DG, et al. Impact of allocation concealment on conclusions drawn from meta-analyses of randomized trials. *Int J Epidemiol* 2007;36:847-857.
89 Moja LP, Telaro E, D'Amico R, Moschetti I, Coe L, et al. Assessment of methodological quality of primary studies by systematic reviews: Results of the metaquality cross sectional study. *BMJ* 2005;330:1053.
90 Moher D, Jadad AR, Tugwell P. Assessing the quality of randomized controlled trials. Current issues and future directions. *Int J Technol Assess Health Care* 1996;12:195-208.
91 Sanderson S, Tatt ID, Higgins JP. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: A systematic review and annotated bibliography. *Int J Epidemiol* 2007;36:666-676.
92 Greenland S. Invited commentary: A critical look at some popular meta-analytic methods. *Am J Epidemiol* 1994;140:290-296.
93 Jüni P, Altman DG, Egger M. Systematic reviews in health care: Assessing the quality of controlled clinical trials. *BMJ* 2001;323:42-46.
94 Kunz R, Oxman AD. The unpredictability paradox: Review of empirical comparisons of randomised and non-randomised clinical trials. *BMJ* 1998;317:1185-1190.
95 Balk EM, Bonis PA, Moskowitz H, Schmid CH, Ioannidis JP, et al. Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. *JAMA* 2002;287:2973-2982.
96 Devereaux PJ, Beattie WS, Choi PT, Badner NH, Guyatt GH, et al. How strong is the evidence for the use of perioperative beta blockers in non-cardiac surgery? Systematic review and meta-analysis of randomised controlled trials. *BMJ* 2005;331:313-321.
97 Devereaux PJ, Bhandari M, Montori VM, Manns BJ, Ghali WA, et al. Double blind, you are the weakest link—Good-bye! *ACP J Club* 2002;136:A11.
98 van Nieuwenhoven CA, Buskens E, van Tiel FH, Bonten MJ. Relationship between methodological trial quality and the effects of selective digestive decontamination on pneumonia and mortality in critically ill patients. *JAMA* 2001;286:335-340.
99 Guyatt GH, Cook D, Devereaux PJ, Meade M, Straus S. Therapy. Users' guides to the medical literature. AMA Press, 2002:55-79.
100 Sackett DL, Gent M. Controversy in counting and attributing events in clinical trials. *N Engl J Med* 1979;301:1410-1412.
101 Montori VM, Devereaux PJ, Adhikari NK, Burns KE, Eggert CH, et al. Randomized trials stopped early for benefit: A systematic review. *JAMA* 2005;294:2203-2209.
102 Guyatt GH, Devereaux PJ. Therapy and validity: The principle of intention-to-treat. In: Guyatt GH, Rennie DR, eds. Users' guides to the medical literature. AMA Press, 2002:267-273.
103 Berlin JA. Does blinding of readers affect the results of meta-analyses? University of Pennsylvania Meta-analysis Blinding Study Group. *Lancet* 1997;350:185-186.
104 Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJ, et al. Assessing the quality of reports of randomized clinical trials: Is blinding necessary? *Control Clin Trials* 1996;17:1-12.
105 Pittas AG, Siegel RD, Lau J. Insulin therapy for critically ill hospitalized patients: A meta-analysis of randomized controlled trials. *Arch Intern Med* 2004;164:2005-2011.
106 Lakhdar R, Al-Mallah MH, Lanfear DE. Safety and tolerability of angiotensin-converting enzyme inhibitor versus the combination of angiotensin-converting enzyme inhibitor and angiotensin receptor blocker in patients with left ventricular dysfunction: A systematic review and meta-analysis of randomized controlled trials. *J Card Fail* 2008;14:181-188.
107 Bobat R, Coovadia H, Stephen C, Naidoo KL, McKerrow N, et al. Safety and efficacy of zinc supplementation for children with HIV-1 infection in South Africa: A randomised double-blind placebo-controlled trial. *Lancet* 2005;366:1862-1867.
108 Deeks JJ, Altman DG. Effect measures for meta-analysis of trials with binary outcomes. In: Egger M, Smith GD, Altman DG, eds. Systematic reviews in healthcare: Meta-analysis in context. 2nd edn. London: BMJ Publishing Group, 2001.
109 Deeks JJ. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. *Stat Med* 2002;21:1575-1600.
110 Engels EA, Schmid CH, Terrin N, Olkin I, Lau J. Heterogeneity and statistical significance in meta-analysis: An empirical study of 125 meta-analyses. *Stat Med* 2000;19:1707-1728.
111 Tierney JF, Stewart LA, Ghersi D, Burdett S, Sydes MR. Practical methods for incorporating summary time-to-event data into meta-analysis. *Trials* 2007;8:16.
112 Michiels S, Piedbois P, Burdett S, Syz N, Stewart L, et al. Meta-analysis

when only the median survival times are known: A comparison with individual patient data results. *Int J Technol Assess Health Care* 2005;21:119-125.
113 Briel M, Studer M, Glass TR, Bucher HC. Effects of statins on stroke prevention in patients with and without coronary heart disease: A meta-analysis of randomized controlled trials. *Am J Med* 2004;117:596-606.
114 Jones M, Schenkel B, Just J, Fallowfield L. Epoetin alfa improves quality of life in patients with cancer: Results of metaanalysis. *Cancer* 2004;101:1720-1732.
115 Elbourne DR, Altman DG, Higgins JP, Curtin F, Worthington HV, et al. Meta-analyses involving cross-over trials: Methodological issues. *Int J Epidemiol* 2002;31:140-149.
116 Follmann D, Elliott P, Suh I, Cutler J. Variance imputation for overviews of clinical trials with continuous response. *J Clin Epidemiol* 1992;45:769-773.
117 Wiebe N, Vandermeer B, Platt RW, Klassen TP, Moher D, et al. A systematic review identifies a lack of standardization in methods for handling missing variance data. *J Clin Epidemiol* 2006;59:342-353.
118 Hrobjartsson A, Gotzsche PC. Placebo interventions for all clinical conditions. *Cochrane Database Syst Rev* 2004;(2):CD003974, doi:10.1002/14651858.CD003974.pub2.
119 Shekelle PG, Morton SC, Maglione M, Suttorp M, Tu W, et al. Pharmacological and surgical treatment of obesity. *Evid Rep Technol Assess (Summ)* 2004:1-6.
120 Chan AW, Altman DG. Identifying outcome reporting bias in randomised trials on PubMed: Review of publications and survey of authors. *BMJ* 2005;330:753.
121 Williamson PR, Gamble C. Identification and impact of outcome selection bias in meta-analysis. *Stat Med* 2005;24:1547-1561.
122 Williamson PR, Gamble C, Altman DG, Hutton JL. Outcome selection bias in meta-analysis. *Stat Methods Med Res* 2005;14:515-524.
123 Ioannidis JP, Trikalinos TA. The appropriateness of asymmetry tests for publication bias in meta-analyses: A large survey. *CMAJ* 2007;176:1091-1096.
124 Briel M, Schwartz GG, Thompson PL, de Lemos JA, Blazing MA, et al. Effects of early treatment with statins on short-term clinical outcomes in acute coronary syndromes: A meta-analysis of randomized controlled trials. *JAMA* 2006;295:2046-2056.
125 Song F, Eastwood AJ, Gilbody S, Duley L, Sutton AJ. Publication and related biases. *Health Technol Assess* 2000;4:1-115.
126 Schmid CH, Stark PC, Berlin JA, Landais P, Lau J. Meta-regression detected associations between heterogeneous treatment effects and study-level, but not patient-level, factors. *J Clin Epidemiol* 2004;57:683-697.
127 Higgins JP, Thompson SG. Controlling the risk of spurious findings from meta-regression. *Stat Med* 2004;23:1663-1682.
128 Thompson SG, Higgins JP. Treating individuals 4: Can meta-analysis help target interventions at individuals most likely to benefit? *Lancet* 2005;365:341-346.
129 Uitterhoeve RJ, Vernooy M, Litjens M, Potting K, Bensing J, et al. Psychosocial interventions for patients with advanced cancer—A systematic review of the literature. *Br J Cancer* 2004;91:1050-1062.
130 Fuccio L, Minardi ME, Zagari RM, Grilli D, Magrini N, et al. Meta-analysis: Duration of first-line proton-pump inhibitor based triple therapy for *Helicobacter pylori* eradication. *Ann Intern Med* 2007;147:553-562.
131 Egger M, Smith GD. Bias in location and selection of studies. *BMJ* 1998;316:61-66.
132 Ravnskov U. Cholesterol lowering trials in coronary heart disease: Frequency of citation and outcome. *BMJ* 1992;305:15-19.
133 Hind D, Booth A. Do health technology assessments comply with QUOROM diagram guidance? An empirical study. *BMC Med Res Methodol* 2007;7:49.
134 Curioni C, Andre C. Rimonabant for overweight or obesity. *Cochrane Database Syst Rev* 2006;(4):CD006162, doi:10.1002/14651858.CD006162.pub2.
135 DeCamp LR, Byerley JS, Doshi N, Steiner MJ. Use of antiemetic agents in acute gastroenteritis: A systematic review and meta-analysis. *Arch Pediatr Adolesc Med* 2008;162:858-865.
136 Pakos EE, Ioannidis JP. Radiotherapy vs. nonsteroidal anti-inflammatory drugs for the prevention of heterotopic ossification after major hip procedures: A meta-analysis of randomized trials. *Int J Radiat Oncol Biol Phys* 2004;60:888-895.
137 Skalsky K, Yahav D, Bishara J, Pitlik S, Leibovici L, et al. Treatment of human brucellosis: Systematic review and meta-analysis of randomised controlled trials. *BMJ* 2008;336:701-704.
138 Altman DG, Cates C. The need for individual trial results in reports of systematic reviews. *BMJ* 2001. Rapid response.
139 Gotzsche PC, Hrobjartsson A, Maric K, Tendal B. Data extraction errors in meta-analyses that use standardized mean differences. *JAMA* 2007;298:430-437.
140 Lewis S, Clarke M. Forest plots: Trying to see the wood and the trees. *BMJ* 2001;322:1479-1480.
141 Papanikolaou PN, Ioannidis JP. Availability of large-scale evidence on specific harms from systematic reviews of randomized trials. *Am J Med* 2004;117:582-589.
142 Duffett M, Choong K, Ng V, Randolph A, Cook DJ. Surfactant therapy for acute respiratory failure in children: A systematic review and meta-analysis. *Crit Care* 2007;11:R66.
143 Balk E, Raman G, Chung M, Ip S, Tatsioni A, et al. Effectiveness of management strategies for renal artery stenosis: A systematic review. *Ann Intern Med* 2006;145:901-912.
144 Palfreyman S, Nelson EA, Michaels JA. Dressings for venous leg ulcers: Systematic review and meta-analysis. *BMJ* 2007;335:244.
145 Ioannidis JP, Patsopoulos NA, Evangelou E. Uncertainty in heterogeneity estimates in meta-analyses. *BMJ* 2007;335:914-916.
146 Appleton KM, Hayward RC, Gunnell D, Peters TJ, Rogers PJ, et al. Effects of n-3 long-chain polyunsaturated fatty acids on depressed mood: systematic review of published trials. *Am J Clin Nutr* 2006;84:1308-1316.
147 Kirsch I, Deacon BJ, Huedo-Medina TB, Scoboria A, Moore TJ, et al. Initial severity and antidepressant benefits: A meta-analysis of data submitted to the Food and Drug Administration. *PLoS Med* 2008;5:e45. doi:10.1371/journal.pmed.0050045
148 Reichenbach S, Sterchi R, Scherer M, Trelle S, Burgi E, et al. Meta-analysis: Chondroitin for osteoarthritis of the knee or hip. *Ann Intern Med* 2007;146:580-590.
149 Hodson EM, Craig JC, Strippoli GF, Webster AC. Antiviral medications for preventing cytomegalovirus disease in solid organ transplant recipients. *Cochrane Database Syst Rev* 2008;(2):CD003774, doi:10.1002/14651858.CD003774.pub3.
150 Thompson SG, Higgins JP. How should meta-regression analyses be undertaken and interpreted? *Stat Med* 2002;21:1559-1573.
151 Chan AW, Krleza-Jeric K, Schmid I, Altman DG. Outcome reporting bias in randomized trials funded by the Canadian Institutes of Health Research. *CMAJ* 2004;171:735-740.
152 Hahn S, Williamson PR, Hutton JL, Garner P, Flynn EV. Assessing the potential for bias in meta-analysis due to selective reporting of subgroup analyses within studies. *Stat Med* 2000;19:3325-3336.
153 Green LW, Glasgow RE. Evaluating the relevance, generalization, and applicability of research: Issues in external validation and translation methodology. *Eval Health Prof* 2006;29:126-153.
154 Liberati A, D'Amico R, Pifferi, Torri V, Brazzi L. Antibiotic prophylaxis to reduce respiratory tract infections and mortality in adults receiving intensive care. *Cochrane Database Syst Rev* 2004;(1):CD000022, doi:10.1002/14651858.CD000022.pub2.
155 Gonzalez R, Zamora J, Gomez-Camarero J, Molinero LM, Banares R, et al. Meta-analysis: Combination endoscopic and drug therapy to prevent variceal rebleeding in cirrhosis. *Ann Intern Med* 2008;149:109-122.
156 D'Amico R, Pifferi S, Leonetti C, Torri V, Tinazzi A, et al. Effectiveness of antibiotic prophylaxis in critically ill adult patients: Systematic review of randomised controlled trials. *BMJ* 1998;316:1275-1285.
157 Olsen O, Middleton P, Ezzo J, Gotzsche PC, Hadhazy V, et al. Quality of Cochrane reviews: Assessment of sample from 1998. *BMJ* 2001;323:829-832.
158 Hopewell S, Wolfenden L, Clarke M. Reporting of adverse events in systematic reviews can be improved: Survey results. *J Clin Epidemiol* 2008;61:597-602.
159 Cook DJ, Reeve BK, Guyatt GH, Heyland DK, Griffith LE, et al. Stress ulcer prophylaxis in critically ill patients. Resolving discordant meta-analyses. *JAMA* 1996;275:308-314.
160 Jadad AR, Cook DJ, Browman GP. A guide to interpreting discordant systematic reviews. *CMAJ* 1997;156:1411-1416.
161 Clarke L, Clarke M, Clarke T. How useful are Cochrane reviews in identifying research needs? *J Health Serv Res Policy* 2007;12:101-103.
162 [No authors listed]. World Medical Association Declaration of Helsinki: Ethical principles for medical research involving human subjects. *JAMA* 2000;284:3043-3045.
163 Clarke M, Hopewell S, Chalmers I. Reports of clinical trials should begin and end with up-to-date systematic reviews of other relevant evidence: A status report. *J R Soc Med* 2007;100:187-190.
164 Dube C, Rostom A, Lewin G, Tsertsvadze A, Barrowman N, et al. The use of aspirin for primary prevention of colorectal cancer: A systematic review prepared for the U.S. Preventive Services Task Force. *Ann Intern Med* 2007;146:365-375.
165 Critchley J, Bates I. Haemoglobin colour scale for anaemia diagnosis where there is no laboratory: A systematic review. *Int J Epidemiol* 2005;34:1425-1434.
166 Lexchin J, Bero LA, Djulbegovic B, Clark O. Pharmaceutical industry sponsorship and research outcome and quality: Systematic review. *BMJ* 2003;326:1167-1170.
167 Als-Nielsen B, Chen W, Gluud C, Kjaergard LL. Association of funding and conclusions in randomized drug trials: A reflection of treatment effect or adverse events? *JAMA* 2003;290:921-928.
168 Peppercorn J, Blood E, Winer E, Partridge A. Association between pharmaceutical involvement and outcomes in breast cancer clinical trials. *Cancer* 2007;109:1239-1246.
169 Yank V, Rennie D, Bero LA. Financial ties and concordance between results and conclusions in meta-analyses: Retrospective cohort study. *BMJ* 2007;335:1202-1205.
170 Jorgensen AW, Hilden J, Gøtzsche PC. Cochrane reviews compared with industry supported meta-analyses and other meta-analyses of the same drugs: Systematic review. *BMJ* 2006;333:782.
171 Gotzsche PC, Hrobjartsson A, Johansen HK, Haahr MT, Altman DG, et al. Ghost authorship in industry-initiated randomised trials. *PLoS Med* 2007;4:e19. doi:10.1371/journal.pmed.0040019
172 Akbari A, Mayhew A, Al-Alawi M, Grimshaw J, Winkens R, et al. Interventions to improve outpatient referrals from primary care to secondary care. *Cochrane Database of Syst Rev* 2008;(2):CD005471,

doi:10.1002/14651858.CD005471.pub2.
173 Davies P, Boruch R. The Campbell Collaboration. *BMJ* 2001;323:294-295.
174 Pawson R, Greenhalgh T, Harvey G, Walshe K. Realist review—A new method of systematic review designed for complex policy interventions. *J Health Serv Res Policy* 2005;10(Suppl 1):21-34.
175 Greenhalgh T, Robert G, Macfarlane F, Bate P, Kyriakidou O, et al. Storylines of research in diffusion of innovation: A meta-narrative approach to systematic review. *Soc Sci Med* 2005;61:417-430.
176 Lumley T. Network meta-analysis for indirect treatment comparisons. *Stat Med* 2002;21:2313-2324.
177 Salanti G, Higgins JP, Ades AE, Ioannidis JP. Evaluation of networks of randomized trials. *Stat Methods Med Res* 2008;17:279-301.
178 Altman DG, Moher D. [Developing guidelines for reporting healthcare research: scientific rationale and procedures.]. *Med Clin (Barc)* 2005;125(Suppl 1):8-13.
179 Delaney A, Bagshaw SM, Ferland A, Manns B, Laupland KB, et al. A systematic evaluation of the quality of meta-analyses in the critical care literature. *Crit Care* 2005;9:R575-582.
180 Altman DG, Simera I, Hoey J, Moher D, Schulz K. EQUATOR: Reporting guidelines for health research. *Lancet* 2008;371:1149-1150.
181 Plint AC, Moher D, Morrison A, Schulz K, Altman DG, et al. Does the CONSORT checklist improve the quality of reports of randomised controlled trials? A systematic review. *Med J Aust* 2006;185:263-267.
182 Simera I, Altman DG, Moher D, Schulz KF, Hoey J. Guidelines for reporting health research: The EQUATOR network's survey of guideline authors. *PLoS Med* 2008;5:e139. doi:10.1371/journal.pmed.0050139
183 Last JM. A dictionary of epidemiology. Oxford: Oxford University Press & International Epidemiological Association, 2001.
184 Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC. A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. Treatments for myocardial infarction. *JAMA* 1992;268:240-248.
185 Oxman AD, Guyatt GH. The science of reviewing research. *Ann N Y Acad Sci* 1993;703:125-133; discussion 133-124.
186 O'Connor D, Green S, Higgins JPT. Chapter 5: Defining the review question and developing criteria for including studies. In: Higgins JPT, Green S, editors. Cochrane handbook for systematic reviews of interventions version 5.0.0 [updated February 2008]. The Cochrane Collaboration, 2008. Available: http://www.cochrane-handbook.org/. Accessed 26 May 2009.
187 McDonagh M, Whiting P, Bradley M, Cooper J, Sutton A, et al. A systematic review of public water fluoridation. Protocol changes (AppendixM). NHS Centre for Reviews and Dissemination. York: University of York, 2000. Available: http://www.york.ac.uk/inst/crd/pdf/appm.pdf.. Accessed 26 May 2009.
188 Moher D, Cook DJ, Jadad AR, Tugwell P, Moher M, et al. Assessing the quality of reports of randomised trials: Implications for the conduct of meta-analyses. *Health Technol Assess* 1999;3:i-iv, 1-98.
189 Devereaux PJ, Choi PT, El-Dika S, Bhandari M, Montori VM, et al. An observational study found that authors of randomized controlled trials frequently use concealment of randomization and blinding, despite the failure to report these methods. *J Clin Epidemiol* 2004;57:1232-1236.
190 Soares HP, Daniels S, Kumar A, Clarke M, Scott C, et al. Bad reporting does not mean bad methods for randomised trials: Observational study of randomised controlled trials performed by the Radiation Therapy Oncology Group. *BMJ* 2004;328:22-24.
191 Liberati A, Himel HN, Chalmers TC. A quality assessment of randomized control trials of primary treatment of breast cancer. *J Clin Oncol* 1986;4:942-951.
192 Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, et al. Assessing the quality of randomized controlled trials: An annotated bibliography of scales and checklists. *Control Clin Trials* 1995;16:62-73.
193 Greenland S, O'Rourke K. On the bias produced by quality scores in meta-analysis, and a hierarchical view of proposed solutions. *Biostatistics* 2001;2:463-471.
194 Jüni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA* 1999;282:1054-1060.
195 Fleiss JL. The statistical basis of meta-analysis. *Stat Methods Med Res* 1993;2:121-145.
196 Villar J, Mackey ME, Carroli G, Donner A. Meta-analyses in systematic reviews of randomized controlled trials in perinatal medicine: Comparison of fixed and random effects models. *Stat Med* 2001;20:3635-3647.
197 Lau J, Ioannidis JP, Schmid CH. Summing up evidence: One answer is not always enough. *Lancet* 1998;351:123-127.
198 DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials* 1986;7:177-188.
199 Hunter JE, Schmidt FL. Fixed effects vs. random effects meta-analysis models: Implications for cumulative research knowledge. *Int J Sel Assess* 2000;8:275-292.
200 Deeks JJ, Altman DG, Bradburn MJ. Statistical methods for examining heterogeneity and combining results from several studies in meta-analysis. In: Egger M, Davey Smith G, Altman DG, eds. Systematic reviews in healthcare: Meta-analysis in context. London: BMJ Publishing Group, 2001:285-312.
201 Warn DE, Thompson SG, Spiegelhalter DJ. Bayesian random effects meta-analysis of trials with binary outcomes: Methods for the absolute risk difference and relative risk scales. *Stat Med* 2002;21:1601-1623.
202 Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ* 2003;327:557-560.
203 Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med* 2002;21:1539-1558.
204 Huedo-Medina TB, Sanchez-Meca J, Marin-Martinez F, Botella J. Assessing heterogeneity in meta-analysis: Q statistic or I2 index? *Psychol Methods* 2006;11:193-206.
205 Thompson SG, Turner RM, Warn DE. Multilevel models for meta-analysis, and their application to absolute risk differences. *Stat Methods Med Res* 2001;10:375-392.
206 Dickersin K. Publication bias: Recognising the problem, understanding its origin and scope, and preventing harm. In: Rothstein HR, Sutton AJ, Borenstein M, eds. Publication bias in meta-analysis—Prevention, assessment and adjustments. West Sussex: John Wiley & Sons, 2005:356.
207 Scherer RW, Langenberg P, von Elm E. Full publication of results initially presented in abstracts. *Cochrane Database Syst Rev* 2007;(2):MR000005. doi:10.1002/14651858.MR000005.pub3.
208 Krzyzanowska MK, Pintilie M, Tannock IF. Factors associated with failure to publish large randomized trials presented at an oncology meeting. *JAMA* 2003;290:495-501.
209 Hopewell S, Clarke M. Methodologists and their methods. Do methodologists write up their conference presentations or is it just 15 minutes of fame? *Int J Technol Assess Health Care* 2001;17:601-603.
210 Ghersi D. Issues in the design, conduct and reporting of clinical trials that impact on the quality of decision making. PhD thesis. Sydney: School of Public Health, Faculty of Medicine, University of Sydney, 2006.
211 von Elm E, Rollin A, Blumle A, Huwiler K, Witschi M, et al. Publication and non-publication of clinical trials: Longitudinal study of applications submitted to a research ethics committee. *Swiss Med Wkly* 2008;138:197-203.
212 Sterne JA, Egger M. Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. *J Clin Epidemiol* 2001;54:1046-1055.
213 Harbord RM, Egger M, Sterne JA. A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints. *Stat Med* 2006;25:3443-3457.
214 Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. Comparison of two methods to detect publication bias in meta-analysis. *JAMA* 2006;295:676-680.
215 Rothstein HR, Sutton AJ, Borenstein M. Publication bias in meta-analysis: Prevention, assessment and adjustments. West Sussex: John Wiley & Sons, 2005.
216 Lau J, Ioannidis JP, Terrin N, Schmid CH, Olkin I. The case of the misleading funnel plot. *BMJ* 2006;333:597-600.
217 Terrin N, Schmid CH, Lau J. In an empirical evaluation of the funnel plot, researchers could not visually identify publication bias. *J Clin Epidemiol* 2005;58:894-901.
218 Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ* 1997;315:629-634.
219 Ioannidis JP, Trikalinos TA. An exploratory test for an excess of significant findings. *Clin Trials* 2007;4:245-253.
220 Sterne JAC, Egger M, Moher D. Chapter 10: Addressing reporting biases. In: Higgins JPT, Green S, eds. Cochrane handbook for systematic reviews of interventions version 5.0.0 [updated February 2008]. The Cochrane Collaboration, 2008. Available: http://www.cochrane-handbook.org/. Accessed 26 May 2009.