UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
FACULDADE DE MEDICINA
PROGRAMA DE PÓS-GRADUAÇÃO EM EPIDEMIOLOGIA

**TESE DE DOUTORADO**

# Modelagem estatística para avaliação de impacto de políticas públicas de saúde no contexto de quase-experimentos longitudinais

Juliana Feliciati Hoffmann

Orientador: Profa. Dra. Suzi Alves Camey

Porto Alegre, setembro de 2019

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
FACULDADE DE MEDICINA
PROGRAMA DE PÓS-GRADUAÇÃO EM EPIDEMIOLOGIA

**TESE DE DOUTORADO**

# Modelagem estatística para avaliação de impacto de políticas públicas de saúde no contexto de quase-experimentos longitudinais

Juliana Feliciati Hoffmann

**Orientador: Profa. Dra. Suzi Alves Camey**

A apresentação desta tese é exigência do Programa de Pós-graduação em Epidemiologia, Universidade Federal do Rio Grande do Sul, para obtenção do título de Doutor.

Porto Alegre, Brasil.
2019

**BANCA EXAMINADORA**

Profa. Dra. Vanessa Bielefeldt Leotti, Programa de Pós-graduação em Epidemiologia, Universidade Federal do Rio Grande do Sul

Profa. Dra. Claunara Schilling Mendonça, Faculdade de Medicina, Universidade Federal do Rio Grande do Sul

Profa. Dra. Marilise Fraga de Souza, Secretaria Estadual de Saúde do Estado do Rio Grande do Sul

**MENSAGEM**

*"Cada segundo é tempo para mudar tudo para sempre."*

Charlie Chaplin

**AGRADECIMENTOS**

Aos meus pais, Ana e José, por terem me ensinado o valor e a importância do conhecimento e por estarem ao meu lado sempre.

Aos meus irmãos, Cris e Márcio, que tanto me confortam com a amizade, parceria e amor incondicional.

A Suzi Alves Camey, com quem estou sempre aprendendo em todas as dimensões da vida, pela dedicação ao longo dessa terceira orientação.

A Natalia Barbieri, por ser meu ombro amigo em todas as horas, seja para rir, para chorar, ou apenas para ter uma conversa que conforta e acalma o coração.

A Monica Oliveira, pela parceria de sempre, na boa e na ruim, por me incentivar e por me mostrar que era possível superar os desafios e chegar até aqui.

A Cristiane Melere, pela amizade de longa data e por ter encarado comigo o desafio de ingressar e concluir um doutorado, sabemos que não foi fácil.

Aos amigos e colegas do Deplan, em especial Antonio Cargnin e Carla Cunha, pelo apoio e pela compreensão nos momentos em que precisei me ausentar.

Às amigas e colegas Fernanda Vargas, Carina Furstenau, Ana Júlia Possamai, Rayssa Araújo, Silvia Lorenzetti, Katiuscia Freitas e Luciana Mieres, pela amizade, pelo apoio e por tornarem meus dias mais alegres.

Às amigas Andreia Fontanella, Paula Sientchkovski, Paula Bracco, Maria Cláudia Schardosim, Bárbara Riboldi e Nicole Utpott, por compartilharem comigo as alegrias e as angústias ao longo dessa construção.

Às colegas da Secretaria Estadual de Saúde, em especial Miriam Bellinaso e Marilise Fraga, por todo auxílio na obtenção das informações necessárias para a avaliação realizada nesse trabalho.

## SUMÁRIO

**ABREVIATURAS E SIGLAS**

ACS - Agente Comunitário de Saúde

AIC - Critério de Informação de Akaike

AIH - Autorização de Internação Hospitalar

ATC - *Average Treatment Effect for the controls* (efeito médio do tratamento nos controles)

ATE - *Average Treatment Effect (*efeito médio de tratamento)

ATT - *Average Treatment Effect on Treated (*efeito médio do tratamento sobre os tratados)

CIA – Conditional Independence Assumption (hipótese da independência condicional)

CIB - Comissão Intergestores Bipartite

CID - Classificação Internacional de Doenças

CNES - Cadastro Nacional de Estabelecimentos de Saúde

DAB - Departamento de Atenção Básica - Ministério da Saúde

DATASUS - Departamento de Informática do SUS

DID - *Diferenças-em-diferenças*

DP – Desvio Padrão

ECR - Ensaio Clínico Randomizado

EP – Erro Padrão

ESF - Estratégia Saúde da Família

EVIPNet - *Evidence-Informed Policy Network (*Rede para Políticas Informadas por Evidências)

FEE - Fundação de Economia e Estatística

GEE - G*eneralized Estimating Equation (*Equações de Estimação Generalizadas)

GLM - G*eneralized Linear Models (*Modelos Lineares Generalizados)

HIA - *Health Impact Assessment*

IBGE - Instituto Brasileiro de Geografia e Estatística

ICSAB - Internações por condições sensíveis à atenção básica

IDESE - Índice de Desenvolvimento Socioeconômico

IPEA - Instituto de Pesquisa Econômica Aplicada

NAAB - Núcleo de Apoio à Atenção Básica

NASF - Núcleo de Apoio à Estratégia de Saúde da Família

OMS - Organização Mundial de Saúde

PIB – Produto Interno Bruto

PMAQ - Programa Nacional de Melhoria do Acesso e da Qualidade da Atenção Básica

PNASH - Programa Nacional de Avaliação de Serviços Hospitalares

PNASS - Programa Nacional de Avaliação de Serviços de Saúde

PNMA-SUS - Política Nacional de Monitoramento e Avaliação do Sistema Único de Saúde

PSF - Programa de Saúde da Família

SES/RS – Secretaria da Saúde do Rio Grande do Sul

SUS - Sistema Único de Saúde

TabWin - Programa tabulador para Windows

TCE/RS - Tribunal de Contas do Estado do Rio Grande do Sul

TI - Tecnologia da Informação

TREND - *Transparent Reporting of Evaluations with Non-randomized Designs*

VIF - *Variance Inflation Factor* (Fator de Inflação da Variância)

**RESUMO**

Esta tese aborda os principais métodos de modelagem estatística direcionados para avaliação de impacto de políticas públicas de saúde, com foco em quase-experimentos longitudinais. É feita uma revisão da literatura sobre o tema e são apresentados dois artigos, considerando sempre o contexto de quase-experimentos longitudinais.

O primeiro artigo busca mostrar, passo-a-passo, cada uma das etapas necessárias para a realização de uma avaliação. Desde a etapa de estimação de escores de propensão, passando por método de pareamento e ponderação chegando, por fim, aos modelos propostos para estimar o impacto do Programa. São apresentados todos os códigos em R, um software livre amplamente conhecido, o que permite replicar as análises realizadas e, assim, contribuir para a disseminação da cultura da avaliação, consolidando a prática de políticas públicas baseadas em evidências. Neste primeiro artigo utilizou-se como exemplo um programa da Secretaria Estadual de Saúde do Rio Grande do Sul, cujos resultados obtidos pelos modelos GEE e GLM foram comparados utilizando as metodologias de análise propostas. Os resultados obtidos pelos dois métodos foram divergentes, ressaltando a atenção necessária na escolha do método.

No segundo artigo a ponderação, metodologia considerada mais adequada a partir das conclusões do primeiro artigo, é aplicada em quatro indicadores de resultado de internações psiquiátricas do mesmo Programa da Secretaria Estadual de Saúde, os Núcleos de Apoio à Atenção Básica. Através da ponderação pelos escores de propensão e modelos GEE para estimação do impacto observou-se que não houve efeito significativo do Programa em nenhum dos quatro indicadores propostos. Entretanto, os indicadores selecionados podem não estar refletindo da melhor forma os resultados esperados pelo programa, dado que a legislação de criação do Programa e os demais registros existentes não definem de forma clara um indicador de resultado. A programação em R apresentada permitiu a

realização de todas as análises necessárias, com a vantagem de viabilizar a disponibilização das rotinas para replicação das mesmas.

**ABSTRACT**

This thesis discusses the main statistical modeling methods used to impact assessment of public health policies, focusing on longitudinal quasi-experiments. The literature is reviewed, and two papers are presented, always considering the context of longitudinal quasi-experiments.

The first paper aims to present, step by step, all necessary phases to perform an impact evaluation. Beginning with the propensity score estimation, through the matching and weighting method, until the proposed models to estimate the Program impact. All R codes are presented, which allows replicating the analysis performed and thus contribute for the evaluation culture dissemination, consolidating the practice of evidence-based public policies. This first article used as an example a program of the State Secretariat of Health of Rio Grande do Sul, which results were compared using the proposed analysis methodologies.

In the second article the methodology considered the most appropriate, based on the conclusions of the first article, is applied to four result indicators, seeking to estimate the impact of the Program on these indicators.

# 1. **APRESENTAÇÃO**

Este trabalho consiste na tese de doutorado intitulada “Modelagem estatística para avaliação de impacto de políticas públicas de saúde no contexto de quase-experimentos longitudinais”, apresentada ao Programa de Pós-Graduação em Epidemiologia da Universidade Federal do Rio Grande do Sul, em 04 de outubro de 2019. O trabalho é apresentado em três partes, na ordem que segue:

1. Introdução, Revisão da Literatura e Objetivos
2. Artigos
3. Conclusões e Considerações Finais

## 2. INTRODUÇÃO

A avaliação vem se consolidando como prática no âmbito da gestão e da administração pública pelo menos nas últimas cinco décadas (Fernandes et al., 2011). O papel das avaliações ganha relevância pela necessidade de se produzir análises criteriosas e mais aprofundadas acerca dos resultados alcançados pelas políticas públicas.

As avaliações são instrumentos poderosos de retroalimentação do ciclo de políticas públicas e, portanto, devem ser pensadas de maneira sistemática, ajustando-se aos ciclos de planejamento e orçamento, de modo que sejam mais tempestivas e, portanto, mais úteis aos gestores. Para tanto, é vital que permeiem todas as etapas das políticas públicas, desde a definição de agenda e concepção das intervenções públicas, até o encerramento de determinada ação do governo, o que demanda diferentes tipos de avaliação (Cavalcanti, 2006).

Apesar de serem um importante meio para a gestão, as avaliações podem ser objeto de disputas institucionais e de controvérsias semânticas e ideológicas, podendo ser percebidas como ameaças à legitimidade política. No entanto, para que seu resultado seja eficaz, faz-se necessária a sua apropriação pela gestão, que deve estar sensibilizada para a sua importância e utilidade (Carneiro, 2013).

Nesse sentido, guias de avaliação têm sido propostos nas diversas esferas de governo, mais comumente no nível federal e estadual (Instituto de Pesquisa Econômica Aplicada, 2018*a*; *b*; Instituto Jones dos Santos Neves, 2018). Tais materiais buscam demonstrar de forma clara e objetiva as etapas e formas de condução de uma avaliação de qualidade, sendo fundamentais para a qualificação de gestores que visam implementar e gerenciar políticas baseadas em evidências.

No que diz respeito a políticas de saúde, a institucionalização da avaliação no SUS ainda é um processo incipiente. A avaliação de políticas

públicas ainda se confunde, muitas vezes, com a etapa de monitoramento. Entretanto, enquanto o monitoramento representa o acompanhamento por meio de indicadores, a avaliação corresponde ao exercício de mensurar, compreender e julgar os efeitos de uma determinada intervenção, de maneira a subsidiar as escolhas no processo de tomada de decisão (Sousa, 2018).

A inserção da avaliação na rotina dos serviços somente se dará por meio da implantação de uma cultura avaliativa. Para isso, é essencial a produção de informações estratégicas para a gestão, como resultados de avaliações bem estruturadas, periódicas e contínuas, destacando a perspectiva útil da avaliação (Dos Reis et al., 2012; Oliveira, A.E.F e Reis, R.S, 2016).

Nesse contexto de avaliação de políticas públicas e geração de informação para uma gestão baseada em evidências, a estatística torna-se uma área de indiscutível relevância, desde o delineamento da avaliação até a estimativa da medida de efeito que capta o impacto da política, garantindo a realização de análises qualificadas e com metodologia adequada. Com isso, esse trabalho se justifica no sentido de apresentar um passo-a-passo para a modelagem estatística de uma avaliação de impacto de políticas públicas no contexto quase-experimental longitudinal. São comparados métodos de análise comumente utilizados e apresentando os resultados obtidos com base na avaliação de um Programa da Secretaria Estadual de Saúde do Rio Grande do Sul. Busca-se demonstrar que é viável a realização de uma avaliação sem custos, com utilização de dados secundários e de um software livre, mesmo que não haja planejamento prévio e alocação aleatória dos grupos, como seria o mais recomendado.

O trabalho apresenta, inicialmente, uma revisão da literatura sobre a temática de avaliação de políticas públicas, descrevendo os tipos de avaliação existentes e contextualizando a avaliação na área da saúde. São também abordadas as metodologias de análise de avaliação de impacto em quase-experimentos, as medidas de efeito causal e a modelagem estatística

para a obtenção dessas medidas. No primeiro artigo apresenta-se um passo-a-passo e comparam-se as diferentes metodologias, pareamento e ponderação. No segundo artigo o método de ponderação é aplicado na avaliação dos Núcleos de Apoio à Atenção Básica, buscando verificar o impacto sobre quatro indicadores de resultado referentes a internações psiquiátrica e aos subgrupos de internações por álcool e outras drogas. Por fim, a última sessão trata das conclusões e considerações finais.

# 3. REVISÃO DE LITERATURA

## 3.1 Avaliação de políticas públicas

Na gestão pública brasileira, apesar de ser ainda muito incipiente, a cultura de avaliação de políticas públicas tem sido cada vez mais internalizada. A Reforma Gerencial, ocorrida na década de 1990, passou a institucionalizar a prática da avaliação, com o objetivo de obter maior eficiência no emprego dos recursos públicos. Tal fato tem estimulado a reunião de considerável arcabouço teórico, múltiplas reflexões metodológicas e destaque acadêmico (Facchini et al., 2008).

Um dos mais recentes fenômenos produzidos por essas transformações ocorridas na administração pública mundial é a ênfase dada a uma gestão voltada para resultados (Cavalcanti, 2006; Facchini et al., 2008). A avaliação de políticas públicas deve ser vista como um recurso que vai fornecer informações e, assim, auxiliar os gestores em situações de tomada de decisão, inclusive sobre alocação dos recursos orçamentários, através do julgamento do valor de cada ação governamental. A avaliação também contribui para a modernização da gestão da administração e dos serviços públicos, proporcionando maior transparência às ações, podendo ser utilizada como uma forma de prestação de contas à sociedade sobre o desempenho dos programas do Governo (Carneiro, 2013).

No caso dos países em desenvolvimento, a discussão sobre políticas públicas insere-se no contexto de recursos escassos e de grande demanda social, o que torna imprescindível comparar os resultados obtidos nos programas, optando por alternativas que maximizem o impacto das melhorias para a sociedade. Nesse sentido, a avaliação é percebida como uma oportunidade de aprimorar a dinâmica das políticas públicas, por meio de instrumentos capazes de oferecer informações mais qualificadas sobre a coerência dos insumos, processos e resultados da ação pública.

No Brasil, a avaliação de desempenho no serviço público, o estabelecimento de metas pela Lei de Responsabilidade Fiscal, a proliferação de conselhos com cadeiras para a sociedade civil na fiscalização de verbas públicas são alguns exemplos da nova dinâmica da gestão pública com vistas a alcançar maior eficiência. A importância da avaliação no relacionamento entre os governos e as agências financiadoras, tais como o Banco Mundial, também vem ganhando importância. Cada vez mais, as agências incluem exigências formais de avaliação em contratos de financiamentos externos (Banco Interamericano de Desenvolvimento, Banco Mundial) ou internos (Banco Nacional de Desenvolvimento Econômico e Social - BNDES, Caixa Econômica Federal), atrelando a apuração periódica ou sistemática de seus resultados aos financiamentos concedidos(Lilia Belluzzo e Rafael Camelo, 2015; Planejamento e avaliação de políticas públicas, 2015).

### 3.1.1 Tipos de Avaliação de políticas públicas

A avaliação pode ser definida como a aplicação de um conjunto de métodos de pesquisa, sendo que todos os métodos disponíveis podem circular em torno dos diferentes tipos de avaliação. Em relação aos tipos, as avaliações podem ser classificadas de acordo com o momento em que ocorrem - antes, durante ou depois da implementação da política ou programa, conforme detalhado a seguir.

#### 3.1.1.1 Avaliação *ex-ante*

A avaliação realizada antes do início do projeto, denominada *ex-ante,* procura medir a viabilidade do programa a ser implementado, no que diz respeito a sua relação de custo-benefício, de custo-efetividade, das taxas de retorno econômico dos investimentos previstos. Esse tipo de avaliação

procura orientar sobre a realização de um dado programa, no que diz respeito a sua formulação e desenvolvimento, através do estudo de seus objetivos, dos beneficiários, de suas necessidades e do seu campo de atuação.

Visando melhorar a formulação inicial do programa, o desenvolvimento do Modelo Lógico, por exemplo, também pode ser utilizado como um instrumento para proceder à avaliação *ex-ante*. O Modelo Lógico pode ser definido como uma ferramenta utilizada para sistematizar e comunicar as relações causais existentes entre recursos disponíveis, atividades desempenhadas e resultados esperados de um programa. Assim, serve como um organizador para desenhar o estudo de avaliação, focalizando nos elementos constitutivos do programa e identificando quais questões de avaliação devem ser colocadas (Cavalcanti, 2006; Planejamento e avaliação de políticas públicas, 2015). Desse modo, evita-se a detecção posterior de erros de formulação e de desenho, que, com maior racionalidade no processo inicial de implantação da política, poderiam ter sido previstos e eliminados (Instituto de Pesquisa Econômica Aplicada, 2018*a*).

#### 3.1.1.2 Avaliação de Processo

A avaliação intermediária é conduzida durante a implementação de um programa como forma de adquirir mais conhecimento quanto ao processo. Este tipo de avaliação não se preocupa com a efetividade do programa, mas tem como foco seus processos e mecanismos de execução. A principal função da avaliação de processo é observar em que medida o programa está sendo implementado conforme foi planejado.

Assim, a avaliação de processos se constitui, basicamente, em um instrumento que se preocupa em diagnosticar as possíveis falhas de um programa, no que diz respeito aos instrumentos, procedimentos, conteúdos e métodos, adequação ao público-alvo, visando o seu aperfeiçoamento,

através da interferência direcionada para seus aspectos intrínsecos. O objetivo é dar suporte e melhorar a gestão, a implementação e o desenvolvimento do programa (Cavalcanti, 2006).

#### 3.1.1.3 Avaliação ex-post

As avaliações posteriores à implementação do programa são chamadas de *ex-post* e visam trabalhar com resultados e impactos obtidos com o programa. Esta categoria de avaliação investiga em que medida o programa atinge os resultados esperados pelos formuladores, sendo realizada ao final da fase de implementação ou após a conclusão de um programa (Cavalcanti, 2006). O objetivo principal dessa modalidade de avaliação é analisar a efetividade de um programa, compreendendo em que medida o mesmo atingiu os resultados ou impactos esperados. Assim, é um instrumento relevante para a tomada de decisões ao longo da execução da política, indicando o que pode ser melhorado, bem como para a melhor alocação de recursos entre as diferentes políticas públicas setoriais (Instituto de Pesquisa Econômica Aplicada, 2018*b*). Dentre os diversos tipos de avaliações *ex-post*, a avaliação de impacto, detalhada a seguir, permite estimar o impacto das políticas públicas implementadas.

##### 3.1.1.3.1 Avaliação de Impacto

A avaliação de impacto busca estabelecer e quantificar estatisticamente as relações causais entre um programa e um conjunto de resultados, tentando excluir outras fontes de variação, verificando se os objetivos ou os impactos desejados estão sendo alcançados (Instituto de Pesquisa Econômica Aplicada, 2018). Esse tipo de avaliação busca analisar se os programas transformam realidades e impactam a sociedade, estando associada à estimação de resultados de médio e longo prazo.

Através dessa metodologia é possível quantificar o impacto das ações de governo, permitindo estimar as melhorias alcançadas com o programa. Para avaliar a mudança no problema a ser melhorado, ao longo do tempo, a avaliação compara a situação antes e depois da implementação do programa. Este modelo também implica a conformação de dois grupos: um grupo que recebe a intervenção proposta pelo programa e outro que não recebe, denominado grupo controle. O grupo controle fornece um parâmetro de comparação, representando a população alvo se não tivesse sido objeto de intervenção do programa. Esse parâmetro de comparação possibilita estimar o impacto do programa (Ramos, 2009).

A escolha do grupo controle deve, em primeiro lugar, circunscrever populações o mais idênticas possíveis ao grupo tratamento, o que permite a comparabilidade entre elas. Em segundo lugar, deve permitir isolar o impacto do programa de outros fatores intervenientes sobre a população alvo e controle, que podem estar influenciando a evolução dos indicadores escolhidos para a avaliação. Dessa forma, quando temos um bom grupo de comparação, a única razão para resultados diferentes entre os grupos é a intervenção oferecida pelo programa.

Na maioria das vezes a seleção dos grupos não é uma tarefa simples. Pensando inicialmente em questões éticas, o ideal seria que toda a população fosse beneficiada pelo Programa. Entretanto, pela necessidade de um grupo de comparação que não tenha recebido a intervenção, é necessário selecionar apenas parte dessa população para ser contemplada.

Outra questão fundamental envolvida na escolha dos componentes de cada grupo é a aleatoriedade. Idealmente todos os elementos da população deveriam ter a mesma probabilidade de pertencer ao grupo controle ou tratamento, o que parece simples, mas na prática apresenta-se como mais um desafio (Schor, A, 2007; Gertler et al., 2010). Esse tipo de estudo, no qual é realizada a alocação aleatória do tratamento, pode receber diferentes denominações, como *avaliação experimental*, *ensaio clínico randomizado* (ECR) ou *avaliação aleatorizada* (Gertler et al., 2010). É o

desenho considerado o padrão-ouro para avaliar causalidade por evitar viés de seleção e permitir a distribuição homogênea de variáveis confundidoras entre os grupos, tanto observáveis quanto não observáveis, de forma que qualquer diferença nos indicadores de resultado pode ser atribuída exclusivamente ao Programa (Handley et al., 2018). Nas situações nas quais não é viável uma seleção aleatória os estudos são denominados *quase-experimentos* (Rosenbaum, 2017).

Além da definição dos grupos, uma das grandes dificuldades da avaliação de impacto está relacionada à disponibilidade de dados na escala geográfica e no período de apuração necessários. Apesar de haver a possibilidade de coleta de dados primários, o custo e o tempo demandados por uma pesquisa de campo são muito elevados, o que acaba inviabilizando esse tipo de estudo. Em função disso, a realização de pesquisas avaliativas sobre políticas que não foram formatadas para serem avaliadas acabam sendo inviabilizadas. Isso demonstra a importância de planejar a avaliação antes de sua implementação, criando as condições necessárias para mapear, de forma precisa, a situação inicial que deveria ser alterada pela política, contrastando-a com a situação final, em dois grupos de comparação (Planejamento e avaliação de políticas públicas, 2015).

### 3.1.2 Avaliação de políticas públicas de saúde

No campo da saúde a avaliação surge vinculada aos avanços da epidemiologia e da estatística, a partir de testes de utilidade de diversas intervenções, particularmente direcionadas ao controle das doenças infecciosas e ao desenvolvimento dos primeiros sistemas de informação que orientassem as políticas sanitárias nos países desenvolvidos (Oliveira, A.E.F e Reis, R.S, 2016). A epidemiologia possui um papel crucial em descrever a situação de saúde de uma população, identificar fatores de risco e analisar as relações de saúde e doença. Entretanto, esse conhecimento tem sido pouco

aplicado para avaliar questões importantes de saúde pública, em especial em relação a políticas públicas de saúde (Gulis e Fujino, 2015).

No Brasil, diante de uma realidade de marcantes desigualdades sociais e de recursos públicos escassos para o financiamento do setor de saúde e, a avaliação em saúde se torna essencial para estabelecer a capacidade de resposta de políticas, programas e serviços às necessidades da população, bem como justificar estratégias implementadas (Facchini et al., 2008; Fernandes et al., 2011). A preocupação com avaliação em saúde tem como marco a criação do Sistema Único de Saúde (SUS), instituído pela Constituição Federal de 1988. Desde então diversas iniciativas de institucionalização da avaliação vem sendo realizadas, buscando incluir a avaliação na rotina das instituições.

Entre elas, algumas foram implementadas pelo Ministério da Saúde com o objetivo de fortalecer a avaliação em saúde. O Programa Nacional de Avaliação de Serviços Hospitalares – PNASH, iniciado em 1998, foi reformulado em 2015 e passou a ser chamado de Programa Nacional de Avaliação de Serviços de Saúde (PNASS). O PNASS foi retomado com o objetivo de avaliar os serviços do Sistema Único de Saúde, em especial estruturas, processos e resultados relacionados ao risco, acesso e satisfação dos cidadãos diante dos serviços e estabelecimentos de saúde (Dos Reis et al., 2012).

O Programa Nacional de Melhoria do Acesso e da Qualidade da Atenção Básica (PMAQ) (Oliveira, A.E.F e Reis, R.S, 2016), lançado em 2011, tem como objetivo incentivar os gestores e as equipes a melhorarem a qualidade dos serviços de saúde oferecidos, além de induzir a ampliação do acesso (Dos Reis et al., 2012). Para isso, foi realizado um processo de certificação de participantes da Estratégia Saúde da Família (ESF), que incluiu indicadores de acompanhamento da Informação da Atenção Básica do Sistema e um conjunto de normas de qualidade comprovadas na fase da avaliação externa do programa (Gonzaga da Matta-Machado et al., 2016). Os municípios participantes que alcançam uma melhoria no padrão de

qualidade no atendimento recebem um aumento no repasse de recursos do incentivo federal. Em 2016 foi iniciado o terceiro ciclo do PMAQ, contando com a adesão de 95,6% dos municípios brasileiros. No total, participam 38.865 (93,9%) equipes de Atenção Básica (Portal do Departamento de Atenção Básica [Internet].). Em 2017 foi iniciada a elaboração da Política Nacional de Monitoramento e Avaliação do Sistema Único de Saúde (PNMA-SUS), a qual pretende institucionalizar e consolidar práticas de monitoramento e avaliação como parte da rotina dos atores do SUS (Ministério da Saúde, 2017).

Embora muitas iniciativas e experiências estejam em curso, processos sistêmicos e periódicos de avaliação e monitoramento do SUS encontram-se pouco desenvolvidos. O uso de dados e indicadores tornou-se frequente entre os gestores do Sistema. No entanto, as informações geradas pouco orientam a tomada das decisões, assim como pouco se prestam para a qualificação dos serviços e ações em saúde (Dos Reis et al., 2012).

É necessário que existam Sistemas de Informação em Saúde capazes de alimentar o planejamento e a tomada de decisão em saúde (de Oliveira Quites, 2016). A Tecnologia da Informação (TI), neste contexto, assumiu nos últimos anos um papel imprescindível nas organizações públicas brasileiras, sendo instrumento fundamental para apoiar esse processo e dar suporte às avaliações do Sistema Único de Saúde (Ministério da Saúde - Secretaria Executiva, 2016).

Além dos avanços de TI, a inclusão da avaliação no planejamento da implementação das políticas é essencial. Caso contrário a avaliação pode se tornar muito complexa ou até mesmo inviável. Considerando o planejamento da avaliação na etapa de formulação de políticas, o conhecimento epidemiológico pode auxiliar a entender a complexidade do problema, definir metas e selecionar intervenções. Na implementação a epidemiologia pode contribuir através de técnicas de monitoramento e vigilância epidemiológica. Por fim, na etapa de avaliação, a epidemiologia

pode ser especialmente útil na análise dos impactos esperados e alcançados (Gulis e Fujino, 2015).

#### 3.1.2.1 Avaliação de impacto de políticas públicas de saúde

O termo *Health Impact Assessment* (HIA) vem sendo utilizado crescentemente pelas principais instituições de saúde do mundo, como a Organização Mundial de Saúde e o Serviço de Saúde Nacional do Reino Unido (N Krieger et al., 2003). HIA pode ser definida como uma combinação de procedimentos, métodos e ferramentas pelos quais uma política, programa ou projeto pode ser julgado em relação aos seus potenciais efeitos na saúde da população e em relação à distribuição desses efeitos nessa população (Dannenberg et al., 2008). A definição é ampla e contempla uma diversidade de técnicas de avaliação que podem incluir desde *checklists* até estudos utilizando técnicas mais sofisticadas (Slotterback et al., 2011). No Brasil ainda há poucas iniciativas de HIA, sendo que as que existem são em sua maioria voltadas ao impacto de questões ambientais sobre a saúde da população, tratando, por exemplo, do impacto da poluição do ar na saúde (Viegas et al., 2011; Abe e Miraglia, 2016), ou de outras questões relacionadas ao desenvolvimento sustentável (Silveira et al., 2012).

Para este trabalho o termo *avaliação de impacto* será aplicado de acordo com a definição descrita anteriormente na seção 3.1.1.3.1: é o tipo de avaliação que busca estabelecer nexos causais entre objetivos traçados no planejamento de uma ação e os impactos alcançados por elas, por intermédio das observáveis alterações da realidade. Em termos metodológicos, essas avaliações demandam a adoção de estratégias que impliquem algum nível de controle sobre o contexto observado, de modo a serem mais frequentes os desenhos quase-experimentais, mas também havendo a possibilidade de pesquisas com desenhos experimentais (Fernandes et al., 2011).

Ao contrário de outras áreas, como a educação, ainda são escassos os trabalhos que buscam estimar o impacto das políticas de saúde em um determinado indicador de resultado (Macinko, 2006). No exterior, estudos sobre avaliação de impacto foram identificados no México (Borja-Aburto et al., 2016), Rwanda (Basinga et al., 2011) e Estados Unidos (Dannenberg et al., 2008; Zeng et al., 2010).

No Brasil, os estudos que de fato avaliam o impacto de políticas de saúde se concentram na avaliação do Programa de Saúde da Família (PSF) e seu impacto em diferentes indicadores de resultado. Considerando períodos distintos, dois desses trabalhos avaliaram o impacto na mortalidade infantil (Macinko, 2006; Aquino et al., 2009), um na redução de mortes sem assistência em menores de 5 anos (Rasella et al., 2010) e outro em diversos indicadores de saúde da criança (Roncalli et al., 2006). Outros estudos verificaram o impacto do PSF sobre a mortalidade, sendo que um deles utilizou dados de 1993 a 2004 (Rocha e Soares, 2010), e outro avaliou o impacto sobre mortalidade por doença cardíaca e cerebrovascular considerando o período de 2000 a 2009 (Rasella et al., 2014). Também foram analisados os efeitos do Programa Mais Médicos sobre indicadores de atendimento básico de saúde, de morbidade e de mortalidade a partir de dados do Programa de 2013 a 2015, mostrando que houve efeitos positivos sobre os indicadores de atendimentos, consultas, encaminhamentos, exames e visitas, e efeitos negativos sobre alguns indicadores de morbidade, sem efeito na mortalidade nos municípios (Mazetto, 2018). Desfechos de saúde foram avaliados em relação ao impacto do Programa Bolsa Família, uma política de proteção social e renda e não estritamente de políticas de saúde. O estudo demostrou que houve efeito positivo no tratamento da tuberculose quando comparados beneficiários e não beneficiários (Oliosi et al., 2018).

## Saúde Pública Baseada em Evidências

O conceito de *medicina baseada em evidências* vem sendo aplicado na prática clínica, tendo por base decisões das melhores intervenções a serem adotadas com base nos resultados de estudos científicos de alta exigência metodológica. Considerando as evidências utilizadas para a prática clínica, a medicina baseada em evidências tem como padrão-ouro os ensaios clínicos randomizados (ECR) (West et al., 2008). Nesse tipo de delineamento ocorre a alocação aleatória dos indivíduos nos grupos de tratamento e controle.

A aleatorização é utilizada para designar os indivíduos aos grupos sem tendenciosidade, esperando-se que os sujeitos de um grupo tenham, em média, a mesma probabilidade de possuir certas características, observáveis ou não observáveis, que os do outro grupo (Robert H. Fletcher et al., 1996). Dessa forma é possível garantir uma estimativa não viesada do verdadeiro efeito causal entre intervenção e desfecho (West et al., 2008), que representa a melhor evidência disponível em pesquisas clínicas.

Os conceitos de medicina baseada em evidências passaram a ser utilizados também no campo das políticas públicas de saúde, um processo que vem sendo denominado *Saúde Pública Baseada em Evidências* (Eriksson, 2000; Des Jarlais et al., 2004; Victora et al., 2004; Brownson et al., 2010). Esse movimento busca utilizar o melhor conhecimento científico disponível para a tomada de decisão em saúde pública. Idealmente, evidências científicas deveriam ser sempre incorporadas na seleção e implementação de programas, desenvolvimento de políticas e processo de avaliação, de forma que sejam implementadas as intervenções que produzem maior retorno em saúde (Brownson et al., 2010).

Trazendo a prática da Medicina Baseada em Evidências para o contexto de políticas públicas, identifica-se que o ECR pode não ser o delineamento mais adequado para avaliar o impacto das intervenções de saúde pública em grande escala (Victora et al., 2004). Novas abordagens metodológicas são necessárias para viabilizar a avaliação dos resultados das políticas públicas de saúde.

Métodos que permitam estimar inferência causal através de delineamentos de avaliação não randomizados têm sido propostos (Des Jarlais et al., 2004; Victora et al., 2004) e recomendados na literatura, incluindo o uso de metodologias quase-experimentais (Barbosa et al., 2018). Nessa direção, também surgiu a iniciativa TREND (*Transparent Reporting of Evaluations with Non-randomized Designs*), que busca padronizar e dar mais transparência às publicações de pesquisas de avaliação de intervenções em saúde pública que utilizam desenhos não aleatórios e que possuem algum tipo de grupo de comparação (Des Jarlais et al., 2004).

Outra iniciativa de destaque é a Rede para Políticas Informadas por Evidências (EVIPNet, do inglês *Evidence-Informed Policy Network*), formada a partir de iniciativa da Organização Mundial de Saúde (OMS), que visa fomentar o uso apropriado de evidências científicas no desenvolvimento e implementação das políticas de saúde (Wichmann et al., 2016). A Rede promove o uso sistemático dos resultados da pesquisa científica na formulação e implementação de políticas e programas de saúde mediante o intercâmbio entre gestores, pesquisadores e representantes da sociedade civil (Wichmann et al., 2016).

Assim, observa-se que diversos esforços têm sido empreendidos com o objetivo de fazer com que as intervenções de saúde pública sejam baseadas em evidências e que não acabem sendo replicadas sem que haja uma avaliação que justifique sua continuidade (Jacobs et al., 2012).

## 3.2 Metodologias de análise para avaliação de impacto em quase-experimentos

Além de definir uma medida de efeito para avaliar o impacto de uma política pública, no caso de quase-experimentos é necessário pensar em formas de tornar as unidades comparáveis. A seguir serão apresentadas as medidas de efeito e, em seguida, os escores de propensão, o pareamento e a ponderação, como alternativas para tornar as unidades comparáveis.

### 3.2.1 Medidas de Efeito Causal

Uma medida de efeito causal não é apenas a diferença do desfecho entre os grupos que receberam e os que não receberam o tratamento, mas sim a diferença esperada caso fosse possível fazer a mesma observação receber e não receber o tratamento. Em termos de notação, $Y$ representa o valor do desfecho onde:

$$Y = \begin{cases} Y^0 tto\text{: } valor\ do\ desfecho\ supondo\ que\ nenhuma\ observação\ recebeu\ o\ tratamento \\ Y^1 tto\text{: } valor\ do\ desfecho\ supondo\ que\ todas\ as\ observações\ receberam\ o\ tratamento \end{cases}$$

São descritas na literatura como medidas de efeitos causal o efeito médio do tratamento (ATE) e o efeito médio do tratamento sobre os tratados (ATT). O ATE, proposto por Rosenbaum e Rubin, pode ser definido por (Paul R. Rosenbaum e Donald B. Rubin, 1983; Wooldridge, 2010):

$$ATE = E[Y^1 - Y^0] \quad (4)$$

Portanto, o ATE representa o efeito esperado do tratamento em um elemento qualquer de uma população, selecionado aleatoriamente. Entretanto, como a média é calculada para toda a população, pode incluir unidades que jamais seriam elegíveis para o tratamento no contexto de políticas públicas (Wooldridge, 2010).

Para a obtenção do ATT, efeito médio do tratamento sobre os tratados, é feita a mesma comparação do ATE, mas são consideradas apenas

as observações que efetivamente receberam o tratamento. Essa é uma medida mais recentemente utilizada na literatura de avaliação de políticas públicas (Caliendo e Kopeinig, 2008; Wooldridge, 2010). Portanto, o ATT é definido por (Wooldridge, 2010) como:

$$ATT = E[Y^1 - Y^0|tto = 1] = \mathrm{E}[Y^1|\mathrm{tto} = 1] - \mathrm{E}[Y^0|\mathrm{tto} = 1] \quad (5)$$

Considerando a hipótese do modelo contrafactual que afirma que o tratamento não teria efeito sobre os não-tratados, isto é:

$$E[Y^0|tto = 1] = E[Y^0|tto = 0] \quad (6)$$

Então

$$ATT = E[Y^1 - Y^0|tto = 1] = \mathrm{E}[Y^1|\mathrm{tto} = 1] - \mathrm{E}[Y^0|\mathrm{tto} = 1] \quad (7)$$

Considerando (6), temos que

$$ATT = E[Y^1 - Y^0|tto = 1] = \mathrm{E}[Y^1|\mathrm{tto} = 1] - \mathrm{E}[Y^0|\mathrm{tto} = 0] \quad (8)$$

Na prática, o ATE indica a eficácia relativa do tratamento em média na população, enquanto o ATT nos diz que resultado teriam os indivíduos do grupo controle caso eles tivessem recebido o tratamento (Angrist e Pischke, 2009).

Outra medida é o efeito médio do tratamento nos que não receberam o tratamento (ATC do inglês *Average Treatment Effect for the Controls*) (Morgan e Winship, 2015), definida como:

$$ATC = E[Y^1 - Y^0|tto = 0] = \mathrm{E}[Y^1|\mathrm{tto} = 0] - \mathrm{E}[Y^0|\mathrm{tto} = 0] \quad (9)$$

Diante das diversas medidas de efeito apresentadas, é necessário que o pesquisador defina qual delas deseja obter antes de iniciar as análises.

### 3.2.2 Escores de Propensão

A utilização de escores de propensão é uma abordagem que busca controlar o confundimento e o viés de seleção em estudos com o objetivo de medir o impacto de um tratamento nos quais não houve alocação aleatória das unidades que recebem a intervenção. A vantagem da estimação dos escores de propensão é permitir a utilização de um conjunto de informações de covariáveis de ajuste simultaneamente (D'Agostino, 1998). É uma alternativa que busca definir *a posteriori* um grupo controle mais parecido possível com o grupo que recebeu o tratamento, a partir da seleção de características observáveis.

O uso de escores para combinar diversas covariáveis em uma única variável vem desde 1976, sendo que em 1983 Rosenbaum e Rubin (Paul R. Rosenbaum e Donald B. Rubin, 1983) criaram o conceito de escore de propensão, estimado na linha de base, para controlar vieses de seleção em estudos de coorte, tornando-se desde então uma técnica popular na epidemiologia (Paul R. Rosenbaum e Donald B. Rubin, 1983; Stürmer et al., 2006).

O escore pode ser definido como sendo a probabilidade (propensão) de um indivíduo ser alocado ao grupo tratamento condicionada aos valores de seus preditores na linha de base (D'Agostino, 1998; West et al., 2008; Schardosim Cotta de Souza, 2010), conforme representado na equação (1) abaixo (Sasha O. Becker e Andrea Ichino, 2002):

$$p(x) = Pr(tto = 1|X = x) = E(tto|X = x) \quad (1)$$

Sendo: $tto = \begin{cases} 0, & \text{se grupo controle} \\ 1, & \textit{se grupo tratamento} \end{cases}$

$X =$ vetor das covariáveis da linha de base

A estimação dos escores pode ser feita através de regressão logística (Schardosim Cotta de Souza, 2010), sendo a variável dependente a participação ou não no grupo tratamento (resposta binária).

Assim, podemos identificar indivíduos que possuem valores próximos de escore de propensão nos dois grupos, tratamento e controle, e eles serão muito similares em relação às covariáveis da linha de base incluídas no modelo de regressão utilizado na estimação dos escores. Dessa forma, o ajustamento usando escores de propensão em média removerá os vieses subjacentes às distribuições das covariáveis (Schardosim Cotta de Souza, 2010). Entretanto, salienta-se que a metodologia de escores de propensão permite reduzir, mas não eliminar, o viés gerado pelos fatores não-observáveis (Resende e Oliveira, 2008).

Entre as alternativas de aplicação dos escores de propensão, encontram-se o pareamento e a ponderação, que serão detalhados a seguir, ou uma combinação deles (Schardosim Cotta de Souza, 2010; Li, 2011).

### 3.2.3 Pareamento

A utilização de métodos de pareamento permite identificar os casos mais similares entre os grupos controle e intervenção. O pareamento é uma técnica de subamostragem na qual é selecionado um elemento controle (não tratado) para cada elemento tratado baseado em características utilizadas na estimação do escore de propensão. Todas as unidades tratadas e os respectivos controles pareados são mantidos na subamostra, enquanto os não pareados são descartados. As diferenças observadas no desfecho são calculadas para casos tratados e controles pareados, sendo a diferença média o efeito estimado do tratamento (Morgan e Winship, 2015).

Diferentes metodologias podem ser utilizadas para identificação dos casos mais parecidos por técnicas de pareamento, algumas delas são: pareamento pelo vizinho mais próximo (*Nearest Neighbor Matching*); pareamento pelo raio (*Radius Matching*); pareamento pelo método de Kernel (*Kernel Matching*); pareamento por estratificação (*Stratified Matching*) e pareamento pela métrica de Mahalanobis (*Mahalanobis Metric*

*Matching*) (Sasha O. Becker e Andrea Ichino, 2002; Schardosim Cotta de Souza, 2010).

Ao realizar o pareamento recomenda-se restringir a comparação aos elementos cujos escores de propensão estão na região de suporte comum. Essa região compreende os valores de escore que estão no intervalo entre o valor mínimo no grupo tratamento e o valor máximo no grupo controle (Lee, 2013; Patorno et al., 2013). Assim, para cada probabilidade de participação estimada para os integrantes do grupo que recebeu tratamento haverá um correspondente no grupo controle com probabilidade semelhante, o que representa um suporte comum entre os dois grupos (Resende e Oliveira, 2008) Na prática, com essa restrição são excluídos os casos que seriam considerados "nunca tratados" (apresentam valor de escore muito baixo) ou aqueles que seriam considerados "sempre tratados" (apresentam valor de escore muito alto), assegurando que para cada indivíduo tratado exista outro não tratado pareado (Resende e Oliveira, 2008). Na ausência de comparabilidade os elementos sem pares são descartados.

Além do suporte comum, a hipótese de independência condicional (CIA, do inglês *Conditional Independence Assumption*) baseada no escore de propensão também é importante no pareamento. Ela assume dado um conjunto de covariáveis que não são afetadas pelo tratamento, os desfechos são independentes da alocação ao grupo tratamento. Ou seja, uma vez controlada pelas covariáveis utilizadas na estimação dos escores de propensão (não são afetadas pelo tratamento), a alocação ao grupo tratamento é dita ser aleatória (Caliendo e Kopeinig, 2008)

Uma vez realizado o pareamento, teremos como resultado a divisão da subamostra em dois grupos: controle e tratamento. Como forma de garantir a comparabilidade dos grupos, objetivo principal dessa análise, a última etapa consiste na avaliação da qualidade do pareamento, chamada de balanceamento (Zhang et al., 2019), que busca verificar a distribuição das covariáveis entre os grupos, na linha de base.

### 3.2.4 Ponderação

Outra forma de aplicação dos escores de propensão, também com o objetivo de eliminar o viés de seleção, pode dar-se através da ponderação (Posner e Ash, 2009). O primeiro passo para utilização da ponderação é calcular os pesos de cada elemento da amostra (denotados por $w$) para estimação do efeito do tratamento. A forma de cálculo varia conforme a medida de efeito a ser estimada, conforme foi abordado em detalhes na sessão 3.2.1.

A fim de obter o efeito médio de tratamento (ATE do inglês *Average Treatment Effect*), o inverso do escore de propensão é usado para ponderar cada observação no grupo tratado e o inverso de um menos o escore de propensão para ponderar cada observação nos controles (Posner e Ash, 2009), conforme apresentado em (2), abaixo.

$$w_{ATE\,i} = \begin{cases} \dfrac{1}{\hat{p}(x_i)} & tto = 1 \\ \dfrac{1}{1-\hat{p}(x_i)} & tto = 0 \end{cases} \qquad (2)$$

$w_{ATEi}$ é a ponderação de cada observação i para obtenção do efeito médio do tratamento

$\hat{p}(x_i)$ é o escore de propensão estimado para cada observação i

Apesar de produzir estimativas não viesadas, a ponderação pelos escores de propensão para estimação do ATE pode não ser o método ideal quando há observações com probabilidade muito baixa de receberem o tratamento. Nesses casos, cujo valor do escore de propensão é próximo de zero, seu valor do inverso será extremamente alto. Com isso, a estimativa do tamanho de efeito ficaria dominada por esse valor produzindo resultados

com variabilidade muito ampla (Posner and Ash, 2009; Schardosim Cotta de Souza, 2010).

Para obtenção dos pesos para estimação do efeito médio do tratamento sobre os tratados (ATT do inglês *Average Treatment Effect on Treated),* todas as observações do grupo tratamento recebem valor 1 e as do grupo controle recebem os pesos calculados conforme abaixo, em (3):

$$w_{ATTi} = \begin{cases} 1 & tto = 1 \\ \dfrac{\hat{p}(x_i)}{1-\hat{p}(x_i)} & tto = 0 \end{cases} \qquad (3)$$

$w_{ATTi}$ é a ponderação de cada observação i para obtenção do efeito médio do tratamento sobre os tratados

$\hat{p}(x_i)$ é o escore de propensão estimado para cada observação i

Através da ponderação espera-se que os grupos tenham características similares, uma vez que o peso aplicado no grupo controle torna a distribuição mais similar ao grupo tratamento. Da mesma forma sugerida no pareamento, recomenda-se verificar o balanceamento dos grupos também na ponderação. Por fim, para estimativa do efeito do tratamento deverá ser utilizado um modelo de regressão ponderado.

Em relação ao pareamento, a ponderação tem a vantagem de não descartar unidades amostrais na análise (Posner and Ash, 2009), além de eliminar a subjetividade de escolha do método de pareamento, que pode variar conforme a preferência do responsável pelas análises (Li, 2011).

## 3.3 Modelagem estatística para obtenção das estimativas de efeito causal

### 3.3.1 Método Diferenças-em-diferenças

O método denominado *Diferenças-em-diferenças* (DID) é utilizado para comparar as mudanças nos desfechos de interesse ao longo do tempo entre o grupo que participou do programa (grupo tratamento) e o grupo que não participou (grupo controle). Essa metodologia é comumente utilizada na área de economia da saúde, mas ainda pouco conhecida em epidemiologia (Fu et al., 2007).

O método DID se baseia no fato de que, considerando somente a mudança no desfecho antes e após a intervenção não é possível obter a estimativa do impacto do programa, uma vez que diversos outros fatores podem influenciar o desfecho ao longo do tempo. Por outro lado, se for utilizada somente a comparação de quem recebeu e quem não recebeu a intervenção, também poderá haver outros fatores não observados que justifiquem porque alguns receberam a intervenção e outros não (Gertler et al., 2010). Entretanto, através da combinação dos dois métodos é possível obter uma medida de impacto do programa mais confiável.

A diferença do desfecho antes e depois da intervenção controla para fatores que são constantes ao longo do tempo naquele grupo, uma vez que a comparação é feita com o próprio grupo. Mas ainda não estariam sendo considerados os fatores que variam com o tempo. Uma forma de captar esses fatores é considerar a mudança no desfecho antes e depois para um grupo que não recebeu o programa, mas que possui as mesmas características do grupo que recebeu. Essa seria a segunda diferença a ser considerada.

Assim, o método *Diferenças-em-diferenças,* que utiliza modelos lineares generalizados, pode ser expresso pela equação (10) (Villa, 2016):

$$E(Y_{ij}) = \beta_0 + \beta_1 t_j + \beta_2 tto_i + \beta_3 t_j tto_i \quad (10)$$

Sendo:

$E(Y_{ij}) =$ esperança do desfecho Y na observação i no momento j

$i = 1, \ldots n \; e \; n = número \; de \; observações;$

$tto_i = \begin{cases} 0, & \text{se grupo controle} \\ 1, & \text{se grupo tratamento} \end{cases}$

$j = 0,1;$

$t_j = \begin{cases} 0, & \text{se j} = 0, \text{ou seja, antes do tratamento} \\ 1, & \text{se j=1, ou seja, após o tratamento} \end{cases}$

A partir de (10) temos que o valor esperado da diferença no grupo tratado, entre o período antes e depois, do desfecho Y é dado por:

$$E(Y_{i1} - Y_{i0}|tto = 1) = (\beta_0 + \beta_1 + \beta_2 + \beta_3) - (\beta_0 + \beta_2) = \beta_1 + \beta_3 (11)$$

e o valor esperado da diferença no grupo controle, entre o período antes e depois, do desfecho Y é dado por:

$$E(Y_{i1} - Y_{i0}|tto = 0) = (\beta_0 + \beta_1) - (\beta_0) = \beta_1 \quad (12)$$

Então, por (11) e (12) temos que o impacto da intervenção é dado por: $DID = (\beta_1 + \beta_3) - \beta_1 = \beta_3$

Assim, o método Diferenças-em-diferenças é uma abordagem comum para estimar o efeito médio do tratamento em quase-experimentos, sendo que a estimativa do coeficiente $\beta_3$ representa a estimativa do ATT (Strezhnev; Athey e Imbens, 2016).

Como pressuposto para a aplicação dessa metodologia e obtenção de estimativas de impacto confiáveis, deve-se verificar se a tendência temporal da variável de desfecho é paralela entre os dois grupos de comparação. Ou seja, na ausência do tratamento, os desfechos deveriam aumentar ou diminuir na mesma taxa em ambos os grupos. Como não é possível saber o que teria acontecido na ausência do programa, deve-se assumir que não existe nenhuma diferença entre os grupos que varie com o tempo (Gertler et al., 2010).

### 3.3.2 Equações de Estimação Generalizadas

Os modelos GEE (do inglês, G*eneralized Estimating Equation e em português* Equações de Estimação Generalizadas) permitem a análise de dados coletados em delineamentos longitudinais, aninhados, ou com medidas repetidas, por permitirem a especificação de uma matriz de correlação das respostas de um sujeito e admite diferentes distribuições de probabilidade (Ballinger, 2004). Com isso, obtêm-se estimativas de variabilidade distintas daquelas obtidas pelo GLM (GLM do inglês, G*eneralized Linear Models e em português* Modelos Lineares Generalizados), em função da matriz de correlação.

A notação utilizada abaixo para o modelo GEE é análoga à do modelo DID, uma vez que são utilizadas as mesmas medidas.

$$E(Y_{ijk}) = \beta_0 + \beta_1 t_j + \beta_2 tto_i + \beta_3 ano_k + \beta_4 t_j tto_i + \beta_5 t_j ano_k + \beta_6 tto_i ano_k + \beta_7 tto_i ano_k t_j \quad (13)$$

Sendo:

$E(Y_{ijk}) = esperança\ do\ desfecho\ na\ observação\ i\ no\ momento\ j\ e\ no\ ano\ k$

$i = 1, \dots n$ e $n = número\ de\ observações$;

$j = 0{,}1$;

$k = número\ de\ anos$

$$t_j = \begin{cases} 0, & \text{se j} = 0, \text{ou seja, antes do tratamento} \\ 1, & \text{se j=1, ou seja, após o tratamento} \end{cases}$$

$$tto_i = \begin{cases} 0, & \text{se a observação i pertence ao grupo controle} \\ 1, & \text{se a observação i pertence ao grupo tratamento} \end{cases}$$

A partir de (13) temos que o valor esperado da diferença no grupo controle, entre o período antes e depois, do desfecho Y será:

$$E(\mathrm{Y_{i1k}} - \mathrm{Y_{i0k}} | tto = 0) = (\beta_0 + \beta_1 + \beta_3 \mathrm{ano_k} + \beta_5 \mathrm{ano_k}) - (\beta_0 + \beta_3 \mathrm{ano_k}) = \beta_1 + \beta_5 \mathrm{ano_k} \quad (14)$$

e o valor esperado da diferença no grupo tratamento, entre o período antes e depois, do desfecho Y será:

$$E(Y_{i1k} - Y_{i0k}|tto = 1) = (\beta_0 + \beta_1 + \beta_2 + \beta_3 ano_k + \beta_4 + \beta_5 ano_k + \beta_6 ano_k + \beta_7 ano_k) - (\beta_0 + \beta_2 + \beta_3 ano_k + \beta_6 ano_k) = \beta_1 + \beta_4 + ano_k(\beta_5 + \beta_7) \quad (15)$$

Então, $DID = \beta_1 + \beta_4 + ano_k(\beta_5 + \beta_7)) - (\beta_1 + \beta_5 ano_k) = \beta_4 + \beta_7 ano_k$

Logo, o impacto será medido por $\beta_4 + \beta_7 ano_k$ e dependerá, portanto, do ano k.

Os mesmos cuidados de qualidade de ajuste de modelos devem ser tomados quando tratamos de avaliação de políticas públicas. Maiores detalhes sobre técnicas de diagnóstico de modelos de regressão podem ser obtidos em Paula, 2003 (Gilberto A. Paula, 2013).

## 4. **OBJETIVOS**

Objetivo Geral:

- Demonstrar e comparar diferentes métodos de avaliação de impacto que sejam aplicáveis a políticas públicas no contexto de um quase-experimento longitudinal.

Objetivos Específicos:

- Construir rotinas no R, software livre, que permitam que os métodos possam ser replicados por gestores de políticas de saúde, favorecendo a cultura da avaliação em saúde e facilitando a tomada de decisão baseada em evidências.
- Verificar o impacto do Programa Núcleo de Apoio à Atenção Básica (NAABs), da Secretaria de Saúde do Estado do Rio Grande do Sul, nas internações psiquiátricas e em especial nas internações por álcool e outras drogas no Rio Grande do Sul, no período de 2012 a 2016.

## 5. REFERÊNCIAS BIBLIOGRÁFICAS


Abe K, Miraglia S. Health Impact Assessment of Air Pollution in São Paulo, Brazil. International Journal of Environmental Research and Public Health. 2016 Jul 11;13(7):694.

Angrist JD, Pischke J-S. Mostly harmless econometrics: an empiricist's companion [Internet]. Princeton: Princeton University Press; 2009 [citado em 31 Ago 2018]. Disponível em: http://www.dawsonera.com/depp/reader/protected/external/AbstractView/S9781400829828

Aquino R, de Oliveira NF, Barreto ML. Impact of the family health program on infant mortality in Brazilian municipalities. American journal of public health. 2009;99(1):87–93.

Athey S, Imbens GW. Identification and Inference in Nonlinear Difference-in-Differences Models. Econometrica. 2016 Mar;74(2):431–497.

Ballinger GA. Using Generalized Estimating Equations for Longitudinal Data Analysis. Organizational Research Methods. 2004 Abr;7(2):127–50.

Barbosa ACQ, Amaral PV, Francesconi GV, Rosales C, Kemper ES, Silva NC da, et al. Programa Mais Médicos: como avaliar o impacto de uma abordagem inovadora para superação de iniquidades em recursos humanos. Rev Panam Salud Publica. 2018 Nov 6;42:e185.

Basinga P, Gertler PJ, Binagwaho A, Soucat AL, Sturdy J, Vermeersch CM. Effect on maternal and child health services in Rwanda of payment to primary health-care providers for performance: an impact evaluation. The Lancet. 2011;377(9775):1421–1428.

Borja-Aburto VH, González-Anaya JA, Dávila-Torres J, Rascón-Pacheco RA, González-León M. Evaluation of the impact on non-communicable chronic diseases of a major integrated primary health care program in Mexico. Family Practice. 2016 Jun;33(3):219–25.

Brownson RC, Baker EA, Leet TL, Gillespie KN, True WR. Evidence-based public health. Oxford University Press; 2010.

Caliendo M, Kopeinig S. Some Practical Guidance for the Implementation of Propensity Score Matching. Journal of Economic Surveys. 2008 Fev;22(1):31–72.

Carneiro F. Avaliação de Políticas Públicas: por um procedimento integrado ao ciclo da gestão. Revista Perspectivas em Políticas Públicas [Internet]. 2013;6(11). Disponível em: http://www.uemg.br/openjournal/index.php/revistappp/article/view/893

Cavalcanti MM de A. Avaliação de Políticas Públicas e Programas Governamentais - Uma abordagem Conceitual. Interfaces de Saberes [Internet]. 2006 [citado em 2 Nov 2016];6(1). Disponível em: https://interfacesdesaberes.fafica-pe.edu.br/index.php/import1/article/view/20

D'Agostino RB. Tutorial in biostatistics: propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat Med. 1998;17(19):2265–2281.

Dannenberg AL, Bhatia R, Cole BL, Heaton SK, Feldman JD, Rutt CD. Use of Health Impact Assessment in the U.S. American Journal of Preventive Medicine. 2008 Mar;34(3):241–56.

Des Jarlais DC, Lyles C, Crepaz N. Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: the TREND statement. American journal of public health. 2004;94(3):361–366.

Dos Reis AT, De Oliveira PDTR, Sellera PE. Sistema de Avaliação para a Qualificação do Sistema Único de Saúde (SUS). RECIIS [Internet]. 2012 Ago 31 [citado em 2016 Nov 2];6(2). Disponível em: http://www.reciis.cict.fiocruz.br/index.php/reciis/article/view/622/1089

Eriksson C. Learning and knowledge-production for public health: a review of approaches to evidence-based public health. Scandinavian Journal of Public Health. 2000 Out;28(4):298–308.

Facchini LA, Piccini RX, Tomasi E, Thumé E, Teixeira VA, Silveira DS da, et al. Avaliação de efetividade da Atenção Básica à Saúde em municípios das regiões Sul e Nordeste do Brasil: contribuições metodológicas. Cad. Saúde Pública. 2008;24(Sup 1):S159–72.

Fernandes FMB, Ribeiro JM, Moreira MR. Reflexões sobre avaliação de políticas. Cad. Saude Publica. 2011;27(9):1667–1677.

Fu AZ, Dow WH, Liu GG. Propensity score and difference-in-difference methods: a study of second-generation antidepressant use in patients with bipolar disorder. Health Services and Outcomes Research Methodology. 2007 Abr 8;7(1–2):23–38.

Gertler PJ, Martinez S, Premand P, Rawlings LB, Vermeersch CMJ. Impact Evaluation in Practice [Internet]. The World Bank; 2010 [citado em 2016 Nov 2]. Disponível em: http://elibrary.worldbank.org/doi/book/10.1596/978-0-8213-8541-8

Gilberto A. Paula. Modelos de Regressão com apoio computacional [Internet]. Universidade de São Paulo; 2013. Disponível em: https://www.ime.usp.br/~giapaula/texto_2013.pdf

Gonzaga da Matta-Machado AT, de Fátima dos Santos A, Xavier de Abreu DM, Oliveira Jorge A, Rodrigues dos Reis CM, Dayrell de Lima AM de L, et al. Asistencia sanitaria, certificación de calidad y apoyo institucional: la atención primaria en Brasil. Salud Pública de México. 2016;358–65.

Gulis G, Fujino Y. Epidemiology, Population Health, and Health Impact Assessment. Journal of Epidemiology. 2015;25(3):179–80.

Handley MA, Lyles CR, McCulloch C, Cattamanchi A. Selecting and Improving Quasi-Experimental Designs in Effectiveness and

Implementation Research. Annual Review of Public Health. 2018 Abr;39(1):5–25.

Instituto de Pesquisa Econômica Aplicada. Avaliação de políticas públicas : guia prático de análise ex ante [Internet]. Brasília: Casa Civil da Presidência da República; 2018. Disponível em: http://www.ipea.gov.br/portal/images/stories/PDFs/livros/livros/180319_avaliacao_de_politicas_publicas.pdf

Instituto de Pesquisa Econômica Aplicada. Avaliação de políticas públicas: guia prático de análise ex post [Internet]. Brasília: Casa Civil da Presidência da República; 2018. Disponível em: http://www.ipea.gov.br/portal/images/stories/PDFs/livros/livros/181218_avaliacao_de_politicas_publicas_vol2_guia_expost.pdf

Instituto Jones dos Santos Neves. Guia para Avaliar Políticas Públicas [Internet]. Vitória, ES.: Instituto Jones dos Santos Neves; 2018. Disponível em: http://www.ijsn.es.gov.br/component/attachments/download/6418

Jacobs J, Jones E, Gabella B, Spring B, Brownson R. Tools for Implementing an Evidence-Based Approach in Public Health Practice. Preventing Chronic Disease [Internet]. 2012 Jun [citado em 2016 Nov 5]; Disponível em: http://www.cdc.gov/pcd/issues/2012/11_0324.htm

Kung-Yee Liang, Scott L Zeger. Longitudinal data analysis using generalized linear models. Biometrika. 1986 Abr;73(1):13–22.

Lee W-S. Propensity score matching and variations on the balancing test. Empir Econ. 2013 Fev 1;44(1):47–80.

Li L. Propensity Score Analysis with Matching Weights. Collection of Biostatistics Research Archive. 2011;18.

Lilia Belluzzo, Rafael Camelo. 1a Análise SEADE. Avaliação de Programas Públicos: Um Percurso na Fundação SEADE [Internet]. 2015 [citado

em 2016 Nov 2];23. Disponível em: http://web01.seade.gov.br/wp-content/uploads/2015/03/Primeira_Analise_n23.pdf

Macinko J. Evaluation of the impact of the Family Health Program on infant mortality in Brazil, 1990-2002. Journal of Epidemiology & Community Health. 2006 Jan 1;60(1):13–9.

Mazetto D. Assessing the impact the "Mais Médicos" program on basic health care indicators. [São Paulo]: Fundação Getúlio Vargas São Paulo School of Economics; 2018.

Ministério da Saúde. PORTARIA Nº 1.535, DE 16 DE JUNHO DE 2017 [Internet]. 1.535 Jun 16, 2017. Disponível em: http://bvsms.saude.gov.br/bvs/saudelegis/gm/2017/prt1535_20_06_2017.html

Ministério da Saúde - Secretaria Executiva. PDTI - Plano Diretor de Tecnologia da Informação 2016 [Internet]. 2016 [citado em 2016 Nov 2]. Disponível em: http://datasus.saude.gov.br/images/0305_PDTI.pdf

Morgan SL, Winship C. Counterfactuals and causal inference: methods and principles for social research. Second Edition. New York, NY: Cambridge University Press; 2015.

N Krieger, M Northridge, S Gruskin, M Quinn, D Kriebel, et al. Assessing health impact assessment: multidisciplinary and international perspectives. J Epidemiol Community Health. 2003;57:659–62.

Oliosi JGN, Reis-Santos B, Locatelli RL, Sales CMM, da Silva Filho WG, da Silva KC, et al. Effect of the Bolsa Familia Programme on the outcome of tuberculosis treatment: a prospective cohort study. The Lancet Global Health [Internet]. 2018 Dez [citado em 2019 Jan 2]; Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S2214109X18304789

Oliveira, A.E.F, Reis, R.S. Gestão pública em saúde: monitoramento e avaliação no planejamento do SUS [Internet]. São Luís:

Universidade Federal do Maranhão. UNA-SUS/UFMA.; 2016. Disponível em: https://ares.unasus.gov.br/acervo/handle/ARES/7408

de Oliveira Quites HF. Barreiras do uso da Informação em Saúde na tomada de decisão municipal: uma Revisão de Literatura. Gestão e Saúde. 2016;(supl.):pág–1011.

Patorno E, Grotta A, Bellocco R, Schneeweiss S. Propensity score methodology for confounding control in health care utilization databases. Epidemiology, Biostatistics and Public Health [Internet]. 2013 [citado em 2016 Nov 15];10(3). Disponível em: http://ebph.it/article/view/8940/0

Paul R. Rosenbaum, Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70:41–55.

Posner MA, Ash AS. Comparing Weighting Methods in Propensity Score Analysis [Internet]. 2009. Disponível em: http://www.stat.columbia.edu/~gelman/stuff_for_blog/posner.pdf

Ramos M. Aspectos Conceituais e Metodológicos da Avaliação de Políticas e Programas Sociais. Planejamento e Políticas Públicas [Internet]. 2009 Ago 18 [citado 2016 Nov 2];1(32). Disponível em: http://www.ipea.gov.br/ppp/index.php/PPP/article/view/11

Rasella D, Aquino R, Barreto ML. Impact of the Family Health Program on the quality of vital information and reduction of child unattended deaths in Brazil: an ecological longitudinal study. 2010;

Rasella D, Harhay MO, Pamponet ML, Aquino R, Barreto ML. Impact of primary health care on mortality from heart and cerebrovascular diseases in Brazil: a nationwide analysis of longitudinal data. BMJ. 2014 Jul 3;349(jul03 5):g4014–g4014.

Resende ACC, Oliveira AMHC de. Avaliando resultados de um programa de transferência de renda: o impacto do Bolsa-Escola sobre os gastos

das famílias brasileiras. Estudos Econômicos (São Paulo). 2008;38(2):235–65.

Robert H. Fletcher, Suzanne W. Fletcher, Edward H. Wagner. Epidemiologia Clínica: Elementos Essenciais. 3a ed. Artes Médicas; 1996.

Rocha R, Soares RR. Evaluating the impact of community-based health interventions: evidence from Brazil's Family Health Program. Health Economics. 2010 Set;19(S1):126–58.

Roncalli AG, Lima KC de, others. Impacto do Programa Saúde da Família sobre indicadores de saúde da criança em municípios de grande porte da região Nordeste do Brasil. Ciênc Saúde Coletiva. 2006;11(3):713–24.

Rosenbaum PR. Observation and experiment: an introduction to causal inference. Cambridge, Massachusetts: Harvard University Press; 2017.

Sasha O. Becker, Andrea Ichino. Estimation of average treatment effects based on propensity scores. The Stata Journal. 2002;2(4):358–77.

Schardosim Cotta de Souza MC. Escores de propensão: aplicações à Epidemiologia. Universidade Federal do Rio Grande do Sul; 2010.

Schor, A A LE. Apostila Avaliacao Economica Itau Social [Internet]. 2007 [citado em 2016 Nov 2]. Disponível em: http://www.redeitausocialdeavaliacao.org.br/wp-content/uploads/2015/01/Apostila_Avaliacao-Economica_06-08-07_final_20150128.pdf

Silveira M, Padilha JD, Schneider M, Amaral PST, Carmo TFM do, Netto GF, et al. Perspectiva da avaliação de impacto à saúde nos projetos de desenvolvimento no Brasil: importância estratégica para a sustentabilidade. Cad Saude Coletiva. 2012;20(1):57–63.

Slotterback CS, Forsyth A, Krizek KJ, Johnson A, Pennucci A. Testing three health impact assessment tools in planning: A process

evaluation. Environmental Impact Assessment Review. 2011 Mar;31(2):144–53.

Sousa AN. Monitoramento e avaliação na atenção básica no Brasil: a experiência recente e desafios para a sua consolidação. Saúde debate. 2018 Set;42(spe1):289–301.

Strezhnev A. Generalized Difference-in-Differences Estimands and Synthetic Controls [Internet]. Disponível em: https://www.antonstrezhnev.com/s/generalized_diff_in_diff.pdf

Stürmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. Journal of Clinical Epidemiology. 2006 Mai;59(5):437.e1-437.e24.

Victora CG, Habicht J-P, Bryce J. Evidence-based public health: moving beyond randomized trials. American journal of public health. 2004;94(3):400–405.

Viegas CV, Bond A, Danilevicz AM, Ribeiro JLD, Selig PM. Health Impact Assessment in Southern Brazilian EIAs: Too Far Away from Recommended Practices. 3º International Workshop Advances in Cleaner Production: cleaner production initiatives and challenges for a sustainable word. São Paulo [Internet]. 2011 [acessado 9 Nov 2016]. Disponível em: https://www.researchgate.net/profile/Angela_Danilevicz/publication/225028221_Health_Impact_Assessment_in_Southern_Brazilian_EIAs_too_far_away_from_recommended_practices/links/5428c0e10cf2e4ce940c534d.pdf

Villa JM. diff: Simplifying the estimation of difference-in-differences treatment effects. The Stata Journal. 2016 Mar;16(1):52–71.

Wang M. Generalized Estimating Equations in Longitudinal Data Analysis: A Review and Recent Developments. Advances in Statistics. 2014;2014:1–11.

West SG, Duan N, Pequegnat W, Gaist P, Des Jarlais DC, Holtgrave D, et al. Alternatives to the randomized controlled trial. American Journal of Public Health. 2008;98(8):1359–1366.

Wichmann RM, Carlan E, Barreto JOM. Consolidação da Rede para Políticas Informadas por Evidências – EVIPNet Brasil: relato da experiência nacional de construção de uma plataforma de tradução do conhecimento para o SUS. Boletim do Instituto de Saúde. 2016;17(1):15.

Wooldridge JM. Econometric analysis of cross section and panel data. 2nd ed. Cambridge, Mass: MIT Press; 2010.

Zeng F, An JJ, Scully R, Barrington C, Patel BV, Nichol MB. The Impact of Value-Based Benefit Design on Adherence to Diabetes Medications: A Propensity Score-Weighted Difference in Difference Evaluation. Value in Health. 2010 Set;13(6):846–52.

Zhang Z, Kim HJ, Lonjon G, Zhu Y. Balance diagnostics after propensity score matching. Ann Transl Med [Internet]. 2019 Jan [acessado em 13 Ago 2019];7(1). Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6351359/

Planejamento e avaliação de políticas públicas. Brasília: Ipea; 2015.

Portal do Departamento de Atenção Básica [Internet]. [acessado em 2 Nov 2016]. Disponível em: http://dab.saude.gov.br/portaldab/ape_pmaq.php

## 6. **ARTIGO 1**

# AVALIANDO O IMPACTO DE POLÍTICAS PÚBLICAS DE SAÚDE: MODELAGEM ESTATÍSTICA EM UM QUASE-EXPERIMENTO LONGITUDINAL

# EVALUATING THE IMPACT OF PUBLIC HEALTH POLICIES: STATISTICAL MODELING IN A LONGITUDINAL QUASI-EXPERIMENT

Juliana Feliciati Hoffmann
Suzi Alves Camey

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL (UFRGS)

**A ser enviado aos Cadernos de Saúde pública**

**Resumo**

**Introdução:** A avaliação de políticas públicas vem se consolidando como prática, auxiliando os gestores na tomada de decisão. Na área da saúde pública diversas iniciativas de avaliação vêm sendo realizadas, mas poucos são os trabalhos que detalham a metodologia de realização de uma avaliação de impacto quase-experimental. Esse trabalho tem o objetivo de apresentar rotinas de um *software* livre de análise estatística, comparando métodos distintos para estimar o impacto do tratamento. **Metodologia:** Foram estimados os escores de propensão e utilizadas as técnicas de pareamento e ponderação para contornar o problema de viés de seleção. Este trabalho sugere um modelo de estimativa de impacto que adapta o DID, permitindo a inclusão de todas as informações disponíveis ao longo do período, com a utilização de modelos de Equações de Estimação Generalizadas para estimar o efeito do tratamento. Para exemplificar a aplicação das metodologias foi considerada uma política pública de saúde mental do Rio Grande do Sul - os Núcleos de Apoio à Atenção Básica, considerando como desfecho as internações psiquiátricas. Para todas as análises foi utilizado o *software* R. **Resultados:** observou-se que o pareamento gerou exclusão de municípios do grupo tratamento e também do grupo controle, em maior número, considerando a proporção de 1:1, enquanto na ponderação não houve exclusão. Os resultados dos métodos foram divergentes, sendo que o modelo com pareamento resultou em um efeito maior quando comparado ao modelo ponderado, apesar de não ser significativo. **Discussão:** observa-se que é necessário ter cautela ao optar por um método ou outro de análise. Nas avaliações nas quais há uma amostra de tamanho grande e as exclusões não afetam o poder da análise, sugere-se a utilização da técnica de pareamento pois não modifica a importância dos casos artificialmente. Neste exemplo, considera-se a ponderação como método mais adequado, por evitar as exclusões que ocorrem no pareamento, as quais parecem forçar um resultado inexistente.

**Palavras-chave:** avaliação de impacto; avaliação de políticas públicas; quase-experimento.

## Introdução

A crescente consolidação da avaliação como prática no âmbito da gestão e da administração pública tem ocorrido no mundo pelo menos nas últimas cinco décadas. Na gestão pública brasileira, apesar de ser ainda muito incipiente, a cultura de avaliação de políticas públicas tem sido cada vez mais internalizada, o que tem estimulado a reunião de considerável arcabouço teórico, múltiplas reflexões metodológicas e destaque acadêmico (1).

Um dos mais recentes fenômenos produzidos por essas transformações é a ênfase dada a uma gestão voltada para resultados (1,2). A avaliação de políticas públicas deve ser vista como um recurso que fornece informações e, assim, auxilia os gestores em situações de tomada de decisão, inclusive sobre alocação dos recursos orçamentários. A avaliação também contribui para a modernização da gestão da administração e dos serviços públicos, proporcionando maior transparência às ações, podendo ser utilizada como uma forma de prestação de contas à sociedade sobre o desempenho dos programas do Governo (3).

A importância da avaliação no relacionamento entre os governos e as agências financiadoras também vem ganhando importância. Cada vez mais as agências incluem exigências formais de avaliação em contratos de financiamentos externos ou internos, atrelando a apuração periódica ou sistemática de seus resultados aos financiamentos concedidos (4,5).

No caso dos países em desenvolvimento, como o Brasil, a discussão sobre políticas públicas insere-se no contexto de recursos escassos e de grande demanda social, o que torna imprescindível comparar os resultados obtidos nos programas, optando por alternativas que maximizem o impacto das melhorias para a sociedade. Nesse sentido, a avaliação é percebida como uma oportunidade de aprimorar a dinâmica das políticas públicas, por meio de instrumentos capazes de oferecer informações mais

qualificadas sobre a coerência dos insumos, processos e resultados da ação pública.

A avaliação de impacto, foco deste trabalho, busca estabelecer e quantificar estatisticamente as relações causais entre um programa e um conjunto de resultados, verificando se os objetivos ou os impactos desejados estão sendo alcançados. Em termos metodológicos, esse tipo de avaliação demanda a adoção de estratégias que impliquem algum nível de controle sobre o contexto observado.

Para avaliar a mudança no problema a ser amenizado, ao longo do tempo, a avaliação compara a situação em dois momentos: antes e depois da implementação do programa. Este modelo também implica a conformação de dois grupos: um grupo que recebe o tratamento proposto pelo programa e outro que não recebe, denominado grupo controle. O grupo controle fornece um parâmetro de comparação (o contrafactual), representando a população-alvo sem ter sido objeto de tratamento (programa). Esse parâmetro de comparação possibilita estimar o impacto do programa (6).

Na maioria dos programas a seleção dos grupos não é uma tarefa simples. Uma questão fundamental envolvida na escolha dos componentes de cada grupo é a alocação aleatória do tratamento, que caracteriza os desenhos experimentais e é considerada o padrão-ouro na avaliação de impacto. Todos os elementos da população deveriam ter a mesma probabilidade de serem alocados ao grupo controle ou tratamento, o que na prática apresenta-se como mais um desafio (7,8). Como maneira de contorná-lo frequentemente utilizam-se os desenhos quase-experimentais (9).

Na área da saúde pública, um processo recente que vem sendo denominado *Saúde Pública Baseada em Evidências* busca aplicar os conceitos de medicina baseada em evidências no campo das políticas públicas de saúde (10–13). Com esse movimento, identifica-se que o tradicional ensaio clínico randomizado pode não ser o delineamento mais factível para avaliar o impacto das intervenções de saúde pública em grande

escala (10), em função da dificuldade de implementação do desenho na prática. Sendo assim, abordagens metodológicas compatíveis com delineamentos alternativos ao ensaio clínico randomizado são necessárias para viabilizar a avaliação dos resultados das políticas públicas de saúde (10,11).

Na mesma direção, também surgiu a iniciativa TREND (*Transparent Reporting of Evaluations with Non-randomized Designs*), que busca padronizar e dar mais transparência às publicações de pesquisas de avaliação de intervenções em saúde pública que utilizam desenhos não aleatórios e que possuem algum tipo de grupo de comparação (11).

A partir de iniciativa da Organização Mundial de Saúde (OMS), surgiu também no Brasil a Rede para Políticas Informadas por Evidências (EVIPNet), com o objetivo de promover o uso apropriado de evidências científicas na formulação e implementação de políticas, programas e serviços de saúde, mediante o intercâmbio entre gestores, pesquisadores e representantes da sociedade civil (14). Idealmente, evidências científicas deveriam ser sempre incorporadas na seleção e implementação de programas, desenvolvimento de políticas e processo de avaliação, de forma que sejam implantadas as intervenções que produzem maior retorno em saúde (13). Entretanto, apesar dos esforços empreendidos e dos benefícios comprovados associados às políticas públicas baseadas em evidências, muitas intervenções de saúde pública no Brasil ainda não seguem essa prática (15).

Apesar de haver muitos artigos de qualidade sobre avaliação de impacto em diversas áreas, existe uma falta de artigos acessíveis que mostrem como conduzir uma avaliação de impacto quase-experimental, detalhando passo-a-passo as etapas necessárias e a forma de realização das respectivas análises de cada etapa. Diante desse contexto e da importância do tema, esse trabalho tem como objetivo apresentar e comparar diferentes métodos de avaliação de impacto que sejam aplicáveis inclusive para políticas públicas que não tiveram um planejamento prévio de sua avaliação. Para isso, utilizando como exemplo a avaliação de uma política de saúde

mental, são apresentadas rotinas de um *software* livre de análise estatística, usando dados do DATASUS. Busca-se, assim, permitir que as análises possam ser replicadas por gestores de políticas de saúde, favorecendo e consolidando a cultura de avaliação em saúde e permitindo, assim, a tomada de decisão baseada em evidências.

## Metodologia

A fim de exemplificar a aplicação das metodologias que serão descritas a seguir, será considerada uma política pública de saúde mental oferecida por adesão a municípios do Rio Grande do Sul: os Núcleos de Apoio à Atenção Básica (NAABs). O Programa foi instituído em 2011 pela Secretaria Estadual de Saúde do Rio Grande do Sul como um dispositivo da Linha de Cuidado em Saúde Mental, Álcool e outras Drogas, com o objetivo de apoiar a inserção do cuidado em saúde mental na atenção básica de municípios com menos de 16 mil habitantes, os quais representam em torno de 76% do total de municípios do Estado. Em 2016, os NAABs estavam presentes em 121 municípios, o que representa 32,2% dos 376 municípios que teriam perfil para sua implementação. Para efeito de definição, será considerado tratado (ou em tratamento) o município que aderiu ao NAAB e controle o que não aderiu ao NAAB. Assim, o grupo tratado é composto de 121 municípios enquanto o grupo controle de 255 municípios.

Os NAABs trabalham junto às equipes de Atenção Básica de modo a compartilhar responsabilidades por ações de promoção e prevenção em saúde no território, tendo como atribuições: ações compartilhadas de promoção da saúde; discussão de casos e atendimento compartilhado entre equipes de atenção básica e NAAB para intervenção interdisciplinar, incluindo articulação com a rede de saúde, intersetorial e rede social; intervenções específicas com usuários e famílias; reunião de equipe e ações de educação permanente.

A legislação de criação do Programa e os demais registros disponibilizados não definem de forma clara um indicador de resultado que permita monitorar e avaliar o Programa. Diante disso, definiu-se como indicador de resultado para essa avaliação o total de internações psiquiátricas, por local de residência, em cada município.

Para essa avaliação, o ano de 2011 está sendo considerado como linha de base, pois antecede o início do tratamento, que ocorreu em 2012. Foram utilizados dados do período de 2008 a 2016, sendo 2008 a 2011 o período que antecede o Programa e 2012 a 2016 o período pós tratamento. Por ser por adesão e, portanto, não haver aleatorização dos municípios que recebem a política, trata-se de uma avaliação em delineamento quase-experimental.

Os dados necessários para as análises foram obtidos do DATASUS (16), e as informações de gestão dos NAABs foram obtidas junto ao coordenador do Programa na Secretaria Estadual de Saúde. A base de dados original, bem como os detalhes sobre sua transposição e manipulação podem ser acessados no material suplementar, bem como a programação no software livre R utilizada.

A seguir serão apresentadas abordagens que visam controlar alguns vieses introduzidos pelo delineamento quase-experimental.

## Avaliação em um quase-experimento

### Eliminar o viés de seleção – Escores de Propensão

Considerando que não foi realizada uma seleção aleatória na implantação da política, uma das alternativas é definir *a posteriori* um grupo controle mais parecido possível com o grupo que aderiu à política, a partir da seleção de características observáveis. Uma abordagem comumente utilizada é a dos escores de propensão, uma técnica que busca controlar o confundimento e o viés de seleção em estudos não aleatórios, uma vez que

nesses casos não há nenhum controle do pesquisador em relação à alocação do tratamento entre os grupos. A grande vantagem da estimação dos escores de propensão é permitir a utilização de um conjunto de covariáveis de ajuste simultaneamente (17).

O uso de escores para combinar diversas covariáveis em uma única existe desde 1976, sendo que em 1983 Rosenbaum e Rubin (18) criaram o conceito de escore de propensão, estimados na linha de base, para controlar vieses de seleção em estudos de coorte, tornando-se desde então uma técnica popular (18,19). O escore pode ser definido como sendo a probabilidade (propensão) de uma observação ser alocada ao grupo tratamento condicionada aos valores de seus preditores na linha de base (17,20,21). A estimação dos escores pode ser feita através de regressão logística, que é a forma mais usualmente utilizada (20).

Assim, considerando os valores dos escores de propensão é possível identificar nos dois grupos, tratamento e controle, aqueles municípios que são mais similares em relação às covariáveis da linha de base incluídas no modelo de regressão utilizado na estimação dos escores. Com isso, o ajustamento usando escores de propensão em média removerá os vieses subjacentes às distribuições das covariáveis (20).

Portanto, usando a probabilidade de um indivíduo ter sido tratado para ajustar a estimativa do efeito do tratamento teremos a situação similar à de um experimento (17). Quase-experimentos tentam reduzir a ambiguidade da explicação da diferença entre dois grupos em relação ao desfecho, buscando efeito de tratamento sem viés de seleção dos grupos (22).

**Tornar os grupos comparáveis – pareamento ou ponderação**

Uma vez estimados os escores de propensão, alguns métodos utilizados para estimar o efeito do tratamento sem viés de seleção incluem pareamento e ponderação (20,23). O pareamento é a técnica que busca selecionar um município controle para cada município tratado, baseando-se

nos valores mais próximos de escores de propensão e permitindo, assim, identificar os municípios com características mais similares entre os grupos.

Ao realizar o pareamento recomenda-se restringir a comparação aos municípios cujos escores de propensão estão na região de suporte comum, a fim de garantir que sempre haja, na distribuição do escores de propensão, observações suficientemente próximas para comparar as observações que receberam o tratamento e as que não receberam (29).

Na prática, com essa restrição são excluídos os casos que seriam considerados “nunca tratados” (apresentam valor de escore muito baixo) ou aqueles que seriam considerados “sempre tratados” (apresentam valor de escore muito alto), facilitando a identificação dos casos mais similares. Para cada probabilidade de participação estimada para os integrantes do grupo que recebeu tratamento haverá um correspondente no grupo controle com probabilidade semelhante, o que representa um suporte comum entre os dois grupos, tratamento e controle (24). Por fim, todos os elementos tratados e os respectivos controles pareados são mantidos em uma subamostra, enquanto os não pareados são descartados (25).

Além do suporte comum, a partir da hipótese de independência condicional baseada no escore de propensão assume-se que dado um conjunto de covariáveis que não são afetadas pelo tratamento, os desfechos são independentes da alocação ao grupo tratamento. Ou seja, uma vez que controlamos pelas covariáveis utilizadas na estimação dos escores de propensão, a alocação ao grupo tratamento é considerada aleatória (26).

Diferentes metodologias podem ser utilizadas para identificação dos municípios mais parecidos por técnicas de pareamento com uso dos escores de propensão, como por exemplo: pareamento pelo vizinho mais próximo (*Nearest Neighbor Matching*); pareamento pelo raio (*Radius Matching*); pareamento pelo método de Kernel (*Kernel Matching*); pareamento por estratificação (*Stratified Matching*) e pareamento pela métrica de Mahalanobis (*Mahalanobis Metric Matching*) (20,27). Neste trabalho optou-se pelo pareamento pelo vizinho mais próximo.

Outra forma de aplicação dos escores de propensão, que também tem o objetivo de eliminar o viés de seleção é a ponderação. A ponderação é um método mais recente, tendo sido proposto incialmente nos anos 2000 (28). O primeiro passo para utilização da ponderação é calcular os pesos ( $w$ ) de cada elemento da amostra para estimação do efeito do tratamento. Todas as observações do grupo tratamento recebem peso 1 enquanto as do grupo controle recebem peso igual ao inverso de 1-escore de propensão.

Assim, através da ponderação espera-se que os grupos tenham características similares, uma vez que o peso aplicado no grupo controle torna a distribuição mais similar ao grupo tratamento. Por fim, para estimativa do efeito do tratamento, deverá ser utilizado um modelo de regressão ponderado. Em relação ao pareamento, a ponderação tem a vantagem de não descartar unidades amostrais na análise (28), além de eliminar a subjetividade de escolha do método de pareamento, que pode variar conforme a preferência do responsável pelas análises (23).

Tanto após o pareamento quanto após a ponderação recomenda-se avaliar a qualidade alcançada, verificando se os grupos se tornaram de fato mais similares em relação às variáveis na linha base. Esta etapa é denominada balanceamento, que busca verificar se após as técnicas de pareamento ou ponderação os dois grupos resultantes são realmente similares.

**Modelos para estimação dos efeitos causais:**

Para medir o efeito de uma política pública, interessa saber a diferença no desfecho entre os grupos que receberam e os que não receberam o tratamento. Neste trabalho será utilizado o efeito médio do tratamento sobre os tratados (ATT do inglês *Average Treatment Effect on Treated),* com a hipótese do modelo contrafactual que afirma que o tratamento não teria efeito sobre os não-tratados, ou seja:

$$E[Y^0|tto = 1] = E[Y^0|tto = 0] \quad (1)$$

onde $Y$ representa o valor do desfecho (internações psiquiátricas),

$$Y = \begin{cases} Y^0 & tto = 0 \\ Y^1 & tto = 1 \end{cases},$$

$$tto = \begin{cases} 0, & \text{se grupo controle (não recebeu NAAB)} \\ 1, & \text{se grupo tratamento (recebeu NAAB)} \end{cases}$$

Assim ATT pode ser definido por (29)

$$ATT = E[Y^1 - Y^0 | tto = 1] = \mathrm{E}[Y^1 | \mathrm{tto} = 1] - \mathrm{E}[Y^0 | \mathrm{tto} = 1] \quad (2)$$

Considerando (1), temos que

$$ATT = \mathrm{E}[Y^1 | \mathrm{tto} = 1] - \mathrm{E}[Y^0 | \mathrm{tto} = 0] \quad (3)$$

A notação $Y = \begin{cases} Y^0 & tto = 0 \\ Y^1 & tto = 1 \end{cases}$ somente foi usada aqui para mostrar o uso da hipótese do modelo contrafactual e não será mais usada para evitar excesso de notação.

O método denominado *Diferenças-em-diferenças* (DID) é o mais utilizado para estimativa de efeito do tratamento, sendo bastante utilizado na área de economia da saúde (30). Ele compara duas diferenças no indicador de resultado: uma ao longo do tempo (antes e depois); e a outra entre os grupos tratamento e controle. Através da combinação dessas duas diferenças é possível obter uma medida de impacto do programa. Lembrando que (3) é válida independente do momento (antes ou depois da aplicação do tratamento), pois se a hipótese do modelo contrafactual é válida na presença da aplicação do tratamento, tem que ser válida antes da aplicação também.

Assim, o método *Diferenças-em-diferenças* pode ser expresso através do modelo descrito na equação (4) abaixo (32):

$$E(Y_{ij}) = \beta_0 + \beta_1 t_j + \beta_2 tto_i + \beta_3 t_j tto_i \quad (4)$$

Sendo:

$E(Y_{ij}) =$ esperança do número de internações psiquiátricas no município i no momento j

$i = 1, \ldots n \; e \; n = número \; de \; municípios;$

$j = 0{,}1;$

$$tto_i = \begin{cases} 0, & \text{se o município i é do grupo controle} \\ 1, & \text{se o município i é do grupo tratamento} \end{cases}$$

$$t_j = \begin{cases} 0, & \text{se j} = 0, \text{ou seja, antes do tratamento} \\ 1, & \text{se j=1, ou seja, após o tratamento} \end{cases}$$

A partir de (4) temos que o valor esperado da diferença no total de internações psiquiátricas no grupo tratado, entre o período antes e depois, é dado por:

$$E(Y_{i1} - Y_{i0}|tto = 1) = (\beta_0 + \beta_1 + \beta_2 + \beta_3) - (\beta_0 + \beta_2) = \beta_1 + \beta_3 \qquad (5)$$

e o valor esperado da diferença no total de internações psiquiátricas no grupo controle, entre o período antes e depois, é dado por:

$$E(Y_{i1} - Y_{i0}|tto = 0) = (\beta_0 + \beta_1) - (\beta_0) = \beta_1 \qquad (6)$$

Então, por (5) e (6) temos que o impacto da intervenção é dado por: $DID = (\beta_1 + \beta_3) - \beta_1 = \beta_3$

O método Diferenças-em-diferenças é uma abordagem que utiliza modelos lineares generalizados (GLM, do inglês G*eneralized Linear Models* e em português Modelos Lineares Generalizados*)* para estimar o efeito médio do tratamento em quase-experimentos. Conforme apresentado acima, o coeficiente $\beta_3$ será a diferença das diferenças ou impacto, ou seja, a estimativa do efeito médio do tratamento nos tratados (ATT) (8,32). Como pressuposto para a aplicação dessa metodologia (DID) e obtenção de estimativas de impacto confiáveis, deve-se confirmar a tendência paralela das variáveis de desfecho entre os dois grupos de comparação antes da intervenção. Ou seja, na ausência do tratamento, espera-se que os desfechos aumentem ou diminuam na mesma taxa em ambos os grupos. Como não é possível saber o que teria acontecido na ausência do programa, deve-se

assumir que não existe nenhuma diferença entre os grupos que varie com o tempo (7).

Esse conceito original do DID considera apenas dois momentos – antes e depois do Programa. Entretanto, quando há disponibilidade de várias medidas antes e várias medidas depois da intervenção, recomenda-se utilizar todas as medidas a fim de obter maior precisão nas estimativas. Nesse caso, sugere-se expandir a ideia do DID para um modelo que considera todas as informações disponíveis ao longo do tempo e a dependência entre elas, neste trabalho utilizou-se os modelos GEE (do inglês G*eneralized Estimating Equation* e em português Equações de Estimação Generalizadas).

Os modelos GEE permitem a análise de dados coletados em delineamentos longitudinais, aninhados, ou com medidas repetida, por permitirem a especificação de uma matriz de correlação entre as respostas repetidas dos sujeitos e diferentes distribuições de probabilidade (33). Com isso, obtêm-se estimativas de variabilidade distintas daquelas obtidas pelo GLM, em função da matriz de correlação.

O modelo especificado abaixo pode ser aplicado para considerar a dependência entre as observações de um mesmo município, nos anos para os quais há informação. A notação utilizada abaixo para o modelo GEE é análoga à do modelo DID, uma vez que são utilizadas as mesmas medidas.

$$E(Y_{ijk}) = \beta_0 + \beta_1 t_j + \beta_2 tto_i + \beta_3 ano_k + \beta_4 t_j tto_i + \beta_5 t_j ano_k + \beta_6 tto_i ano_k + \beta_7 tto_i ano_k t_j \quad (7)$$

Sendo:

$E(Y_{ijk}) = esperança\ do\ número\ de\ internações$

$psiquiátricas\ no\ município\ i\ no\ momento\ j\ e\ no\ ano\ k$

$i = 1, \ldots n$ e $n = número\ de\ municípios$;

$j = 0{,}1$;

$k = 1, \ldots 9$

$$t_j = \begin{cases} 0, & \text{se j} = 0, \text{ou seja, antes do tratamento} - 2008 \text{ a } 2011 \\ 1, & \text{se j=1, ou seja, após o tratamento} - 2012 \text{ a } 2016 \end{cases}$$

$$tto_i = \begin{cases} 0, & \text{se o município i pertence ao grupo controle} \\ 1, & \text{se o município i pertence ao grupo tratamento} \end{cases}$$

$$ano_k = 2008, 2009, \ldots, 2016$$

A partir de (7) temos que o valor esperado da diferença no grupo controle, entre o período antes e depois, do total de internações psiquiátricas será:

$$E(\mathrm{Y_{i1k}} - \mathrm{Y_{i0k}}|tto = 0) = (\beta_0 + \beta_1 + \beta_3 \mathrm{ano_k} + \beta_5 \mathrm{ano_k}) - (\beta_0 + \beta_3 \mathrm{ano_k}) = \beta_1 + \beta_5 \mathrm{ano_k} \quad (8)$$

e o valor esperado da diferença no grupo tratamento, entre o período antes e depois, do total de internações psiquiátricas será:

$$E(Y_{i1k} - Y_{i0k}|tto = 1) = (\beta_0 + \beta_1 + \beta_2 + \beta_3 ano_k + \beta_4 + \beta_5 ano_k + \beta_6 ano_k + \beta_7 ano_k) - (\beta_0 + \beta_2 + \beta_3 ano_k + \beta_6 ano_k) = \beta_1 + \beta_4 + ano_k(\beta_5 + \beta_7) \quad (9)$$

Então, $DID = \beta_1 + \beta_4 + ano_k(\beta_5 + \beta_7)) - (\beta_1 + \beta_5 ano_k) = \beta_4 + \beta_7 ano_k$

Logo, o impacto será medido por $\beta_4 + \beta_7 ano_k$ e dependerá, portanto, do ano *k*.

Os modelos GEE nesse contexto estão sendo apresentados para um quase-experimento, mas poderiam ser aplicados da mesma forma em estudos experimentais. Os mesmos cuidados de qualidade de ajuste de modelos devem ser tomados quando tratamos de avaliação de políticas públicas, embora esse ponto não seja abordado aqui. Maiores detalhes sobre técnicas de diagnóstico de modelos de regressão podem ser obtidos em Paula, 2003 (34).

Quando forem consideradas mais de uma observação de cada município é necessário utilizar a base de dados longa (também chamada de dados em painel). Neste formato os municípios são replicados nas linhas,

tantas vezes quantas forem o número de anos nos quais o município foi repetidamente avaliado.

Podemos definir, a partir disso, que:

$m_i = 9 = número\ de\ observações\ do\ município\ i$;

$m = total\ de\ observações = \sum_{i=1}^{n} m_i$ e

$n = número\ de\ municípios$;

Com isso, na base de dados longa para o modelo GEE, cada um dos municípios terá nove observações, referentes aos anos de 2008 a 2016. Assim, o grupo tratado que é composto de 121 municípios fica com um total de 1089 observações e o grupo controle, composto de 255 municípios, fica com 2295 observações. Salienta-se que os modelos GEE permitem observações incompletas, ou seja, os municípios poderiam ter um número diferente de observações em razão de perda de informação em algum ano (37).

**Códigos R para estimação dos efeitos causais**

Para todas as análises foi utilizado o *software* R, por ser livre e, portanto, facilmente replicável. O quadro 1 apresenta a descrição das variáveis utilizadas para o exemplo tratado neste trabalho, enquanto o quadro 2 mostra os códigos utilizados para a obtenção dos escores de propensão, para o pareamento e para o cálculo dos pesos utilizados na ponderação. As bases de dados estão sendo denominadas “base_curta” e “base_longa”.

Os escores de propensão podem ser estimados por regressão logística, conforme apresentado no quadro, a partir de um GLM com especificação de função de ligação binomial. Neste caso o desfecho que está sendo considerado é o grupo (1=tratamento e 0=controle), uma vez que se pretende obter a probabilidade de alocação ao grupo tratamento.

Em relação ao pareamento, foi utilizado o pacote `Matchit` do R utilizando o método do vizinho mais próximo. O comando `summary` (objeto do `Matchit`) apresenta os resultados do balanceamento para todo o conjunto de dados e também para a subamostra de dados pareados (médias e desvios de cada grupo), além do percentual de melhoria no balanceamento obtido com o pareamento, considerando a diferença média entre os grupos para cada uma das variáveis do modelo ajustado, bem como para a distância. O percentual de melhoria pode ser um dos critérios utilizados para definição do melhor método de pareamento a ser utilizado. A saída também fornece os tamanhos de amostra de cada grupo: inicial, pareado, não pareado e descartado. Outro pacote do R é o `Cobalt`, que indica de forma clara e simples o balanceamento das covariáveis. Nesse pacote é possível estabelecer a distância máxima entre as médias padronizadas, que devem ser próximas de zero, sendo o ideal menor que 0,10 (38).

O *software* R também permite grande flexibilidade na elaboração de gráficos, auxiliando a apresentação e compreensão dos resultados, desde a etapa da estimação dos escores de propensão até as estimativas de efeitos causais. O gráfico do tipo *jitter* permite identificar as observações clicando sobre elas - apresenta em um mesmo gráfico a distribuição dos valores dos escores de propensão estimados na amostra, divididos em quatro categorias: elementos do grupo tratamento não pareados; elementos do grupo tratamento pareados; elementos do grupo controle pareados; elementos do grupo controle não pareados. A opção de histograma, por sua vez, apresenta quatro gráficos, estando no eixo horizontal os valores dos escores de propensão e no vertical a frequência. Cada um dos histogramas mostra: elementos do grupo tratamento não pareados; elementos do grupo tratamento pareados; elementos do grupo controle não pareados; elementos do grupo controle pareados.

Para a ponderação é apresentada a forma de cálculo dos pesos, considerando a estimativa do ATT. Os pesos devem ser salvos como novas variáveis na base de dados para que sejam utilizados no passo seguinte de estimação dos modelos por ponderação. Assim como recomendado no

pareamento, após a ponderação também deve-se verificar o balanceamento dos grupos. Com esse propósito pode ser utilizado o pacote `twang` do R, que fornece opções específicas para essa verificação, mostrando os valores de médias e desvio padrão originais e ponderados, distância entre as médias padronizadas e valor-p.

Quadro 1– Descrição das variáveis utilizadas

| **Variável** | **Descrição** |
|---|---|
| NAAB | 0=controle (sem NAAB);<br>1=tratamento (com NAAB) |
| CODUFMUN | código do município |
| ano | 2008 a 2016 |
| peso.ATT | pesos construídos para estimar o ATT |
| y1_Int_psiq_total | internações psiquiátricas totais |
| tempo | 0=antes; 1=depois |
| Expect_vida_nasc_2010 | expectativa de vida ao nascer |
| Bloco_Saude_cond_gerais_2011 | componente de condições gerais de saúde do bloco de saúde do IDESE |
| perc_Saude_2011 | % orçamento aplicado em saúde |
| tx_mort_inf_2011 | taxa de mortalidade infantil |
| perc_VINC_SUS_total | % de estabelecimentos SUS em relação ao total de estabelecimentos |
| escore_pr13 | escores de propensão |

Quadro 2 – Códigos R para obtenção dos Escores de propensão, pareamento e ponderação

| Etapa | Código R |
|---|---|
| **1. Escores de Propensão** | |
| 1.1 Estimar os escores de propensão através de regressão logística | `ps_mf13<- glm(NAAB ~Expect_vida_nasc_2010 + Bloco_Saude_cond_gerais_2011 + perc_Saude_2011 + tx_mort_inf_2011 + perc_VINC_SUS_total, family = binomial(link = "logit"), data = Base_curta)` |
| 1.2 Salvar os valores dos escores de propensão no *dataframe* | `Base_curta$escore_pr13 <- predict(ps_mf13, type = "response")` |
| 1.3 Verificar área de suporte comum pelo *boxplot* | `boxplot(Base_curta$escore_pr~Base_curta$NAAB,data=Base_curta, main="Verificar suporte comum", xlab="NAAB", ylab="Escore de Propensão")` |
| **2. Pareamento** | |
| 2.1 Vizinho mais próximo (*nearest*)* | `match.it_13.2 <- matchit(NAAB ~ Expect_vida_nasc_2010 + Bloco_Saude_cond_gerais_2011 + perc_Saude_2011 + tx_mort_inf_2011 + perc_VINC_SUS_total, data = Base_curta, method="nearest", ratio=1,discard="both", caliper = 0.1)` |
| 2.2 Avaliar a qualidade do pareamento – balanceamento** | `bal.tab(match.it_13.2,m.threshold=0.1)` |
| 2.3 Salvar a amostra pareada | `Base_curta_nearest <- match.data (match.it_13.2, group="all")` |
| 2.4 Histograma dos Escores de Propensão | `plot(match.it_13.2, type = "hist")` |
| 2.5 *Jitter* - Distribuição dos Escores de Propensão por resultado do pareamento | `plot(match.it_13.2, type = "jitter")` |
| **3. Ponderação** | |
| 3.1 Cálculo dos pesos para estimar o ATT | `Base_curta$peso.ATT<- ifelse(Base_curta$NAAB == 0, (Base_curta$escore_pr13)/(1 - Base_curta$escore_pr13), 1)` |
| 3.2 Avaliar a qualidade da ponderação – balanceamento*** | `bal.pslog1 <- dx.wts(x = Base_curta$peso.ATT, data= Base_curta, vars=c("Expect_vida_nasc_2010" , "Bloco_Saude_cond_gerais_2011" , "perc_Saude_2011" , "tx_mort_inf_2011" , "perc_VINC_SUS_total"), treat.var="NAAB", estimand = "ATT")`<br>`bal.table(bal.pslog1)` |

*usando pacote `matchit`; **usando pacote `cobalt` ***usando pacote `twang`;

O quadro 3, por sua vez, mostra os códigos utilizados para obtenção das estimativas do efeito do programa (ATT). Para estimação dos modelos que consideram a dependência entre os municípios, considerando a base de dados original ou pareada, aplica-se a função `geeglm`, disponível no pacote `geepack`, acrescentando "`id=CODUFMUN`". A função `geeglm` segue basicamente a mesma sintaxe da função `glm` e muitos dos métodos disponíveis para objetos de `glm` estão também disponíveis para objetivos do `geeglm` (39). Para estimativas ponderadas também pode ser utilizado o comando `geeglm` acrescentando a opção `weights`. Resultados obtidos pelo método DID foram estimados para fins de comparação e são apresentados no material suplementar.

Quadro 3. Códigos R utilizados para a estimação dos efeitos causais* através dos modelos GEE.

| **Modelos** | **Código R do modelo**** |
|---|---|
| GEE | `fit_Y1_GEE = geeglm(y1_Int_psiq_total ~ NAAB + ano + tempo + tempo:ano + NAAB:tempo + NAAB:ano + ano:NAAB:tempo, id=CODUFMUN, data = Base_longa, family = gaussian(link = "identity"), corstr="exchangeable")` |
| GEE com pareamento | `fit_Y1_GEE_par = geeglm(y1_Int_psiq_total ~ NAAB + ano + tempo + tempo:ano + NAAB:tempo + NAAB:ano + ano:NAAB:tempo, id=CODUFMUN, data = Base_longa_nearest, family = gaussian(link = "identity"), corstr="exchangeable")` |
| GEE com ponderação | `fit_Y1_GEE_pond <- geeglm(y1_Int_psiq_total ~ NAAB + ano + tempo + tempo:ano + NAAB:tempo + NAAB:ano + ano:NAAB:tempo, id=CODUFMUN, data = Base_longa, family = gaussian(link = "identity"), corstr="exchangeable", weights = peso.ATT)` |

*medida de efeito: ATT; ** usando pacote geepack.

## Resultados

A figura 1 apresenta um comparativo da estrutura amostral resultante de cada um dos métodos utilizados. No primeiro histograma observa-se que com o pareamento há exclusão de municípios do grupo tratamento com maiores valores de propensão, por não haver similares no grupo controle, e também de municípios do grupo controle, em maior número, considerando a proporção de 1:1 e a obtenção de maior similaridade entre os grupos. No caso da ponderação não ocorre exclusão de casos. Por outro lado, ganham menor peso as observações do grupo controle que apresentam valores muito baixos de escores de propensão e maior peso observações com escores mais altos, enquanto não há alteração no grupo tratamento, por este receber peso de valor um.

Os gráficos da figura 2 mostram as linhas de tendência da média de internações psiquiátricas, de 2008 a 2016, conforme a base de dados de cada método utilizado, separadas por grupo (tratamento e controle). O gráfico (a) apresenta a amostra inicial, sem qualquer método que busque tornar os grupos comparáveis. No gráfico (b), obtido após o pareamento, observa-se que as médias dos grupos se aproximam, com alteração nos valores das médias (redução em todos os anos) em ambos os grupos em função da exclusão de casos, e que as retas se tornam paralelas. Por último, em (c), apresenta-se o gráfico obtido após a ponderação, que também torna as médias dos grupos mais próximas, mantendo os valores originais no grupo tratamento, uma vez que a ponderação neste grupo é igual a um.

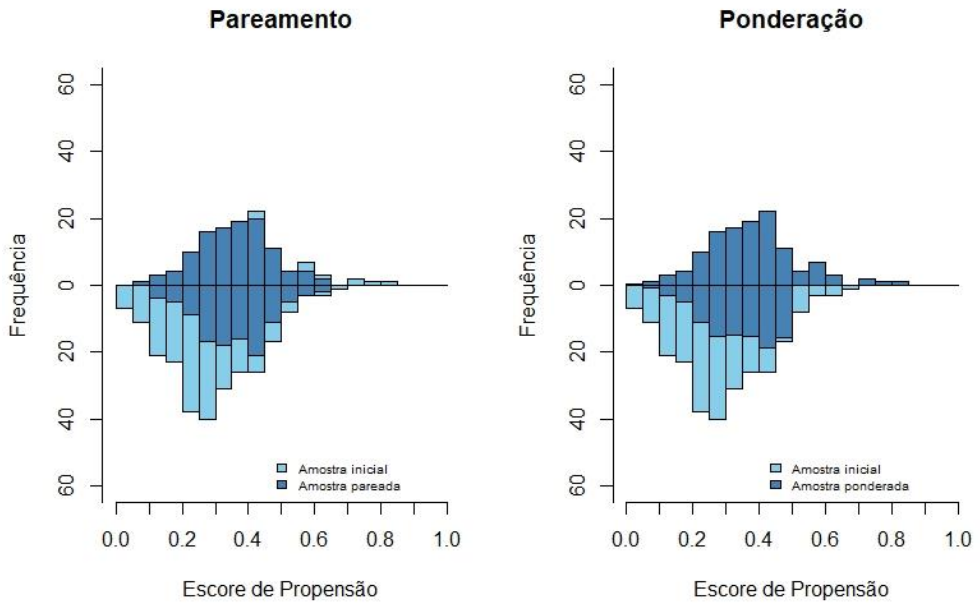



Figura 1: Histograma espelhado mostrando os escores de propensão da amostra inicial e das amostras pareadas (a esquerda) e ponderada (a direita). Acima da linha horizontal do zero: histograma dos municípios com tratamento (com NAAB); abaixo da linha horizontal do zero histograma dos municípios controle (sem NAAB)

Com o pareamento, do total de121 municípios do grupo tratamento 111 foram pareados e 10 excluídos. Como optou-se por grupos de tamanhos iguais (razão 1:1), o grupo controle ficou com o mesmo tamanho do tratamento e, portanto, 144 municípios foram excluídos. A maioria dos municípios do grupo tratamento que não foram pareados são aqueles com maiores valores de escore de propensão.

A tabela 1 apresenta os resultados obtidos quando se utiliza a base de dados considerando todos os anos disponíveis para a análise (2008 a 2016). O modelo tem três coeficientes que permitem uma estimativa que considere as variações ano a ano, conforme descrito na metodologia. A partir desse modelo, a estimativa de impacto do programa será representada pela soma: $\beta_4 + \beta_7 \times ano_k$, sendo as estimativas de $\beta_4$ e $\beta_7$ apresentados na tabela 1.

A tabela 2, por sua vez, utiliza os valores das duas últimas linhas da tabela 1, modelos com pareamento e com ponderação, para estimar o impacto do programa ao longo dos anos, comparando o método de pareamento e o de ponderação, ambos para estimativa do efeito ano a ano. Essa tabela serve apenas como forma de exemplificação, uma vez que os coeficientes não foram significativos na tabela 1. $\beta_4$ é o efeito médio do tratamento após intervenção e o $\beta_7$ é a mudança no efeito médio a cada ano. Sendo assim, caso o resultado fosse significativo, seria possível afirmar que o aumento médio esperado em 2012 é de 0,65 internações nos municípios que aderiram ao NAAB, considerando o modelo GEE com ponderação. O aumento esperado das internações ocorre com menor intensidade a cada ano, sendo que em 2016 é de apenas 0,54 internações. Por outro lado, o resultado do modelo GEE com pareamento apresenta redução no total de internações a partir de 2013, seguindo essa tendência até 2016.

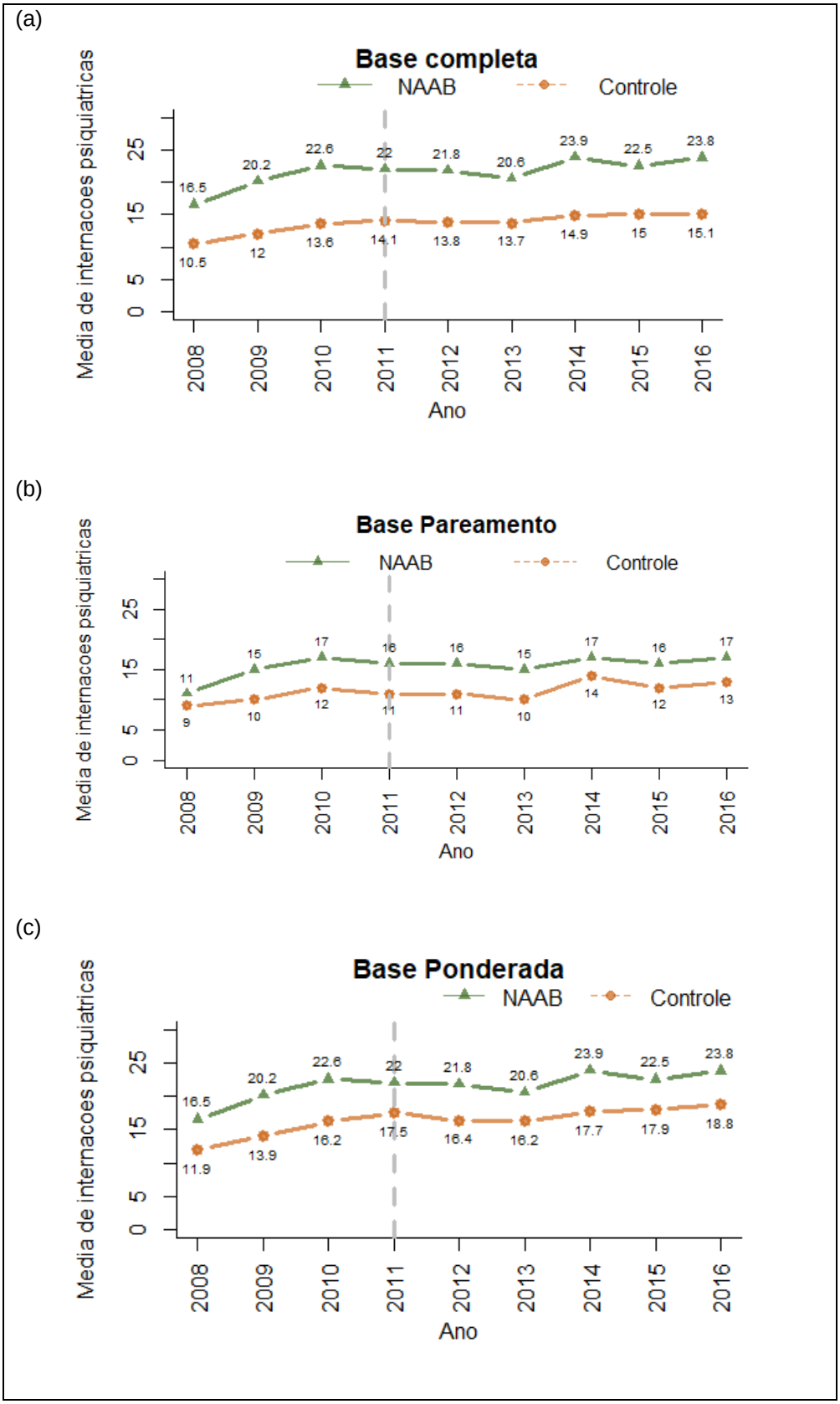



Figura 2. Tendências da média de internações psiquiátricas, por grupo, conforme a base de dados utilizada em cada um dos métodos de análise – RS, 2008 a 2016.

Tabela 1. Coeficientes dos modelos GEE para estimar efeito do tratamento* incluindo um parâmetro para ano

| | m | $\widehat{\beta}_4$ | EP | p | $\widehat{\beta}_7$ | EP | p |
|---|---|---|---|---|---|---|---|
| GEE | 3384 | 873,56 | 2184,62 | 0,689 | -0,44 | 1,09 | 0,689 |
| GEE com pareamento | 1998 | 1126,95 | 2638,86 | 0,669 | -0,56 | 1,31 | 0,669 |
| GEE com ponderação | 3384 | 59,00 | 2550,00 | 0,982 | -0,03 | 1,27 | 0,982 |

*medida de efeito: ATT; m: total de observações; EP: erro padrão

Tabela 2. Efeito do tratamento* para cada ano após o início do tratamento, para pareamento e ponderação $(\beta_4 + \beta_7 \times ano_k)$

| Ano | GEE com pareamento | GEE com ponderação |
|---|---|---|
| 2012 | 0,23 | 0,65 |
| 2013 | -0,33 | 0,62 |
| 2014 | -0,89 | 0,59 |
| 2015 | -1,45 | 0,56 |
| 2016 | -2,01 | 0,54 |

*medida de efeito: ATT

## Discussão

Diante das metodologias apresentadas e respectivos resultados obtidos, observa-se que é necessário ter cautela ao optar por um método ou outro de análise. No exemplo aqui apresentado o pareamento provocou redução na amostra, buscando deixar a distribuição dos dois grupos parecidas. Olhando para o número de internações psiquiátricas em 2011, dos municípios pareados e não pareados, observa-se que no grupo que recebeu o programa a média do total de internações dos dez casos excluídos é superior à média dos pareados, sendo 30,1 (DP=39,9) versus 21,3 (DP=13,9), respectivamente. Por outro lado, entre os municípios que não receberam o tratamento, a média do total de internações psiquiátricas é maior nos que se mantiveram na análise, sendo 17,2 (DP=18,7) versus 11,7 (DP=13,9). Ou

seja, a exclusão reduziu o valor médio inicial do grupo tratamento e aumentou a média inicial no grupo controle.

Em relação à ponderação, houve alteração apenas no grupo controle, uma vez que todas as observações do tratamento recebem peso igual a um. No grupo controle quanto menor o escore de propensão menor a ponderação, sempre com o objetivo de tornar a distribuição desses escores mais similar à do grupo tratamento. Visualmente, com base na figura 1, identifica-se pouca diferença em relação ao pareamento, com exceção dos valores extremos à direita da distribuição dos escores de propensão do tratamento, que foram excluídos no pareamento.

Os resultados dos dois métodos foram divergentes, sendo que o modelo GEE com pareamento resultou em um efeito maior quando comparado ao modelo ponderado. A ponderação parece diluir o efeito do tratamento. Em função disso, é necessário sempre fazer uma análise do que está sendo excluído.

Nas avaliações nas quais dispõe-se de uma amostra de tamanho grande e as exclusões não afetarem o poder da análise, sugere-se a utilização da técnica de pareamento pois, assim, a importância dos casos não é modificada artificialmente como acontece na ponderação. Para isso o ideal é realizar um cálculo de tamanho amostra para verificação do poder. No caso utilizado como exemplo neste trabalho, no entanto, não seria viável aumentar o número de municípios no grupo tratamento, uma vez que a participação no programa ocorre por adesão.

Neste caso, considera-se que a ponderação seria o método mais adequado, por evitar as exclusões que ocorrem no pareamento, as quais parecem forçar um resultado que não existe. Uma alternativa para aumentar o tamanho de amostra neste exemplo seria aumentar o número de anos observados após a implantação dos NAABs.

Os resultados obtidos através das duas técnicas tendem a ser muito próximos quando não há muitos outliers. Ressalta-se, assim, a importância da qualidade da coleta de dados, incluindo a checagem dos valores

extremos, a fim de identificar se são de fato valores reais e devem ser mantidos, ou se são erros de aferição e podem ser corrigidos.

Este trabalho sugere um modelo de estimativa de impacto que adapta o DID, permitindo a inclusão de todas as informações disponíveis ao longo do período, e não apenas valores médios do que foi medido antes e depois. Assim, é possível agregar informação sempre que ela estiver disponível, e não apenas descartá-la das análises ou agregá-las (na forma de média, total ou outra medida resumo), reduzindo a qualidade da informação. Para isso, foi proposta a utilização de modelos GEE com dados em painel, uma vez que qualifica a análise por a dependência de todas as observações dos municípios, ano a ano (33). Neste trabalho foi utilizado um modelo supondo relação linear e distribuição normal, mas o modelo GEE permite utilizar outras distribuições e outras funções de ligação. Outra opção seria a utilização de modelos mistos.

Todas as etapas anteriores às estimativas dos efeitos, incluindo o ajuste do modelo de regressão logística para obtenção dos valores dos escores de propensão, o pareamento ou a ponderação e o balanceamento obtido com cada método, poderiam não ter sido necessários caso os dados estivessem no contexto de um ensaio clínico randomizado. Isso ocorreria se a avaliação do programa tivesse sido planejada anteriormente, no início de sua implantação. Nesse caso, através de uma seleção aleatória dos municípios (com exceção das situações em que não é eticamente viável a seleção aleatória), teríamos uma avaliação experimental, facilitando muito as análises para obtenção do efeito do programa. Isso reforça a importância da cultura de avaliação entre os gestores e planejadores das políticas de saúde antes de sua implementação, criando as condições necessárias para mapear, de forma precisa, a situação inicial que deveria ser alterada pela política, contrastando-a com a situação final, nos dois grupos de comparação (5), gerando um ganho tanto na implementação, quanto na qualidade e facilidade da avaliação de impacto.

No mesmo sentido de disseminar a importância da avaliação de políticas públicas, esse trabalho apresenta todos os códigos e pacotes necessários para a realização das análises em um software livre utilizado mundialmente. Isso permite a replicação por todos que tiverem interesse, tornando a informação mais acessível e de fácil utilização.

Embora muitas iniciativas e experiências estejam em curso, processos sistêmicos e periódicos de avaliação e monitoramento do SUS encontram-se pouco desenvolvidos. O uso de dados e indicadores tornou-se frequente entre os gestores do Sistema. No entanto, as informações geradas pouco orientam a tomada das decisões, assim como pouco auxiliam na qualificação dos serviços e ações em saúde (40). A inserção da avaliação na rotina dos serviços somente se dará por meio da implantação de uma cultura avaliativa, a ser alcançada com a produção de informações estratégicas para a gestão, como resultados de avaliações bem estruturadas, periódicas e contínuas, destacando a perspectiva útil da avaliação (40,41).

Com as informações e o passo-a-passo descritos neste trabalho, espera-se mostrar aos gestores de políticas de saúde a viabilidade da avaliação de impacto quase-experimental, com a utilização de dados públicos do DATASUS e de um software livre de fácil acesso a todos. Por fim, busca-se contribuir para a disseminação de uma gestão de política públicas de saúde baseada em evidências (14).

## Referências


1. Facchini LA, Piccini RX, Tomasi E, Thumé E, Teixeira VA, Silveira DS da, et al. Avaliação de efetividade da Atenção Básica à Saúde em municípios das regiões Sul e Nordeste do Brasil: contribuições metodológicas. Cad Saúde Pública. 2008;24(Sup 1):S159–72.

2. Cavalcanti MM de A. Avaliação de Políticas Públicas e Programas Governamentais - Uma abordagem Conceitual. Interfaces de Saberes [Internet]. 2006 [citado 2 de novembro de 2016];6(1). Disponível em: https://interfacesdesaberes.fafica-pe.edu.br/index.php/import1/article/view/20

3. Carneiro F. Avaliação de Políticas Públicas: por um procedimento integrado ao ciclo da gestão. Revista Perspectivas em Políticas Públicas [Internet]. 2013 [citado 2 de novembro de 2016];6(11). Disponível em: http://www.uemg.br/openjournal/index.php/revistappp/article/view/893

4. Lilia Belluzzo, Rafael Camelo. 1a Análise SEADE. Avaliação de Programas Públicos: Um Percurso na Fundação SEADE [Internet]. 2015 [citado 2 de novembro de 2016];23. Disponível em: http://web01.seade.gov.br/wp-content/uploads/2015/03/Primeira_Analise_n23.pdf

5. Planejamento e avaliação de políticas públicas. Brasília: Ipea; 2015. 473 p. (Pensamento estratégico, planejamento governamental & desenvolvimento no Brasil Contemporâneo).

6. Ramos M. Aspectos Conceituais e Metodológicos da Avaliação de Políticas e Programas Sociais. Planejamento e Políticas Públicas [Internet]. 18 de agosto de 2009 [citado 2 de novembro de 2016];1(32). Disponível em: http://www.ipea.gov.br/ppp/index.php/PPP/article/view/11

7. Paul J. Gertler, Patrick Premand, Christel M. J. Vermeersch, Laura B. Rawlings, Sebastián Martínez. Avaliação de Impacto na Prática [Internet]. 2ª Edição. Banco Internacional para Reconstrução e Desenvolvimento/Banco Mundial; 2018. 375 p. Disponível em: https://openknowledge.worldbank.org/bitstream/handle/10986/25030/9781464808890.pdf

8. Strezhnev A. Generalized Difference-in-Differences Estimands and Synthetic Controls [Internet]. Disponível em: https://www.antonstrezhnev.com/s/generalized_diff_in_diff.pdf

9. Fernandes FMB, Ribeiro JM, Moreira MR. Reflexões sobre avaliação de políticas. Cad Saude Publica. 2011;27(9):1667–1677.

10. Victora CG, Habicht J-P, Bryce J. Evidence-based public health: moving beyond randomized trials. American journal of public health. 2004;94(3):400–405.

11. Des Jarlais DC, Lyles C, Crepaz N. Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: the TREND statement. American journal of public health. 2004;94(3):361–366.

12. Eriksson C. Learning and knowledge-production for public health: a review of approaches to evidence-based public health. Scandinavian Journal of Public Health. outubro de 2000;28(4):298–308.

13. Brownson RC, Baker EA, Leet TL, Gillespie KN, True WR. Evidence-based public health [Internet]. Oxford University Press; 2010 [citado 5 de novembro de 2016]. Disponível em: https://books.google.com/books?hl=en&lr=&id=9fxzvhVoD2cC&oi=fnd&pg=PR17&dq=%22When+we+hear+the+word+evidence,+most+of+us+conjure+up+the%22+%22with+executive+and+managerial+responsibilities+and+their+many%22+%22in+improving+health+in+the+populations+we+are+serving%3F+This%22+&ots=kPGgwBwlZm&sig=S5BT1mxijAFqFw_7MfXZXXLrwFs

14. Wichmann RM, Carlan E, Barreto JOM. Consolidação da Rede para Políticas Informadas por Evidências – EVIPNet Brasil: relato da experiência nacional de construção de uma plataforma de tradução do conhecimento para o SUS. Boletim do Instituto de Saúde. 2016;17(1):15.

15. Jacobs J, Jones E, Gabella B, Spring B, Brownson R. Tools for Implementing an Evidence-Based Approach in Public Health Practice. Preventing Chronic Disease [Internet]. junho de 2012 [citado 5 de novembro de 2016]; Disponível em: http://www.cdc.gov/pcd/issues/2012/11_0324.htm

16. DATASUS [Internet]. Disponível em: http://datasus.saude.gov.br/

17. D'Agostino RB. Tutorial in biostatistics: propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat Med. 1998;17(19):2265–2281.

18. Paul R. Rosenbaum, Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70:41–55.

19. Stürmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. Journal of Clinical Epidemiology. maio de 2006;59(5):437.e1-437.e24.

20. Schardosim Cotta de Souza MC. Escores de propensão: aplicações à Epidemiologia. Universidade Federal do Rio Grande do Sul; 2010.

21. West SG, Duan N, Pequegnat W, Gaist P, Des Jarlais DC, Holtgrave D, et al. Alternatives to the randomized controlled trial. American Journal of Public Health. 2008;98(8):1359–1366.

22. Rosenbaum PR. Observation and experiment: an introduction to causal inference. Cambridge, Massachusetts: Harvard University Press; 2017. 374 p.

23. Li L. Propensity Score Analysis with Matching Weights. Collection of Biostatistics Research Archive. 2011;18.

24. Resende ACC, Oliveira AMHC de. Avaliando resultados de um programa de transferência de renda: o impacto do Bolsa-Escola sobre os gastos das famílias brasileiras. Estudos Econômicos (São Paulo). 2008;38(2):235–65.

25. Morgan SL, Winship C. Counterfactuals and causal inference: methods and principles for social research. Second Edition. New York, NY: Cambridge University Press; 2015. 499 p. (Analytical methods for social research).

26. Caliendo M, Kopeinig S. Some Practical Guidance for the Implementation of Propensity Score Matching. Journal of Economic Surveys. fevereiro de 2008;22(1):31–72.

27. Sasha O. Becker, Andrea Ichino. Estimation of average treatment effects based on propensity scores. The Stata Journal. 2002;2(4):358–77.

28. Posner MA, Ash AS. Comparing Weighting Methods in Propensity Score Analysis [Internet]. Disponível em: http://www.stat.columbia.edu/~gelman/stuff_for_blog/posner.pdf

29. Wooldridge JM. Econometric analysis of cross section and panel data. 2nd ed. Cambridge, Mass: MIT Press; 2010. 1064 p.

30. Fu AZ, Dow WH, Liu GG. Propensity score and difference-in-difference methods: a study of second-generation antidepressant use in patients with bipolar disorder. Health Services and Outcomes Research Methodology. 8 de abril de 2007;7(1–2):23–38.

31. Villa JM. diff: Simplifying the estimation of difference-in-differences treatment effects. The Stata Journal. março de 2016;16(1):52–71.

32. Athey S, Imbens GW. Identification and Inference in Nonlinear Difference-in-Differences Models. Econometrica. março de 2016;74(2):431–497.

33. Ballinger GA. Using Generalized Estimating Equations for Longitudinal Data Analysis. Organizational Research Methods. abril de 2004;7(2):127–50.

34. Gilberto A. Paula. Modelos de Regressão com apoio computacional [Internet]. Universidade de São Paulo; 2013. Disponível em: https://www.ime.usp.br/~giapaula/texto_2013.pdf

35. Wang M. Generalized Estimating Equations in Longitudinal Data Analysis: A Review and Recent Developments. Advances in Statistics. 2014;2014:1–11.

36. Kung-Yee Liang, Scott L Zeger. Longitudinal data analysis using generalized linear models. Biometrika. abril de 1986;73(1):13–22.

37. Salazar A, Ojeda B, Dueñas M, Fernández F, Failde I. Simple generalized estimating equations (GEEs) and weighted generalized estimating equations (WGEEs) in longitudinal studies with dropouts: guidelines and implementation in R. Stat Med. 30 de 2016;35(19):3424–48.

38. Zhang Z, Kim HJ, Lonjon G, Zhu Y. Balance diagnostics after propensity score matching. Ann Transl Med [Internet]. janeiro de 2019 [citado 13 de agosto de 2019];7(1). Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6351359/

39. Halekoh U, Højsgaard S, Yan J. The R Package geepack for Generalized Estimating Equations. Journal of Statistical Software [Internet]. 2006 [citado 29 de outubro de 2018];15(2). Disponível em: http://www.jstatsoft.org/v15/i02/

40. Dos Reis AT, De Oliveira PDTR, Sellera PE. Sistema de Avaliação para a Qualificação do Sistema Único de Saúde (SUS). RECIIS [Internet]. 31 de agosto de 2012 [citado 2 de novembro de 2016];6(2). Disponível em: http://www.reciis.cict.fiocruz.br/index.php/reciis/article/view/622/1089

41. Oliveira, A.E.F, Reis, R.S. Gestão pública em saúde: monitoramento e avaliação no planejamento do SUS [Internet]. São Luís: Universidade Federal do Maranhão. UNA-SUS/UFMA.; 2016. Disponível em: https://ares.unasus.gov.br/acervo/handle/ARES/7408

**Material Suplementar**

1. Link para base de dados, no formato "`Base_curta`":

https://drive.google.com/open?id=1RKbG8zrdu4nPHrAIseA4iC4Zi7sM1aOx

2. Código R para transposição das bases, do formato base curta para o formato base longa*:

```
Base_longa <- reshape(Base_curta,
varying = c("Int_Grupo_F_n_08",
"Int_Grupo_F_n_09", "Int_Grupo_F_n_10",
"Int_Grupo_F_n_11", "Int_Grupo_F_n_12",
"Int_Grupo_F_n_13", "Int_Grupo_F_n_14",
"Int_Grupo_F_n_15","Int_Grupo_F_n_16"),
v.names = "y1_Int_psiq_total",
timevar = "ano",
times = c(2008:2016),
direction = "long")


Base_longa_nearest <- reshape(Base_curta_nearest,
varying = c("Int_Grupo_F_n_08",
"Int_Grupo_F_n_09", "Int_Grupo_F_n_10",
"Int_Grupo_F_n_11", "Int_Grupo_F_n_12",
"Int_Grupo_F_n_13", "Int_Grupo_F_n_14",
"Int_Grupo_F_n_15","Int_Grupo_F_n_16"),
v.names = "y1_Int_psiq_total",
timevar = "ano",
times = c(2008:2016),
direction = "long")
```

Criação da variável `tempo`:

```
Base_longa$tempo = ifelse(Base_longa$ano > 2011, 1, 0)
Base_longa_nearest$tempo =
ifelse(Base_longa_nearest$ano > 2011,1,0)
```

3. Resultados dos modelos GLM, com pareamento e com ponderação, utilizando a base de dados longa com 2 anos (antes:2011 e depois:2016).

| | Código R do modelo | β1 | EP | P | m |
|---|---|---|---|---|---|
| GLM com pareamento | `diff_diff_2tempos_par = glm(y1_Int_psiq_total ~ NAAB+ tempo + NAAB:tempo, data = Base_longa_nearest_2tempos)` | -0,45 | 1,68 | 0,790 | 468 |
| GLM com ponderação | `diff_diff2_2tempos_pond = glm(y1_Int_psiq_total ~ NAAB + tempo + NAAB:tempo, data = Base_longa_2tempos, weights = peso.ATT)` | 0,54 | 2,99 | 0,856 | 752 |

m: total de observações; EP: erro padrão; GLM: G*eneralized Linear Models*

## 7. **ARTIGO 2**

# AVALIAÇÃO DO IMPACTO DO PROGRAMA NÚCLEOS DE APOIO À ATENÇÃO BÁSICA NAS INTERNAÇÕES PSIQUIÁTRICAS NO RIO GRANDE DO SUL

# IMPACT EVALUATION OF THE PROGRAM NÚCLEOS DE APOIO À ATENÇÃO BÁSICA ON PSYCHIATRIC HOSPITALIZATIONS IN RIO GRANDE DO SUL

Juliana Feliciati Hoffmann
Suzi Alves Camey

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL (UFRGS)

**A ser enviado ao Cadernos de Saúde pública**

**Resumo**

**Introdução**: A importância e a demanda por avaliação de políticas públicas têm crescido por viabilizar uma gestão baseada em evidências. Os Núcleos de Apoio à Atenção Básica (NAAB), criados em 2011, têm o objetivo de apoiar a inserção do cuidado em saúde mental na atenção básica de municípios com menos 17 mil habitantes. Esse trabalho busca avaliar o impacto da implantação dos NAABs nas internações por saúde mental e também internações por álcool e outras drogas no Rio Grande do Sul. **Metodologia:** Foram comparados os municípios com e sem NAAB, considerando o período de 2008 até 2016. Para contornar a ausência de aleatoriedade foi utilizada a metodologia de ponderação a partir dos escores de propensão como forma de lidar com viés de seleção. Para a estimação do impacto foi utilizado o modelo de Equações de Estimação Generalizadas incluindo todos os anos para os quais há informação. Para todas as análises foi utilizado o software livre R. **Resultados**: A ponderação demonstrou bons resultados, tornando os grupos comparáveis. O grupo com NAAB apresenta valores superiores ao grupo controle em todos os anos para qualquer um dos indicadores de resultado, com exceção do ano de 2009. Observou-se que não houve efeito do programa para nenhum dos quatro indicadores de resultado utilizados. **Discussão:** A ponderação tornou os grupos comparáveis, sendo uma boa alternativa para a avaliação de impacto no caso de Programas que não foram desenhados para serem avaliados. Os indicadores selecionados podem não estar refletindo da melhor forma os resultados esperados pelo programa, dado que a legislação de criação do Programa e os demais registros existentes não definem de forma clara um indicador de resultado que permita monitorar e avaliar o Programa. A programação em R apresentada permitiu a realização de todas as análises necessárias, com a vantagem de viabilizar a disponibilização das rotinas para replicação das mesmas.

**Palavras-chave:** gestão baseada em evidências; avaliação de impacto, quase-experimento

## Introdução

A importância e a demanda por avaliação de políticas públicas têm crescido no Brasil e no exterior (1), como forma de proporcionar maior transparência às ações, podendo ser utilizada para prestação de contas à sociedade sobre o desempenho dos programas do Governo (2) e, principalmente, como alternativa para uma gestão baseada em evidências. As informações fornecidas pelos diversos tipos de avaliação auxiliam os gestores em situações de tomada de decisão, inclusive sobre alocação dos recursos orçamentários em contexto de escassez de recursos.

No campo das políticas de saúde a avaliação surge vinculada aos avanços da epidemiologia e da estatística, a partir de testes de utilidade de diversas intervenções, particularmente direcionadas ao controle das doenças infecciosas e ao desenvolvimento dos primeiros sistemas de informação que orientassem as políticas sanitárias nos países desenvolvidos (3). Entretanto, esse conhecimento tem sido pouco aplicado para avaliar questões importantes de saúde pública, em especial em relação a políticas públicas de saúde (4), que afetam grande parte da população.

No Brasil, diante de marcantes desigualdades sociais e de recursos públicos escassos para o financiamento do setor de saúde, a avaliação é fundamental para estabelecer a capacidade de resposta de políticas, programas e serviços às necessidades da população, bem como justificar estratégias implementadas (5,6). A preocupação com avaliação em saúde vem desde a criação do Sistema Único de Saúde (SUS), em 1988, quando diversas iniciativas passaram a ser realizadas, buscando incluir a avaliação na rotina das instituições.

A avaliação de impacto, foco deste trabalho, busca estabelecer e quantificar estatisticamente as relações causais entre um programa e um conjunto de resultados, verificando se os objetivos ou os impactos desejados estão sendo alcançados no médio e longo prazo. Esse tipo de avaliação

demanda estratégias que impliquem algum nível de controle sobre o contexto observado.

Para avaliar a mudança no cenário de interesse, ao longo do tempo, a avaliação de impacto compara a situação em dois momentos do tempo: antes e depois da implementação do programa. Este modelo também implica a conformação de dois grupos: um grupo que recebe o tratamento proposto pelo programa e outro que não recebe, denominado grupo controle. O grupo controle fornece um parâmetro de comparação, representando a população-alvo caso não tivesse sido objeto de tratamento (programa). Esse parâmetro de comparação possibilita estimar o impacto do programa (7).

Uma questão fundamental envolvida na escolha dos elementos de cada grupo é a aleatorização da seleção, considerada o padrão-ouro da avaliação de impacto. Entretanto, na prática, a alocação aleatória dos grupos apresenta-se como mais um desafio (8,9), muitas vezes inviável no contexto de políticas públicas, inclusive por questões éticas. Assim, tornam-se mais frequentes os desenhos quase-experimentais (6), nos quais não ocorre a aleatorização dos elementos.

Os Núcleos de Apoio à Atenção Básica (NAAB) são um exemplo de programa cujo planejamento inicial não incluiu a previsão de uma avaliação de impacto a ser realizada posteriormente. Criados através da Resolução CIB 403/11, de outubro de 2011, como um dispositivo da Linha de Cuidado em Saúde Mental, Álcool e outras Drogas, a implantação dos NAABs ocorre por adesão dos municípios ao programa. Os Núcleos têm o objetivo de apoiar a inserção do cuidado em saúde mental na atenção básica de municípios com menos de dezesseis mil habitantes, os quais representam em torno de 76% do total de municípios do Estado. Estes municípios não estavam, naquela época, incluídos nos critérios estabelecidos para implantação do Núcleo de Apoio à Estratégia de Saúde da Família (NASF). Em 2016, os NAABs estavam presentes em 121 municípios, 32,2% dos que teriam perfil para sua implementação.

Os NAABs trabalham junto às equipes de Atenção Básica compartilhando responsabilidades por ações de promoção e prevenção em saúde, como discussão de casos e atendimento compartilhado entre equipes de atenção básica e NAAB para intervenção interdisciplinar, incluindo articulação com a rede de saúde, intersetorial e rede social, bem como intervenções específicas com usuários e famílias, além de reunião de equipe e ações de educação permanente. A implantação dos Núcleos objetiva alcançar mudanças nos processos de trabalho e no modelo de atenção, territorialização e regionalização, ações que qualifiquem a atenção em saúde mental, álcool e outras drogas, tendo a atenção básica como ordenadora do sistema.

O Programa se dá através de repasse de recursos financeiros aos municípios para contratação de profissionais com experiência em saúde metal, sendo 10 mil reais para implantação dos NAABs, mais um valor mensal de custeio de 8 mil reais para o município que aderir ao PMAQ ou de 6 mil reais para o que não aderir. Caso a equipe do NAAB tenha um ou mais profissionais com residência multiprofissional em saúde, será acrescido 20% sobre o valor do incentivo financeiro mensal a ser repassado ao município (10). As equipes dos NAABs são compostas por profissionais de diferentes áreas de conhecimento, sendo dois profissionais de nível superior (assistente social, médico, terapeuta ocupacional, educador físico, fonoaudiólogo, pedagogo, bacharel ou licenciado em artes ou psicólogo) e um de nível médio (preferencialmente acompanhante terapêutico, redutor de danos ou artesão).

Diante da relevância da temática de avaliação e da relevância do Programa nas políticas estaduais de saúde mental, esse trabalho tem como objetivo avaliar o impacto da implantação dos NAABs, buscando verificar se o Programa provocou mudanças, tanto em números absolutos quanto em números relativos, nas internações por saúde mental e também internações por álcool e outras drogas no Rio Grande do Sul, desde seu início, em 2012, até o ano de 2016.

## Metodologia

A fim de verificar o impacto da implantação dos NAABs foram comparados os municípios que aderiram com aqueles que não aderiram ao Programa, mantendo na análise somente os municípios que possuem até dezesseis mil habitantes. O período considerado foi de 2008 até 2016, sendo os anos de 2008 a 2011 a etapa anterior aos NAABs, cuja implantação iniciou-se em 2012, e 2012 a 2016 posterior. Os dados foram utilizados a partir de 2008 pois nesse ano foi instituída uma nova tabela de procedimentos para informação da produção SUS, tanto ambulatorial quanto hospitalar (11), o que poderia prejudicar a qualidade das informações caso fossem utilizados dados anteriores e posteriores a 2008, em função da conversão necessária na tabulação para verificar equivalências dos procedimentos, sendo que alguns novos foram criados.

Como a implantação dos NAABs ocorre por adesão dos municípios ao Programa, e não por seleção aleatória, trata-se de uma avaliação quase-experimental. Foi utilizada a técnica de ponderação por escores de propensão com o objetivo de eliminar o viés de seleção decorrente da falta de planejamento de uma seleção aleatória dos grupos na implantação da política.

Foram utilizados dados secundários de fontes diversas, sendo o município a unidade de análise, tratando-se, portanto, de um estudo ecológico longitudinal. As informações de gestão dos NAABs foram obtidas junto à Secretaria Estadual de Saúde no ano de 2016. Os dados das internações foram obtidos do SIHSUS - Sistema de Informações Hospitalares do SUS, os quais estão disponíveis no DATASUS (12). Também do DATASUS foram obtidos os dados do Cadastro Nacional de Estabelecimentos de Saúde (CNES), incluindo as características do estabelecimento, da equipe e dos profissionais, mensalmente, dado que essa é a forma de organização dos dados disponíveis na plataforma. As variáveis utilizadas, com suas respectivas fontes, estão apresentadas no quadro 1.

Para a estimação dos escores de propensão foram utilizadas como covariáveis todas as variáveis do quadro 1, com exceção das informações de gestão do Programa, referentes à linha de base, através de um modelo de regressão logística. Dessas, permaneceram no modelo final: expectativa de vida ao nascer, bloco de condições gerais do bloco de saúde do Índice de Desenvolvimento Socioeconômico (IDESE), percentual do orçamento aplicado em saúde, taxa de mortalidade infantil, carga horária semanal média por 1000 habitantes de Psicólogo e percentual de estabelecimentos SUS em relação ao total de estabelecimentos.

Para ajuste do modelo de regressão logística multivariável todas as variáveis da análise univariável que tiveram $p<0,30$ foram incluídas na etapa inicial, sendo excluídas uma a uma de acordo com o maior valor de VIF (do inglês *Variance Inflation Factor*) em função da multicolinearidade, até obter todas com VIF < 1,7. Na etapa seguinte foram excluídas as variáveis em ordem decrescente de p, até que todas fossem significativas no modelo. O modelo com menor valor de AIC foi considerado o modelo de melhor ajuste para obtenção dos escores de propensão.

Quadro 1. Variáveis utilizadas nas análises e respectivas fontes de dados

| Descrição da Variável | Fonte | Ano de referência |
|---|---|---|
| **Informações de Gestão sobre o Programa** | | |
| Município | SES/RS | Dados obtidos em 2016 |
| CNES | SES/RS | |
| Data de adesão ao Programa | SES/RS | |
| Valor repasse mensal | SES/RS | |
| Valor repasse implantação | SES/RS | |
| Município possui NASF tipo 3 | SES/RS | |
| **Informações sobre os municípios** | | |
| Bloco de Renda do IDESE (e seus componentes) | FEE | 2011 |
| Bloco de Saúde do IDESE (e seus componentes) | FEE | 2011 |
| Bloco de Educação do IDESE (e seus componentes) | FEE | 2011 |
| IDESE | FEE | 2011 |
| População estimada | FEE | 2011 |
| PIB per capita em R$ | FEE | 2011 |
| Expectativa de vida ao nascer | FEE | 2010 |

| | | |
|---|---|---|
| Proporção de ICSAB (N internações condições sensíveis à atenção básica/total de internações) | SES/RS | 2011 |
| % orçamento do município aplicado em saúde | TCE/RS | 2011 |
| Estimativa da População coberta por ACS | DATASUS | 2011 |
| Estimativa da População coberta por ESF | DATASUS | 2011 |
| Estimativa da Cobertura populacional por ACS | DATASUS | 2011 |
| Estimativa da Cobertura populacional por ESF | DATASUS | 2011 |
| Taxa de mortalidade infantil | IBGE Cidades | 2011 |
| % domicílios com saneamento adequado | IBGE - Censo | 2010 |
| % pessoas com renda per capita de até meio salário Mínimo | IBGE - Censo | 2010 |
| Taxa de analfabetismo em maiores de 15 anos | IBGE - Censo | 2010 |
| **Informações sobre os estabelecimentos de saúde** | | |
| Tipo de estabelecimento | DATASUS | 2011 |
| Número de estabelecimentos do município que são postos de saúde ou UBS para cada 10.000 habitantes | DATASUS | 2011 |
| Número de estabelecimentos do município com atendimento em 1 ou 2 turnos ou intermitente para cada 10.000 habitantes | DATASUS | 2011 |
| Número de estabelecimentos do município com atendimento 24hs com plantão para cada 10.000 habitantes | DATASUS | 2011 |
| Número de estabelecimentos do município com atendimento em três turnos para cada 10.000 habitantes | DATASUS | 2011 |
| Número de estabelecimentos do município para cada 10.000 habitantes | DATASUS | 2011 |
| Número de estabelecimentos do município com vínculo SUS para cada 10.000 habitantes | DATASUS | 2011 |
| % de estabelecimentos SUS em relação ao total de estabelecimentos | DATASUS | 2011 |
| % de UBS ou posto em relação ao total de estabelecimentos SUS | DATASUS | 2011 |
| **Informações sobre os profissionais dos estabelecimentos de saúde** | | |
| Número de Médicos por habitante | DATASUS | 2011 |
| Carga horária semanal média por 1000 habitantes: | | |
| Psicólogo | DATASUS | 2011 |
| Neuropsicólogo | DATASUS | 2011 |
| Psicólogo psicanalista | DATASUS | 2011 |
| Técnico de Enfermagem Psiquiátrica | DATASUS | 2011 |
| Agente Comunitário de Saúde | DATASUS | 2011 |
| Auxiliar de Enfermagem Saúde Família | DATASUS | 2011 |
| Médico | DATASUS | 2011 |
| Médico Psiquiatra ou psicanalista | DATASUS | 2011 |
| Médico de Saúde da Família | DATASUS | 2011 |
| Enfermeiro | DATASUS | 2011 |
| Enfermeiro Saúde Pública | DATASUS | 2011 |

Para a estimação do impacto dos NAABs foi utilizado o modelo GEE adaptado da metodologia de diferenças em diferenças (DID). A

abordagem modelo GEE (do inglês, G*eneralized Estimating Equation e em* português Equações de estimação Generalizadas) especificado abaixo, permite incluir todos os anos para os quais há informação considerando a dependência entre as observações. Maiores detalhes sobre o modelo e suas especificações podem ser obtidos em Hoffmann e Camey (13).

$$E(Y_{ijk}) = \beta_0 + \beta_1 t_j + \beta_2 tto_i + \beta_3 ano_k + \beta_4 t_j tto_i + \beta_5 t_j ano_k + \beta_6 tto_i ano_k + \beta_7 tto_i ano_k t_j \quad (1)$$

Sendo:

$E(Y_{ijk}) = esperaça\ do\ número\ de\ internaçõe s\ psiquiátricas\ no\ município\ i\ no\ momento\ j\ e\ no\ ano\ k$

$i = 1, \ldots n;\ \ n = número\ de\ municípios;\ \ j = 0{,}1;\quad k = 1, \ldots 9;\ ano_k = 2008, 2009, \ldots, 2016$

$$t_j = \begin{cases} 0, & \text{se j} = 0, \text{ou seja, antes do tratamento} - 2008 \text{ a } 2011 \\ 1, & \text{se j=1, ou seja, após o tratamento - 2012 a 2016} \end{cases}$$

$$tto_i = \begin{cases} 0, & \text{se o município i pertence ao grupo controle} \\ 1, & \text{se o município i pertence ao grupo tratamento} \end{cases}$$

A partir de (1), o impacto será medido por $\beta_4 + \beta_7 ano_k$ sendo dependente, portanto, do ano k.

A legislação de criação do Programa e os demais registros disponibilizados não definem de forma clara um indicador de resultado que permita monitorá-lo e avaliá-lo. Para fins dessa avaliação foram escolhidos quatro indicadores de resultado, referentes a internações por saúde mental e também internações por álcool e outras drogas, em números absolutos e em números relativos. Os dados referentes às internações foram obtidos das AIHs (Autorização de Internação Hospitalar), disponíveis no DATASUS, por ano e local de residência.

Quadro 2 – indicadores de resultado, por município, por ano.

| **Indicadores de resultado*** | **Descrição** |
|---|---|
| Y1. internações psiquiátricas totais (n) | Número de internações psiquiátricas (Grupo F do CID 10) |
| Y2. total internações psiquiátricas/total de internações (%) | Número de internações psiquiátricas (Grupo F do CID 10) / Número de internações por qualquer causa x 100 |
| Y3. internações psiquiátricas por álcool e outras drogas(n) | Número de internações psiquiátricas por álcool e outras drogas (Subgrupo F10 a F19 do CID 10) |
| Y4. internações psiquiátricas por álcool e outras drogas/total de internações psiquiátricas (%) | Número de internações psiquiátricas por álcool e outras drogas (Subgrupo F10 a F19 do CID 10) / Número de internações psiquiátricas (Grupo F do CID 10) x 100 |

*internações por local de residência

Para todas as análises e manipulação de bases de dados foi utilizado o software livre R, considerando um nível de significância de 5%. A opção pelo R se justifica por ser um software livre de código aberto utilizado por uma comunidade de usuários ativos no mundo todo, o que torna o conhecimento bastante colaborativo e acessível (14). No material suplementar apresenta-se a programação em R utilizada.

## Resultados

A tabela 1 mostra o balanceamento das variáveis utilizadas na regressão logística para obtenção dos escores de propensão. São apresentados os valores de média e desvio padrão para cada uma das variáveis utilizadas no modelo final para estimação dos escores de propensão, antes e depois da ponderação. O p das tabelas refere-se ao teste t, dado que todas as variáveis são contínuas e que assumiu-se que os dados têm distribuição normal. Observa-se que antes da ponderação os grupos apresentavam médias bastante diferentes em todas as variáveis, o que pode ser confirmado com a significância do teste t. Após a ponderação deixou de existir diferença significativa entre as médias, dado que as médias ficaram iguais ou muito próximas.

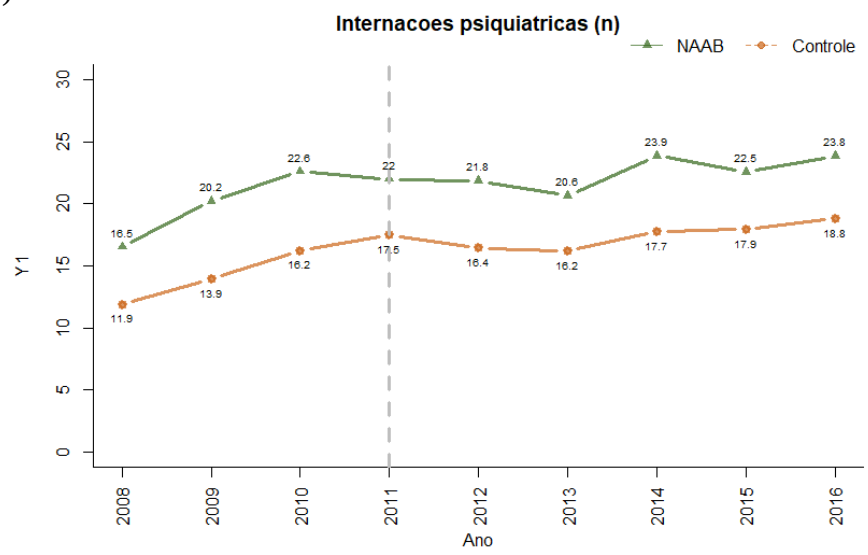
Tabela 1. Média e desvio-padrão das covariáveis usadas no escore de propensão, com e sem ponderação.

| | Não ponderado | | | | | Ponderado | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Tratamento | | Controle | | | Tratamento | | Controle | | |
| | Média | DP | Média | DP | P | Média | DP | Média | DP | P |
| Expectativa de vida ao nascer | 74,97 | 1,44 | 75,39 | 1,34 | 0,007 | 74,97 | 1,44 | 74,99 | 1,32 | 0,901 |
| Condições gerais de saúde* | 0,76 | 0,05 | 0,77 | 0,05 | 0,004 | 0,76 | 0,05 | 0,76 | 0,05 | 0,958 |
| % orçamento aplicado em saúde | 19,01 | 2,78 | 18,36 | 2,26 | 0,026 | 19,01 | 2,78 | 18,99 | 2,53 | 0,960 |
| Taxa de mortalidade infantil | 7,88 | 14,06 | 14,52 | 23,43 | 0,001 | 7,88 | 14,06 | 8,19 | 14,04 | 0,838 |
| % estabelecimentos SUS em relação ao total de estabelecimentos | 75,18 | 24,14 | 81,26 | 22,15 | 0,020 | 75,18 | 24,14 | 75,15 | 23,68 | 0,992 |

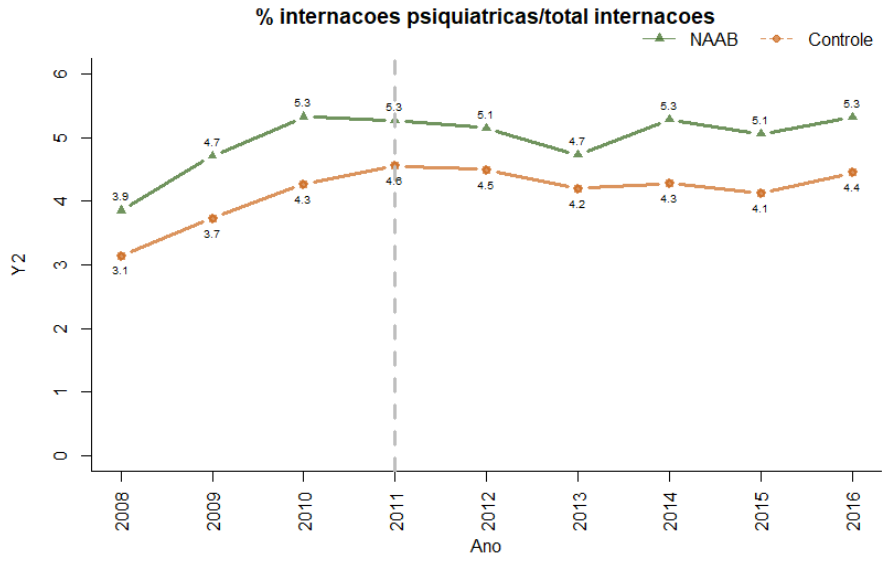
DP: desvio padrão. *Bloco de Saúde do IDESE

A figura 1 mostra a evolução dos valores médios dos desfechos considerados no Rio Grande do Sul, de 2008 até 2016, utilizando a amostra ponderada pelos escores de propensão. Na linha verde estão representados os valores do grupo tratamento e na laranja os do grupo controle. Observa-se que o grupo com NAAB apresenta valores superiores ao grupo controle em todos os anos para qualquer um dos indicadores de resultado, com exceção do ano de 2009 na figura d, quando o grupo controle tem média superior ao grupo tratamento.

(a)


Internacoes psiquiatricas (n)
NAAB
Controle
Y1
Ano
2008
2009
2010
2011
2012
2013
2014
2015
2016
16.5
20.2
22.6
22
21.8
20.6
23.9
22.5
23.8
11.9
13.9
16.2
17.5
16.4
16.2
17.7
17.9
18.8

(b)


% internacoes psiquiatricas/total internacoes
NAAB
Controle
Y2
Ano
2008
2009
2010
2011
2012
2013
2014
2015
2016
3.9
4.7
5.3
5.3
5.1
4.7
5.3
5.1
5.3
3.1
3.7
4.3
4.6
4.5
4.2
4.3
4.1
4.4

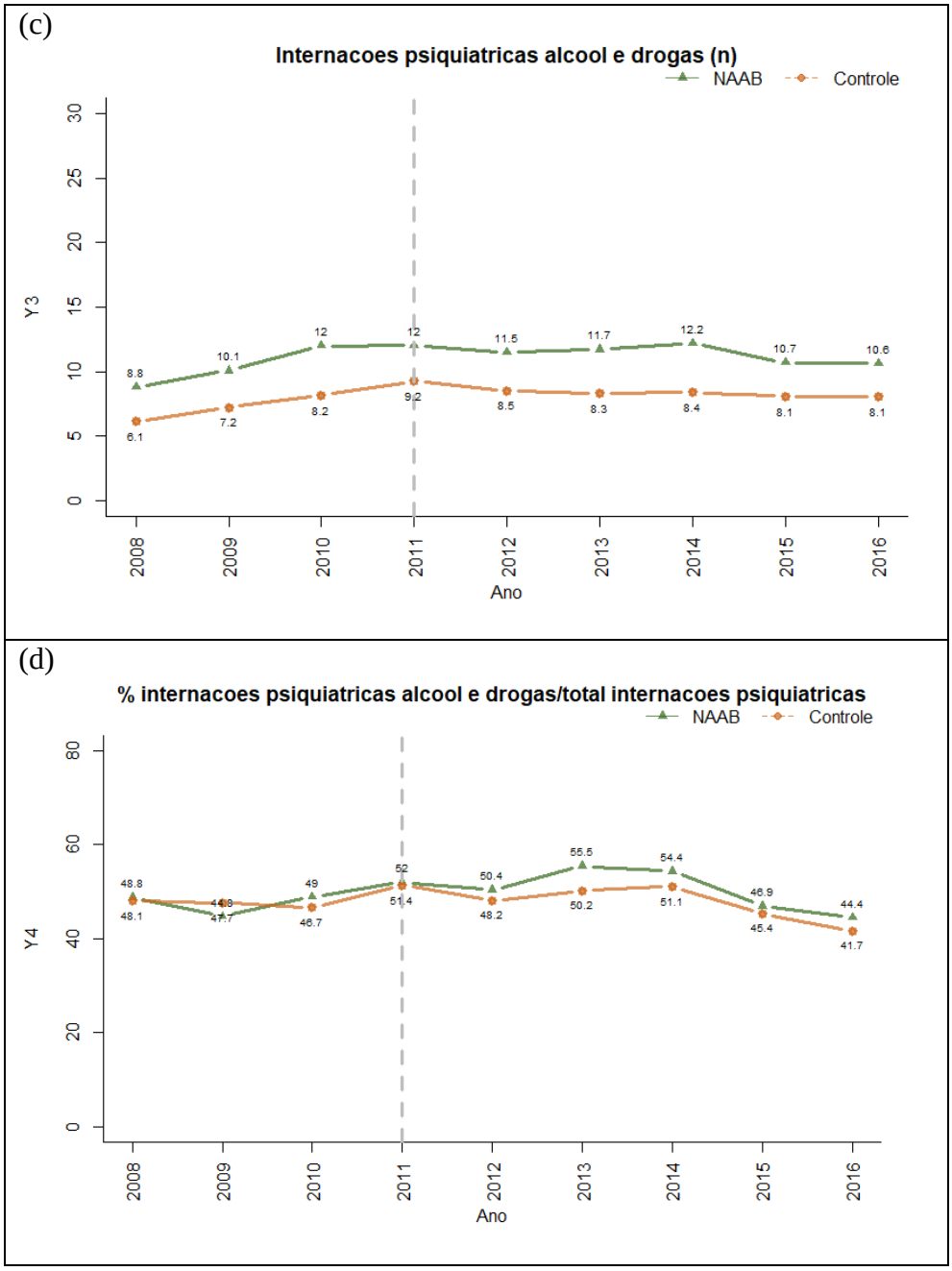



Figura 1. Valores médios dos indicadores de resultado no Rio Grande do Sul, por grupo, 2008 a 2016, com amostra ponderada pelos escores de propensão.

A tabela 2 mostra as estimativas dos dois coeficientes ($\beta_4$ e $\beta_7$) que indicam o impacto do Programa obtidos pelo modelo GEE para cada um dos quatro indicadores. Observa-se que não há efeito do programa para nenhum dos quatro indicadores de resultado utilizados.

Tabela 2. Coeficientes estimados do modelo GEE com ponderação* para cada um dos indicadores de resultado (m=3384)

| Indicador de resultado | $\widehat{\beta}_4$ | EP | p | $\widehat{\beta}_7$ | EP | p |
|---|---|---|---|---|---|---|
| Y1. internações psiquiátricas (n) | 59,00 | 2555,00 | 0,982 | -0,03 | 1,27 | 0,982 |
| Y2. % internações psiquiátricas/total internações | -1,50 | 4,07 | 0,713 | 0,00 | 0,00 | 0,713 |
| Y3. internações psiquiátricas por álcool e outras drogas (n) | 556,49 | 1522,57 | 0,715 | -0,28 | 0,757 | 0,715 |
| Y4. % internações psiquiátricas por álcool e outras drogas/total de internações psiquiátricas | 15,40 | 32,94 | 0,640 | -0,01 | 0,02 | 0,640 |

*Modelo para estimar ATT; EP: Erro Padrão; m= total de observações.

## Discussão

As estimativas de impacto do Programa encontradas nessa avaliação não foram significativas, indicando que não há qualquer impacto do Programa nas internações por saúde mental, mesmo quando foram consideradas as internações de forma geral ou aquelas por álcool e outras drogas especificamente, tanto em valor absoluto quanto em percentual.

Para contornar a ausência de aleatoriedade foi utilizada a metodologia de ponderação a partir dos escores de propensão como forma de lidar com viés de seleção. O resultado do balanceamento indica que a utilização da ponderação demonstrou bons resultados, tornando os grupos comparáveis em relação às covariáveis que mostraram ter relação com a adesão dos municípios ao Programa. Assim, é uma alternativa para permitir a avaliação de impacto no caso de Programas que não foram desenhados para serem avaliados, como é o caso dos NAABs, se aproximando ao que ocorre num delineamento experimental no que diz respeito à comparabilidade dos grupos.

Ressalta-se que os NAABs foram criados como sendo uma alternativa aos municípios menores, com menos de dezesseis mil habitantes, que não tinham perfil para aderir aos Núcleos de Apoio à Saúde da Família

(NASF), do nível federal, que tem o objetivo de apoiar a consolidação da Atenção Primária no Brasil, ampliando as ofertas de saúde na rede de serviços, além da resolutividade, abrangência e alvo das ações. Entretanto, em dezembro de 2012, a portaria 3.124 abriu a possibilidade de qualquer município do Brasil implantar equipes NASF. Desde então, a opção de permanecer com o recurso estadual (NAAB), migrar para o financiamento federal (NASF) ou manter as duas modalidades de financiamento é uma escolha dos gestores municipais, desde que atendam aos critérios descritos nas respectivas legislações. A possibilidade de adesão ao NASF ao longo do período também é uma limitação do estudo, pois pode ser um dos fatores que influenciaram os resultados. De qualquer forma, sabe-se que, em 2016, apenas 18% (n=22) dos municípios do grupo tratamento e 14% (n=36) do grupo controle tinham NASF tipo 3.

Como limitação desse estudo destaca-se que a legislação de criação do Programa e os demais registros existentes não definem de forma clara um indicador de resultado que permita monitorar e avaliar o Programa, nem metas de alcance desses objetivos. Além disso, a falta de uma avaliação *ex-ante*, definindo claramente uma cadeia causal do programa no seu planejamento, que auxiliasse a descrever os resultados esperados de cada ação planejada, também prejudica a definição dos indicadores de resultado. Para essa avaliação considerou-se a hipótese que as ações dos profissionais junto à atenção básica permitiriam a prevenção do desenvolvimento ou agravo de doenças mentais, em especial aquelas causadas por álcool e outras drogas, o que provocaria uma redução nas internações nos municípios que aderiram ao Programa. Entretanto, os indicadores aqui selecionados podem não estar refletindo da melhor forma os resultados esperados pelo programa.

Por outro lado, caso o Programa tivesse sido preparado para a avaliação e a seleção dos municípios fosse aleatória, a análise dos dados seria muito mais simples. O método de análise proposto, com modelos GEE que permitem a utilização de mais de uma medida antes e depois da implantação do Programa, pode ser utilizado em qualquer uma das situações. De qualquer forma, mesmo que haja métodos que viabilizem a

avaliação, ressalta-se que o padrão-ouro é a avaliação experimental, o que reforça a importância do planejamento da avaliação desde o início do planejamento da política pública.

Outro ponto fundamental para a avaliação de políticas de saúde é a disponibilidade de dados. No caso dessa avaliação foram utilizados apenas dados secundários, tendo o município como unidade de medida, conforme apresentado na metodologia. No caso dos dados do DATASUS, apesar de estarem disponíveis *online*, ainda são de difícil localização, identificação, *download* e manipulação, sendo esse um problema recorrente percebido há mais de uma década (15). É necessária uma pesquisa ampla e contato com profissionais que lidam diariamente com as informações para que seja possível compreender a organização utilizada, tanto para arquivos de dados quanto para arquivos de descrição dos dados do DATASUS. No caso de arquivos de dados, não estão preparados para o *download* agregado por períodos maiores, apenas mês a mês. Isso dificulta a utilização da informação, dado que demanda bastante tempo e *expertise* em manipulação de bases de dados para formatação de uma estrutura que permita uma avaliação como a que foi desenvolvida neste trabalho.

O Programa *TabWin* do DATASUS, para tabulação e tratamento dos dados, permite poucas análises e cruzamentos, de forma que não facilita a criação de uma base de dados com maior quantidade de informações vinculadas. Além disso, os arquivos de descrição dos dados, disponibilizados no DATASUS, estão formatados para utilização no *TabWin* apenas, não havendo outra opção de formato que possa ser utilizada diretamente em outros softwares de análise estatística.

Uma vez organizada a base de dados, a programação em R aqui apresentada permitiu a realização de todas as análises necessárias, com a vantagem de permitir a disponibilização das rotinas para replicação das mesmas. Dessa forma, apresenta-se uma alternativa sem custos aos gestores dos programas, por ser um *software* livre, auxiliando a disseminar a cultura de avaliação de impacto em saúde.

De qualquer forma, ressalta-se que o Programa vem repassando recurso aos municípios desde 2012, quando foram iniciadas as adesões, sem qualquer tipo de avaliação dos seus impactos. Os resultados aqui apresentados permitem aos gestores repensarem as ações que estão sendo desenvolvidas, bem como o recurso que vem sendo investido, dado que o Programa parece não ter impacto. Assim, busca-se contribuir para uma gestão baseada nas evidências aqui apresentadas, o que poderia servir de subsídio para justificar a aplicação do recurso em outra inciativa que resulte em maior impacto aos beneficiários.

**Referências**


1. Wholey JS, Hatry HP, Newcomer KE, organizadores. Handbook of Practical Program Evaluation. 2nd edition. JossryBass Inc Pub,2004; 2004.

2. Carneiro F. Avaliação de Políticas Públicas: por um procedimento integrado ao ciclo da gestão. Revista Perspectivas em Políticas Públicas [Internet]. 2013 [citado 2 de novembro de 2016];6(11). Disponível em: http://www.uemg.br/openjournal/index.php/revistappp/article/view/893

3. Oliveira, A.E.F, Reis, R.S. Gestão pública em saúde: monitoramento e avaliação no planejamento do SUS [Internet]. São Luís: Universidade Federal do Maranhão. UNA-SUS/UFMA.; 2016. Disponível em: https://ares.unasus.gov.br/acervo/handle/ARES/7408

4. Gulis G, Fujino Y. Epidemiology, Population Health, and Health Impact Assessment. Journal of Epidemiology. 2015;25(3):179–80.

5. Facchini LA, Piccini RX, Tomasi E, Thumé E, Teixeira VA, Silveira DS da, et al. Avaliação de efetividade da Atenção Básica à Saúde em municípios das regiões Sul e Nordeste do Brasil: contribuições metodológicas. Cad Saúde Pública. 2008;24(Sup 1):S159–72.

6. Fernandes FMB, Ribeiro JM, Moreira MR. Reflexões sobre avaliação de políticas. Cad Saude Publica. 2011;27(9):1667–1677.

7. Ramos M. Aspectos Conceituais e Metodológicos da Avaliação de Políticas e Programas Sociais. Planejamento e Políticas Públicas [Internet]. 18 de agosto de 2009 [citado 2 de novembro de

2016];1(32). Disponível em: http://www.ipea.gov.br/ppp/index.php/PPP/article/view/11

8. Paul J. Gertler, Patrick Premand, Christel M. J. Vermeersch, Laura B. Rawlings, Sebastián Martínez. Avaliação de Impacto na Prática [Internet]. 2ª Edição. Banco Internacional para Reconstrução e Desenvolvimento/Banco Mundial; 2018. 375 p. Disponível em: https://openknowledge.worldbank.org/bitstream/handle/10986/25030/9781464808890.pdf

9. Schor, A A LE. Apostila Avaliacao Economica Itau Social [Internet]. 2007 [citado 2 de novembro de 2016]. Disponível em: http://www.redeitausocialdeavaliacao.org.br/wp-content/uploads/2015/01/Apostila_Avaliacao-Economica_06-08-07_final_20150128.pdf

10. Secretaria da Saúde do Estado do Rio Grande do Sul. Resolução Nº 403/11 – CIB/RS [Internet]. 403/11 out, 2011. Disponível em: https://atencaobasica.saude.rs.gov.br/upload/arquivos/201510/01114725-20141105174549rs-res-403-2011-naab.pdf

11. Ministério da Saúde. Portaria Nº 321. 321 fev 8, 2007.

12. DATASUS [Internet]. Disponível em: http://datasus.saude.gov.br/

13. Hoffmann JF, Camey SA. Avaliando o impacto de políticas públicas de saúde - Modelagem estatística em um quase-experimento longitudinal. setembro de 2019;

14. Randolph JJ, Falbe K, Manuel AK, Balloun JL. A step-by-step guide to propensity score matching in R. Practical Assessment, Research & Evaluation. 2014;19(18):2.

15. Candiago RH, Abreu PB de. Use of Datasus to evaluate psychiatric inpatient care patterns in Southern Brazil. Revista de Saúde Pública. 2007;41(5):821–829.

## Material Suplementar

Tabela 3. Códigos R utilizados nesse artigo, para cada uma das etapas de análise

| Etapa | Código R |
|---|---|
| **1. Escores de Propensão** | |
| 1.1. Estimar os escores de propensão através de regressão logística | `ps_mf13<- glm(NAAB ~Expect_vida_nasc_2010 + Bloco_Saude_cond_gerais_2011 + perc_Saude_2011 + tx_mort_inf_2011 + perc_VINC_SUS_total, family = binomial(link = "logit"), data = Base_curta)` |
| 1.2. Salvar os valores dos escores de propensão no *dataframe* | `Base_curta$escore_pr13 <- predict(ps_mf13, type = "response")` |
| 1.3. Verificar área de suporte comum pelo boxplot: | `boxplot(Base_curta$escore_pr~Base_curta$NAAB,data=Base_curta, main="Verificar suporte comum", xlab="NAAB", ylab="Escore de Propensão")` |
| **2. Ponderação** | |
| 2.1 Calcular os pesos para estimar o ATT | `Base_curta$peso.ATT<- ifelse(Base_curta$NAAB == 0, (Base_curta$escore_pr13)/(1 - Base_curta$escore_pr13), 1)` |
| 2.2 Avaliar a qualidade da ponderação – balanceamento* | `bal.pslog1 <- dx.wts(x = Base_curta$peso.ATT, data= Base_curta, vars=c("Expect_vida_nasc_2010" , "Bloco_Saude_cond_gerais_2011" , "perc_Saude_2011" , "tx_mort_inf_2011" , "perc_VINC_SUS_total"), treat.var="NAAB", estimand = "ATT")`<br>`bal.table(bal.pslog1)` |
| **3. Modelo para estimar o efeito** do Programa sobre Y1** | |
| 3.1 Estimar os efeitos através de GEE com ponderação*** | `fit_Y1_GEE_pond <- geeglm(y1_Int_psiq_total ~ NAAB + ano + tempo + tempo:ano + NAAB:tempo + NAAB:ano + ano:NAAB:tempo, id=CODUFMUN, data = Base_longa,`<br>`family = gaussian(link = "identity"), corstr="exchangeable", weights = peso.ATT)` |

*usando pacote `twang`; **ATT; ***usando pacote `geepack`

# 8. CONCLUSÕES E CONSIDERAÇÕES FINAIS

Este trabalhou mostrou que são diversos os tipos de avaliação, bem como também são muitos os desafios a serem enfrentados para a condução de uma avaliação de impacto de política pública na prática. A disponibilidade de dados, os interesses e as barreira políticas, o receio dos gestores ao serem avaliados, e a utilização dos resultados obtidos são exemplos de dificuldades a serem contornadas.

Considerando o ciclo de uma política pública, a avaliação deve, idealmente, percorrer todas as suas etapas. No planejamento inicial, a avaliação *ex-ante* permite estruturar as cadeias causais das ações que serão desenvolvidas, verificando a lógica construída para o alcance dos resultados que se espera obter. Durante a implementação, a avaliação de processo busca verificar se o que foi planejado está sendo de fato realizado. Após o início do programa, a avaliação *ex-post* verifica a obtenção dos resultados e o impacto da política sobre indicadores de resultado selecionados conforme os objetivos planejados.

O desenho de avaliação de impacto de uma política pública tem origem em delineamentos desenvolvidos para pesquisas clínicas, que embasam a medicina baseada em evidências. Entretanto, apesar desse histórico e dos esforços que vêm sendo realizados, ainda há muito a avançar nessa temática quando se trata de avaliação de políticas de saúde. O DATASUS, por exemplo, apesar de disponibilizar uma grande quantidade de dados, ainda precisa de uma gestão da informação mais moderna e qualificada.

Quanto à análise, o método *Diferenças-em-diferenças* é o mais tradicionalmente utilizado para estimação do impacto das políticas públicas. Entretanto, foi possível demonstrar que existem outras abordagens de modelagem estatística que permitem a estimação das medidas de efeito do tratamento agregando a informação de todo o período no tempo, e não apenas de dois momentos (antes e depois), como ocorre no DID. Os

modelos GEE, utilizados neste trabalho, permitem incluir na análise a correlação entre as observações ao longo dos anos.

Os Núcleos de Apoio à Atenção Básica não demostraram impacto sobre as internações psiquiátricas, tanto de forma geral quanto quando consideradas apenas as internações por álcool e outras drogas. Os resultados encontrados e o levantamento de dados realizado sobre a legislação e documentação dos NAABs sugerem que é necessário repensar o Programa. A utilização de ferramentas de análise *ex-ante*, como modelo lógico ou teoria da mudança, poderiam qualificar a definição de ações que possam ser realizadas buscando alcançar os objetivos almejados.

Por fim, a utilização do software R para a manipulação das bases de dados e para todas as análises realizadas mostrou-se plenamente satisfatória. A documentação disponível referente aos pacotes e funções existentes facilita muito a compreensão e o desenvolvimento da programação necessária. Com a disponibilização dos códigos utilizados nas análises espera-se que os métodos aqui utilizados sejam replicados por gestores e demais interessados na ampliação do conhecimento em avaliação de impacto de políticas públicas. Somente assim, com uma gestão baseadas em evidências, será possível qualificar as ações de saúde públicas entregues aos cidadãos.