EDITORIAL

DYSPLASIA IN THE BARRETT'S ESOPHAGUS: utopia, difficult or impossible concordance

HEADING – Barrett esophagus.

Once upon a time, 12 well-known pathologists and specialists in gastrointestinal tract histopathology examined a case of Barrett's esophagus, and were asked for their opinion twice (giving a total of 24 answers); their answers show how much utopia would be to expect unanimity: absence of dysplasia - three, indefinite dysplasia - three, low-grade dysplasia - eight, high-grade dysplasia - nine, invasive carcinoma - one. Actually there were 250 cases, and this one was only one of the most emblematic[5].

Endoscopic surveillance of patients with Barrett's esophagus is considered as standard of care, and ethically accepted, because it detects the cases with pre-malignant changes (dysplasia), removal of which would prevent development of cancer. Definition of Barrett's esophagus by the presence of columnar epithelium is too broad, and would make surveillance impracticable, due to the great number of persons to be followed, and by the very low risk of cancer in such a group. It is a general consensus that only patients with specialized columnar epithelium, e.g. with incomplete intestinal metaplasia should be followed. We call dysplasia the sum of cytological changes (mucin depletion, cytoplasmic basophilia, and nuclear atypias), architectural changes (irregularity of the glands), and when possible to evaluate, genomic instability.

But, what is the variability of the histopathological diagnosis of dysplasia in the Barrett's esophagus?

Variability in the diagnosis or interpretation by different observers of the same medical phenomenon is a well-known fact, but relatively less studied. Also known is the variability of one observer when analyzing the same phenomenon some time later on. Pathologists do not evade this rule. The same histopathological lesion frequently receives distinct diagnoses when examined by different pathologists. And, when solicited to grade a change, even using the same frame of reference or scale, there is large variability in the grading, variability that will be larger in systems with many categories to choose from. The diagnosis and categorization of the precursor lesions of cervical cancer is a well-known example.

In this issue of ARQUIVOS de GASTROENTEROLOGIA, LOPES et al.[3] studied this variability problem in the histopathological diagnosis and grading of dysplasias in Barrett's esophagus. Using MONTGOMERY's criteria[5], from Johns Hopkins University, they studied reproducibility and variability of dysplasia, studying biopsies of 40 patients with Barrett's esophagus, seen by three different pathologists, and applying the kappa (K) of COHEN[2]. The results, even being not surprising, merit considerable thought and further studies, because of the very large degree of variability, seen by the low K values reported. Evaluating absence of dysplasia, the K scores were between 0.07 and 0.20. In the diagnosis of low-grade dysplasia, the K scores were between 0.05 and 0.6. The intra-observer scores were somewhat higher, but still low.

When concordance occurs by chance alone K is 0 (zero). When concordance is total, K is 1 (one). When concordance is worse than expected by chance, K has a negative value. Sometimes, it is necessary to weight differently one situation: for example, to misclassify a lesion of little significance is totally different than to commit an error with a lesion of grave or lethal prognosis: it is necessary to create a weighted K to analyze the observer performance[4]. But, for each situation, where to put an acceptable level of K is always an open question of scale interpretation. Low values of K, in general, are due to two types of problems: subjective diagnostic criteria, or too great differences in the observer's experience. This second cause, can be remedied by joint study of several cases by an expert and the naïve observer, until this latter is made a new expert. This is what we do with our residents in daily life. Presence of too subjective or difficult to apply criteria is identified by low K values, when the topic is studied by several experts seeing the same material. In this application of kappa statistics, there are serious divergences between bio-statisticians. In the calculation of K, the term "chance proportion" or "expected concordance". But, this is valid only under the condition of statistical independence of the raters. In the practice of medicine this is seldom seen. As a statistical test, K can verify that

concordance between observers exceeds chance. But as a measure of the level of concordance between the observers K is not corrected by chance. In reality, we do not know how chance, or external factors such as emotional status, stress, fatigue, involvement with the case, previous experiences, etc. influence the decision making of the observers, and how to correct or evaluate its importance. K can be quite low, and even then high levels of concordance can occur[1]. To determine that a given value of K signifies a bad or good system of grading or observing a give phenomenon, depends on the assumed model about the decision making of the raters/observers[6].

We do not agree with the use (actually abuse), because of its lack of statistical reality, of the categories Excellent (K between 0.93 and 1), very good, good, reasonable, weak, poor (K between 0.01 and 0.20). Besides, K is influenced by prevalence, making comparisons between different series or studies, practically impossible. So, the comparison of the value of K (very high for high grade dysplasia, and somewhat smaller for low grade dysplasia) seen in MONTGOMERY's study[5] is tempting but without statistical value.

In the original evaluation by the pathologists, 23 patients had dysplasia, 22 low grade and one high grade, or 54.8% of the sample. Using the criterion of concordance of three in four observations, the diagnosis of dysplasia would be reduced to six cases (low grade in five and high grade in one case) or 14.3% of the sample. The study could have had more statistical power if done in a larger series of cases, after sample size determination. The authors quite judiciously, conclude by the necessity of having a second opinion even in low grade dysplasias in Barrett's esophagus, and not only high grade lesions, as usually recommended by the consensus on Barrett's esophagus.

Obviously, this will have significant additional cost, involving many experts, not always at the disposition of laboratories receiving the meager sums of Brazilian National Health Service (SUS).

João Carlos **PROLLA**[1]

Prolla JC. Displasia no esôfago de Barrett: utopia, concordância difícil ou impossível. Arq Gastroenterol 2004;41(2): 77-8.
DESCRITOR - Esôfago de Barrett.

REFERENCES

1. Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. J Clin Epidemiol 1990;43:551-8.
2. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20:37-46.
3. Lopes CV, Pereira-Lima J, Hartmann AA, Tonelotto E, Salgado K. Displasia no esôfago de Barrett. Concordância intra e interobservador no diagnóstico histopatológico. Arq Gastroentrol 2004;41:79-83.
4. Maclure M, Willett WC. Misinterpretation and misuse of the kappa statistic. Am J Epidemiol 1987;126:161-9.
5. Montgomery E, Bronner MP, Goldblum JR, Greenson JK, Haber MM, Hart J, Lamps LW, Lauwers GY, Lazenby AJ, Lewin DN, Robert ME, Toledano AY, Shyr Y, Washington K. Reproducibility of the diagnosis of dysplasia in Barrett esophagus: a reaffirmation. Human Pathol 2001;32:368-78.
6. Uebersax JS. Diversity of decision-making models and the measurement of interrater agreement. Psychol Bull 1987;101:140-6.

[1]Full Professor, Internal Medicine Department, Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil.