International Journal of Epidemiology 2003;**32**:314–315
DOI: 10.1093/ije/dyg065

Commentary: Black and white or shades of grey?

Markus Battaglia and Daniel Pewsner

In the quest for diagnostic certainty we should welcome anything that improves our ability to interpret diagnostic tests. In this issue of the *International Journal of Epidemiology* Joël Coste and Jacques Pouchot[1] describe a method for constructing a three-zone division for continuously measured diagnostic test results. The concept of three-zone diagnostic decision making, coined by Feinstein in 1990,[2] strengthens the explanatory power of our customary 'yes–no' reasoning by including a grey zone of intermediate values in which a disease cannot be said to be present or absent. Coste and Pouchot illustrate their proposal with examples of tuberculin testing and markers of anaemia in children. They use empirical data from the literature to analyse the distribution of test results in diseased and non-diseased populations and define upper and lower limits of a grey zone beyond which the post-test probabilities would allow the target disease to be safely ruled in or out. This method makes assumptions about the pre-test probability and the estimated size of the likelihood ratios needed to achieve the required post-test probabilities.

It is tempting to replace the 'black–white division' of continuous test results, where black is diseased and white non-diseased, and include a grey zone. Coste and Pouchot's method implicitly uses the principle of decision thresholds[3] to estimate the bounds of the grey zone. The 'treatment-test threshold' denotes the probability of a disease when treating (or invasive testing) and further testing are of equal value and equates to a test result around the grey–black limit. The 'no treatment-test threshold' defines the disease probability at which one would neither treat nor conduct further tests and includes test results around the grey–white limit. The calculation of such limits is complex and depends on elements of the disease, on values of the society, the doctor, and the patient. The difficulties encountered when applying this concept to decision making in individual patients should not be underestimated, and is not adequately addressed by Coste and Pouchot.

According to Bayes' theorem a test is a transformer of pre-test probability.[4] Test results are only interpretable if this probability can be estimated, however roughly. The definition of any partition—be it two zone or three zone—has to consider the pre-test probabilities, which usually vary considerably within different settings or even in between single patients. In exercise electrocardiograms (ECG), for example, pre-test-probabilities for coronary artery disease in a young woman and an elderly man with chest pain may differ 50-fold—and so will the post-test probability of an identical test result.[5]

Division of Epidemiology and Biostatistics, Department of Social and Preventive Medicine, University of Bern, Finkenhubelweg 11, CH-3012 Bern, Switzerland. E-mail: battag@ispm.unibe.ch

The distribution of disease stages and co-morbidity within a population influences test performance. Thresholds are applicable only to populations similar to the one in which they have been 'calibrated'.[6] The example of tuberculin skin testing used by Coste and Pouchot themselves illustrates this problem. Can the properties of this test be applied to people with human immunodeficiency virus (HIV) if it has been calibrated on a population without impaired immune systems? It is therefore important to calibrate test parameters in comparable settings to minimize the problem of different distributions of disease stages and co-morbidities.[7] Furthermore, a single test can often be used in the diagnosis of different diseases so the usefulness and width of the grey zone varies according to the condition the test is used for.

In practice clinicians hardly ever interpret results of continuous tests as being only 'normal' or 'pathological'. They always take into consideration 'how positive' or 'how negative' the result is. Will the suggested grey zone improve diagnosis or may it even mislead less-experienced clinicians into a false diagnostic security? By defining grey zones that take into account the reservations made above the limits may be unacceptably widened. In the two examples used by the authors one-third or even more of the possible test results could come to lie in the grey zone. This jeopardizes the explanatory power of this concept.

In spite of all these drawbacks we think that the proposal to enrich the interpretation of test results by measuring continuous parameters in shades of grey is an important step in the right direction. In their graphs Coste and Pouchot illustrate the role of likelihood ratios in defining multiple threshold-calibrations. They show how likelihood ratios allow for the estimation of a test result to enhance or reduce pre-test probabilities. Why not limit ourselves to defining likelihood ratios—measures not influenced by pre-test probability—for different partitions of test results? For example, it may be helpful to know how much the pre-test probability of coronary artery disease is modified by different degrees of ST-segment depression in stress ECG testing.

Likelihood ratios for different thresholds do not solve all the problems mentioned so far but they allow for a more easily interpretable differentiation of the black–white concept. We agree with Coste and Pouchot that the presentation of quantitative test results in 'black and white' categories is unnecessarily restricting. The introduction of different shades of grey may help to improve the interpretation of diagnostic test results and, more importantly, improve clinical outcomes.

Acknowledgement

We are grateful to Nicola Low and Matthias Egger for helpful comments on earlier drafts of this commentary.

References

[1] Coste J, Pouchot J. A grey zone for quantitative diagnostic and screening tests. *Int J Epidemiol* 2003;**32:**304–13.

[2] Feinstein AR. The inadequacy of binary models for the clinical reality of three-zone diagnostic decisions. *J Clin Epidemiol* 1990;**43:**109–13.

[3] Pauker SG, Kassirer JP. The threshold approach to clinical decision making. *N Engl J Med* 1980;**302:**1109–17.

[4] Sox HC, Blatt MA, Hinggins MC, Marton KI. *Medical Decision Making*. Boston: Butterworth-Heinemann, 1987.

[5] Diamond GA, Forrester JS. Analysis of probability as an aid in the clinical diagnosis of coronary-artery disease. *N Engl J Med* 1979;**300:** 1350–58.

[6] Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med* 1978;**299:** 926–30.

[7] Irwig L, Bossuyt P, Glasziou P, Gatsonis C, Lijmer J. Designing studies to ensure that estimates of test accuracy are transferable. *BMJ* 2002; **324:**669–71.