not be seriously considered. However, adopting scenario situations based on specified sensitivity and specificity values can be misleading and pessimistic; specificity and sensitivity are not absolutes, they are restricted in their range of possible values, depending on the measured exposure prevalence. The key to considering whether bias is present is not in terms of ‘differentiality’ of specificity and sensitivity but in terms of differentiality of their chance-corrected counterparts.

A Stata program to draw graphs like the one in the Figure is available from me on request.

Reply to Roger Marshall
From SANDER GREENLAND,1,2 MATTHEW P FOX3,4* AND TIMOTHY L LASH4,5,6

While appreciating Dr Marshall’s comments1 on our manuscript, we are concerned that his approach understates the potential distortions engendered by misclassification. In particular, while we see no flaw in Dr Marshall’s algebra, we disagree strongly with his conclusion that

adopting scenario situations based on specified sensitivity and specificity values can be misleading and pessimistic; specificity and sensitivity are not absolutes, they are restricted in their range of possible values, depending on the measured exposure prevalence. The key to considering whether bias is present is not in terms of differentiality of specificity and sensitivity but in differentiality of their chance-corrected counterparts.

Sensitivity and specificity are indeed not absolutes, which is one reason we treated them as random variables.2 Nonetheless, Dr Marshall’s model1,3 and conclusion focus on algebraic rather than contextual (substantive) properties. We believe this focus is potentially misleading and leads to overly optimistic conclusions. In particular, we think he has overlooked optimistic biases in his own literature analyses.4

To say that a classification method is nondifferential with respect to disease means it has identical operating characteristics among cases and controls, so that sensitivity and specificity do not vary with disease status. We expect this property to hold when the mechanisms that determine the classification are identical among cases and controls. In particular, we expect nondifferentiality when the disease is unrelated to exposure measurement. This expectation is reasonable when the mechanisms that determine the classification precede the disease occurrence and are not affected by uncontrolled risk factors. Thus to say there is no recall bias (as when exposure data are collected from records that predate the outcome) means that neither disease nor uncontrolled risk factors prompt more false-positive responses from cases than from controls. Put more abstractly, nondifferentiality means that the classification X is independent of the outcome Y (i.e. Y conveys no information about X) conditional on the true exposure E and whatever has been controlled, not that some ‘chance-corrected’ measure is the same in cases and controls. This condition may be rarely met, but at least it can be criticized based on qualitative mechanistic considerations.

It is intuitions and judgments about the role of the outcome Y in exposure classification errors that are the basis for priors about measurement behaviour. These judgements provide one reason to express such priors in terms of sensitivity and specificity, despite the greater clumsiness of algebra and statistics when compared to expressions based on predictive values5 or on ‘chance-corrected’ measures like Q1 and Q0. Furthermore, statisticians have traditionally focused on sensitivity and specificity because, when combined with other assumptions (such as independence of errors across variables), their nondifferentiality can preserve validity of the test of the null,6 although this result is often over-interpreted.7

With regard to Dr Marshall’s empirical meta-study of bias,4 we think he has over-generalized limited findings from a highly select group of 16 reports found in the original source.8 First, we should expect that the very presence of a validation substudy (required for inclusion in the meta-study) would mark the study as atypical, e.g. conducted with greater care and in a population amenable to such a substudy. Second, studies of exposures in which a near-gold standard

References

Advance Access publication 14 November 2006

 doi:10.1093/ije/dyl226

LETTERS TO THE EDITOR 1589

1 Department of Epidemiology, University of California, Los Angeles, CA, USA
2 Department of Statistics, University of California, Los Angeles, CA, USA
3 Department of International Health, Boston University School of Public Health, Boston, MA, USA
4 Department of Epidemiology, Boston University School of Medicine, Boston, MA, USA
5 Geriatrics Section, Boston Medical Center, Boston, MA, USA
6 Department of Medicine, Boston University School of Medicine, Boston, MA, USA

* Corresponding author. E-mail: mfox@bu.edu
may be available (e.g. prescription drugs and medical treatments), which included most of the studies, are no guide to general practice, since the precise nature of the exposure involved facilitates accurate recall. Third, ‘validation’ studies of other measurements compare two flawed measures and are therefore really studies of agreement. The compared measures often have correlated error components, and will share any bias in selection for the ‘validation’ substudy. We thus expect such studies to overstate validity, even when analysed as agreement studies (rather than as validation studies), and find it unsurprising that the results from such studies appear optimistic. This concern is especially worrisome for exposures for which genuine validation data have never been and probably never will be available, such as lifetime histories of nutrient intake, exercise, and environmental and occupational exposures.

Regardless, to claim that the relation of \( Q_1 \) and \( Q_0 \) to bias should lead us to express nondifferentiality in terms of \( Q_1 \) and \( Q_0 \) is almost circular: independence of \( Q_1 \) and \( Q_0 \) from disease status \( Y \) directly implies no bias. The essential scientific question is as follows: what are the substantive bases for such an assumption? For example, what exposure-assessment mechanisms (as opposed to algebraic relations) should lead us to expect this independence condition to hold? One might base an analysis using the studies in Marshall\(^4\) as a prior for \( Q_1 \) and \( Q_0 \), but this usage could be quite misleading, given the limitations of the studies and that both \( Q_1 \) and \( Q_0 \) depend on prevalence of \( X \), which is bound to vary across study populations, and even across strata within a study. The mistake here is similar to naively applying predictive values from one population to another.

In summary, we doubt that Marshall’s theoretical and empirical observations provide valid guidance to the impact of unknown classification error in typical studies, especially in lifestyle, environmental, and occupational epidemiology. When no genuine validation data are available, sensitivity and specificity may often be the most natural expressions on which to base prior expectations, including expectations about the impact of the outcome on classification errors. When building a prior for errors (such as illness effects on memory), the exposure prevalence in \( Q_1 \) and \( Q_0 \) or in predictive values renders these measures less generalizable and removes them further from the mechanisms affecting those errors.

References


doi:10.1093/ije/dyl227
Advance Access publication 26 October 2006

Dimensional errors of metaphorical measurements. Can they be resolved?

From GIO BATTA GORI

A recent commentary by James Marshall in this journal condenses a number of reports published in the last few years and is concerned, at long last, that imprecision ‘...may well consign epidemiological inquiry to the scientific sidelines.’\(^1\) The central unease is that precision and accuracy of measurement are essential to scientific qualification, coupled with the realization that well-behaved imprecision, of the sort Marshall links with postulations of Bross\(^2\) half a century ago, is virtually unknown in today’s epidemiologic practice. At large, imprecision is seen as hanging on the synthesis of the random and systematic errors of primary data (e.g. exposures) and of estimates of and corrections for biases and confounders.

The problem is that many of the primary data cannot actually be measured and their values have to be simply estimated, especially in observational studies of chronic conditions, which has led to recommendations that study conclusions should be complemented with sensitivity analysis using Monte Carlo and Bayesian techniques. Yet, the question remains of whether a range of uncertainty is more useful or more confusing for decision makers confronted with choices. For instance, the Fox et al. sensitivity analysis that Marshall quotes presents a range of odds ratios from 1.11 to 10.7, namely an hypothetical span that could be wider or narrower depending on the arbitrary choice of sensitivity and specificity parameters.\(^1,3\) Thus the