Reviewer's report

Title: Extreme Verification Bias in Paired Continuous Tests Can Cause Researchers to Choose the Wrong Screening Modality

Version: 1 Date: 21 July 2008

Reviewer: Jørgen Hilden

Reviewer's report:


Extreme verification bias in paired continuous tests can cause researchers to choose the wrong screening modality, by Glueck et al.

This is a nicely written, well-designed paper, which amply fulfills the 8 requirements listed in the Guidance for Reviewers. I enjoyed receiving it for review. I have only a few minor quibbles concerning terminology and one principal worry emanating from how I would have attacked the problem if I were to analyze the same problem.

The comments all fall in the category Discretionary Revisions (apart from the error pointed out in (M7) below). However, I hope the authors will feel obliged to help me resolve the problem summed up in the final lines of (P2).

Terminology

T1) 'Extreme verification bias’ = (extreme verification) bias, or extreme (verification bias)? The novice reader may wonder whether we are examining the bias associated with an extreme(-ly skewed) verification procedure or it is just verification bias of an extreme kind. In fact it is both – so here the quibbles come to a halt – but suppose it had be so that the most extreme bias was obtained when the verificative asymmetry was not quite extreme: then the grammatical difference between the two readings would make a difference. For such reasons the authors owe it to the reader to make the reading explicit, e.g. by writing: Abstract … . Previous authors have described the concept of ‘extreme verification bias’, extreme in the sense that disease status is verified … only when … .

T2) And look at ‘paired extreme verification bias,’ which inherits some of the ambiguities of ‘King George V Square Car Park’ (in Brisbane, Queensland) (5 square = 25? – or is it where George V used to park his square cars?).

T3) Returning to a serious note, the word ‘verification’ is also sometimes dangerous in screening contexts, and I think the authors should make sure that it is used in an unambiguous manner. The word can have three meanings: (1) being subjected to gold-standard testing (here one refers to the process, without hinting at its outcome); (2) confirmation of suspected disease (a screen-positive
individual is found in fact to have cancer); and (3) confirmation of whatever was
the initial impression (screen-negatives are proved true-negative, those
screen-positive are proved true-positive). At least, this is how I read the word,
and others may have it the same way.

T4) However, the authors use true about the actual underlying disease status (p.
5, top), and I shall follow that practice hereinafter. Just 8-10 lines further down we
have ‘true positives’ and ‘true negatives’ in the new sense, different from the
standard one I employed in (T3) above. Some readers may be pushed off the
track. Why not simply write ‘true cancer’ and ‘true non-cancer’ now that most
sections are written with cancer screening in mind?

T5) P. 6, top: the epidemiological term “rate” is misused.

Minor points
M1) Model section, second parag., and later: the same cutoff notation (beta) as
used for both tests. This is fine – and must have made the programming easier,
but the reader needs a warning that this is just a programming trick, made
possible by the fact a lowering of beta can be programmed as an increase in the
Corresponding mu. Hope I got this right! Anyhow please go over the text and
make sure that the modeling interrelationships between the two mu’s and the
beta’s is clear.

(Conceptually there are five locational quantities, of which one is redundant due
the translational invariance, and the elimination of one parameter can be
described in several ways; intelligibility depends on how one describes the
elimination.)

A second thought: Readers may have less difficulty following the mathematical
set-up if the symbol beta (but not beta_B) is replaced with a letter like x or c. The
point is that the imagined situation is then described by Greek-letter parameters,
whereas a purely auxiliary variable, viz. the cutoff variable generating the ROC
trajectory, which is essentially a variable-of-integration, is denoted by a mundane
x (or, as it serves as a cutoff, c).

M2) Model section, first parag., line ~6: it is true that the Normality assumption,
and its bivariate version in particular, is an idealization, but the text forgets to
mention that it suffices that there exists a (monotonely increasing) transformation
of the observation scale that ensures a Normally distributed transform in the
one-test case; similarly, the analysis programmed in the paper applies whenever
there exist a transformation for screening measurement T1 and another for
screening measurement T2 so that the two transforms end up having a joint
bivariate normal distribution. – This applies whether we are confined to the
equal-variance setting as here or not.

M3) Apropos, given the free choice of a common SD, the authors chose it to be
1, with variance=1 as a consequence. The text on p. 5 (mid-page) suggests that
the common variance was chosen, with SD = 1 as a consequence. Less natural,
isn’t it? The math-stat giant’s footprint, I guess. Simply rephrase in terms of
standard deviation.
M4) P. 6, mid-page: “cutoff value for the confirmatory test” would be understood differently in other contexts; please prevent misunderstandings by writing “cutoff value for referral to confirmatory testing.”

M5) P. 6: Bivariate cumulative distribution functions involve a convention that not all readers may be familiar with. Spell out the defining inequalities.

M6) The headings of Tables 1 and 2: here I would prefer to see the beta inequalities spelled out in plain words after an “i.e.” and the critical “with” replaced with “. The situations listed in this table are those where”. Incidentally, the authors may consider combining the two tables into an “upper” and a “lower part” of a single table. Also, to prevent another misunderstanding (some readers skim table captions), change “presence” of signs and symptoms so as to refer to what happens during the follow-up period. By assumption, signs and symptom are never present at screening time.

M7) Table 4 contains a couple of editing errors (Test 1 or Test 2? Whence 1.3?).

M8) In a printed version of the paper the two final figures would have to be deleted. They simply hold too little information.

Issues of principle

P1) Suppose an investigator, on behalf of a healthcare authority, carries out a study of the kind described, in the hope to obtain input to deliberations concerning a proposed type of screening. Let us disregard problems associated with non-participants. An infallible confirmatory test is carried out when one or both screening tests is positive. To make things simple, assume data on interval cases are not available. If the tests are binary, (s)he will obtain frequency data of the form shown below [square brackets: not available].

\[
\begin{array}{cccc}
T_2 (+) & T_2 (–) & T_2 (+) & T_2 (–) \\
T_1 (+) & A & B & T_1 (+) & a & b \\
T_1 (–) & C & [D] & T_1 (–) & c & [d] \\
\end{array}
\]

(Sum = 1)

With continuous or ordinal-scaled tests, the dataset takes this form for any pair of cut-points \((x_1, x_2)\) for the two tests.

Observed is also the total fraction, \(D+d\), of individuals with doubly negative screening results and hence unknown diseases status, though each of the terms is unknown. Obviously \((x_1, x_2)\) is also known.

Is the sensitivity of \(T_1\) > that of \(T2\)? Yes, if (and only if) \(B > C\). It is unnecessary to know \(D\) to answer that question. The same holds true of the specificity comparison.

In fact, if \((FN)\) and \((FP)\) denote the regrets associated with a false negative and a false positive outcome, respectively, then \(T_1\) superior to \(T2\) if and only if the
expected regret(T1) < expected regret(T2), or

\[(C - B)(FN) + (b - c)(FP) < 0.\]

It is unnecessary here to know D and d. Like fractions A and a, they cancel out because, whatever test is adopted, the contingent D will receive the wrong treatment (delayed cancer diagnosis, say) and the contingent d will receive the null treatment that is appropriate for the non-diseased.

Likewise, the expected flow through the confirmatory service (often a bottleneck in screening projects) is available to the analyst as \(A + B + C + a + b + c = 1 - (D + d)\). Admittedly, some of the wider ramifications of screening being introduced may depend on the D:d ratio being estimated, e.g. from past records. They will allow the investigator to estimate the paper’s \(r = A + B + C + D\), and hence D.

Apart from this, it appears that the statistical input to a clear and justified comparison is available. Why then does the ms. find situations in which the wrong test is preferred? (The ‘utilistic’ input, (FN) and (FP), must be had from external sources; but in the ms. they also remain external to the test comparison. Nor does access to interval-case data make a difference, as the paradox of the ms. does not depend on psi being > 0, if I read the text correctly.)

P2) As far as I can see the paradox rests on the fact that (P1) above considers each adoptable cutoff pair separately, whereas the criterion in the paper, expressed as the ROC AUC, rests on what is essentially an integration over all possible cutoffs (beta values) – despite the fact that only one, hopefully a near optimal one, is to be selected by the healthcare authority involved.

[The screen decision is binary, so perhaps one should devise a binarized version of the AUC, i.e., something like \(\text{INTEGRAL}((\text{sens} + \text{spec}/2) w(\text{cutoff})) d(\text{cutoff})\) for some novel weight function \(w(.)\), since the AUC of a binary test is \((\text{sens} + \text{spec})/2\). But even that trick would not make the paradox go away, I believe.]

These considerations prompted me to go through the arguments that the authors use in deciding on the AUC as their key statistic.

P. 14: “The area under the curve (AUC) is a measure of the diagnostic accuracy of the test.” Yes, and a popular one, but not necessarily the appropriate one in a specific context. This, then, may be one of the instances where the AUC is misleading – and provably so, perhaps. [See also (P3) below.]

P. 3: “When differentially biased estimates … are used … the resulting areas under the ROC … are also incorrect.” This is the first mention of the AUC concept – and one that gives the reader the impression that the AUC is the only worthwhile statistic to calculate from an ROC curve. Are the authors biased against other measures of screening efficacy than the AUC? That was my first thought when I read this passage – and I still have a nagging feeling that my suspicion is right.

P. 4: The study design section imagines an investigator in the situation of (P1) above. But the text immediately says that he or she will focus on AUCs. No argument is given for this choice, except possibly the implicit one that some investigators have done so and had their fingertips burnt. P. 8 (mid-page)
reiterates, without qualification, that the “study investigator makes the decision as to which test is better based on the difference in AUC.”

Returning to the title of the ms. (which cannot be criticized), it has the form ‘Study design XYZ can cause researchers to make the wrong choice.’ Implicit is the use of AUCs, so perhaps a tail should be appended to the title: “... when they trust AUCs.” The authors probably wouldn’t object to that (except that the title then becomes fairly long). To me, however, the title would then be of the form: ‘Study design XYZ can cause researchers to make the wrong choice if they use the wrong method.’ Now, anything can go wrong with a wrong method, so the latter statement is a simple truism. Its message is void. Therefore:

It is absolutely essential to figure out whether the unfortunate fact the authors are pointing out is due to an inherent shortcoming of the design envisaged – or due to the mode of analysis envisaged. I.e., due to people mindlessly continuing to use a method that (has served us well but) is inappropriate for the purpose at hand.

P3) Examples probably exist in the literature where the AUC has produced a patently wrong answer. Assuming it turns out that the present analysis offers a new example, it would be extremely nice if someone would collect the relevant literature (or has that been done?). The simplest example I am aware of is one presented in the 1991 Festschrift to Lee B. Lusted (ref.: J. Hilden, Medical Decision Making, vol. 11, p. 95-).

P4) The AUC analyses presented here do not involve any considerations of gain and loss (cf. (FN) and (FP) above). Suppose one extra item were added to the hypothetical investigator’s agenda (on p. 4), namely that, in making the test comparison, the investigator will be aware that overlooking a cancer is much worse (and much more costly to all parties) than follow-up on a false suspicion of cancer. I.e., (FN) >> (FP). What consequences would that have? Or is that entirely external to the philosophy of the ms.?

**Level of interest:** An article whose findings are important to those with closely related research interests

**Quality of written English:** Acceptable

**Statistical review:** Yes, and I have assessed the statistics in my report.

**Declaration of competing interests:**

I declare that I have no competing interests.