Author's response to reviews

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

Authors:

Deborah H Glueck (Deborah.Glueck@ucdenver.edu)
Molly M Lamb (hgl9@cdc.gov)
Colin I Odonnell (Colin.Odonnell@ucdenver.edu)
Brandy M Ringham (Brandy.Ringham@ucdenver.edu)
John T Brinton (John.Brinton@ucdenver.edu)
Keith E Muller (Keith.Muller@biostat.ufl.edu)
John M Lewin (jlewinn@divrad.com)
Todd A Alonzo (talonzo@childrensoncologygroup.org)
Etta D Pisano (etpisano@med.unc.edu)

Version: 2 Date: 11 November 2008

Author's response to reviews: see over
Dear Dr. Norton:

My co-authors and I very greatly appreciate the detailed and thorough reviews that we received from all three reviewers. The conscientiousness and attention to details of the reviewers was outstanding. I also appreciate your patience during the time it took me to revise the manuscript. The suggestions of the reviewers have greatly strengthened the manuscript.

After I had made all the revisions requested by the reviewers, my two radiologist co-authors, Drs. John Lewin and Etta Pisano, made a further round of suggestions. Their suggestions made the results more clinically relevant, and thus more acceptable to study designers.

The manuscript number is 3318613332079587. The manuscript used to have the title “Extreme Verification Bias in Paired Continuous Tests Can Cause Researchers to Choose the Wrong Screening Modality”. Following the suggestions of the reviewers, the manuscript title has now been changed to “Bias in trials comparing paired continuous tests can cause researchers to choose the wrong screening modality”.

A detailed note giving a point-by-point response to the concerns of the reviewers follows this cover letter. As per the request of the editor, I have also restructured the manuscript to conform to BMC style, reformatted the abstract, included a competing interest statement, and added an authors’ contributions section.
Three new authors have been added to the manuscript. All made strong intellectual contributions to the revised manuscript.

I appreciate your consideration of this revised manuscript. If I can be of any further assistance, please feel free to contact me. I look forward to hearing from you.

Sincerely,

Deborah Glueck, Ph.D.

Assistant Professor
Department of Biostatistics
Colorado School of Public Health
University of Colorado Denver
Tel: 303-315-7227
Fax: 303-315-3183
E-mail: Deborah.Glueck@ucdenver.edu
RESPONSE TO THE REVIEWERS

We very much appreciate all the comments by the reviewers, and have attempted to make the requested clarifications and corrections. We feel that the comments of the reviewers have greatly improved the manuscript. For the convenience of the editors and the reviewers, we include below the comments of the editors and reviewers (in italics), and our response (in standard text).

Editor:

Abstract - Abstracts must be structured into Background, Methods, Results, Conclusions. Please remember to also update the Abstract details on the submission page.

The abstract has now been restructured.

Competing interests - Please include a Competing interests section between the Conclusions and Authors' contributions.

A competing interests section has now been added, as requested.

Authors' contributions: Please include an Authors' contributions after the Competing interests.

We have now added an authors' contributions section.

Reviewer 1: Jørgen Hilden

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?).

We agree with Professor Hilden's points, and have now renamed the bias to "paired screening trial bias", which should hopefully be both more grammatical and clearer. We
have also included in the abstract the following paragraph, which should explain the definitions of the various components of paired screening trial bias.

The limitations of this study design lead to a bias in the ROC curves we call paired screening trial bias. This bias reflects the synergistic effects of inappropriate reference standard bias, differential verification bias, and partial verification bias. The absence of a gold reference standard leads to inappropriate reference standard bias. When different reference standards are used to ascertain disease status, it creates differential verification bias. When only suspicious screening test scores trigger an accurate secondary test, the result is a form of partial verification bias.

_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.

We agree again, and have made strenuous efforts to remove the words "confirmatory" and "verification" wherever possible. They do appear in the text when we cite from Whiting et al, 2004, as below, as we did not want to change standard nomenclature. We have tried to clarify in what sense we mean verification, shown in the underlined text below.

The limitations of this design leads to a previously undescribed bias we call paired screening trial bias. This bias results from the synergistic effects of inappropriate reference standard bias, differential verification bias, and partial verification bias [4]. Here, verification is used to describe the process of ascertaining the disease status. In classical partial verification bias, only some participants undergo determination of disease status. A variant of partial verification bias is extreme verification bias, when only strongly abnormal results on one of the screening tests lead to secondary testing [5]. In the paired screening trial design we discuss here, an effect similar to partial verification bias operates. A disease status is assigned for all participants, but determined with great sensitivity and specificity only for those with strongly abnormal results on
an initial screening test. Because different methods are used to ascertain disease status, depending on the results of the initial screening tests, the trial is subject to differential verification bias. Finally, paired screening trials often yield fewer observed than true cases of disease. Some cases of disease are missed because the ascertainment of disease status is not perfect. Thus, the trial is subject to inappropriate reference standard bias. All three of these biases interact to inflate the sensitivity and to slightly deflate the specificity, in potentially differential amounts for each screening test.

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?

We agree. We have changed the text to read 'participants with disease' and 'participants without disease' as shown below.

First, we assume that the results of screening Test 1 and screening Test 2 have a bivariate normal distribution for the participants with disease, and a potentially different bivariate normal distribution for the participants without disease.

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

We have changed the text to "prevalence" instead, wherever it appeared.

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.)
We agree. We added in the following paragraph, which we hope is intelligible, and useful for the reader. Let us know whether you think the addition of the paragraph clarifies the issue, or confuses the reader.

For convenience in the derivation, we use the same value of the threshold for recall, $\theta$, for both screening tests. Because ROC analysis is invariant to translation, choosing the same values of $\theta$ for each screening test, and then shifting the means of the screening test scores has the same mathematical result as choosing different values of $\theta$ for each screening test.

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, $c$).

A good notational point. We have made the change, as suggested. We used $x$ instead of $c$ as $C$ was already in use in the paper. At that point, it made sense to switch from $\beta_B$ to $\theta$, since $\beta$ no longer had a meaning, and we could save the reader a subscript.

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 $T_1$ and another for screening measurement $T_2$ 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.

Thanks for pointing this out. We have added the reference to the paper by Hanley, and some discussion of this point to the discussion section, as follows.

In real paired cancer trials, the scores have a conditional probability structure driven by the fact that real observers miss cancers (and score a screening test as if no disease were present), and see cancer where there is none (and then score a non-cancerous finding as abnormal). The resulting distribution of scores is far from the bivariate normal distribution we assumed.

There is some theoretical justification that our results will still hold even if the data are non-normal.
Hanley [12] points out that single test ROC analysis is robust to the violation of the normality assumption if there exists a monotonely increasing transformation of the test scores that yields a normally distributed result. Thus, the results described in the paper should hold whenever there is a transformation for screening Test 1, and another for screening Test 2 so that the transformed data has a bivariate normal distribution.

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.

We rephrased as follows.

We write $\Phi(x)$ to indicate the cumulative distribution function of a normal distribution with mean 0 and standard deviation 1, evaluated at the point $x$, and $\Phi(x, y, \rho)$ to indicate the cumulative distribution function of a bivariate normal distribution with mean vector $[0, 0]$, standard deviations both 1 and correlation $\rho$, evaluated at the points $x$ and $y$.

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.”

Thanks. The line now reads

Each test score could fall above or below $\theta$, the threshold value for referral to the invasive, yet sensitive and specific secondary test.

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

The line now reads.

That is, if $X$ and $Y$ have a bivariate normal distribution, we write $\Phi(x, y, \rho)$ to indicate

$$\Pr(X \leq x \text{ and } Y \leq y | \sigma_X^2, \sigma_Y^2 = 1, \rho).$$

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 symptoms are never present at screening time.

Combining the tables results in an enormous table that we reckon would disturb the copy-editor. We do not want to bother the copy-editor, and left the tables separate.

The caption for both tables has been changed. The new captions follow.

Table 1: Scores and results for Test 1 and Test 2, presence of signs and symptoms during the follow-up period, observed disease status and true disease status. The situations listed in this table are those where $x < \theta$, i.e. that the test score is less than the threshold that leads to referral to the invasive, yet sensitive and specific secondary test.

Table 2: Scores and results for Test 1 and Test 2, presence of signs and symptoms during the follow-up period, observed disease status, and true disease status. The situations listed in this table are those where $x > \theta$, i.e. that the test score is greater than the threshold that leads to referral to the invasive, yet sensitive and specific secondary test.

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

We have corrected the errors. The revised caption now reads as follows.

Table 3: True disease status and Test 2 results. Test 2 is + if the score on Test 1 is greater than $x$.

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

We have now deleted them.

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].

<table>
<thead>
<tr>
<th>True disease</th>
<th>= Not diseased</th>
</tr>
</thead>
<tbody>
<tr>
<td>T2 (+)</td>
<td>T2 (–)</td>
</tr>
<tr>
<td>T1 (+)</td>
<td>A</td>
</tr>
<tr>
<td>T1 (–)</td>
<td>C</td>
</tr>
<tr>
<td>(Sum = 1)</td>
<td></td>
</tr>
</tbody>
</table>

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 $T_2$? 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 $T_2$ if and only if the expected regret($T_1$) < expected regret($T_2$), 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.)

Professor Hilden considers a situation with the following assumptions.

1. Each screening test is binary
2. A positive result on one or both screening tests leads to an infallible secondary test, which enables the researcher to determine the true disease status.

He points out correctly that in this situation, one can make correct inference about the sensitivity of $T_1$ and $T_2$. 

He then asks "Why then does the ms. find situations in which the wrong test is preferred?"

We thank Professor Hilden for his comments. His question, and his suggestion of an alternative trial design pointed out a weakness in our discussion. We have now substantially rewritten the discussion, to explain why we chose to focus on the trial design and testing procedure considered in the manuscript. The pertinent paragraphs follow.

In this paper, we define a new type of bias that is a result of the interaction between a particular design for a paired screening trial, and the choice of a particular statistical test. Specifically, the bias occurs when the diagnostic accuracy of two continuous tests are compared using area under the ROC curve in a design with two limitations. First, different methods are used to ascertain disease status, depending on the results of the initial screening tests. Secondly, only some subjects undergo an invasive, yet sensitive and specific secondary test. Thus, some cases of disease are missed because the method used to ascertain disease status for those who test negative on both initial screening tests may not be 100% sensitive.

Both the statistical test and the trial design we considered were modeled closely after recently completed and published trials [1, 2, 3]. These trials compared the diagnostic accuracy of two modalities for breast cancer detection. Although authors have suggested the use of other statistical approaches to compare screening modalities [10, 11], the area under the full ROC curve remains the most commonly used test for paired screening trials in major American journals [1, 2, 3].

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 INTEGRAL{(sens+spec/2)w(cutoff)}d(cutoff) for some novel weight function w(.), since the AUC of a binary test is (sens+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.

No, not biased, but simply inarticulate. We failed to explain why we focused on this trial design, and this measure of screening efficacy.

We focused on both the trial design and the statistical test because they are commonly used.

We have hopefully now explained this in the text, as in this section in the Background.

Paired trials designed to compare the diagnostic accuracy of screening tests using area under the receiver operating characteristic (ROC) curve may fall victim to a strong bias that renders the conclusions of the trial incorrect. In English, "bias" often has a pejorative connotation, implying that those who conduct the study prefer one scientific conclusion, rather than another. We use the term "bias" in the epidemiological and statistical sense, as the difference between the results obtained in a study, and the true results.

The bias occurs because limitations in the trial design may differentially affect the area under the ROC curve for each screening test. Many competing statistical approaches have been suggested for comparing the diagnostic accuracy of two continuous tests - consider. The area under the ROC curve, because it continues to be used as the standard in prominent medical journals [2, 3, 9].

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.

We tried, but could not get AUC in the title under the word limit. Hopefully, the changes in the manuscript will make it clear why we focused on AUC.

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-).

We agree, although this is beyond the scope of the present paper.

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.?

An interesting question for our future research. Thanks for the suggestion.

For this paper, rather than considerations of gain and loss, we chose to focus instead on area under the curve, because this is the metric currently in use in most recently published paired cancer screening trials.
Reviewer 2: Corne Biesheuvel

The only comment I have is that it would be extremely helpful for the reader if the authors provide an extra figure that explains this phenomenon. I suggest something like figure 1 in the paper 'Case-control and two-gate designs in diagnostic accuracy studies' by Rutjes AW, Reitsma JB, Vandenbroucke JP, Glas AS, Bossuyt PM. Clin Chem. 2005 Aug;51(8):1335-41 or figure 2 in the paper 'Randomised comparisons of medical tests: sometimes invalid, not always efficient' by Bossuyt PM, Lijmer JG, Mol BW. Lancet. 2000 Nov 25;356(9244):1844-7.

We agree. We have added a new figure with a flow chart similar to those in the papers suggested. The new figure, Figure 1, now shows a flow chart of the hypothetical study.
Reviewer 3: Mariska M. G. Leeflang

1. The paper addresses a very relevant point, although the authors only investigate a small part of the problem and they do it in a very simplistic way (which is not necessarily bad, but addresses in this case the problem not to its full extent). They do for example not show how the situation is when clinicians would base their decisions on the optimal sensitivity and specificity for each test, in stead of the complete ROC curve. Furthermore, I would like to know what happens when the (true) ROC curves of both tests cross each other in stead of running parallel.

We did not explain well why we chose area under the ROC curve. The abstract, background, and discussion section have all been rewritten to stress why this metric was chosen.

Furthermore, I would like to know what happens when the (true) ROC curves of both tests cross each other in stead of running parallel.

The true ROC curves of the tests cross each other under the following circumstances. For Test 1, define the variance of the test results for the cases to be $\sigma_{C1}^2$, and for the non cases, let the variance be $\sigma_{N1}^2$. Similarly, for Test 2, define the variance of the test results for the cases to be $\sigma_{C2}^2$, and for the non cases, let the variance be $\sigma_{N2}^2$. In a paired design, the receiver operating characteristic curves will cross when $\sigma_{C1}^2 / \sigma_{N1}^2 \neq \sigma_{C2}^2 / \sigma_{N2}^2$.

As pointed out in Zhou et al., 2002, p.172, "if two ROC curves cross, one diagnostic test may be superior for some sets of false positive rates, and inferior for others, even though the areas may be similar". They also warn that if receiver operating characteristic curves cross, " comparing the area under the entire curve may give misleading results " (Zhou et al., 2002, p.188).

Zhou reflects the widely held consensus that when the curves cross, using area under the receiver operating characteristic curves as the metric is a poor idea.

We made the assumption in the manuscript that the variances of all the scores (those for both Test 1 and Test 2, for participants with disease and participants without disease) were equal, in order to prevent the curves from crossing.

We added the underlined sentence to the text to clarify the results of this assumption.

First, we assume that the results of screening Test 1 and screening Test 2 have a bivariate normal distribution for the participants with disease, and a potentially different bivariate normal distribution for the participants without disease. …

Suppose that the variance is the same for both distributions, $\sigma^2$. The equal variance assumption prevents the ROC curves from crossing.

Even though it is outside the scope of our manuscript, we were curious about the answer to your question. What does happen to the bias when the curves cross?
Reviewing the derivation, it became clear that one could rewrite the formulas in the end of the paper, with appropriate scaling to correct for the variance. That is, now assume a binormal model for participants with disease, with means $\mu_{C1}$ and $\mu_{C2}$. For Test 1, define the variance of the test results for the cases to be $\sigma_{C1}^2$, and for the non cases, let the variance be $\sigma_{N1}^2$. Similarly, for Test 2, define the variance of the test results for the cases to be $\sigma_{C2}^2$, and for the non cases, let the variance be $\sigma_{N2}^2$. Finally, suppose the covariance between Test 1 and Test 2 scores for participants with disease is $\rho_C \sigma_{C1} \sigma_{C2}$, and the covariance between Test 1 and Test 2 scores for participants without disease is $\rho_N \sigma_{N1} \sigma_{N2}$. Using these assumptions, one can rewrite the formulae in Table 3 as follows.

<table>
<thead>
<tr>
<th>True Disease Status</th>
<th>Test 1</th>
<th>+</th>
<th>(1 - $r$){1 - $\Phi$[(x - $\mu_{N1}$)/$\sigma_{N1}$]}</th>
<th>Test 2</th>
<th>-</th>
<th>(1 - $r$){1 - $\Phi$[(x - $\mu_{N1}$)/$\sigma_{N1}$]}</th>
</tr>
</thead>
<tbody>
<tr>
<td>(+) r{1 - $\Phi$[(x - $\mu_{C1}$)/$\sigma_{C1}$]}</td>
<td>- r$\Phi$[(x - $\mu_{C1}$)/$\sigma_{C1}$]</td>
<td>(+) (1 - $r$){1 - $\Phi$[(x - $\mu_{N1}$)/$\sigma_{N1}$]}</td>
<td>- (1 - $r$){1 - $\Phi$[(x - $\mu_{N1}$)/$\sigma_{N1}$]}</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The formulae in Tables 4, 5 and 6 can be rewritten similarly.

One can see that the bias remains strong in this case, as shown in the figure below.

2 In the introduction, the authors do not explain what paired extreme verification bias is. They should do so in either the first or the second paragraph of the introduction. They expect the readers to know or read the Begg-1987-paper, but the paper should be generally understandable without having to check the cited papers. Not all readers will immediately know what is meant with this bias, so a description of the corresponding (biased) study design may help. The authors should at least explain that in the case of verification bias, only a (selective) part of the tested individuals will be verified.
We agree. We have rephrased both the abstract and the introduction. Both now include the following lines:

The limitations of this study design lead to a bias in the ROC curves we call paired screening trial bias. This bias combines the synergistic effects of inappropriate reference standard bias, differential verification bias, and partial verification bias. When a gold reference standard cannot be used, it leads to inappropriate reference standard bias. When different reference standards are used to ascertain disease status, depending on the results of the screening tests, it creates differential verification bias. When only suspicious scores are verified, the result is partial verification bias.

3. In the Study design section, the authors state that the true disease status is determined by a confirmatory test or follow up. This is what I would call differential verification bias (Whiting, 2004), while the previous text implicates that the paper was about partial verification bias (see also Whiting, 2004). These two biases may have different effects and that is not addressed in this paper.

The Whiting et al., 2004 reference was very helpful, and led us to rephrase the description of the bias, as shown in the response to comment number 2 above.

*Can the authors tell us something about the results when follow-up was not included in the analyses at all?*

When follow-up is not included in the analysis at all, the assumptions shift strongly from those used in the breast cancer screening trials that inspired this work (Pisano et al., 2005; Berg et al., 2008).

A complete analysis of the case without follow-up for binary tests is presented in Alonzo and Kittelson, 2006. In this paper, the disease status of the subjects who do not have verification is considered unknown.


4. *A big factor in these analyses is the correlation between the two index tests. Could the authors please explain how the correlation between these two tests influences their results? What happens when there is much stronger correlation (which may often be the reality)? Or when there is no correlation at all?*
This is an important point which we had failed to address. We added the following text to the end of the results section:

*Paired screening trial bias* increases with the increase in correlation between the results of the index tests for participants with disease, $\rho_C$. The bias in the observed ROC curves increases because as the two index tests become more highly correlated, the number of observed cases of disease becomes smaller relative to the number of true cases of disease. When the two index tests are highly correlated, they essentially produce the same information as to whether a participant has disease. When the index tests are independent, each test makes diagnoses on its own that the other test misses. Thus, when the tests are independent, and $\rho_C$ is 0, the number of observed cases is highest, relative to the number of true cases. The percentage of participants receiving the infallible secondary test increases as $\rho_C$ decreases. The bias lessens as the true disease status is ascertained for more participants.

The series of three pictures shown below show the effect of the increase in correlation between the tests. We did not add the figure to the paper because of length restrictions. However, we did include the figure here because we felt that it illustrated our response to your question well.

5. The “Simpson’s rule trapezoidal numerical integration method” may not be the appropriate method for calculating AUCs, because it does not give confidence intervals around the AUCs. There are better, more formal ways of calculating AUCs in for example SAS that do give confidence intervals.

We used the Simpson's rule trapezoidal numerical integration method because we had derived the exact probabilities, and wished to know the area under the curve. The closed form integral of the expressions we had derived was not immediately obvious.
These missing confidence intervals are another crucial limitation in the paper. When researchers formally want to test or investigate whether one AUC is bigger than the other one, they will need confidence intervals to say something about the precision of their decision. In the example in this paper, the ROC curves of the two tests swap. And the AUC that is in the observed situation the biggest, is in reality the smallest. But confidence intervals may have shown that there was no significant difference between the two AUC at all. However, that is a question the authors do not address.

We agree that this was an oversight. In practice, a researcher would observe data, and then use a test to determine if the curves were different.

To clarify this important point, we added two paragraphs to the methods, and one to the results section.

The paragraphs in the methods sections now read as follows:

We calculate the theoretically correct ROC curves and AUC's (ignoring the error of integration), using our mathematical derivations. In a real trial, the study investigator would use a hypothesis test and a p-value to compare the difference in AUC's. Depending on the sample size chosen for the trial, the precision of the estimates and the accuracy of the decision may change.

To illustrate the effect of the bias, we present the theoretical results. To illustrate the effect of sample size on the precision of the estimates, we conduct a simulation. For the simulation, we suppose that the study investigator decided to test the null hypothesis of no difference between the areas under the curves, using a non-parametric AUC test for paired data [8], and fixing the Type I error rate at 0.05. To ensure adequate power, for a fixed set of parameters, we set the sample size so that 90% of the time, if the true state of disease were known, the null hypothesis would be rejected. For that fixed set of parameters and sample size, we simulate 10,000 sets of data. For both the true state of disease, and the observed state of disease, we record the magnitude of the differences in AUC's, and the decision whether to reject the null. The proportion of rejections for the true and observed data is estimated by the number of rejections, divided by 10,000. Ten thousand is chosen so the maximum half width for the confidence interval for the proportion rejected is no more than 0.01.

In the results section, we added the following paragraph:
Study investigators never observe the true state of nature. They observe data, and make estimates, the precision of which depends on the sample size. They decide which screening test is better using hypothesis tests. To see which conclusion the hypothesis tests would suggest, both for the true and observed disease status, we conducted a simulation. For the parameters of Figure 2, for a Type 1 error rate of 0.05, if the true disease status were known, a non-parametric test [8] would have 90% power with 33,000 participants. With the true disease status known, we would reject the null roughly 90% of the time. The remaining 10% of the time, we would conclude no difference in AUC between Test 1 and Test 2. If the true disease status were known, every time we rejected the null, we would conclude correctly that Test 2 is better than Test 1.

If we conduct the same simulation experiment from the point of view of the study investigator, for the experimental situation of Figure 2, we see only the observed state of disease. In that case, the study investigator will reject the null hypothesis only 71% of the time. The remaining 29% of the time, the study investigator will conclude that there is no difference in AUC between Test 1 and Test 2. The lower power is due to more variance in the observed data, compared to the true data. When the study investigator rejects the null, every time, she concludes incorrectly that Test 1 is better than Test 2.

7. The observed curve for test for test 1 shows a “point of inflection”. Do the authors have an explanation for this effect? And why do we see it in the curve for test 1 and not for test 2? And why not in the true curves?

These are all good questions. We added this paragraph as the second paragraph in the results section, to answer these questions.

As shown in Figure 2 and Figure 4, the observed curves have inflection points, where the slope changes. There is no inflection point in the true ROC curves for either test, because the formulae that govern the sensitivity and specificity for the true curves are the same no matter what the ROC cutoff points are (see Tables 3 and 4). By contrast, as shown in Tables 5 and 6, the formulae for the observed ROC curves change depending on whether the cutpoint is above or below $\theta$. This causes a change in slope for the observed ROC curve. The inflection point is more
obvious for Test 2 than for Test 1. The inflection point for Test 1 occurs at specificity of about 0.80, and is obscured in Figure 2. In general, as $\theta$ increases relative to the mean of the test score distribution, the point of inflection occurs at higher values of specificity.

Although we didn't have room in the manuscript, we added a series of pictures here to illustrate how the point of inflection moves left as $\theta$ increases.

**Minor Essential Revisions**

1. The abstract is a bit obscure for readers who are not familiar with the phenomenon of paired extreme verification bias. So please rewrite it. It may already help if the second sentence is put up front.

   We agree. The abstract has been rewritten.

2. I do not agree that in the case of verification bias the sensitivity will always be inflated and the specificity will always be deflated (as is stated in the introduction). Could that perhaps be illustrated with an example? Or do the authors have a reference to this statement?

   Sensitivity will be inflated and specificity will be deflated for the design described in this manuscript if those with more suspicious index test results are more likely to receive accurate disease status ascertainment. A detailed example with four two by two tables has been added to the results section. In addition, a citation to the Begg and Greenes, 1983 paper has been added in support.

3. The authors state that “a paired comparison of ROC curves is the most common trial design used to compare screening modalities”. Do they have a citation for this statement?

   We have added citations, as follows.
Although authors have suggested the use of other statistical approaches to compare screening modalities (Baker and Pinsky, 2001; Li et al., 2008), the area under the full ROC curve remains the most commonly used test for paired screening trials in major American journals (Lewin et al., 2002, Pisano et al., 2005, Berg et al., 2008).

4. Furthermore, do they have a citation for the “Simpson’s rule trapezoidal numerical integration method”?
We have added a citation. The revised sentence now reads

Simpson's rule numerical integration methods (p. 608, Apostol, 1969) with accuracy of 0.001 are used to calculate the area under the ROC curve (AUC) for each screening test.

5. The first sentence of the Results section is a conclusion, not a result.
Agreed. The first two sentences of the results section have now been deleted. The new first sentence reads

Our derivations demonstrate that observed ROC curve differs from the true ROC curve, with the amount of bias depending on the correlation between the screening tests for participants with disease, \( \rho_C \), the rate of signs and symptoms, \( \psi \), and the threshold for recall, \( \theta \).

6. On page 8, on top: “The slope of the ROC curve increases strongly at the cutoff point above which all test results are confirmed”. If you read the figure from left to right, the slope decreases instead of increases. So please reword or explain.
We reworded this, as shown in our response to comment 7 above.

7. If all patients undergo all tests, how is it then possible that “one test was verified more often than the other test” (page 8, bottom)?
This writing was unclear, and has been removed.

We added instead a few paragraphs, and one figure to attempt to explain the source of the bias. They follow.

To understand how and why this bias occurs, consider a single specificity value on the true and observed ROC curves shown in Figure 2. Choose the value of specificity where there is the greatest increase in observed sensitivity relative to true sensitivity, for Test 1. This occurs when specificity is 0.82. For a hypothetical study of 10,000
participants, and specificity of 0.82, the observed and true

2 × 2 tables for Test 1 and Test 2 are shown in Figure 3.

Figure 3: True and observed 2 × 2 tables for Test 1 and Test 2 for the point of maximum bias in sensitivity for Test 1†

<table>
<thead>
<tr>
<th>True Disease Status</th>
<th>True Disease Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test 1 + 34 1783 1817</td>
<td>Test 2 + 43 1783 1826</td>
</tr>
<tr>
<td>Test 1 − 66 8117 8173</td>
<td>Test 2 − 57 8117 8174</td>
</tr>
<tr>
<td>100 9900 10,000</td>
<td>100 9900 10,000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Observed Disease Status</th>
<th>Observed Disease Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test 1 + 34 1793 1827</td>
<td>Test 2 + 23 1793 1816</td>
</tr>
<tr>
<td>Test 1 − 11 8162 8173</td>
<td>Test 2 − 22 8162 8184</td>
</tr>
<tr>
<td>45 9955 10,000</td>
<td>45 9955 10,000</td>
</tr>
</tbody>
</table>

†Numbers were rounded to the nearest whole number.

Specificity for all four tables is about 0.82. However, the sensitivities vary.

For Test 1, the true sensitivity is 0.34, with observed sensitivity at 0.76. For Test 2, the true sensitivity is 0.43, with observed sensitivity at 0.51.
For this hypothetical example, the disease rate, \( r = 0.01 \); the chance that participants with disease would experience signs and symptoms within the year of follow-up, \( \psi = 0.1 \); the variance, \( \sigma^2 = 1 \). The means of the ROC distributions for cases for Test 1 and Test 2 were 2.1 and 1.1, respectively, and the means for non-cases for Test 1 and Test 2 were 1.6 and 0.35. The correlation between test scores for cases was fixed at 0.1, as was the correlation for non-cases. All test scores above 2.5 on either test, or participants who had signs or symptoms had an infallible secondary test to determine disease status. For participants with scores below 2.5 on both tests, a less accurate method was used to approximate disease status.

To understand how and why paired screening trial bias occurs, consider a single specificity value on the true and observed ROC curves shown in Figure 2. Choose the value of specificity where there is the greatest increase in observed sensitivity relative to true sensitivity, for Test 1. This occurs when specificity is 0.82. For a hypothetical study of 10,000 participants, and specificity of 0.82, the observed and true \( 2 \times 2 \) tables for Test 1 and Test 2 are shown in Figure 3.

Each one of the four tables uses a slightly different ROC cutpoint. For the observed table, Test 1 is positive if it exceeds 2.511; for the true table, Test 1 is positive if it exceeds 2.515. For the observed table, Test 2 is positive if it exceeds 1.269; for the true table, Test 2 is positive if it exceeds 1.265. The tables have different ROC cutpoints because they were chosen to have the same specificity, not the same cutpoint.

Also, the number of cases of disease observed in the study, 45, is much smaller than the true number of cases of disease in the population, 100. The observed number of cases of disease is smaller than the true number because not every participant undergoes the invasive, yet sensitive and specific secondary test, and thus some cases of disease are missed. The observed number of cases of disease is the denominator of the observed sensitivity. Because the denominator is smaller for observed sensitivity than for true sensitivity, the observed sensitivity is strongly inflated for both tests. When specificity is 0.82, the observed sensitivity of Test 1 is 0.72, with true sensitivity of 0.33.
For Test 2, the observed sensitivity is 0.52, with true sensitivity of 0.43.
Yet if the bias only affected the denominator, the inflation in sensitivity would be the same for both tests. After all, the same number of observed cases is used as the denominator for both tests. The differential inflation for Test 1 compared to that for Test 2 must be due to the numerator of the observed sensitivity.

For Test 2, the numerator of the observed sensitivity is the number of study participants who are positive on Test 2, and who are observed to have disease in the study. For Test 2, the numerator for observed sensitivity, 23, is smaller than the true numerator, 43. The difference occurs because disease can only be observed if the invasive, yet sensitive and specific secondary test is used. Even though the participants have a score that exceeds the ROC cutpoint for Test 2, they do not all undergo the invasive, yet sensitive and specific secondary test. Thus, they do not yield observed cases of disease. By contrast, for Test 1, because the ROC cutpoint is higher than the threshold which leads to the invasive, yet sensitive and specific secondary test, every participant positive on Test 1 undergoes the secondary test, and is shown to have disease. For each test, there is a different proportion of participants who exceed the cutpoint, who truly have disease, and who proceed to secondary testing. This is the source of the differential bias that causes the curves to reverse order in Figure 2.

8. Can the author explain what “differential inflation in sensitivity” is and why it is unique to paired extreme verification bias? Please add some citations to this statement,

We have removed the term "differential inflation in sensitivity", and instead added the paragraphs shown above, in the response to comment seven.

9. Sometimes the term paired extreme verification bias is written in Italic and sometimes it is not. Please be consistent in this.

We have now italicized every occurrence.

10. The authors state that in “many situations, complete verification ... is impossible, as there is no accurate, [etc] test”. I think the problem arises only when the reference standard (confirmatory test) is too invasive, too costly or otherwise too big a burden. It has nothing to do with a reference test not being accurate. That gives other problems (imperfect reference standard bias, verification bias in general).
We rephrased the sentence to read as follows

Cancer screening trials in particular are susceptible to paired screening trial bias, because the secondary test is typically biopsy. Negative screening results cannot lead to biopsy because there is no visible lesion to be biopsied. Because biopsy is painful and invasive, it is infeasible and unethical to do a biopsy unless there are suspicious screening test results. Also, one can only biopsy what one can see: one cannot put a needle in an invisible lesion. Negative screening test results are verified, but typically by follow-up, which has lower sensitivity than biopsy.

11. Could the authors elaborate a bit more on the problem of follow-up as second confirmatory ‘test’? Not only may new cases occur in time, but (especially in cancer), cases may also resolve or never develop to disease (depending on how well the tests measure the disease).

We added the following paragraph to the discussion section

Although we modeled our trial design on real trials, we made simplifying assumptions, which may not be correct. We assumed that there was a method for determining disease status which was infallible. In reality, all methods of determining disease status may be fallible. In breast cancer, diagnostic mammography, biopsy and follow-up all make errors. Too short a follow-up time may miss cases of disease. While longer follow-up time will reveal a larger fraction of occult disease, it may also reveal increasing numbers of cases of disease that developed after the initial screening period, thus confusing the results. We assumed that all cases of disease are harmful. In screening studies, many cases of disease may resolve, or proceed so slowly as to be benign.

12. What is an ROC score? How is it calculated?

We removed the reference to ROC scores.

13. In the legend with figure 1: “The means of the ROC distributions for cases..... 1.6 and 0.35.”. How have the authors calculated the means of the ROC distributions and what does it mean?

Figure 1 now is Figure 2. We re-wrote the caption and legend as follows to clarify our example. We used "means the distributions of test results"
instead of "means of ROC distributions". We added the words "hypothetical example", and "fixed" to emphasize that we had chosen, not calculated the parameters.

Figure 1: True and Observed ROC curves for Test 1 and Test 2 for a hypothetical example where one screening test leads to a higher chance of recalls than the other screening test.†

†The parameters for this example were chosen to illustrate a case where paired screening trial bias may cause an incorrect scientific conclusion. The chance of recall for Test 2 for a participant who had disease was 34%, while for Test 1 it was 8%.

For this example, we fixed the disease rate, \( r = 0.01 \); the chance that participants with disease would experience signs and symptoms within the year of follow-up, \( \psi = 0.1 \); the variance, \( \sigma^2 = 1 \). The means of the distributions of test results for cases for Test 1 and Test 2 were 2.1 and 1.1, respectively, and the means for non-cases for Test 1 and Test 2 were 1.6 and 0.35. The correlation between test scores for cases was fixed at 0.1, as was the correlation for non-cases. All test scores above 2.5 on either test, or participants who had signs or symptoms had an infallible secondary test to determine disease status. For participants with scores below 2.5 on both tests, a less sensitive method was used to approximate disease status.

**Discretionary Revisions**

1. The authors use the term “observational trial”. This may be confusing, because the term trial often refers to an experimental (and thus not observational) design. Furthermore, the design that the authors describe (study participants undergo tests), may not be really observational. Alternatives may be diagnostic study, diagnostic accuracy study.

   We agree. We got rid of the words observational trial.

2. The same may be true for the term “screening trial”, which may implicate a trial to study the effect of screening programs on patient outcomes.

   After a thorough literature review, we concluded that "screening trial" is often used in trials that compare the diagnostic accuracy of two tests. Thus, we did continue to use the words "screening trial" in our manuscript.

3. The authors may want to consider some more recent publications on diagnostic test accuracy studies:


Bossuyt et al., BMJ, 2006 (about test comparisons)

Book: Knottnerus (ed.) The Evidence Base of Clinical Diagnosis.

Thanks for the references. We now have added the reference to Whiting et al., 2004 to the paper. It helped clarify the different sources of bias that interact to form paired screening trial bias.
**References**


6. Metz C, Wang P, Kronman HA: New approach for testing the significance of
differences between ROC curves measured from correlated data. In Information
Processing In Medical Imaging. Edited by Deconinck F. The Hague, the Netherlands:

7. Apostol TM: Calculus:Volume II. Multivariable Calculus and Linear Algebra, With
Wiley and Sons; 1969.

8. DeLong ER, DeLong DM, Clarke-Pearson DL: Comparing the areas under two or
more correlated receiver operating characteristic curves: a nonparametric


10. Baker SG, Pinsky P: A proposed design and analysis for comparing digital and
analog mammography: special ROC methods for cancer screening. J Am Stat Assoc


12. Hanley JA: The robustness of the 'binormal' assumptions used in fitting ROC
13. Begg CB, Greenes RA: **Assessment of diagnostic tests when disease verification is subject to selection bias.** *Biometrics* 1983, **39**:207-215.


15. Gray R, Begg CB, Greenes RA: **Construction of receiver operating characteristic curves when disease verification is subject to selection bias.** *Med Decis Making* 1984, **4**:151-164.


