Reviewer's report

Title: The impact of imprecisely measured covariates on estimating gene-environment interactions

Version: 1 Date: 20 September 2005

Reviewer: Robert Lyles

Reviewer's report:

General

This paper makes two interesting simulation-based observations regarding the impact of measurement error when estimating a regression parameter that represents gene-environment interaction. In Scenario 1, it is shown that measurement error in a confounder correlated with exposure has no impact on estimation of the exposure-genotype interaction parameter. In Scenario 2, it is shown that measurement error in the exposure that interacts with genotype does potentially bias the usual estimator of the interaction parameter.

The simulation presented is simple to apply and readily replicated, and some of the comments below are based on my own findings when replicating some of the scenarios to address questions about standard errors. Conclusions based solely on simulation studies are necessarily somewhat limited, but I believe they can be of use here.

--------------------------------------------------------------------

Major Compulsory Revisions (that the author must respond to before a decision on publication can be reached)

I find the result in Scenario 1 to be interesting and potentially useful, in that if the primary goal of a study is to estimate and make inference about the usual gene-environment interaction parameter, there need be no concern about confounder measurement error despite the fact that (as is known) such error impacts estimation of the "main" exposure effect. I think this may be the most important finding and that the authors should highlight it as such. I have one question and one concern based on replicating many of the simulation settings summarized in Table 1:

1) It is apparently assumed that genotype (G) is independent of both the exposure (X) and the confounder (C) in the authors' simulations. If they believe that these independence assumptions are generally reasonable in practice, then this should be stated clearly and a rationale given. Otherwise, some commentary should be made about whether the findings remain the same if association between G and C (and hence between G and X, since C and X are correlated) is introduced. From limited experimentation with this situation, I believe bias will be introduced in the estimator of beta_1 in addition to the bias already known to impact estimation of beta_2 (and beta_4) in this event. Interestingly, it appears that there may still be no effect upon the estimator of beta_3, which would keep the author's main conclusion for Scenario 1 intact. However, again my experimentation was limited and I suggest the authors look more carefully to confirm and comment on this issue unless they choose to argue that G and C will not be associated in practice.

2) I replicated some of the simulation settings in part because of some surprise at the comment on pg. 8 indicating that the standard error of the estimator of beta_2 was constant (0.08) regardless of values of sigma_v and the correlation between X and C. My findings were in agreement with the authors in terms of the magnitude of bias in estimating beta_2 and the lack of bias in estimating beta_3. However, I found a couple of things in contrast to the results reported. First, the standard
deviation of the ‘naïve’ estimates of beta_2 were smallest when sigma_v=0. Second, although I did find little variation in this standard deviation across positive values of sigma_v, they were approximately 0.06 rather than 0.08 in all cases based on n=1,000. These relatively small discrepancies nevertheless gave me some pause and may bear checking by the authors. The empirical standard deviations that I obtained matched well with the mean of the estimated standard errors based on the naïve ("surrogate") regression in each case for Scenario 1, and the authors should comment on this issue based on their simulations as well.

Regarding Scenario 2, where the confounder is eliminated and there is measurement error in X, the finding that the relative bias introduced into the usual estimates of beta_2 and beta_3 is the same is quite interesting. I was able to confirm this finding and the specific values for the mean beta_2 and beta_3 estimates given in Table 2. However, once again I noticed some discrepancies between the empirical standard deviations of these estimates and the “se’s” quoted by the authors in Table 2. A further interesting thing that I found in attempting to replicate the results should be looked into and, if confirmed, commented on. Under Scenario 2, I found that the empirical standard deviations of the 1,000 estimates of beta_2 and beta_3 no longer matched well with the means of the corresponding estimated standard errors. This may be quite important in terms of the validity of standard regression calibration under Scenario 2, because I believe the usual inferences that accompany that technique assume that the standard error estimates associated with the naïve estimators are valid. By this I mean that the average estimated standard error should match the standard deviation of the naïve regression parameter estimate over many simulations, even though the naïve estimator is biased. If this is not the case, then standard regression calibration may not be reliable (at least inferentially) for Scenario 2, and this directly impacts the authors’ example. Indeed, I believe standard regression calibration may not necessarily be tailored to the case considered in Scenario 2, where there is measurement error in X and the model of interest also includes a second covariate (G) and the interaction term G*X. This point deserves close attention. Perhaps if the authors could implement in their simulations the same regression calibration correction that they applied in the example corresponding to Scenario 2, with attention to standard errors, some light could be shed upon this issue. If standard regression calibration (and perhaps SIMEX as well) is suspect under Scenario 2 for this reason, then the example must be presented in a way that acknowledges this fact.

-------------------------------------------------------------------------------

Minor Essential Revisions (such as missing labels on figures, or the wrong use of a term, which the author can be trusted to correct)

a) In Scenario 1, there will clearly be bias in the estimates of beta_4 as well as beta_2. The authors should comment on their simulation-based findings regarding beta_4, even though this bias is expected.

b) I believe there are existing references regarding the effect that measurement error in a confounder has upon the estimated regression parameter for exposure as well as power for assessing the impact of exposure on the outcome. Please see if one or more references can be cited to this effect on pg. 8.

c) In describing the simulations made under Scenarios 1 and 2, it would be helpful to spell out for the reader exactly what model the data were generated from in each case, as well as exactly what model was fit to each simulated dataset to obtain the results in Tables 1 and 2. My assumption is that for Scenario 1, the data were generated under the model Y=beta_0 + beta_1*G + beta_2*X + beta_3*XG + beta_4*C + e, and the regression model fit replaced C with its surrogate, D. For Scenario 2, I assume the data are generated by Y=beta_0 + beta_1*G + beta_2*X + beta_3*XG + e, and the model fit replaced X by its surrogate (W) and replaced XG by WG. If these assumptions are correct, they are implied in the paper but could be stated more clearly to aid the reader.
d) Tables 1—3 could benefit from a bit more detail in footnotes (e.g., the sample size assumed for each simulation, the # of simulations in each case, etc.). It would help to make the tables as self-explanatory as possible.

Discretionary Revisions (which the author can choose to ignore)

i). It would be a simple matter to produce simulations like those summarized in Tables 1 and 2 for the identical scenarios, except where Y is a binary response (e.g., a logistic model rather than linear). If the authors examined a limited number of simulation settings in this case, even without reporting them in detail, it might allow them to inform the reader as to whether similar findings held for logistic regression.

ii). The result regarding the ratio estimator in Scenario 2 being unaffected by measurement error in X is interesting. However, I am not certain how often the ratio estimator is used in practice? In particular, inference about that ratio parameter would be somewhat complicated and I assume one must resort to a delta method approximation for its variance, even without measurement error? To me this makes it less appealing than the standard interaction term coefficient, at least as a basis for inference. Further, the comment above about standard errors in Scenario 2 may suggest that even though the ratio estimate itself is unaffected, standard inference about it may be hampered by the exposure measurement error. This may merit commentary in the discussion.

What next?: Unable to decide on acceptance or rejection until the authors have responded to the major compulsory revisions

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

Quality of written English: Acceptable

Statistical review: Yes

Declaration of competing interests:

I declare that I have no competing interests.