

Author's response to reviews

Title: Bias Magnification in Ecologic Studies: A Methodological Investigation

Authors:

Thomas F Webster (twebster@bu.edu)

Version: 3 **Date:** 2 June 2007

Author's response to reviews:

2 June 2007: After consultation with editor David Ozonoff, additional changes were made in the revised manuscript.

Title : Bias Magnification in Ecologic Studies: A Methodological Investigation

Authors: Thomas F Webster

Your revised manuscript should be in strict accordance with our instructions for authors, cf. the pre-acceptance checklist at

<http://www.biomedcentral.com/info/edgr-preacceptcheck.asp><http://www.biomedcentral.com/info/edgr-prea>

Please make the use of indented first lines of paragraphs uniform, preferably by avoiding the indents. Also, please make the line spacing after equations uniform. On p. 7, please rephrase the sentence 'Let's regress...' to fit the style of the remainder of the manuscript. Please check the explanation of NIEHS abbreviation in the Acknowledgements. The references must be in strict accordance with the instructions to facilitate automatic linkage. Please remove 'and' between authors, and commas before author initials. All tables must be in portrait format, or they will need to be uploaded as 'additional files'. Please redesign the tables so that they will fit this format as described in the instructions at <http://www.ehjournal.net/info/instructions/>. You may want to look at one of the previously published papers for guidance.

Response: The stylistic comments have all been addressed in the revised manuscript.

Charles Poole:

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

1. (general) How many ecologic analyses attempt to estimate risk differences? Don't they usually work with rates and log transformations and attempt to estimate rate ratios? To what extent would the results of this paper not apply to that setting?

Response: The manuscript has been revised to address this comment. Pg. 4 and 21 point out that while ecologic analysis of risk differences may be uncommon in practice, this framework shows important aspects of ecologic bias with simple tools. An extension to rates is provided (pg. 6). Log transformations and the RR are briefly discussed (pg. 19, 21-22). Risk/rate diagrams can still useful qualitative information about RR (pg. 19). I'm working on quantitatively applying bias magnification in these situations and anticipate a future paper on this topic.

2. (pp.5-6) I'm confused as to which risk function is being referred to here. Since x is defined as being truly binary, it would seem that the function relating y to x , expression (3), must be linear. But since x is truly binary, values of x between 0 and 1 are impossible. So only the open circles can depict individual data and the annotations "individual data" pointing toward the straight lines in Figures 1 and 3 are incorrect. These

straight lines, it seems to me, are lines on which, if every X-Y pair fell on them, a linear regression of Y on X (i.e., a linear ecologic regression) would give an unbiased estimate of the individual-level slope. If I'm right about this, then I have no idea what the "upper bound on error" annotation in Figure 2 means, as it is pointing toward one such straight line. This leaves me wondering as well what the curved line in Figure 2 represents. It doesn't seem that it can represent a regression of y on x because, as noted above, x cannot take any value other than 0 or 1. It would then seem to be some kind of non-linear function relating Y to X, such that a linear regression through points on that function would yield a biased estimate of the slope of the necessarily linear relation by y and x.

Response: Page 5-7 and 21-22 have been extensively revised to address this comment, as well as the notation on several figures. The risk function on the individual level is assumed to be linear in equations 1-3 but need not be (An exponential model can also be fit to a 2x2 table, for example). On the other hand, the relationship between Y and X (ecologic) must be linear for 2x2 tables.

The line in the risk diagram is meant to convey the important individual-level information (for 2x2 tables, we're here most interested in the risk in the unexposed and the RD), information that is lost upon aggregation. The term information is now used instead of data (as the comment points out, for 2x2 tables, the actual data occur only at the end points). The same model can also be used to describe continuous exposures where data are not limited to the endpoints. Figure 2 was revised to show this case in order to try to avoid confusion.

3. (p.10) I wonder about the conclusion in the last paragraph on p.10 that, in the circumstances of Table 2 (different baseline risks, constant individual-level slope or RD), if there is no individual-level confounding, the linear ecologic regression of Y on X won't be biased. For confounding to be absent under these conditions, it would have to be the case that $X_0 = X_1$. If $X_0 = X_1 = 0.4$, $Y_0 = 0.28$ and $Y_1 = 0.70$. If $X_0 = X_1 = 0.6$, $Y_0 = 0.32$ and $Y_1 = 0.72$. Thus, if confounding is absent at the individual level, the solid circles in Figure 2c would lie on a vertical line and a linear ecologic regression of Y on X would yield a maximally biased slope of infinity. In this example, the ecologic bias would be positive if $X_1 > X_0$ and negative if $X_1 < X_0$. The bias would reach a minimum of $|0.4|$ when $X_1 = 1$ and $X_0 = 0$ ($be = 0.8 - 0.2 = 0.6$, bias = 0.4), and when $X_1 = 0$ and $X_0 = 1$ ($be = 0.4 - 0.6 = -0.2$, bias = -0.4).

Response: The text was revised (page 12) to address this exception to equation 10. When $X_0 = X_1$, an assumption used in the derivation of equation 10 (non-zero variance of the X_i) is violated. be is then undefined. A new figure 5 was inserted (with text on pg. 12) to further explain the effects on confounding by group of changes in q_i or X_i .

4. (pp.11-12) bw is an individual-level regression estimate obtained when a modeling assumption (constant b across groups) is violated. Although many epidemiologists routinely fail to check this assumption, if pressed most of them would probably agree that they should. From this standpoint, bw doesn't seem to measure up very well as a gold standard. We're in essence examining the bias in ecologically trying to estimate an individual-level parameter we shouldn't be trying to estimate (a presumptively constant RD when the RD is, in fact, not constant).

Response: The text was revised (pg. 13-14) to point out that bw is not truly a gold standard. However, it is still useful as a reference point for understanding ecologic bias due to effect modification of the RD.

Discretionary Revisions (which the author can choose to ignore)

1. (p.3) It was unclear from the Background section what the standard would be for the assessment of bias (i.e., what effect we are trying to estimate). This was ultimately revealed on p.5. It might be helpful to clarify that important point from the outset.

Response: I was not sure exactly what was unclear, but a revision was made to page 3 to try to be clearer.

2. (pp.4-5) I'm a little confused by the notion that each individual's outcome is that individual's risk plus a residual. Couldn't each individual's risk be that individual's outcome (i.e., $r_j = y_j$)? If so, expression (1) could

be dropped and the exposition could start with expression (2).

Response: Pg. 4-5 and 20-21 have been revised to make this clearer.

3. (p.6) Is expression (4) just another way of saying that an OLS linear regression must pass through the mean of x and the mean of y ?

Response: Additional text (pg. 21) was added to explain the derivation and meaning of equation (4).

4. (p.6) Is it always implicitly understood that "concave" means "concave up"?

Response: The wording was changed to "concave up" to prevent any confusion.

5. (p.11) Perhaps the proof attributed to reference 11 should be given (e.g., in an appendix).

Response: Added (pg. 25).

Ulf Stromberg:

General

The author addresses bias magnification in ecologic studies. The paper is clearly written. In particular, the author explains clearly the magnification of individual-level bias due to ignoring groups. I think the merit of the paper is the presentation and explanation of the magnification factor and its impact.

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

1. I think that the paper should be more focused. I suggest the author to concentrate on the impact of the magnification factor under the confounding-by-group and effect-measure-modification-by-group scenarios. Nondifferential misclassification of binary exposure, which has been discussed several times in the literature, can be withdrawn from Results. It is enough to mention that particular source of bias in Discussion, together with other potential sources of bias not considered in the methodological investigation (systematic exposure assessment errors; confounding within groups or effect modification within groups).

Response: Nondifferential exposure misclassification has indeed been discussed before (and the seminal paper by Brenner et al was referenced). I discuss this topic in a somewhat different way, e.g., using risk diagrams and the relationship to M (pg. 27-28). I think this it worth including.

2. I suggest that the author provides a concrete ecologic study example, in order to demonstrate (a) how to accomplish a favorable design by trying to minimize the magnification factor and (b) how to post-study-estimate the magnification factor. An example with air pollution as the environmental exposure would be illustrative.

Response: The examples in the tables and risk diagrams were meant to provide examples of otherwise abstract ideas. Concrete ideas on the implications for study design and the ability to compute M for certain study designs are provided on pages 17-18. Application of these ideas to real world data is very worthwhile as the text notes on page 19, although I think doing this is beyond the scope of this paper.

3. The comments on page 6 concerning nonlinearity of the risk function are somewhat confusing. I realize that, for a continuous or polytomous exposure variable, the underlying (true) risk function can be nonlinear. Consequently, model misspecification bias may arise. On the other hand, for a dichotomous exposure variable on the individual-level, isn't the risk function on the group-level always linear under a bias-free scenario? Please clarify Figure 2 and the statement "The amount of error depends on the curvature of the risk function and the exposure distribution within groups, reaching a maximum when exposure is

dichotomous".

Response: See response to comment 2 of reviewer 1.

Jonathan Wakefield:

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

Page 5. Clarification of the models assumed would be useful here and later. If y_j is the response of individual j then y_j takes the values 0 and 1 and so is Bernoulli with $E[y_j] = q + b x_j$. I think it is confusing to write the model in terms of (2), at least without further explanation. Weighted least squares can still be used for the fractions with disease (for example), though I would first present a more reasonable statistical model.

Response: The text was revised (Pg. 4-5 and 20-21) to make this clearer.

Appendix 2. It is worth referencing Plummer and Clayton (1996, JRSS, Series B, 113-126) here. These authors also discuss a model with the variance of the exposure included. Richardson and Montfort (2000, in Spatial Epidemiology: Methods and Applications, eds Elliott et al) also discuss this model.

It wasn't clear to me what the second half of appendix 2 was suggesting. I would be inclined to stay with the log-linear model. One could then assume that the aggregate count followed a Poisson distribution and fit a model with mean $\exp(q + b X_i + b^2 \sigma_i^2/2)$.

A slightly more complex version of this model was fitted by Best et al (2001, JRSS, Series A, 155-174), see equation (12).

Response: The references to Plummer and Clayton, Richardson & Montfort, and Best et al were added in appendix. The second half of appendix 2 was rewritten to make it clearer.

Ecological regression, as discussed by Goodman, is extensively discussed by Wakefield (2004, JRSS, Series A, 385-441), along with a number of other issues discussed in this paper. Linking the two presentations would be very useful.

Response: The text was revised to include references to Wakefield 2004 on pages 8, 20.

The crude analysis in Table 2 illustrates Simpson's paradox - it would be useful to note this.

Response: I believe Simpson's paradox refers to a reversal of direction due to confounding. In this example the ecologic estimate of the RD is in the same direction (positive) as both the true value and the crude individual-level value.

Page 16. Wakefield (2003, Biometrics, 9-17) discusses sensitivity studies for ecological analysis, it would be interesting to try to link that with this paper's suggested sensitivity analysis.

Response: A reference to Wakefield's method of sensitivity analysis was added on page 18. I agree that a more detailed comparison would be useful, but I would like to defer this to the future.

Page 16. I found the last full paragraph on this page confusing - if the fraction exposed in each area is known, then as long as there are at least 2 areas, and the baseline risk and risk difference remain constant across areas, then an appropriate ecological model can be fitted. This model has $E[Y_i/n_i] = (1-p_i)\theta_0 + p_i\theta_1$ where n_i is the number of individuals in area i (and Y_i is the number of diseased individuals) p_i is the fraction exposed in area i and θ_0 and θ_1 are the risks for unexposed and exposed individuals) and will provide an unbiased estimator. How does this relate to the authors comments concerning the maximization of the magnification factor? Is this assuming that the wrong model is fitted, i.e. the naive model $\exp(\alpha_0 + \alpha_1 p_i)$?

Response: The paragraph was rewritten to make it clearer, i.e., that the magnification factor matters if there is bias due to confounding by group or effect measure modification by group. In your notation, this means that θ_0 and/or $(\theta_1 - \theta_0)$ differ between groups and covary with exposure.

Appendix 8. Wakefield (2003, Biometrics) also shows when within-group confounding does not cause problems for ecological studies. I think the problems of within-area confounding should be emphasized more - it is highly unlikely that the risk model is linear in terms of all confounders, and so in practice this will be a big problem.

Response: The claim made in the text is that confounding within groups alone does not bias the ecologic results when the relationship between outcome and exposure is linear (the relationship between outcome and covariate may be nonlinear). The text was revised in on page 18 to emphasize that the linearity assumption must be carefully considered. It was also revised on page 29 (the appendix) to make clear that confounding within groups does matter when the relationship with outcome is nonlinear for both exposure and covariates. A reference to Wakefield (2003) was also added here.

Discretionary Revisions (which the author can choose to ignore) The bias due to aggregating a non-linear model was called "pure specification bias" by Greenland (1992), and it might be useful to state this at least once.

Response: Use of the terminology "pure specification bias" was added on pages 7 and 21-22.