Peer Review in Epidemiology Cannot Accomplish Its Ostensible Goals Due to Incomplete Reporting and Unverifiable Analyses

Karyn Heavner, MSPH, PhD
Carl V. Phillips, MPP, PhD
University of Alberta, School of Public Health
Suite 215, 8215 112th Street
Edmonton, Alberta, T6G2L9 Canada

Brad Rodu, DDS
James Graham Brown Cancer Center, University of Louisville
529 South Jackson Street, 4th Floor
Louisville, KY 40202 USA

ABSTRACT

The accuracy and appropriateness of an analysis often cannot be verified by the contemporary peer review process. Peer review is also unlikely to identify possible publication bias *in situ* (PBIS) (running many different statistical models but only reporting the results of one, often an outlier result). A review of articles reporting analyses of a cohort of Swedish construction workers revealed unacknowledged and unexplained variations in methodology, including the use of different variables measuring tobacco use, age and body mass index and a failure to adequately and accurately reference previous related articles about the cohort. Seemingly minor changes in methodology, such as the cutoffs used to convert continuous variables to categorical variables may result in significant changes to the results. These inconsistencies were likely not discovered during the peer review process as it would have required that reviewers conduct a systematic review of previous analyses of the dataset. Practical solutions to this dilemma include enhanced post publication review and ensuring that data are available for secondary analysis.

Keywords: peer review, publication bias, epidemiological methods, data analysis, data mining

INTRODUCTION

Peer review supposedly vouches for the accuracy and appropriateness of an analysis. But in epidemiology and related health sciences, reviewers typically comment on everything except the analysis. Comments focus on the background, discussion, and writing style. Reviewers make suggestions or demands about elements that cannot be changed after the study has been conducted, such as the overall usefulness of the study, the goal of the research agenda, and data collection methods. At best, these can be taken as suggestions for future research. By contrast, reviewers typically offer few comments on the accuracy and appropriateness of the statistical analysis that attempts to estimate the effect of an exposure on disease. To some extent, the emphasis on discussion and motives represents the insufficient skills or political biases of the reviewers, but even the best intentioned reviewers generally are not able to evaluate the actual analysis even if they have the time and skills: They seldom have access to the data and the reported methods are often inadequate to replicate the analysis.

Beyond the simple inability to check the analysis as conducted, the inaccessibility of the data and complete methodology also makes it impossible for reviewers to assess "publication bias *in situ*" (PBIS) [1]. PBIS exists within a research report when authors run many different

statistical analyses but report the results of only one. These various analyses may use different inclusion criteria, exposure or outcome definitions, covariates or regression models. It is unknown how often the reported result is one of only many analyses that were run or is an outlier estimate that is not representative of the true exposure-disease relationship found in the data. It is fairly easy for seemingly minor changes, such as the cutoff used to convert a continuous variable into a dichotomous variable, to produce an outlier effect estimate [2, 3] (http://epiereview.com/wp1.php). The results that are reported may reflect the authors' preferences in support of worldly political goals, which are common in public health science, or perhaps just to publish a result that is further from the null and more dramatic. Due to a very active anti-tobacco and anti-nicotine lobby, this is particularly problematic in studies exploring low risk nicotine alternatives to cigarette smoking, otherwise known as tobacco harm reduction (THR). We provide a recent example of a series of articles with clues that suggest that peer review failed to prevent PBIS in the publication of analyses of a cohort of Swedish construction workers. These articles are increasingly cited as the basis for anti-harm-reduction activism (which some of the authors of the articles participate in) and regulatory policies regarding smokeless tobacco use [4, 5].

INCONSISTENCIES IN THE SWEDISH CONSTRUCTION WORKERS COHORT STUDY

We reviewed a series of publications based on analyses of a cohort of more than 330,000 Swedish construction workers to look for inconsistencies that might be indicative of PBIS. Many of these methodological differences have been noted before [6-8] (tobaccoharmreduction.org/papers/heavner-phillips-heffernan-rodu-jun08.pdf, tobaccoharmreduction.org/wpapers/008.htm), so only a few examples are presented here to focus this analysis on the challenges to peer review that these examples represent.

Articles included in this review are those that reported the results of data analyses from the cohort study, included tobacco use variables as an exposure or covariate, and were published in English. Most of the 36 articles that were included focused on occupational exposures or tobacco use (including smoking and snus use) and outcomes related to cancer or heart disease. Snus is the Swedish term for smokeless tobacco, which is approximately 99% less risky than cigarettes because the users do not inhale smoke [9]. We investigated consistency in methodological choices, including eligibility criteria and variables measuring tobacco use, age and body mass index (BMI). The results of the literature review are summarized here; additional information is available at http://www.tobaccoharmreduction.org/wpapers/008.htm.

One of the most notable differences in the studies is the years from which snus use data were used. There were 3 phases of enrolment: 1971-1974/1975, 1975-1977 and 1978-1992/1993 during which tobacco use was assessed differently. (No tobacco use information was collected between 1975-1977.) The recent (2005 and later) studies of snus use were inconsistent in their use of the information on snus use collected from 1971-1974. The 1971-1974 snus use data was excluded from most of the studies (e.g., [10-13]). In these studies of snus use, exclusion of participants for whom snus use data was not collected after 1977 resulted in elimination of more than 30,000 male participants from the analysis (this is a rough estimate as this aspect of the methodology is very unclear). Luo et al. cited "ambiguities in the coding of smoking status in the questionnaires used during 1971–75" as the reason for excluding these data and cited unpublished data by Zendehdel et. al. as a source for this information [13]. However, Zendehedel et al. included the 1971-1974 snus use data in their analyses but provided no justification for their inclusion or even acknowledged that their eligibility criteria substantially differed from those in the other contemporary non-occupational studies [14].

There are also inconsistencies and omissions in the references to the previous articles about the cohort. For example, the majority of the recent studies cited a 1996 study by Nyren et. al. [15] regarding the quality of the smoking data. However, some of the studies cited the article as a reference for the quality of the tobacco use or exposure data [13, 14, 16], which is inappropriate in studies of snus use because Nyren et. al. evaluated the consistency of the smoking variables between participants' first and second visits. The 1996 study by Nyren et al. did not publish any information about the consistency or validity of the snus use variables.

Age and BMI were included as covariates in many of these analyses. However, the cutoffs (values used as thresholds to convert continuous variables to categorical variables) and number of categories used for age and BMI vary greatly between studies. These variations were not acknowledged, much less justified, by the authors and the consequent effects on the measures of association are unknown.

These are examples of some of the many inconsistencies and sources of uncertainty in this series of articles. Others include 1) how snus use was measured between 1971 and 1974; 2) whether amount and duration of snus use were included in the analysis; 3) variation in the smoking variables included in the analysis; and 4) a failure to adequately reference previous related articles. Despite our repeated requests, the authors have refused to make the data publicly accessible for reanalysis. Thus, it is unknown how the results would have changed if the data were analyzed consistently. Even if the data become available for secondary analysis, the methods in many of the articles are confusing and it is unlikely that they would serve as a straightforward template for re-analysis.

THE REVIEWERS HAD NO WAY TO ENSURE THAT THESE ARTICLES INCLUDED ACCURATE AND UNBIASED ANALYSES

Many of the Swedish construction worker articles are, in effect, different parts of a single study. But the methodological inconsistencies among them are rarely acknowledged by the authors, and in many cases previously published studies were not cited, even when they had analyzed the same data and had similar exposures or outcomes. This created an unrecognized and therefore impossible challenge for peer reviewers. As most of these articles were published in different journals, it is unlikely that the manuscripts were reviewed by the same researchers. Given the cursory attention often paid to peer reviews in epidemiology, it is likely that the reviewers did not even know about other articles in the series when writing their reviews.

This illustrates a great challenge for peer review in the health sciences: Exactly what are reviewers supposed to be reviewing, and is it even possible for them to conduct an informed and comprehensive review?

The peer review process typically involves reviewers judging an article based on their expertise and experience. This expertise does not necessarily extend to knowledge of other published articles using the same dataset, let alone the content of the dataset. Reviewers of the Swedish construction workers studies were not even offered a complete list of references to previous analyses to inform their judgment.

In theory, data collection and analysis methods are reported, but in practice there are always ambiguities in the reporting and there are almost never explanations for critical choices. In the present example, choices that may have had critical effects on the results were not justified or the explanations were inconsistent. The reviewers likely assumed that there were legitimate theory-based reasons for these choices, though perhaps they should not have

done so. Reviewers (and readers) often take articles at face value, treating them as if the authors presented all of their methods and the results of all analyses. They have no way of knowing how minor variations in the methods can affect the results. A complete reporting of the data analysis methods might reveal this effect (e.g., "we analyzed the data in the 1,024 different ways shown in Table 2 and reported model #503 because it gave the largest statistically significant odds ratio"), but such disclosure of data mining is never reported.

A rare chance to identify potential cases of PBIS or conduct a thorough peer review comes in a case like this, when a reviewer could theoretically review all of the articles about a particular dataset and then identify discrepancies that the authors fail to mention. This is almost impossible given the time constraints on peer review and is clearly beyond what can reasonably be expected of the reviewers. Authors presumably realize this and take advantage of the fact that each manuscript is reviewed in isolation with little or no consideration of how the data were previously analyzed and interpreted. Our review revealed a pattern that suggests that the methodological choices were not theory driven (or else they would not change between analyses of similar outcomes) and thus were quite possibly data-driven (which is to say, the choices are made to produce the results that the authors preferred). But this required far more research by us than is typically done by a journal reviewer.

And even if reviewers of this series of articles did manage to notice inconsistencies and were sophisticated enough methodologically to recognize their implications, they would have no way to check their implications without access to the data and the exact statistical programs that were used. Without access to the data, there is no way to test a suspicion that the data analysis was designed to bias the result. The only recourse would have been for a reviewer to insist that the authors report different statistical models, a suggestion or demand that our experience suggests is typically ignored.

PRACTICAL SOLUTIONS

In the above example, the methodology had to be pieced together from numerous articles, published theses and responses to letters to the editor (which is an ongoing process). In fact, the authors provided much valuable information about methodological problems in response to an inquiry from a Wall Street Journal writer. Such information should have been disclosed in the methods sections of the published articles. Many of these problems could have been resolved if the methodology had been published in detail (in the peer reviewed literature or online) and consistently referenced in all the articles and if the data were available for re-analysis. After piecing this together, we learned information that the reviewers (and most readers) were unlikely to have known, even as they claimed to scrutinize the methodology. But even if readers and reviewers had complete information about the methods used to obtain the reported results, it would not answer the questions about whether the analysis was done correctly and whether biased choices were reported.

It is difficult to know what to do about potentially inappropriate or incorrect data analysis, since exhaustive re-analyses by reviewers is unlikely to occur in most cases, even if reviewers have access to the data. In terms of bias, peer review can only be useful in epidemiology if expectations change to demand the reporting of more methods and results. When a single effect estimate is presented, reviewers have little information about the sensitivity of the estimate to changes in covariate or exposure definitions. Reviewers typically accept that this estimate is “the” best estimate with no knowledge of the alternative models and on blind faith that the authors have not chosen an outlier or some other misleading estimate.

Graphing the distribution of possible results to demonstrate the sensitivity of "the" result to modeling choices, as we and others have suggested elsewhere (epiereview.com/wp1.php) [2, 3, 17], is one possible way in which researchers can report a collection of candidate results. This allows reviewers and readers to easily assess at least one dimension of model choice and know whether the result the authors prefer is really representative of the data. Additional partial solutions are needed to combat these problems. For example, the inconsistencies in the Swedish studies were reported by us and others after publication. However, the primary forum for reporting such problems is the letters pages of the same journals that judged the articles as worthy of publication in their present form, creating an obvious conflict of interest. Indeed, some of our letters have been rejected from publication. In response to this, some of us have developed a website which will facilitate more robust post-publication peer review in epidemiology and related sciences (details of which are included in another report in this collection).

Unfortunately, assessments of biased analyses rely on the cooperation of researchers who control secret datasets, requiring that they run all requested alternative analyses or release the data. Investigators trying to protect biased analyses are unlikely to volunteer either of these, or to report the distribution of results that we propose. Thus, the problem comes back to requiring more robust and sophisticated reviews by journals in this field, which in turn depends on educating consumers of this information on the biases that currently exist (or can easily exist without detection). So long as journal editors and authors are comfortable with the current system, peer reviewers (who are largely the same people, of course) cannot be held responsible for the accuracy and representativeness of results in epidemiology and related fields. Likewise, the designation of “peer reviewed” publication cannot be seen as vouching for the accuracy or appropriateness of the analysis or the representativeness of the reported results.

REFERENCES

[1] C.V. Phillips, "Publication Bias in Situ", **BMC Medical Research Methodology**, Vol. 4, 2004, pp. 20

[2] M. Cheung, J. Gould, W. Hare, K.K. Heavner, M. Khorvash, D. Li, K.K. Lock, C.V. Phillips, and M. White, "Ideas for Reducing Data Interpretation Bias in Epidemiological Studies", **Technical report prepared for the 11th PIMS Industrial Problem Solving Workshop**, 2007,

[3] C.V. Phillips, K.K. Heavner, and W. Hare, "Categorization of Continuous Variables: Recognizing and Reducing the Bias Caused by the Choice of Cutpoints", **American Journal of Epidemiology**, Vol. 167, No. Suppl, 2008, pp. S115

[4] International Agency for Research on Cancer (IARC)., "Smokeless Tobacco and Some Tobacco-Specific N-Nitrosamines", **IARC Monographs on the Evaluation of Carcinogenic Risks to Humans**, Vol. 89, 2007.

[5] Scientific Committee on Emerging and Newly Identified Health Risks (SCENIHR)., "Health Effects of Smokeless Tobacco Products", 2007, (http://ec.europa.eu/health/ph_risk/committees/04_scenihr/docs/scenihr_o_009.pdf).

[6] K.K. Heavner, C. Heffernan, C.V. Phillips, and B. Rodu, "Methodologic and Ethical Failures in Epidemiologic Research, As Illustrated by Research Relating to Tobacco Harm Reduction (THR)", **American Journal Epidemiology**, Vol. 167, No. Suppl, 2008, pp. S115.

[7] B. Rodu, K.K. Heavner, and C.V. Phillips, "Snuff Use and Stroke", **Epidemiology**, Vol. 20, No. 3, 2009, pp. 468-469.

[8] B. Rodu, K.K. Heavner, "Errors and Omissions in the Study of Snuff Use and Hypertension", **Journal of Internal Medicine**, Vol. 265, No. 4, 2009, pp. 507-508.

[9] C.V. Phillips, D. Rabiu, and B. Rodu, "Calculating the Comparative Mortality Risk From Smokeless Tobacco Versus Smoking", **American Journal of Epidemiology**, Vol. 163, No. Suppl, 2006, pp. S189.

[10] M.P. Hergens, L. Alfredsson, G. Bolinder, M. Lambe, G. Pershagen, and W. Ye, "Long-Term Use of Swedish Moist Snuff and the Risk of Myocardial Infarction Amongst Men", **Journal of Internal Medicine**, Vol. 262, No. 3, 2007, pp. 351-359.

[11] M.P. Hergens, M. Lambe, G. Pershagen, A. Terent, and W. Ye, "Smokeless Tobacco and the Risk of Stroke", **Epidemiology**, Vol. 19, No. 6, 2008, pp. 794-799.

[12] M.P. Hergens, M. Lambe, G. Pershagen, and W. Ye, "Risk of Hypertension Amongst Swedish Male Snuff Users: a Prospective Study", **Journal of Internal Medicine**, Vol. 264, No. 2, 2008, pp. 187-194.

[13] J. Luo, W. Ye, K. Zendehdel, J. Adami, H.O. Adami, P. Boffetta, and O. Nyren, "Oral Use of Swedish Moist Snuff (Snus) and Risk for Cancer of the Mouth, Lung, and Pancreas in Male Construction Workers: a Retrospective Cohort Study", **Lancet**, Vol. 369, No. 9578, 2007, pp. 2015-2020.

[14] K. Zendehdel, O. Nyren, J. Luo, P.W. Dickman, P. Boffetta, A. Englund, and W. Ye, "Risk of Gastroesophageal Cancer Among Smokers and Users of Scandinavian Moist Snuff", **International Journal of Cancer**, Vol. 122, No. 5, 2008, pp. 1095-1099.

[15] O. Nyren, R. Bergstrom, L. Nystrom, G. Engholm, A. Ekbom, H.O. Adami, A. Knutsson, and N. Stjernberg, "Smoking and Colorectal Cancer: a 20-Year Follow-Up Study of Swedish Construction Workers", **Journal of the National Cancer Institute**, Vol. 88, No. 18, 1996, pp. 1302-1307.

[16] A. Odenbro, R. Bellocco, P. Boffetta, B. Lindelof, and J. Adami, "Tobacco Smoking, Snuff Dipping and the Risk of Cutaneous Squamous Cell Carcinoma: a Nationwide Cohort Study in Sweden", **British Journal of Cancer**, Vol. 92, No. 7, 2005, pp. 1326-1328.

[17] D. Wartenberg, M. Northridge, "Defining Exposure in Case-Control Studies: a New Approach", **American Journal of Epidemiology**, Vol. 133, No. 10, 1991, pp. 1058-1071.

DISCLOSURE

Detailed conflict of interest disclosures for the authors can be found at tobaccoharmreduction.org.