Commentary: Toward systematic reviews in epidemiology

Michael B Bracken

One does not need to agree with the premise of Swaen et al., who examine the issue of false positive outcomes in this issue of the International Journal of Epidemiology, or have to ignore some methodological weakness in their study, or even to think their conclusions simply reaffirm some very basic scientific principles, to see considerable merit in the approach taken by these investigators and to believe their paper commands attention.

Taubes' paper, which highlighted the discrepancies between different study results that arise in epidemiology and the effect this has on public opinion, was famously read by some as predicting the imminent demise of epidemiology but it also permitted a broader examination of the state of the discipline. To be sure, epidemiology produces conflicting results but so does any research enterprise. It is only because the public has such a keen interest in the results of epidemiological studies that they are seen to be at the sharp end of this particular stick. Climatologists, nuclear physicists and students of the fall of the Roman Empire are all seen to produce their share of discrepant observations when the spotlight of public scrutiny falls on them. In fact, epidemiologists know a lot about the correct ways to conduct a research study but less about how to review and synthesize data from multiple studies, and this, I suggest, is a principal source of the public's confusion when faced with a new result from an epidemiological study.

The paper by Swaen et al. incurred several methodological difficulties of its own. How does one define a true positive result (i.e. what is the gold standard)? If investigators set up a hypothesis and test it on data already collected for another purpose is that necessarily a 'fishing expedition'? Indeed, is all secondary data analysis a fishing expedition? Should Swaen et al. have looked at the issue of false negative studies to derive a completely balanced picture and one might mischievously ask whether their own paper is itself an example of a false positive result.

When looking at the issue of false negative studies to derive a completely balanced picture and one might mischievously ask whether their own paper is itself an example of a false positive result. When looking at the issue of false negative studies to derive a completely balanced picture and one might mischievously ask whether their own paper is itself an example of a false positive result. When looking at the issue of false negative studies to derive a completely balanced picture and one might mischievously ask whether their own paper is itself an example of a false positive result. When looking at the issue of false negative studies to derive a completely balanced picture and one might mischievously ask whether their own paper is itself an example of a false positive result.
science of review, these studies have led to a better understanding of the role of publication and citation bias, the importance of different data base searching strategies, the validity of abstracts in accumulating research evidence, and the validity of different methods used for quantifying the quality of studies.

Reassuringly, Swaen et al.1 find the largest factor in a false positive study to be the absence of an a priori hypothesis as that is arguably the most fundamental of all scientific principles. Similarly, a dose-response relationship and adjustment for a major confounder (smoking), as expected, lead to fewer false positive results. It is interesting that study design is not itself a factor but there is little reason it should be; there is nothing intrinsically in error with case-control studies once problems with confounding have been adjusted, as they have here. Many of these issues are a matter of faith in epidemiology and it is reassuring to have some empirical evidence for them.

It is a great paradox in epidemiology that while the profession is very conversant with the requirements for conducting valid studies, it has generally neglected the need for rigorous, objective and hypothesis-driven summaries of the totality of epidemiological evidence when reviewing a particular topic. Early critics of the lack of scientific rigor in literature reviews focused on medicine, but in one recent analysis, over 60% of epidemiology reviews were considered to not meet the standards of a systematic review, and several specific biases in epidemiology reviews have been reported for chronic fatigue syndrome and passive smoking. While calls for more quantitative reviews in epidemiology are starting to be made, the overall poor quality of current epidemiology reviews is in marked contrast to the field of evidence-based medicine and healthcare which over the last 12 years has made remarkable strides in developing a methodology and strict standards for systematically reviewing and analysing a body of literature. While some epidemiologists have played a major role in these developments, by and large it appears that epidemiologists still review evidence using traditional and potentially biased methods.

Tröhler has recently provided a comprehensive account of the origins of evidence-based medicine, focusing on its early history in Britain. Of particular note are the ‘arithmetic observationists’ who sought to quantify the mass of new observations being made in medicine in the late 18th century, and exemplified by William Black who in his text ‘An Arithmetical and Medical Analysis of the Diseases and Mortality of the Human Species’ wrote in 1789:

‘... however it may be slighted as an heretical innovation, I would strenuously recommend Medical Arithmetic, as a guide and compass through the labyrinth of the therapeutick.‘

The preparation of systematic reviews in epidemiology goes back at least 100 years. Chalmers et al.21 remind us of an early review and meta-analysis of 11 studies by Karl Pearson who in 1904 reviewed evidence of typhoid vaccines using many of the strategies expected in modern systematic reviews. Winkelstein22 has also brought to our attention the early work of Joseph Goldberger who in 1907 reviewed 26 studies concerning the frequency of urinary infection in cases of typhoid fever. Goldberger also followed many of the maxims of modern research synthesis. It remains an interesting question why epidemiologists today have only rarely continued in the early tradition of more empirical research review.

To test (in an admittedly simple manner) the hypothesis that epidemiology reviews are not meeting modern standards of research synthesis, I analysed 39 reviews in 5 recent issues of Epidemiology Reviews, the pre-eminent source for reviews of the epidemiology literature. I asked three questions of each review, all reflecting some but by no means all of the principles frequently used to characterize a high quality systematic review within evidence-based medicine, as promulgated by the Cochrane Collaboration. First, did the review address a focused research question based on well-defined a priori exposures being related to a defined pattern of disease. Second, was the method of locating evidence described in detail in the review. Third, were explicit criteria prespecified to indicate the rationale for including or excluding a study. These criteria are also the first three in a larger set of criteria used by Mulrow and colleagues to examine the quality of medical reviews. Importantly, the use of meta-analysis was not a criterion for a systematic review. Systematic reviews do not require a meta-analysis, which may be deemed inappropriate because of sparse or heterogeneous results, and not all reviews which include meta-analyses follow the requirements of a systematic review. The choice of studies meta-analysed may be serendipitous rather than being based on a well-defined protocol.

Table 1 shows the result of the analysis of the epidemiology reviews and compares them to the results of Mulrow and colleagues who recently updated their earlier examination of medical review articles. The single criterion that epidemiology reviews most commonly meet is to have the review address a focused well-defined question although this was still only met in about half of reviews (49%). Providing a description of the methods used to locate the evidence being reviewed in the form of prespecified criteria for data base searching (15%) and using explicit criteria to select studies included in the review (10%) were rarely met criteria. Reviews in epidemiology show a similar lack of rigor to those in medicine generally and are methodologically inferior to meta-analyses, systematic reviews and overviews.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Review addressed a focused research question</td>
<td>80</td>
<td>14</td>
<td>49</td>
<td>95</td>
</tr>
<tr>
<td>Described method for locating evidence</td>
<td>2</td>
<td>28</td>
<td>15</td>
<td>95</td>
</tr>
<tr>
<td>Used explicit criteria to select studies</td>
<td>2</td>
<td>14</td>
<td>10</td>
<td>68</td>
</tr>
</tbody>
</table>

a From Epidemiologic Reviews Vols 19–21 omitting methodology reviews, editorials, and a historical review of one major trial.
b A subset of the 158 Medicine reviews described as ‘meta-analyses’, ‘systematic reviews’ or ‘overviews’.26
Epidemiologists are not alone in having neglected the need to construct methodologically rigorous and unbiased reviews of research evidence. Chalmers et al. document calls for systematic reviews in physics, education, psychology, and the social sciences. They suggest:

‘Many, if not most people working within academia, have not yet recognized (let alone come to grips with) the rationale for and methodological challenges presented by research synthesis. Research synthesis is only now beginning to be seen as ‘proper research’ in the corridors of academic power. As much as anything else, this change seems to reflect the fact that the consumers of research are pointing out more forcibly that the ‘atomised’, unsynthesized products of the research enterprise are of little help to people who wish to use research to inform their decisions.

If epidemiologists fail to use modern methods of scientific review to derive unbiased syntheses of study results, is it any surprise that journalists do not do so either? It has always been a premise of scientific reporting that after describing a study’s new findings, the investigator has a duty to synthesize the new results into the extant body of evidence on the topic. This aspect of epidemiology reporting may be occurring less frequently, perhaps because of editorial pressures to reduce the length of articles or perhaps because students are not being trained in this aspect of report writing; this itself is an area of research. Clinical trial reports have been found to inadequately synthesize their results within the current body of comparable evidence. A systematic review should validly reflect the current state of knowledge on a given topic and should form the basis for scientific reporting. If there were more concurrent systematic reviews in epidemiology, and new research findings were routinely discussed within the context of a systematic review, it would be a relatively easy task to refer the inquiring journalist or policy maker to the discussion section of a paper for an explanation of how the new report had changed the totality of evidence, if at all.

The study by Swaen et al. uses an innovative, albeit imperfect, research design to investigate sources of bias in the epidemiology literature. In doing so, it joins a growing body of literature on the science of systematically reviewing and analysing research evidence. The Cochrane Library includes a methodological data base of some 1350 titles. It is worth noting that scholars of evidence-based medicine have largely focused their attention on randomized trials, the methodology widely considered to be the gold standard of study design, but even here there remain concerns about reviewing evidence based on trials. How much more difficult will be the review of areas of research based on observational study designs and how much more likely the chance for bias, confusion and error?

The limits of epidemiology are most likely faced when studying associations of rare disease with rare exposures, and some of the characteristics of these studies are found in the occupational studies forming the basis for Swaen et al.’s analysis. Would a comparable review of a more common exposure with a common outcome lead to similar conclusions? Only more study will tell. However, it is the rare exposure-rare outcome that increasingly tests epidemiology. As individual studies become more challenging then systematically reviewing the evidence from these studies will pose its own increasing difficulty. It may be that unsystematic and poorly conducted reviews of the smoking-lung cancer association would still correctly conclude that an association existed simply because of the strength of the relationship under study. This is less likely to happen as epidemiologists focus on rare disease-rare exposure associations. In these instances, the science of conducting high quality evidence-based reviews becomes increasingly critical if epidemiology is to credibly inform the public of the current risks to health to which it believes it may be exposed.

Acknowledgement

I am grateful to Iain Chalmers for comments on an early draft of this paper.

References

4 Laupacis A. Methodologic studies of systematic reviews: is there publication bias. Arch Intern Med 1997;157:357–58.
6 Ojasoo T, Doré JC. Citation bias in medical journals. Scientometrics 1999;45:81–84.
Commentary: Prior specification of hypotheses: cause or just a correlate of informative studies?

David A Savitz

Published epidemiological studies are prone to spurious positive findings. This is not just an issue bearing on the discipline’s credibility to outsiders but a fundamental methodological concern. Epidemiologists must accept the challenge to improve research methods, publish findings regardless of their implications, and objectively appraise the validity of our results and those of our colleagues. Results are often dichotomized as ‘positive’ or ‘negative’, despite the loss of quantitative information resulting from this practice. The aetiology of false positive reports is surely multifactorial. Some of this falls to the media and the public for overinterpretation. Some results from the exuberance of investigators who advertise their most surprising, dramatic findings, despite the fact that results that run counter to the conventional wisdom are most likely to be erroneous. Human beings (not just epidemiologists) can become enamoured with their own achievements, lose objectivity, and seek the fame and fortune that result from startling discoveries. We need to improve the resolution of our methods and devote greater energy to helping to ensure appropriate use (or lack of use, in many cases) of our findings by policy makers and the public.

Swaen and colleagues1 have taken on the important goal of improving understanding of the aetiology of false positive studies, searching for causes based on past research that could be applied to future studies to help distinguish between true positive and false positive findings. Such identifiers would enable us to place a more appropriate level of confidence in study findings, discounting some and paying more attention to others. The authors deserve credit for attempting to bring some empirical, quantitative evidence to bear on this important issue, but some practical and conceptual barriers constrain the effectiveness of the search and threaten to introduce false positive predictors of false positive studies.

Formal specification of prior hypotheses, while empirically predictive of more valid positive findings, is an artefact, not a cause of such accuracy. In order for the hypothesis to be defined in advance and narrowly focused, for few statistical tests to be conducted, and for the study not to be categorized as a ‘fishing expedition’, the prior evidence in support of the hypothesis must be sufficiently strong. The biological context, experimental support, or prior epidemiological studies presumably lay the foundation that enables the researcher to specify a hypothesis for evaluation. The act of articulating the hypothesis obviously does not magically confer improved quality to the study. The prior evidence in support of the hypothesis simultaneously...