The shortfalls of observational epidemiology in terms of the generation of contradictory and spurious findings have been highlighted by many commentators—perhaps most memorably by pop-science journalist and general practitioner, James Le Fanu, who stated that ‘the simple expedient of closing down most University departments of Epidemiology could both extinguish this endlessly fertile source of anxiety-mongering while simultaneously releasing funds for serious research’.1 Serious critiques of observational epidemiology have highlighted its vulnerability to confounding, reverse causation, measurement error and selection bias.2,3 However, the closure of University departments and the abandonment of observational epidemiology might be premature without first trying some remedial steps.

One important step would be to improve the reporting of observational epidemiology studies. Lessons may be learned from the early days of randomized controlled trials where it was obvious that not all trials were equal—some were clearly better done than others and consequently produced more trustworthy results. Exploration of the sources of bias in the conduct of trials identified characteristics that were linked with the validity of estimates of effect sizes, and led to criteria that could be used to improve trial quality and, importantly, should help users to distinguish well conducted from poorly conducted trials.4–6 An interesting issue became apparent as more and more energy was put into abstracting data from published trials for systematic reviews—the quality of a trial is often difficult to assess from the published report of its findings. Published reports are not research protocols. Methods sections are often too brief for adequate understanding of how the trial was done. The vagaries of editors anxious to save space for the more exciting findings over the details of what was actually done, make large cuts to methodological detail but this detail is essential in assessing whether the findings are valid or not. With these concerns in mind the CONSORT—CONsolidated Standards Of Reporting Trials—statement was derived.8 CONSORT has required revisions to take into account reporting of adverse events in trials,7 non-inferiority and equivalence trials,8 and has been updated.9 Moreover, application of CONSORT appears to have improved the quality of reporting of trials.10,11

In an effort to improve reporting of observational epidemiology studies, the STROBE (STrengthening the Reporting Of Observational studies in Epidemiology)12 initiative was established in 2004, including several of the people involved in CONSORT. The group will publish its main statement in several international journals simultaneously and a longer explanatory paper giving a detailed rationale will be published in Epidemiology. STROBE decided to limit its initial work to three major designs—cohort, case-control and cross-sectional. This was wise. If the approach proves applicable in practice and improves reporting of the major study designs, its application to other designs such as case series, audits and database studies can be considered in updates.

The authors of STROBE state many things that the checklist of 22 items is not. It is not intended to be prescriptive, to regulate terminology, to be used to assess study quality or to guide the planning or design of studies. So what is it actually for? It is meant to assist authors in writing up their work, help editors and reviewers as part of the peer review process, and make it easier for readers to critically appraise published papers. These purposes and some of the items do seem to overlap with assessing study quality as indicated by the similarity of items identified in a systematic review of tools to assess quality and susceptibility to bias.13

So how successfully have the STROBE group risen to the challenges of the varied nature of observational epidemiology? The effort to be non-prescriptive has resulted in rather general suggestions for authors, some of which will be found in any epidemiology text targeted at Masters students in the first term of their first year. Despite the companion explanatory long paper, the investigator seeking guidance may end up confused. For example the item ‘State the scientific background and rationale for the investigation being reported’ sounds unambiguous but in long-term cohort studies the original rationale may have been understanding the causes of cardiovascular disease but the purpose of the report is study of a quite different condition—say incontinence of urine. The original purpose of the cohort will have a bearing on the types of covariates available for investigation and will illuminate what is feasible for the investigators to attempt in analyses but ultimately the availability of the latter are the important factor.

1 Co-Editor IJE and Coordinating Editor, Cochrane Heart Group, LSHTM.
2 Director UK Cochrane Centre, Oxford.
* Corresponding author, London School of Hygiene & Tropical Medicine, Keppel Street, London WC1E 7HT. E-mail: Shah.ebrahim@lshtm.ac.uk
The STROBE group recommend ‘cautious overall interpretation of results.’ If observational studies are largely incapable of making definitive conclusions on the basis of robust findings perhaps Le Fanu is correct and it is now time to close the enterprise down! A better approach would be to give an interpretation of the results appropriate to the available data. Did Richard Doll and Ernest Wynder make only a tentative interpretation from the series of observational studies conducted from 1948 on smoking and lung cancer? Certainly conclusions from the initial studies were cautious given that the associations between lung cancer and smoking were somewhat unexpected at the time but stronger inferences were possible as evidence from both retrospective and prospective studies emerged and were confirmed by others. Doll’s important statement ‘if you find something that is unexpected and is going to be of social significance you have a responsibility to be sure that you’re right before you publicize your results to the rest of the world. This does at least require repeating some of your observations’ should find a place in the first update of STROBE.

Asking investigators to make clear ‘which confounders were adjusted for and why they were included’ might deal with the curious case of the changing effects of folic acid on stroke risk published by one research group in the same year. In the first study, participants in the physicians health study enjoyed stroke protection with high folate levels. The second study published later that year using data from the nurses health study demonstrated no important effect of folate on stroke risk. One contribution to the apparent discrepancy may be that the ‘correct’ null answer for the association of dietary folate and stroke derived from a randomized controlled trial was published between the two observational studies. The trial was actually cited in the second observational study in support of the negative finding. Both observational studies showed similar effects in unadjusted analyses but in the second study the effects attenuated on adjustment for potential confounders. A major difference between the two studies was the approach taken to adjusting for confounders. The second study adjusted for more covariates, and included Vitamin E adjustment (which is strongly correlated with socio-economic position) and might therefore contain better adjustment for confounding by socio-economic position. A repeat analysis of the initial positive study with Vitamin E adjustment would be interesting. It remains to be seen whether a further attempt to get the new ‘correct’ answer will result from the recent meta-analysis of randomised trials of folic acid supplementation which showed almost a 20% reduction in stroke risk.

STROBE is perhaps missing the point in suggesting ‘a discussion of the existing external evidence is particularly important for studies reporting small increases in risk.’ Reports of small increases in risk generally evoke a ‘so what’ response and are not nearly so dangerous as reports of large increases (or decreases) in risk. For example, the meta-analysis of observational epidemiological studies of the association between hormone replacement therapy (HRT) and coronary heart disease showed a ‘too good to be true’ reduction of about 50%, prompting the authors to conclude ‘overall, the bulk of the evidence strongly supports a protective effect of estrogens that is unlikely to be explained by confounding factors.’ Women took heed and started taking or continuing HRT for its alleged cardio-protective effects. It is in precisely these circumstances that more external evidence is required in order to protect the public from false claims. In this case the clue had been reported earlier by Pettiti and colleagues who found a lack of specificity of effect of HRT—it ‘protected’ against accidents and violent deaths too. The same confounding structure probably explained the associations with accidents and violent deaths and also with CHD, as indicated by the subsequent trials of HRT.

One of the really useful aspects of CONSORT to us as Cochrane systematic reviewers has been the vast improvement in reporting which makes the assessment of the quality of new randomized trials so much easier when updating a review, than the assessment of the old trials when the reviews were first done. With systematic reviewing in mind, it is disappointing that STROBE does not recommend that new findings are put into context by conducting a systematic review of other similar studies. This would alert us to the consistency of findings, allow exploration of sources of heterogeneity and would ensure, in the same way as it does for randomized trials, that research effort is not spent on rediscovering the same finding again and again.

STROBE will likely make a strong contribution to improving the quality of reporting of observational studies. In this issue Debbie Lawlor considers a different approach to improving the quality of epidemiological research by reducing publication bias—submitting papers for editorial appraisal and peer review without the results, which would permit papers to be assessed on the importance of the objectives and quality of the methods and statistical approaches to be used. In an accompanying commentary Sander Greenland offers some support for this suggestion and notes the error in assuming, as many epidemiologists do, that publication bias (in observational epidemiology or RCTs) is nearly always of null studies: trying to conceal ‘positive’ findings from publication also introduces important bias to scientific understanding.

We would like to see more discussion on how observational epidemiology can be made more robust—and on the feasibility of implementing these new options for authors and editors.

References
Commentary: Strengthening the reporting of observational epidemiology—the STROBE statement

Matthias Egger,1* Douglas G Altman2 and Jan P Vandenbroucke3 of the STROBE group

Accepted 29 August 2007

We welcome Ebrahim and Clarke’s comments1 on the STROBE Statement and are grateful for the opportunity to clarify some of the issues they raise in their editorial. What is STROBE all about? The STROBE Statement is a checklist of items that should be addressed in articles reporting cohort studies, case-control studies or cross-sectional studies, to STrengthen the Reporting of OBServational studies in Epidemiology. A short paper that presents the checklist and explains how it was developed will be published in several journals2 in October 2007, and will be freely available on the websites of these journals (see www.strobe-statement.org for links to the paper). The intention is to provide guidance on how to report observational research well: the recommendations are not prescriptions for designing or conducting studies—these decisions must be made by investigators who know the subject matter. Also, while clarity of reporting is a prerequisite to evaluation, the checklist should not be seen as an instrument to evaluate the quality of observational research. Good reporting does not necessarily

1 Institute for Social & Preventive Medicine (ISPM), University of Bern, Bern, Switzerland.
2 Centre for Statistics in Medicine, Oxford, UK.
3 Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, The Netherlands.

* Corresponding author. Institute for Social & Preventive Medicine (ISPM), Finkenschiebelweg 11, CH-3012 Bern, Switzerland.
E-mail: egger@ispm.unibe.ch