Reviewer's report

Title: Efficiency of two-phase methods with focus on a planned population-based case-control study on air pollution and stroke

Version: 3 Date: 29 June 2007

Reviewer: Roseanne McNamee

Reviewer's report:

General
The manuscript is now much easier to read. The Objectives of the paper are now clearer. There are still areas that need improvement.

1. The Background/Introduction does not really link up with the substantial literature on 2-phase design and analysis: it mainly reviews the literature on air pollution instead. But this is a methodology paper! An additional document is entitled “overview of 2-phase methods”: why is material from this document not incorporated into the background? This is the traditional way of explaining what is already known so that the new method in a research paper can be justified. Without that we cannot tell what is new.

2. Importance of design

The issue of how subjects come to be included in the 2nd phase of the simulations is not clear nor is the general importance of this question acknowledged.

(a) The introduction should clarify that, in general, a non-random design or even stratified random designs with incorrect analysis will lead to bias. Also the issue of self-selection needs to be differentiated from active selection by researchers as part of a study design. In the latter case some data is missing (randomly) by design and bias can be avoided; in the former it is not and bias may be unavoidable. ‘Participation’ seems to be used to mean both things in this paper – leading to confusion. Mention of ‘participation’ in the Abstract and p 14 seems misleading – it seems to imply that two–phase studies could overcome the problem of bias caused by self-selection.

The discussion does eventually pick up these issues (cf participation bias, p13) but it would be clearer if the general principles of 2-phase designs had already been set out in the Introduction - ie which combinations of design and analysis method are valid? In general a 2-phase design, where data is MAR - see below-accompanied by an analysis which acknowledges the design will be unbiased.

(b) The simulations

Is the random sampling plan (p10) stratified by disease status only or by disease status and area? The abstract Conclusion refers to varying participation rates
across areas. This is nowhere mentioned in the Background. The Methods seem to imply that selection probabilities vary across areas but they do not give any information.

Discussion (p13) ‘Missing at random (MAR)’ assumption is NOT violated if selection probabilities vary by area although the Missing Completely at Random (MCAR) assumption will be (cf Little and Rubin usage of these terms).

It is earlier stated that all of the analysis methods are unbiased for the design used in the simulations but the discussion refers to the potential bias of Method 1 (presumably for designs other than those simulated) – more details are needed.

3. Description of the design of the study (as implied by the simulations) should precede the description of the analysis methods since choice of correct methods depends on design.

It would be helpful to distinguish between 2-phase design and 2-phase analysis. The idea of a 2-phase design is not new; the authors’ proposed ‘method’ is a method for analysis of certain 2-phase study designs.

P5, line 10 should read: “we simulate data from a two-phase design”.

P10, paragraph 5, 1st sentence. Distributions of what?

P12, last part of paragraph 1 beginning “In Table 3….”. I cannot understand the point being made here.

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

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

Discretionary Revisions (which the author can choose to ignore)