Reviewer’s report

Title: Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study

Version: 1 Date: 16 September 2009

Reviewer: Agus Salim

Reviewer’s report:

This article is well-written and addresses important practical questions for researchers who want to use better, but yet relatively simple, alternatives to complete-case (CC) analysis when faced with missing covariates data. However, I have the article suffers from lack of clarity at various points, especially in relation to the description of the simulation setup and data generation. I am also concerned the conclusion that CC being unbiased under missing-at-random (MAR) mechanism cannot be generalized to other settings and the authors need to discuss more about the reasons for this discrepancy between their finding and previously published report (eg, Demissie et al., 2003).

Major Revisions
1. The Methods section needs to be explained more carefully. For example, it would seem to be all continuous covariates are Normally distributed (p.7), however later on p.11 it was revealed that log-transformation was needed to make these covariates more Normally distributed. What was generated at the first place, was it the logarithm of these covariates? What transformation was used to transform the generated data (p.7), was it an antilog (exponential) transformation?

2. Why do you need to generate moderately independent simulation? (p.7). I am worried that this strategy introduces strong dependence between subsequent simulations. Is it possible to generate each simulation independently as the usual norm in simulation studies?

3. I don’t really understand the explanation on the amount of missingness imposed on each of the four covariates (p.7). Why the percentages do not sum to 100%? Also, how do you keep these percentages fixed while varying the overall rates of missingness per case?

4. p.9-10 and Table 2: it may be useful to refine the class of missing data mechanisms by introducing another class called ‘covariate-dependent’ (Little, 1995). Covariate-dependent (CD) mechanism occurs when the missingness depends on the observed covariates only. MAR is then defined as missing mechanism where missingness depends on both observed covariates and outcomes. This differentiation is important because CC is expected to work under CD, but generally not under MAR.

5. Results section: I did not expect CC to be almost unbiased for X5 under MAR (Figure 2). The results displayed here demands a closer inspection to rule out
possibility of bugs in the R codes used. It is even more puzzling to see that the bias in CC estimate is getting smaller as the amount of missingness increases. Can this due to the ‘moderately independent’ simulation setup?

6. Discussion section: the authors claim that using CC results in loss of efficiency (p.17), and thus MI is preferred with increasing amount of missingness (p.22). However, their results show that MI can produce substantially more biased estimates then CC. Surely, it is crucial to ensure we have unbiased or almost unbiased estimates first, before we can talk about efficiency/power?

Additional references

**Level of interest:** An article whose findings are important to those with closely related research interests

**Quality of written English:** Acceptable

**Statistical review:** No, the manuscript does not need to be seen by a statistician.

**Declaration of competing interests:**

I declare that I have no competing interests