Reviewer's report

Title: Fitting parametric random effects models in very large data sets: An application to VHA national data

Version: 1 Date: 21 May 2012

Reviewer: Mariel McKenzie Finucane

Reviewer's report:

- Major Compulsory Revisions

1. The primary outcome is HbA1c, and thus the primary model is a simple linear mixed model. In this model, you can integrate out the subject-specific random effects $b$ to obtain $Y_{ij} | # = X_{ij} \beta + e_{ij}$, with $e_{ij} \sim N(0, \Sigma)$, where $\Sigma$ is a covariance matrix that accounts for both within- and between-subject variance. See, for example, equation 12.14 on page 265 of Gelman and Hill's “Data Analysis Using Regression and Multilevel/Hierarchical Models” (2007). It is computationally feasible to fit this model to the full dataset. This option should be addressed in the manuscript.

- Minor Essential Revisions

1. In paragraph #6 of the background, you describe an approach due to Owen and the Huang and Gelman. In what way is REMR an improvement over their method?

2. In the last paragraph of Section 3.1, please give the variance of the $b$'s and please define the matrix $G$.

3. In section 3.3.2 you introduce the terms $v_i$ and $\epsilon_i$. Please define these terms. (For example, as you defined $\tau$ above: “$\tau$ is the pooled mean.”)

4. The first sentence of Section 3.4.1 is unclear. “Obtaining pooled estimates assumes homogeneity of VISN level effects.” Please clarify that your goal is to estimate a single global effect summarized across VISNs.

5. Last sentence of 3.4.2.: “It is only useful to compare AIC/BIC values from the sampling models to the full data models.” Do you mean that it is not useful to compare AIC/BIC values from the sampling models to the full data models, since they differ in their pseudo data?

6. In Section 3.4.3, please describe the design of your simulation study more clearly. Did you resample 1000 datasets of each size? In Tables 1, 2, and 3 and in Figure 3, are you summarizing across datasets of each size? How are you performing this summarization?

7. In the first paragraph of Results, please clarify what the “four sampling levels” are.

- Discretionary Revisions
1. The impact of this paper would be greatly heightened if you included code for implementing REMR.

- Minor issues not for publication

1. Please define “VISN” in the abstract.
2. Please clarify the following sentence in the abstract. “Our results indicate that REMR provides parameter estimates that are less likely to be biased with tighter confidence intervals when the VISN level estimates are homogenous.” When I read “biased with tighter confidence intervals”, I thought the bias you were referring to was an underestimation of uncertainty.
3. Please define “FEMR” in Section 3.3.1.
4. In paragraph #2 of Section 3.2, you refer to Equation #1. Which equation is this?

**Level of interest:** An article of importance in its field

**Quality of written English:** Acceptable

**Statistical review:** Yes, and I have assessed the statistics in my report.

**Declaration of competing interests:**

I declare that I have no competing interests.