Reviewer's report

Title: Using Bonferroni, BIC and AIC to assess evidence for alternative biological pathways: Covariate selection for the multilevel Embryo-Uterus model.

Version: 1 Date: 18 March 2013

Reviewer: Chad C Brown

Reviewer's report:

The authors provide a well written and logical description of a simulation study aimed at exploring various hypothesis testing and model selection methods for covariate selection in embryo-uterus models. Overall, the quality of the manuscript is very good. However several important items could further improve the quality substantially.

Minor Essential Revisions:

1. Embryo, uterus and Type (as in Type I error) are not proper nouns. These should not be capitalized.

2. Page 8, Equation 2: Please explain what delta(k) is.

3. Page 6, Background: You state, "it is clearly anti-conservative to test the parameter only in the sub-model selected (by whatever procedure)." The reason(s) for this anti-conservative nature is not clear to me. Please explain.

4. Page 7, Section 1.2: In "This present work presents a series of simulation studies which aimed", aimed should be aim.

5. Section 2.1 needs a carriage return in the title.

6. Page 8, Equation 2: Here, Pi is the probability of pregnancy for patient i, and ui is the uterine receptivity for the ith cycle. This is confusing, since i represents two different things. Please use different notation.

7. Page 9, Section 2.2: Please change "The model can readily fitted to data", to "The model can be readily fit to data".

8. Page 11, top: The phrase: "giving estimates giving" is awkward. Please reword.

9. Page 14, Section 3.1: "The two cases where the true effect is in the embryo sub-model embryo", the second "embryo" might be a typo.

10. Page 15, Section 3.1: Please change, "However if the model specification is incorrect there is a non-insignificant" to "However if the model specification is incorrect there is a significant"

11. Page 16, Section 3.2: The phrase, "From the table we also observe that the
BIC differences proportions of correct classifications increases slightly as sample size increases" is awkward. Please reword.

12. Page 17, Section 3.3: The statement, "As the dataset was large, all the potential factors were included in the model, but a mixture of pre-specification and selection on the basis of AIC was used to determine which of the sub-models (E, U or both) would include each factor." is awkward and hard to follow. Please rephrase.

13. Page 23, Figure 1 caption: In "Statistical Power", power should not be capitalized.

Major Compulsory Revisions

1. Page 4, Background: You state, "Analysis of data arising from IVF treatment often includes prognostic factors observed at an embryo level." Later you mention that, "but that identification of the appropriate level may have a biological interpretation with clinical consequences." but give little further discussion about why studying models for IVF is important. It's not clear how exactly a better EU model leads to improvements in IVF treatment. It seems logical that there would be some connection, but this should be stated explicitly.

2. Page 9-11, Section 2.3: The exact algorithm that you used to generate simulated data needs to be made more clear. It appears that, U, Ep and Ee are generated using the described distributions, these values are used to calculate ui and eij using Equation 3, and then Equation 2 is used generatively to create a realization for a k-fold pregnancy. However, these steps are only my best guess, and they need to be explicitly stated.

3. Page 9-11, Section 2.3: On page 5, you mention that, "As noted by Roberts et al. [10] the level at which a covariate acts is identifiable (albeit not strongly) through the twin rate if there are multiple embryos transferred per cycle." Intuitively, this makes sense. It appears that this factor would then be critical for model selection in Section 2.5. The twin rate for simulated data is not reported. This value should be reported, and checked to make sure that it matches real data. Ideally, simulation conditions would be altered such that a few reasonable values for the twin rate were generated, and the performance of model selection for each could be compared. Similarly, the simulated pregnancy rate should be reported and checked to make sure it's reasonably representative of real data.

4. Page 9, Section 2.3: You mention that, "The U, Ep the distributions in the multi-centre dataset:”, but never describe what these variables actually in the towardSET data set. Please give a short description. In addition, how were the distributions for U, Ep and Ee determined? If they were modeled directly from relevant variables from the towardSET data set, then why were do all of the beta coefficients equal one? Also, how were the intercept terms decided? I think this section needs to be described more carefully, because the exact choices in simulation conditions may have a large impact on the results.
5. Page 20-21, Section 3.5: A major limitation of the current simulation was that data were simulated assuming that the true model was known. This is acknowledged with the statement, "The major caveat is the underlying assumption in the simulation work that the EU model does reflect the true behaviour and further work is required to relax this assumption." This is followed with the statement, "Given its rich structure it is unlikely that any realistic dataset will have sufficient power to detect departures from the EU model." However, no evidence is given to support this latter statement.

On the other hand, several strong assumptions are made in the EU models. Just as an example, the assumption of independence between the embryo viability and uterine receptivity, might be questioned by some. It is not clear how well the proposed methods perform when modeling assumptions are violated. Ideally, simulations would include models that are slightly different from the assumed model, and performance under model misspecification could be evaluated. At the very least, the authors should identify the obvious possible sources of misspecification (such as E-U independence), and describe how these may lead to a loss of performance, in practice, if real data did not behave according to the proposed model.

6. Figure 4 did not appear in the manuscript (the caption appeared, just not the actual figure). I also sent a statement to the editor, in case this was an editing mistake.

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

**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