Response to Referee's report of SIM submission 13-0382.R1 entitled "Comparing and combing biomarkers as principle surrogates for time-to-event clinical endpoints"

Erin E. Gabriel, Michael C. Sachs, Peter B. Gilbert

July 17, 2014

We are happy to have the opportunity to revise our manuscript. We have put forth considerable effort to address reviewer 2's comments and we feel that the paper has improved greatly.

We have made some critical revisions to the methods section, and have modified the real-data example to better illustrate our proposed methods. We have developed a new summary statistic that does not require the selection of a specific time. During the development of this new summary statistic, we noticed a minor convergence issue for the multivariate PS estimands, that can lead to lower efficiency in practice. This issue was previously masked by our truncation of the STG for a given time point. We now suggest using a multivariate BIP, one BIP for each candidate PS, this appears to remedy the convergence issues and improve efficiency over a univariate BIP.

We have spent several months attempting to locate an adequate new example data set. Although we were unable to do so, we tried several, including the Mr. FIT data, and several small, unpublished immune marker trials sponsored by NIAID. In each case there was inadequate baseline information to support the estimation of the counterfactual estimands. However, in running each of these examples, we determined that the use of truncated normal in the simulations did not fit what we were suggesting in practice for the example. We have modified these simulations to use censored normal generation and censored imputation. Below we outline the specific changes.

1 Comments from Reviewer 2

> I have to admit that I am somewhat disappointed with this revision.
>
> Only minor changes in the essential section 2 have been made.
>
> The paper is still not sufficiently self-contained and one has to read at least [1] and [10] to have a chance at understanding what's going on in the paper. Both [1] and [10] are very nice and interesting papers by the way.
>
> The extension from [10] in terms of modelling is straightforward and minor.

The tools developed to assess and select one or more surrogates are problematic at least to me as they include the subjective choice of one or more time points at which performance is evaluated. Also, the actual selection of an optimal subset of surrogates is not well described.

It would be ideal with an example where these extensions really mattered.

2 Response to Reviewer 2

1. We reorganized the methods into sections 2 and 3. Section 2 now contains the notation and background information on the concept of principal surrogacy. This section provides a more thorough introduction to the methods so that uninitiated readers may better understand without referring to [1] or [10].

2. The remaining subsections of the methods have been reorganized so that the concepts flow more naturally from defining estimands of interest, to assumptions, to modeling, to estimation, to summary statistics, and finally to PS comparison.

3. We have added an additional subsection where we introduce the integrated standardized total gain to address reviewer 2's concerns about the subjective choice of time points. The integrated STG does not require the analyst to specify a time point, but rather averages the time dependent STG with respect to the distribution of event times. We suggest estimating the marginal distribution of event times with the Kaplan-Meier estimator. We evaluate the integrated STG in the simulations and demonstrate its operating characteristics. The integrated STG tends to have slightly higher power for comparison of candidate PS as compared to the STG at a single time point.

4. In regard to the reviewer's comment about selecting the optimal subset of surrogates, we have revised section 3.7 to more clearly describe the hypothesis tests that are used to rule out useless PS from consideration and for comparing and ranking candidates with some value.

5. We have modified the example to better illustrate our extensions. We created a third candidate PS from a linear combination of the two real candidate PS such that the PS has some value as a surrogate, but with a significantly smaller STG. In this case the test for the difference in STG yields evidence of a difference between the two surrogates where standard hypothesis testing could not. This makes the example more like a simulation, in that we modified the data to generate the desired results, but we feel that it does provide a better illustration of where our proposed methods actually matter.

6. We have added to the multivariate simulations the investigation of the use of a multivariate BIP. We find that the use of the vector BIP improves convergence and efficiency over the use of a univariate BIP for multivariate surrogates. Although it was suggested in [1] that the EML method applies directly to a vector BIP, we do

not believe simulations have ever been performed to demonstrate this approach. We now include simulations using the vector BIP for the multivariate candidate PS in the simulations. We include the simulations using a univariate BIP in the appendix for comparison.

7. We have updated the simulations to reflect our suggested censoring of a candidate surrogate, as done in the example, rather than truncation. We also now make it clear that when censoring the candidate surrogate, we use censored imputation.

We thank you again for the thoughtful review; the paper has improved greatly thanks to your efforts.

Sincerely,

Erin E. Gabriel