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

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

November 19, 2013

We are happy to have the opportunity to revise our manuscript. We are very pleased that reviewer 1 felt that the ideas presented in our manuscript are of potential interest and value and that our theoretical developments seem solid. We apologize for the slow turn-around on this response; we waited until our cited other work, (Gabriel and Gilbert 2013), citation number [10] in the revised manuscript, had been accepted for publication. We have however attempted to make this manuscript as stand-alone as possible. We have reorganized the sections to improve readability and reduce the technical feel of the manuscript.

1 Response to Reviewer 1

We felt that Reviewer one had two main points and two minor points, and we hope we have addressed all in the revised manuscript.

- Main point

    1. In its present form, the paper is hard to appreciate as a stand alone. Moreover, the application is hard going on impossible to understand for me at least without the details apparently provided in [10]. Much of the intuition and rationale behind the ideas presented is omitted with reference to another paper [10] by two of the co-authors. This paper is to my knowledge presently not available for a general reader. In conclusion, for the paper to be readable, the authors need to provide a much more self-contained presentation of their ideas and how they in effect differ from the developments presented in [10]. In particular, we find this to be very problematic in the crucial section 2 where an in detail motivation for using F1 and F0 as the cornerstone of evaluating surrogates lacks.
    We have added slightly more detail about Gabriel and Gilbert 2013,citation [10], that has now been accepted for publication in

Biostatistics. This manuscript in its accepted form has been included in the revisions submission materials. In addition, we have taken care to make sure the explanation of the methods is as standalone as possible throughout the manuscript. We have more clearly pointed out the novel methods outlined in the manuscript in the introduction. We have also reorganized the sub-sections of the methods section to reduce the technical feel of the manuscript and make the methods easier to follow. We believe the manuscript is now much more self-contained and the advancements are easier to see and less weighed down in technical detail.
We feel that the manuscript introduces a set of tools that can be used in practice to estimate the causal risk estimands and evaluate and compare univariate and multivariate candidate principle surrogates for continuous time-to-event clinical endpoints. Our proposed estimation method for the risks was previously introduced in the literature, but has only been explored for use with binary clinical endpoints. What is more, similar estimation methods have been investigated for continuous time-to-event endpoints, but only for univariate candidate surrogates. Many of the suggested summary statistics were introduced in previous works to evaluating diagnostic tests in a time-to-event setting, however they have not been suggested previously for evaluating principle surrogates. The standardized total gain was suggested in the PS literature as a means of comparison of candidate PS, but only in binary-clinical endpoint settings.

2. A more in detail discussion of the assumptions A1-A5 and how they provide a link between the fundamental causal quantities sought and the observational data at hand is lacking. **We have added what we hope to be a more clear statement about what the assumptions A1-A4 imply, directly following their outline in the revised text.**

- few minor details

1. page 4, line 13: perfect, **corrected, thank you**
2. page 4, line 56: we think you're lacking an 1 before the sum. **I have changed this equation for clarity**

2 Response Reviewer 2

1. The presentation is highly technical, with numerous references to your own work, some of which is not published yet. We had a very difficult time reading the manuscript without consulting the references. The paper should be written such that it can mostly be understood on its own. This should include a clarification of

which parts are novel, as surrogate combinations for binary endpoints and surrogate endpoints in a time-to-event setting already seem to have been addressed in other manuscripts.
We have taken care to make sure the explanation of the methods is as stand-alone as possible throughout the manuscript. We have reorganized the sub-sections of the methods section reduce the technical feel of the manuscript and make the methods easier to follow. We believe the manuscript is now much more self-contained and less weighed down in technical detail.

2. The application example is somewhat disappointing, as no real differences between the surrogates (and their combination) can be seen. This might either mean that there are no differences, or that the proposed tools are not adequate. You should at least present one real application, where the proposed tools are useful for identifying a promising surrogate combination.
We found an error in the analysis dataset we were using. Now that we have corrected this issue the two candidate surrogates do not perform similarly, however the combination surrogate is still not an improvement over HBA1C alone. We selected these data to use as the example because HBA1C had been identified in our previous work as having surrogate value and it was hypothesized that EGFR might have some surrogate value and add to a combination surrogate of the two. EGFR was also selected because a baseline predictor could be found that was adequately correlated with both it and HBA1C. As with any real data analysis we did not know the results a priori and we do not know the true values of the coefficients. The simulations are used to show the adequacy of our proposed tools under known scenarios, we believe the example is of clinical interest and not run simply as an illustration of our methods.

3. One specific aspect of the application example that needs to be clarified is why a combination of the surrogates can perform worse than its components. A reasonable estimation procedure might set the parameter for one of the components to a value close to zero in such a situation. Why does this not happen here? Is this potentially an instance where the Weibull model does not fit very well?
The combination surrogate being estimated to have less value as a surrogate than either of the components in our example was an error that has been corrected; the same error that caused HBA1C and EGFR to perform similarly caused this issue. We should have caught this error and we apologize and thank you for your question that allowed us to find the error. Although it is no longer the case in our example, we do not rule out the that in practice one could find that the estimated STG or other summaries might be non-significantly lower for a combination surrogate than for one of its components. This may be caused

by the addition of biomarkers to the risk model that are not adequately correlated with the joint BIP used or the addition of useless surrogates that only add noise to the model. As suggested, it could also be the case that the Weibull model does not fit the data well. All of these cases reflect a reduction in the precision of the estimated risks and not a true reduction in PS value. If it is believed that the Weibull model is a poor fit for the data other parametric models can be considered for the clinical outcome and can be easily accommodated by the suggested semi-parametric EML method. The reasons for a reduced estimated PS value of the summaries for a combination surrogate over its components should be investigated, but can be considered reasonable evidence of no improvement given the data available and the risk model assumed.

4. While a naive analysis, such as separately evaluating the prognostic potential of surrogate variables separately in the treatment arms (using regression models or even Kaplan-Meier estimators) and then picking promising candidates by hand, will be obviously biased, but might be the first approach considered by practitioners. Therefore, the proposed approach should be contrasted with such a naive (biased) approach, e.g. to highlight the severity of the bias. In particular, it would be interesting to see the size of the bias relative to problems potentially introduced by the more restrictive model required by your proposal.
We have performed a naive analysis of both the candidate surrogates separately in the arms of the trial using a Cox model. We find that although the ranking of the candidate surrogates would be the same based on the magnitude of the HRs in the treatment arm alone, we didn't have the power to show significance for either candidate's HR. This may be due to our use of the Cox model. In the control arm the ranking of the candidates would be the opposite, showing that HBA1C under control is less associated with outcome than is EGFR under control. The naive analysis addresses a different question than our analysis, mainly the association of the observed candidate and observed outcome in either arm of the trial. Our method evaluates the association of treatment-efficacy and the candidate surrogate under treatment. Therefore, we cannot discuss the bias of the naive analysis versus our method. Although for one of the naive analyses we find some evidence that HBA1C is a better candidate surrogate than is EGFR, this will not always be the case. The literature is full of many strong correlates of risk, biomarkers that are associated with outcome in the treatment arm alone, that are not associated with treatment efficacy and are therefore not useful PSs (Fleming and Demets, 1996). We summarize these analyses in the revised manuscript.
With regard to the Weibull model used, the detection of time-variation in the CEP curve is not the focus of this work, as it was in Gabriel and Gilbert (2013), and therefore the restricted model does not impede

the illustration of our methods. The restricted model is used to reduce the complexity of the technical detail needed to understand the more general concepts that are proposed in the manuscript. We use the restricted Weibull model, which is actually a standard Weibull model in most settings, to help focus the manuscript on the evaluation and comparison of multivariate candidate surrogates in the time-to-event setting and our proposed summary statistics. We believe we have now made this point clear in the revised manuscript.

5. Time-to-event endpoints sometimes are analyzed with respect to a certain time horizon, such as 1-year survival. The proposed criteria for judging surrogates reflect such a fixed horizon reasoning. However, time-to-event endpoints will often be judged over the whole course of time, e.g. using logrank tests. To reflect this, criteria for judging surrogate endpoints should not be limited to a certain time horizon. Therefore, you should extend your proposal accordingly, or at least present results for several time horizons.
We looked at several different time points for the example outcomes and figures. We have added an additional figure to the supplementary materials with these new figures at 3, 4 and 9 year time points. Also, although we illustrate our methods at particular fixed time points t, average risk over time based on the CDF can be estimated using the same methods. We define average risk by,

$$\textbf{avergrisk}_z(s_{1,1},\ldots,s_{1,J}) = 1/k \sum_{t=t_1}^{t_k} risk_z(t|s_{1,1},\ldots,s_{1,J}).$$

Average risk can then be used in any of the suggested summary curves or statistics; those statistics are then time-independent but still account for a time-to-event clinical outcome with censoring. We have added a similar statement to the manuscript.