Author's response to reviews

Title: Baseline Hospital Performance and the Impact of Medical Emergency Teams: Modelling vs. Conventional Subgroup Analysis

Authors:

Jack Chen (jackchen@unsw.edu.au)
Arthas Flabouris (aflabour@mail.rah.sa.gov.au)
Rinaldo Bellomo (rinaldo.bellomo@austin.org.au)
Ken Hillman (ken.hillman@swsahs.nsw.gov.au)
Simon finfer (sfinfer@med.usyd.edu.au)

Version: 3 Date: 30 November 2009

Author's response to reviews: see over
Response to the reviewer’s comment

We acknowledge the different perspectives that the reviewer adopted and thank her for elaborate the views again. Our responses are outlined below.

The reviewer’s comments:

In the MERIT study the primary outcome was the “aggregate incidence of the three adverse events”, however, for the authors analysis they consider the “change in incidence of adverse events”. I do not agree that a change score is more interpretable - it would surely be more appropriate to mimic the primary outcome used in the main study. Also, from a statistical point of view, using an analysis of covariance approach where follow up scores are compared adjusting for baseline is more powerful, producing more precise estimates that are unbiased (Frison & Pocock 1992).

Response: We would like to further explain the rationally underlying our approach and to demonstrate why our approach was acceptable statistically and preferable policy-wise and why, in this case, an ANCOVA approach may make the results hard to interpret for clinicians and policy-makers.

1. Our approach is more policy-friendly and can be interpreted in a intuitive and non-statistical manner

Our approach was designed for ease of policy interpretation as presented in our paper because the results have policy implications. However, if we were to use an ANCOVA approach and the primary endpoint during the study period as our dependent variable, testing the effect modifier gives rise to the following results:

```
.xi:reg pnewcoto2 i.met10*pnewcoto0 i.met10*pnewcotosq [aw=admission2]
i.met10 _met10_0-1 (naturally coded; _met10_0 omitted)
i.met10*pnewc-0 _metXpnewc_# (coded as above)
i.met10*pnewc-q _metXpnewc# (coded as above)
(sum of wgt is 3.6409e+05)

Source | SS df MS Number of obs = 23
-------------+------------------------------ F( 5, 17) = 8.97
Model | 95.9908881 5 19.1981776 Prob > F = 0.0003
Residual | 36.3650467 17 2.13912039 R-squared = 0.7252
-------------+------------------------------ Adj R-squared = 0.6444
Total | 132.355935 22 6.0161785 Root MSE = 1.4626

pnewcoto2 | Coef. Std. Err. t P>|t| [95% Conf. Interval]
-------------+--------------------------------------------------
_met10_1 | -4.460827 3.096788 -1.44 0.168 -10.99448 2.072824
_pnewcoto0 | -1.168329 .762397 -1.53 0.144 -2.776846 .4401885
_metXpnewc-1 | 1.79707 .8235093 2.18 0.043 .0596175 3.534523
(dropped)
_pnewcotosq | .1251794 .0456695 2.74 0.014 .0288251 .2215337
_metXpnewc-1 | -1.389184 .0490843 -2.83 0.012 -.2424772 -.0353595
_cons | 6.632312 2.893968 2.29 0.035 .5265729 12.73805
```

The results indicated a significant complex quadratic interaction effects between the baseline primary endpoint and the treatment groups. Thus, we would interpret this as that the outcome relationships before and after introducing a MET system were
significantly different between two arms. We could leave these findings as they are and present them with full statistical efficiency. However, what are the implications of this broad statement for policy-makers and clinicians? How can we make a sense of such difference? What is the nature of this difference? Can this ‘significant difference’ inform any policy? We found that, for clinicians and policy makers, it is very hard to interpret these findings and then make clinically meaningful interpretations. If we adopted the same regression approach we proposed, the following results are found (Table 1 below).

Table 1. The before and after relationship of the primary endpoint by treatment group

<table>
<thead>
<tr>
<th></th>
<th>Control hospitals</th>
<th>Control hospitals</th>
<th>MET hospitals</th>
<th>MET hospitals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary endpoint</td>
<td>-1.150</td>
<td>0.860</td>
<td>0.642</td>
<td>0.439</td>
</tr>
<tr>
<td>Baseline incidence</td>
<td>(0.145)</td>
<td>(0.004)**</td>
<td>(0.071)</td>
<td>(0.003)**</td>
</tr>
<tr>
<td>Baseline incidence squared</td>
<td>0.124</td>
<td>-0.012</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>6.578</td>
<td>-0.454</td>
<td>2.089</td>
<td>2.684</td>
</tr>
<tr>
<td>Observations</td>
<td>11</td>
<td>11</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.817</td>
<td>0.625</td>
<td>0.616</td>
<td>0.595</td>
</tr>
</tbody>
</table>

*p values in parentheses

* significant at 5%; ** significant at 1%

The results showed that there was a complex quadratic relationship for control hospitals and a positive linear relationship for MET hospitals.

Again, we found that it was difficult to make policy-related interpretations given the above results. The clinicians in the study group could not appreciate or intuitively understand the findings.

On the other hand, using the change as outcome in our setting is intuitively clearer (as discussed in the paper) than when using the ANCOVA approach; although both approaches showed that there were qualitative differences in terms of before and after relationships between MET and control hospitals. As confirmed by clinician investigators in our group, one clearly and intuitively conveys the message to clinicians and policy makers; the other does not.

2. Using change as outcome has sufficient power as using ANCOVA approach in our study

The well-know limitation of using change as outcome instead of ANCOVA is that it may lose some power if the before and after correlation is low (Andrew Vickers, 2001). However, when the correlation is high, it is acceptable to use change as the outcome instead of the ANCOVA approach. As shown by Vicker in the paper using simulation, that when a correlation is as high as 0.80, using change as the outcome has 97.7% power in comparison to 98.6% using the ANCOVA approach. The power is much higher than the nominal 80% or 90% level and the efficiency loss (0.9%) is
negligible. In our study, the correlation between the before and after values for the primary outcome is 0.793. This indicates that our approach would be essentially as powerful as an ANCOVA-based approach. Moreover, the significant difference we showed in the results may only be more conservative and increases the robustness of our findings.

3. The circumstances were different between the MERIT publication and the current study, so did the methodology

The MERIT publication was based on an agreed pre-specified statistical plan. Such plan did not exclude the ability to subsequently examine any effect modifier. Thus, we adopted the ANCOVA approach as the most appropriate way to analyse the data. However, in current study, we felt that a research question that directly explores the factors contributing to the magnitude of ‘improvement’ or the degree of ‘change’ after implementing a MET system, was more closely related to the question that lies in the hearts of clinicians and policy-makers. That effect modifier we explored happened to be the baseline outcome. This is because, for clinicians, baseline performance is intuitively believed to be likely to affect how much improvement can be achieved with the introduction of a MET system. According to this intuitive paradigm, baseline outcomes may represent the magnitude of problem of adverse events in a given hospital before a “system change”. Furthermore, according to the same clinical paradigm, the magnitude of ‘improvement’ or the degree of ‘change’ should be measured by the difference in the before and after value for the outcomes. Given all the above clinical and statistical considerations, we think that both the MERIT approach used in the initial seminal paper and the current approach are valid with each to be preferred in its own setting.

The reviewer: “Moreover, the authors not only use change scores as the outcome they also adjust for baseline. This is an over correction of baseline differences and could lead to an over-estimate of treatment effect.”

Response: We feel this criticism is not justified. We did not attempt to estimate any ‘treatment effect’ in the current study, based on our regression approach. When there is a continuous effect modifier, this is no single treatment effect to be estimated. As such, it is hard to understand how our approach would lead to an over-estimate of a treatment effect that has never been estimated. Furthermore, using change as an endpoint will only result in an under-estimate of treatment effect. Accordingly, the significant results that we have demonstrated would be more robust.

The reviewer: “As I also explained previously the authors are comparing a formal test of interaction between treatment and a continuous effect modifier with subgroup specific tests, which dichotomise the effect modifier. The latter does not directly test the differences in treatment effect across the two subgroups and so the authors are not comparing like with like. There is much literature demonstrating that subgroup-specific tests are misleading and should not be performed, so I do not see it’s value here (Brookes, JCE 2004; Rothwell, Lancet 2005).”

Response: We are totally in agreement with the reviewer that the subgroup analysis approach has serious limitations. When there is a continuous modifier, there is little value in further testing the effect modifier based on an arbitrary cut-off value.
Moreover, given a cluster randomized controlled trial such as MERIT with only 23 hospitals in total, splitting them into two groups and estimating the corresponding treatment effect in each group is unsound and as an approach, it is doomed to fail. This was made even more evident by the fact that the MERIT investigators estimated that it may need well above 100 hospitals to have sufficient power to demonstrate significance. So why we still want to compare our approach with such subgroup analysis? The reason is that, we have been asked by many clinicians and colleagues at multiple meetings why we had not tried to do a subgroup analysis when there was a continuous effect modifier. Given such widespread clinical concerns, it seemed to us that the non-statistical readers would benefit from the seemingly obvious appreciation that a subgroup analysis would not be able to demonstrate any treatment effect.

**The reviewer:** “If there is any comparison of methods to be made here it should perhaps be between two formal tests of interaction: one which treats the effect modifier as continuous (as they do) and one which dichotomises it but incorporates it as a proper interaction term in a regression model.”

**Response:** Thank you for the suggestion. We included such an interaction effect term in our model and it was not significant. Logically, there should be no need to conduct a subgroup analysis. However, we added the following into the text “There was no significant interaction effect (p=0.081) for a dichotomized modifier using baseline median value as cut-off. Logically, there was no need to conduct further subgroup analysis. Nevertheless, the following subgroup analysis was presented for demonstrative purpose.” (p13, ln21-ln24).

**Summary:** We felt that our approach is statistically acceptable and able to be understood by non-statistical experts. It showed that choosing analytical approach should take into account both substantive research questions and the statistical properties of the technique.

**Reference:**

Vickers, A. The use of percentage change from baseline as an outcome in a controlled trial is statistically inefficient: a simulation study. BMC Medical Research Methodology. 2001;1:6