PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below. Some articles will have been accepted based in part or entirely on reviews undertaken for other BMJ Group journals. These will be reproduced where possible.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Benchmarks for detecting "breakthroughs" in clinical trials: empirical assessment of the probability of large treatment effects using kernel density estimation.
AUTHORS	Miladinovic, Branko; Kumar, Ambuj; Mhaskar, Rahul; Djulbegovic, Benjamin

VERSION 1 - REVIEW

REVIEWER	Derek Ward National Institute for Health Research (NIHR) Horizon Scanning Centre School of Health and Population Sciences University of Birmingham Birmingham, B15 2TT United Kingdom
REVIEW RETURNED	07-Jul-2014

GENERAL COMMENTS	This is an interesting study that addresses a clearly articulated and current need using an appropriate and rigorous methodological approach. However, there is no discussion of the potential limitations that could arise from the approach to data collection. For example, there may be systematic differences in effect size (and primary outcome) that result from therapeutic/disease areas specific to the included study cohorts or the inclusion of different therapeutic modalities in different study cohorts which requires some discussion, even if only to dismiss this as a potential source of bias. In addition, studies funded through the four major UK and US research councils are unlikely to be representative of the full breadth of phase III RCTs (in terms of demonstrated equipoise or methodological rigour) funded through a mix of industry, charitable and other funding streams, as demonstrated by the inclusion of the results from a commercially funded cohort of studies. In addition to the specific comment above, the manuscript would benefit from further elaboration of the following issues. A brief description/summary of the methods (overall approach) - the casual reader who did not follow up on the reference would not automatically be aware that the studies that form the dataset are derived from a set of cohorts, themselves chosen according to a set of inclusion criteria. The methods section describing size of treatment effects mentions accounting for unpublished data, but the methods used to achieve this are not described - this should be amended. The authors note that in 15% of trials, the original researchers

	recommend immediate adoption of the treatment. It would be helpful to know how this relates to reported effect size and funding source. Finally, the introduction rightly notes the growth and currency of biomarker outcome trials, highlighting the differences in typical reported effect sizes. In some fields, such trials have been reported for some years now, and it is likely that these have been included in at least some of the included study cohorts. The authors may like to consider the extent to which this may influence their reported overall results and the stability of their findings over time. No additional comments.

REVIEWER	Conrad Kabali McMaster University, Ontario, Canada University of Toronto, Ontario, Canada
REVIEW RETURNED	03-Sep-2014

GENERAL COMMENTS	Although the authors referred readers to their Cochrane review, it is still important that the search strategy elaborated in this paper, to help readers judge whether selection bias did not play a role. Large effects depend on the susceptibility of the study population to develop an outcome. This factor is not considered in the computation of weights. On page 6 (line 3-line 8), how were the published and unpublished data accounted for? Authors do n ot elaborate. For investigators performing sample size and power calculations how will this method help them? From the way the paper is written, I can't see how. Investigators in phase III trials are usually interested in the efficacy of a specific treatment for specific conditions. Inferring from different treatments and different conditions seems illogical. How were the qualities of included studies evaluated? Did the trials with high risk of bias being given a higher weight simply because the recruited more patients? In Figures 1 & 2 very large protective effectivenes have almost a zero probability of being detected while very large detrimental effects have almost probability one of being detected. Is there something wrong with the figures? It doesn't seem logical.

VERSION 1 – AUTHOR RESPONSE

Reviewer Name Derek Ward
Institution and Country National Institute for Health Research (NIHR) Horizon Scanning Centre School of Health and Population Sciences University of Birmingham Birmingham, B15 2TT United Kingdom
Please state any competing interests or state 'None declared': None declared.

This is an interesting study that addresses a clearly articulated and current need using an appropriate and rigorous methodological approach. However, there is no discussion of the potential limitations that could arise from the approach to data collection.

1. For example, there may be systematic differences in effect size (and primary outcome) that result from therapeutic/disease areas specific to the included study cohorts or the inclusion of different therapeutic modalities in different study cohorts which requires some discussion, even if only to dismiss this as a potential source of bias. In addition, studies funded through the four major UK and US research councils are unlikely to be representative of the full breadth of phase III RCTs (in terms of demonstrated equipoise or methodological rigour) funded through a mix of industry, charitable and other funding streams, as demonstrated by the inclusion of the results from a commercially funded cohort of studies.

Response: We thank the reviewer for the comment. Under the Limitations in our original submission, we had already noted the following:

• Access to only one industry funded cohort and generalizability of the results to privately funded trials.
• The distribution of observed treatment effects could have been affected by bias in the form of inferior established treatments or other well documented biases plaguing randomized trials

However, the biases mentioned by the reviewer (systematic differences in effect size, differences in therapeutic modalities,…etc), and indeed all other possible sources of bias (generation of random sequence, allocation concealment, description of withdrawals\dropouts, blinding, intention-to-treat analysis for benefits, per protocol analysis for harms, pre-specified outcomes reported) and random error (expected difference in primary outcome pre-specified, alpha/beta, sample size calculations performed) were addressed in our Cochrane review and PlosOne manuscript. We have added the following to the manuscript to make this message clear:

In the Methods section we stated that "The risk of bias and random error in the included studies was assessed systematically according to established methods (cited Higgins JPT, Altman DG, Gotzsche PC, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. Brit Med J 2011;34)." In the Results section we stated that "All cohorts included high quality RCTs with low risk for bias."

2. In addition to the specific comment above, the manuscript would benefit from further elaboration of the following issues. A brief description/summary of the methods (overall approach) - the casual reader who did not follow up on the reference would not automatically be aware that the studies that form the dataset are derived from a set of cohorts, themselves chosen according to a set of inclusion criteria. The methods section describing size of treatment effects mentions accounting for unpublished data, but the methods used to achieve this are not described - this should be amended.

Response: In the Methods section we have listed all six cohorts. We have also clarified the Search strategy. We have clarified that "All studies in which the publication bias could not be ruled out were

excluded." For the unpublished studies, we clarified that "In the case of unpublished publicly funded studies, data were obtained from the cooperative groups. In the case of GlaxoSmithKline, data were extracted from trial summary reports."

3. The authors note that in 15% of trials, the original researchers recommend immediate adoption of the treatment. It would be helpful to know how this relates to reported effect size and funding source.

Response: Indeed, we have added in the Results section that "(15% (n =124) public versus 35% (n = 14) privately funded RCTs, $p < 0.001$)." Please note that the original 15% was from the publically funded cohorts only and with the inclusion of privately funded trials, the revised number is 16%.

4. Finally, the introduction rightly notes the growth and currency of biomarker outcome trials, highlighting the differences in typical reported effect sizes. In some fields, such trials have been reported for some years now, and it is likely that these have been included in at least some of the included study cohorts. The authors may like to consider the extent to which this may influence their reported overall results and the stability of their findings over time.

Response: We agree with the reviewer that the next step in the assessment will be to systematically look at the success rate in biomarker primary outcome studies. None of the RCTs in our cohort were biomarker primary outcome studies. We clarify in the Methods that "Only the trials that compared new versus established treatments in humans were eligible."

Reviewer Name Conrad Kabali
Institution and Country McMaster University, Ontario, Canada
University of Toronto, Ontario, Canada
Please state any competing interests or state 'None declared': None declared

1. Although the authors referred readers to their Cochrane review, it is still important that the search strategy elaborated in this paper, to help readers judge whether selection bias did not play a role.

Response: The Search strategy and issues related to bias and random error have been elaborated upon in the Methods. Thanks.

2. Large effects depend on the susceptibility of the study population to develop an outcome. This factor is not considered in the computation of weights.

Response: It was not possible to assess the susceptibility that the reviewer brings up. The reviewer raises an important issue, which will need to be taken into account for future biomarker primary outcome trials.

3. On page 6 (line 3-line 8), how were the published and unpublished data accounted for? Authors do n ot elaborate.

Response: We have clarified that "In the case of unpublished publically funded studies, data were obtained from the cooperative groups. In the case of GlaxoSmithKline, data were extracted from trial summary reports."

4. For investigators performing sample size and power calculations how will this method help them? From the way the paper is written, I can't see how. Investigators in phase III trials are usually interested in the efficacy of a specific treatment for specific conditions. Inferring from different

treatments and different conditions seems illogical.

Response: The aim of our paper was to provide benchmarks for future studies based on our cohorts of trials. As we noted on page 2 of our PloS One manuscript, “A fundamental premise in the overall assessment of treatment success is that investigators make their ‘bets’ regarding the superiority of one treatment over another before a trial is conducted.” This goes for biomarker primary outcome trials as well. We hope that our empirical results will be used to help future investigators place their bets in a more informed fashion.

5. How were the qualities of included studies evaluated? Did the trials with high risk of bias being given a higher weight simply because the recruited more patients?

Response: The quality of the studies was evaluated using the accepted Cochrane methods (Higgins JPT, Green S, Cochrane Collaboration (2008) Cochrane handbook for systematic reviews of interventions. Chichester, England;Hoboken, NJ: Wiley-Blackwell. xxi, 649 p. p.). We have noted in the Results that “All cohorts included high quality RCTs with low risk for bias.” The trials with lower variance of treatment effect would be given more weight according to the random effects model.

6. In Figures 1 & 2 very large protective effectivenes have almost a zero probability of being detected while very large detrimental effects have almost probability one of being detected. Is there something wrong with the figures? It doesn’t seem logical.

Response: Figures 1A and 1B give cumulative probabilities of treatment success, which start at zero and sum up to one. In other words, the lines give probability of observing up to the specified ln(HR) or ln(OR). These were used to estimate probability of observing treatment success in Figure 2. We suspect the reviewer is refering to probability density functions, which is Figure 1 from Djulbegovic et al. Medical research: Trial unpredictability yields predictable therapy gains. Nature 2013;500(7463):395-6.

VERSION 2 – REVIEW

REVIEWER	Derek Ward National Institute for Health Research (NIHR) Horizon Scanning Centre School of Health and Population Sciences University of Birmingham Birmingham, B15 2TT United Kingdom
REVIEW RETURNED	23-Sep-2014

GENERAL COMMENTS	The authors have addressed my original comments in their revised manuscript and cover letter, and I have no further comments to make.