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.

ARTICLE DETAILS

TITLE (PROVISIONAL)	A feasibility study to examine discrepancy rates in pre-specified and reported outcomes in articles submitted to The BMJ.
AUTHORS	Weston, Jennifer; Dwan, Kerry; Altman, Doug; Clarke, Mike; Gamble, Carrol; Schroter, Sara; Williamson, Paula; Kirkham, Jamie

VERSION 1 - REVIEW

REVIEWER	George Peat Keele University, UK Authored a paper on improving transparency for prognosis research with Schroter and Altman as co-authors (doi: 10.1371/journal.pmed.1001671). I was a member of PROGRESS group with Schroter and Altman and which published a series of 4 position papers on prognosis research.
REVIEW RETURNED	14-Oct-2015

GENERAL COMMENTS	This is a very well-written, succinct, and interesting study from leaders in this area. It is a small but detailed study providing some new insights from an important perspective and arguing convincingly for a larger study. I believe it is important, less in terms of highlighting the deficiencies in trial reporting, but in beginning a process for how discrepancies in outcome reporting (and other aspects) in the registry, protocol, and report ought to be sensibly handled. 1. Feasibility study. The aim is stated as “to gain an understanding of the processes involved in comparing pre-specified outcomes from the trial protocol and reported outcomes…” (p5). Presumably as a prelude to a large study. If so, suggest stating that clearly. The Conclusions rather drift away from that aim towards the more substantive implications of their findings of the nature and extent of discrepancies. This results in quite a lot of ‘should’ statements in the Discussion which feels out of step with the designation of this as a feasibility study and its stated aim. 2. Bias and intent. The authors appear to conflate these two concepts whereas I might be inclined to regard them as related but distinct. For example, in the Introduction “…if reporting is driven by the statistical significance or direction of the estimated effect” (my italics) (p5), and upon the reliance placed on authors’ reasons for discrepancies. But outcome reporting bias, like bias in all its forms, can also arise from random error, from unconscious systematic processes, and from intentions motivated by factors other than the p value. Yet if these processes result in systematic deviations in the findings and interpretation of a trial they have the common effect of producing bias. I would encourage the authors to consider

	rephrasing the term 'driven by' in the opening sentence and also to reflect this distinction in the Discussion. 3. Importance of discrepancies. It may be unreasonable in this feasibility study to address this but while the frequency of discrepancies is well-covered I was less clear on their importance. The term 'major discrepancies' is used in the context of previous studies (p12). Do the authors have a sense of how important or trivial their identified discrepancies were or how this might be judged in a future large study (including the very valid point in their Discussion re implications for future meta-analyses)? 4. Changing practice in registration and protocol specification. Practice and policy in this area has been evolving. The recent study period reflects the date of submission but these submissions may relate to trials whose pre-enrollment registration entry and protocols were many years ago (in one case pre-1999). It would be helpful to provide the date range of trial registration (and of published protocols) for the 21 trials accepted and 21 rejected. 5. Implications. Consider using the opportunity to promote and encourage use of SPIRIT and highlighting the role of TSC/DMC in this. Is one implication of this work that either the CONSORT checklist is inadequately specified or it is poorly completed at the time of submission? Minor • Consider providing the rationale for the study period chosen. • There were what felt like a couple of blind alleys: firstly, information on trial submissions rejected by the BMJ for not having a protocol. It is important to have an estimate of this number (136/275 eligible submitted RCTs!) but the argument "This approach allows a comparison between articles that were suitable to be published in the BMJ and those that were not." (p7) was not really followed through in the Results. Consider deleting that sentence. Secondly, the purpose and value in this study of identifying whether manuscripts rejected by BMJ were published elsewhere was not clear to me. Consider removing this. • The 'next steps towards transparency' felt rather too BMJ-centric and, while potentially relevant, not obviously arising specifically in response to selective outcome reporting. Consider rephrasing or removing this paragraph. • P10. "…there was a tendency for more trials that were accepted…" Over-interpretation? Omit. p13 Typo. "term"

REVIEWER	Ian Saldanha Johns Hopkins Bloomberg School of Public Health USA
REVIEW RETURNED	23-Nov-2015

GENERAL COMMENTS	The authors have done a nice job comparing submissions to the BMJ to what was eventually published for a sample of RCTs. While I don't see major problems with the methodology, the authors can do a better job of clarifying the novelty and the value of the work to the

field. The article can be shortened and the tables can be condensed. I've detailed some thoughts and suggestions for strengthening the work below.

Substantive Comments

- The wording of the aim of the study is not clear. I'm still not convinced I understand what it says.
- Page 7 – 1st para – Could you always tell whether a publication was "primary"? How did you define primary?
- Page 8 – 2nd para – Your definition of outcome "domain" is not conventional. I'd refer you to Zarin 2011 NEJM (PMID 21366476) and Saldanha 2014 PLoS One (PMID 25329377), who describe how outcomes are defined. You should use their naming convention or explain why you chose to define things differently from them.
- Page 8 – 3rd para – I worry about your decision to not classify as discrepancies instances where the protocol only mentioned the "domain" (outcome category). Surely under-specification, as you admit later in the Discussion can lead to bias. Imagine scenarios where the researchers are broad at the outset and then report only the specific outcome within a "domain" that is statistically significant.
- Page 9 – 2nd para – You should consider using the language "upgrading" and "downgrading" of outcomes. See Pandis 2015 PLoS One (PMID 26368938). Admittedly, their work was for systematic reviews and not trials, but you could cite/use their language.
- How about adding a comparison of the published outcomes versus the combination of protocol and registered outcomes? Wouldn't that provide an additional, perhaps useful, picture of the differences between the "pre" and "post" that you are trying to get at?
- What kind of instructions do BMJ peer reviewers receive about comparing pre-specified outcomes with published outcomes? Could that have impacted your results?
- Are you sure the reasons for identified discrepancies that you listed in the results are "not considered to indicate bias"? See my comment above regarding under-reporting.
- I was curious about how the protocols faired as regards the SPIRIT items. I realize these protocols were too soon after SPIRIT came out, but it might be interesting to look at other aspects about the beyond outcome reporting that might shed a light on the quality.
- What might be the resource implications of asking peer reviewers to check published outcomes versus protocols versus registry entries for all manuscripts? Is that broadly feasible? Should journals have staff do this and write a summary for the peer reviewers?
- I did not find each row of Table 2 and Supplementary Tables 1 and 2 to be informative. I think only the Totals and percentages are. Maybe because the totals for each row are different and you only report absolute numbers for each row, I got weary by the time I got to row 6. You could also show a spread for each column. Alternatively, you could perhaps draw a graph?

	Minor Comments • Page 5 – 1st para - I would avoid the use of the word "suppressed in the opening paragraph. I would say, "not reported". Later in the paragraph, clarify what you mean by "fully reported". • Page 5 – 2nd para – Insert…. considered "for publication" if the trial… • Page 8 – 1st para – Cite the ICMJE definition. • Page 8 – 3rd para – Repeat what the four possible source documents were. • Page 9 – 1st sentence – Change "assurance" to "control". • Page 11 – 1st para – You should provide the reader some interpretation of the information in this para. • Page 11 – 3rd para – subjective "outcome" measures. • Page 12 – 1st para – I would refrain from using the strong word "protect". I would say, "reduce". • Para 12 – 3rd para - I would delete this para that talks about systematic reviews. It's not relevant to your discussion. • Figure 1 – The "RCTs identified", "RCTs excluded", and "RCTs included" boxes should have the word "articles" in them. These were articles, and not RCTs, correct? Was it always one unique RCT per article?

REVIEWER	Padhraig Fleming Barts and The London School of Medicine and Denristry, Queen Mary University of London
REVIEW RETURNED	28-Dec-2015

GENERAL COMMENTS	This is a very clearly presented piece of work evaluating discrepancies which has benefitted from access to submissions to a leading journal. It does represent (as the authors suggest) a feasibility study and uses a leading medical journal, doubtless attracting the best research submissions, as an exemplar. As such the generalisability of the findings could be disputed but the authors do address this clearly in their discussion and limitations section. Overall this is worthy of publication and is likely to stimulate further valuable work.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1

George Peat
Keele University, UK

Please state any competing interests or state 'None declared':

Authored a paper on improving transparency for prognosis research with Schroter and Altman as co-authors (doi: 10.1371/journal.pmed.1001671). I was a member of PROGRESS group with Schroter

and Altman and which published a series of 4 position papers on prognosis research.

Please leave your comments for the authors below This is a very well-written, succinct, and interesting study from leaders in this area.

>>Thanks you for this positive comment about our research.

It is a small but detailed study providing some new insights from an important perspective and arguing convincingly for a larger study. I believe it is important, less in terms of highlighting the deficiencies in trial reporting, but in beginning a process for how discrepancies in outcome reporting (and other aspects) in the registry, protocol, and report ought to be sensibly handled.

1. Feasibility study. The aim is stated as "to gain an understanding of the processes involved in comparing pre-specified outcomes from the trial protocol and reported outcomes…" (p5). Presumably as a prelude to a large study. If so, suggest stating that clearly. The Conclusions rather drift away from that aim towards the more substantive implications of their findings of the nature and extent of discrepancies. This results in quite a lot of 'should' statements in the Discussion which feels out of step with the designation of this as a feasibility study and its stated aim.

>>The aim has now been made clearer to show that our intention is to inform a larger future study. This larger study is mentioned in the future work section of our Discussion section.

Our discussion section 'conclusions and policy' section do reflect some of the findings from this feasibility study (for example, the need to write sufficiently detailed protocols was clearly evident from the difficulty in being able to identify what the trial outcomes actually were in some cases).

Our 'should' recommendations reflect mostly those already in current guidance (e.g. describing reasons for outcome discrepancies is in CONSORT guidance – this guidance is defined in the Introduction) or common sense recommendations that could possibly reduce the number of discrepancies found and subsequently eliminate suspicions of possible biased reporting (or non-reporting). A full recommendation and implementation strategy is in place for the larger study which will be agreed with relevant stakeholders. This will include implementable tools to detect discrepancies.

2. Bias and intent. The authors appear to conflate these two concepts whereas I might be inclined to regard them as related but distinct. For example, in the Introduction "…if reporting is driven by the statistical significance or direction of the estimated effect" (my italics) (p5), and upon the reliance placed on authors' reasons for discrepancies. But outcome reporting bias, like bias in all its forms, can also arise from random error, from unconscious systematic processes, and from intentions motivated by factors other than the p value. Yet if these processes result in systematic deviations in the findings and interpretation of a trial they have the common effect of producing bias. I would encourage the authors to consider rephrasing the term 'driven by' in the opening sentence and also to reflect this distinction in the Discussion.

>>The distinction to be made here is between selective reporting and selective reporting bias (or outcome reporting bias). Selective reporting can occur for any reason but when the reason is driven by knowledge of the results (often based on knowledge of the significance of an outcome then this defines the bias). We change the text slightly to

"Selective outcome reporting occurs when a subset of the originally recorded outcome variables in a trial are selectively reported in a publication based on their results. When outcome reporting is informed by the statistical significance and/or effect size (e.g., outcomes where the results are not

statistically significant are suppressed or reported only as p>0.05), we refer to this as outcome reporting bias". This is a standard definition that is now well accepted and cited- referenced in the manuscript [ref 1].

We realise that there are many other forms of bias that may affect the medical literature but other forms of bias are outside the scope of this research; here we focus only on outcome reporting bias. We now clarify this in the discussion.

3. Importance of discrepancies. It may be unreasonable in this feasibility study to address this but while the frequency of discrepancies is well-covered I was less clear on their importance. The term 'major discrepancies' is used in the context of previous studies (p12). Do the authors have a sense of how important or trivial their identified discrepancies were or how this might be judged in a future large study (including the very valid point in their Discussion re implications for future meta-analyses)?

>>The purpose of this feasibility study was to identify and describe the discrepancies and to report the reasons (if provided) for these. We have an impact strategy for the larger study to assess the importance of the discrepancies which requires some verification from the study authors about what was actually done (this was not part of the feasibility study).

4. Changing practice in registration and protocol specification. Practice and policy in this area has been evolving. The recent study period reflects the date of submission but these submissions may relate to trials whose pre-enrollment registration entry and protocols were many years ago (in one case pre-1999). It would be helpful to provide the date range of trial registration (and of published protocols) for the 21 trials accepted and 21 rejected.

>>We have included the date ranges for trial registrations and protocols for both accepted and rejected RCTs in the results section. However in the case where the trial started recruiting participants in 1999, the trial was only registered in 2010 and the protocol was not dated.

5. Implications. Consider using the opportunity to promote and encourage use of SPIRIT and highlighting the role of TSC/DMC in this. Is one implication of this work that either the CONSORT checklist is inadequately specified or it is poorly completed at the time of submission?

>>We've included a few sentences on the role of TSC and the use of SPIRIT guidelines in the discussion.

The CONSORT checklist was presented in all manuscripts (it's a requirement at the submission stage for The BMJ at least). However, it is well documented that endorsement and proper adherence to a guideline to not necessarily equate to the same thing. We need to get handle on this and this will form part of the recommendations in our final larger study, i.e. we need to question authors why they say they have addressed the relevant CONSORT item when clearly there are undocumented outcome discrepancies.

Minor
• Consider providing the rationale for the study period chosen.

>>The BMJ's manuscript tracking system stores articles for 12 months only. As the feasibility study was based on historic submissions, we chose a 10 month time frame to ensure we didn't lose any articles that were to be included in the assessment. A sentence has been added to reflect this.

• There were what felt like a couple of blind alleys: firstly, information on trial submissions rejected by the BMJ for not having a protocol. It is important to have an estimate of this number (136/275 eligible submitted RCTs!) but the argument “This approach allows a comparison between articles that were suitable to be published in the BMJ and those that were not.” (p7) was not really followed through in the Results. Consider deleting that sentence. Secondly, the purpose and value in this study of identifying whether manuscripts rejected by BMJ were published elsewhere was not clear to me. Consider removing this.

>>Agreed – both sentences removed and information removed from the results section and Table 2.

• The ‘next steps towards transparency’ felt rather too BMJ-centric and, while potentially relevant, not obviously arising specifically in response to selective outcome reporting. Consider rephrasing or removing this paragraph.

>>Agreed and removed. Generic guidance for all journals will be provided in the larger study.

• P10. “…there was a tendency for more trials that were accepted…” Over-interpretation? Omit.

>>We believe this was a valid observation to note. Statistical significance testing of the difference was avoided to so as not to over-interpret.

p13 Typo. "term"

>>Changed

Reviewer: 2

Ian Saldanha
Johns Hopkins Bloomberg School of Public Health Ian Saldanha

The authors have done a nice job comparing submissions to the BMJ to what was eventually published for a sample of RCTs. While I don’t see major problems with the methodology, the authors can do a better job of clarifying the novelty and the value of the work to the field. The article can be shortened and the tables can be condensed. I’ve detailed some thoughts and suggestions for strengthening the work below.

Substantive Comments
• The wording of the aim of the study is not clear. I’m still not convinced I understand what it says.

>>The aim has been slightly amended for clarity, changed the word ‘processes’ to ‘practicalities’.

• Page 7 – 1st para – Could you always tell whether a publication was “primary”? How did you define primary?

>>A primary study publication was defined as a study reporting on the main clinical outcomes of the trial. This was often clear from the manuscript/covering letter from the authors. If this was not clear the decision was made in discussion with two experienced study authors. For example, if the article reported on only the economic evaluations of the trial, then this was unlikely to be the primary publication. We realise that this method has its limitations but we re-iterate that this is a feasibility study. In the larger study where contact with trial authors will be made – this level of detail will be

confirmed with the trial authors.

Text updated to reflect how the decision regarding a primary study was made.

• Page 8 – 2nd para – Your definition of outcome "domain" is not conventional. I'd refer you to Zarin 2011 NEJM (PMID 21366476) and Saldanha 2014 PLoS One (PMID 25329377), who describe how outcomes are defined. You should use their naming convention or explain why you chose to define things differently from them.

>>Thanks for finding a suitable reference. We have changed this so that outcomes/domains and specific measurements are coherent with these studies.

• Page 8 – 3rd para – I worry about your decision to not classify as discrepancies instances where the protocol only mentioned the "domain" (outcome category). Surely under-specification, as you admit later in the Discussion can lead to bias. Imagine scenarios where the researchers are broad at the outset and then report only the specific outcome within a "domain" that is statistically significant.

>>This is a limitation of the feasibility study which is why this was highlighted in the discussion although we didn't want to over-estimate the number of discrepancies. We suspect that there could be a lack of detail in registries/protocols on the 'specific measurement' with a domain for reasons unrelated to the bias. For example, there is much debate as to when the analysis plan is written (as often this is not part of the protocol), which may ultimately determine how you list your domains / specific measurements within a protocol. The main larger study aims to unravel this as we plan to contact trialists.

• Page 9 – 2nd para – You should consider using the language "upgrading" and "downgrading" of outcomes. See Pandis 2015 PLoS One (PMID 26368938). Admittedly, their work was for systematic reviews and not trials, but you could cite/use their language.

>>Amended language. We actually used this terminology ourselves within reviews in 2010 (published in PLoS ONE).

Kirkham¬¬ JJ, Altman DG, Williamson, PR. Bias due to changes in specified outcomes during the systematic review process. PLoS ONE 2010; 5(3): e9810.

• How about adding a comparison of the published outcomes versus the combination of protocol and registered outcomes? Wouldn't that provide an additional, perhaps useful, picture of the differences between the "pre" and "post" that you are trying to get at?

>>This was not something that we planned to do. We mention as a limitation in the discussion that some of the trials were retrospectively registered and some protocols were not date stamped. Therefore you cannot be sure in some cases that you are doing a true 'pre'/ 'post' comparison.

In the larger study we plan to clarify these issues with trial authors and these limitations will form part of our guidance.

• What kind of instructions do BMJ peer reviewers receive about comparing pre-specified outcomes with published outcomes? Could that have impacted your results?

>>In conjunction with reviewer 1 we have taken all descriptions out of the discussion that refers to The BMJ policies (to avoid becoming too BMJ centric). We used the BMJ manuscripts as a source of data to conduct this research but ultimately we wish the guidance to be generic to all journals. The larger study does involve more journals.

• Are you sure the reasons for identified discrepancies that you listed in the results are “not considered to indicate bias”? See my comment above regarding under-reporting.

>>Obtaining the reasons for discrepancies was not an objective of this feasibility study although we decided to record/report them for completeness if reviewers/authors did comment on the reason.

We have to adhere strictly to our definition of bias which is clearly defined in the manuscript. In the feasibility study, it appeared that the reasons provided for the discrepancies were unrelated to the results and hence we did not suspect any bias. In fact this may be expected as peer reviewers / editors may have been less willing to accept the author responses to the discrepancy if it did appear to be bias related.

We cannot always be certain if the reason is bias related or not - hence the need to contact the trialists in the larger study to find out more information.

See previous response regarding under-reporting.

• I was curious about how the protocols faired as regards the SPIRIT items. I realize these protocols were too soon after SPIRIT came out, but it might be interesting to look at other aspects about the beyond outcome reporting that might shed a light on the quality.

>>This was not an objective of this feasibility study.

• What might be the resource implications of asking peer reviewers to check published outcomes versus protocols versus registry entries for all manuscripts? Is that broadly feasible? Should journals have staff do this and write a summary for the peer reviewers?

>>We will be developing guidance and tools for detecting outcome discrepancies to authors/peer reviewers and journal editors in the larger study. As part of the larger study, we also have an implementation strategy that will be agreed with all the important stakeholders. The matrix of outcomes that was used in the feasibility study is one potential tool to demark discrepancies. This is relatively straightforward and quick to complete if a template is provided.

• I did not find each row of Table 2 and Supplementary Tables 1 and 2 to be informative. I think only the Totals and percentages are. Maybe because the totals for each row are different and you only report absolute numbers for each row, I got weary by the time I got to row 6. You could also show a spread for each column. Alternatively, you could perhaps draw a graph?

>>In the larger study we would not present the results in this way as there will be too many studies. You would also lose some of the information by representing this graphically and we feel there are too few studies to graphically represent the data here. However, as a feasibility study, we feel this table clearly demonstrates at the individual study level, the number and types of discrepancy (the main aim of the feasibility study), which is informative for assessing the magnitude of the problem within studies. For example, from the table you can clearly see whether studies are affected by a single type of discrepancy (e.g. new outcome introduced), or a battery of different type of discrepancy. Distinguishing between the discrepancies in retrospectively and prospectively registered trials may be important to some readers. Essentially we provide here the raw data.

We'd be happy to change this based on the Editors advice.

Minor Comments
• Page 5 – 1st para - I would avoid the use of the word "suppressed in the opening paragraph. I would say, "not reported". Later in the paragraph, clarify what you mean by "fully reported".

>>Amended word. Definition of full reporting provided.

• Page 5 – 2nd para – InsertD. considered "for publication" if the trialID

>>Amended sentence.

• Page 8 – 1st para – Cite the ICMJE definition.

>>Done.

• Page 8 – 3rd para – Repeat what the four possible source documents were.

>>Four documents already listed at the start of this 3rd paragraph.

• Page 9 – 1st sentence – Change "assurance" to "control".

>>Amended word.

• Page 11 – 1st para – You should provide the reader some interpretation of the information in this para.

>>Not an objective of this feasibility study – which is mainly descriptive. Interpretations will come with the larger study when reasons for this type of discrepancy are obtained. Too few cases (of downgrades) in the feasibility study to interpret.

• Page 11 – 3rd para – subjective "outcome" measures.

>>Amended error.

• Page 12 – 1st para – I would refrain from using the strong word "protect". I would say, "reduce".

>>Amended sentence.

• Para 12 – 3rd para - I would delete this para that talks about systematic reviews. It's not relevant to your discussion.

>>Deleted this paragraph.

• Figure 1 – The "RCTs identified", "RCTs excluded", and "RCTs included" boxes should have the word "articles" in them. These were articles, and not RCTs, correct? Was it always one unique RCT per article?

>>Amended figure to include articles. One unique RCT per article.

Reviewer: 3

Padhraig Fleming
Barts and The London School of Medicine and Denristry, Queen Mary University of London

Please leave your comments for the authors below This is a very clearly presented piece of work evaluating discrepancies which has benefitted from access to submissions to a leading journal. It does represent (as the authors suggest) a feasibility study and uses a leading medical journal, doubtless attracting the best research submissions, as an exemplar. As such the generalisability of the findings could be disputed but the authors do address this clearly in their discussion and limitations section.

Overall this is worthy of publication and is likely to stimulate further valuable work.

VERSION 2 – REVIEW

REVIEWER	George Peat Keele University, England
REVIEW RETURNED	26-Jan-2016

GENERAL COMMENTS	I thank the authors for their considered response and am satisfied with these, One minor typo:" trial steering committee"s" in new sentence added to the Discussion.

REVIEWER	Padhraig Fleming Barts and The London School of Medicine and Dentistry, QMUL
REVIEW RETURNED	27-Jan-2016

GENERAL COMMENTS	I am satisfied with the revisions.