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)	Do Non-inferiority Trials of Reduced Intensity Therapies Show Reduced Effects? A Descriptive Analysis.
AUTHORS	aberegg, Scott; Hersh, Andrew; Samore, Matthew

VERSION 1 – REVIEW

REVIEWER	Ben Ewald University of Newcastle NSW Australia
REVIEW RETURNED	14-Sep-2017

GENERAL COMMENTS	Details Line 22-23 of introduction. In each described comparison the lower intensity is listed first. Unclear if this applies to androgen deprivation. Is Continuous lower intensity than intermittent? Line 29. The a priori assumption that reduced intensity will be less efficacious is not obvious to me. There are plenty of treatments where there is a markedly diminished return for increased dose. Examples that come to mind are inhaled steroids for asthma, or statins for CVD prevention. At least some of the topics chosen for this kind of research will be ones where clinicians suspect this to be the case. For others such as the example given of ABVD without the B the assumption seems well founded, and figure 3 supports the assumption. Introduction: There are 2 questions here. One is the basic design of non inferiority trials, the other is biocreep. It would be good to identify these as separate problems, and check in the analysis if biocreep was a possibility in any of the trails reported. “Clinicians may be advised to carefully inspect the results with an emphasis on the delta margin utilised ….” This is good advice for the reader of any non inferiority trial, but I fear it is little understood. Figure 1: Caption should indicate that this is hypothetical. Overall This is an interesting exploration of a novel facet of the methods of this increasingly important study design.

REVIEWER	Philippe Flandre INSERM France
REVIEW RETURNED	16-Oct-2017

GENERAL COMMENTS	Aberegg and colleagues discussed the bio-creep phenomenon in noninferiority trials of reduced intensity therapies. Through non-

inferiority clinical trials, a new therapy may be approved even if it is less effective than the previous therapy. This raises the possibility that, after a series of non-inferiority trials with each new drug being a little worse than the previous drug, an ineffective or harmful therapy may falsely be deemed efficacious. This phenomenon is known as 'bio-creep'. It is true that this phenomenon is likely to happen with a 'successful' series of trials of reduced intensity therapies. Aberegg and colleagues draw their conclusions based on the review of trials published in the five highest impact general medical journals during a 5-year recent period. The topic is interesting and should be highlighted to clinicians. However, there are few major issues with this paper and the manuscript is not carefully written.

1. The manuscript should be re-read to remove vagueness or mistake. For example, Introduction (page 3, line 21) what is TAP ? (line 22) what is ABVD ?; Results section (page 5, line 5), Figure 1 should be Figure 2; page 5 line 31 Figure 2 should be figure 3; line 24 what is RIT ?,I guess Reduced Intensity Therapies but it is not written. In the figure 1 all horizontal lines representing 95% CI cross the vertical (not dashed) line so noninferiority is not demonstrated whereas the legend of the Figure 1 and the text say the opposite. Overall, more care must be brought to the writing. In the paragraph starting page 5 line 17, the comparison 58.3% vs 82.2% p=0.002 is mentioned two times…..
2. From the introduction it is not clear to understand that the comparison is between trials using reduced intensity therapy and other trials. I wrongly understood that the focus was on former trials. So what are the exact inclusion criteria mentioned in the beginning of the Results section?
3. Much more details of the 31 trials and the 36 comparisons should be given either in the table 2 or e-table in an appendix.
4. As introduced by Figure 1 the bio-creep phenomenon is due to a chain of noninferiority trials. So results based on the 36 comparisons of a reduced intensity therapy are somewhat paradoxical. The authors concluded that such trials increase the risk of bio-creep but less trials of RIT conclude to noninferiority (58.3% vs 82.2%) breaking the chain of RIT trials. In fact, the rank of the trial in the chain is very important. The authors should provide the rank of the RIT trials used in their analysis because the first RTI trial (#2 in Figure 1) as a different impact in the bio-creep phenomenon than the fourth trial (#5 in Figure 1). If noninferiority is not demonstrated in the first trial there is no 'bio-creep' while if noninferiority is not demonstrated in the fourth trial it is likely that there is already a 'bio-creep'. The authors should discussed this point.
5. The analysis that provides Figure 3 has no meaning to me. I understand that the 151 absolute differences come from the 182 noninferiority comparisons mentioned page 5 line 9. Such a comparison is like comparing apples and strawberries. I guess that primary endpoints of these trials are quite different. For example, in noninferiority trials involving HIV-1 infected patients the primary endpoint is often virologic failure. In many oncology trials the primary endpoint is still mortality. That one of the reason noninferiority margins are quite different according to the primary endpoint. In Figure 3 what is the meaning of comparing a 0.5% mortality difference with a 2% virologic failure difference or with a 2% rate of adverse events….
6. Although I agree with the fact that the sample size would increase after a series of noninferiority trials there is a confusion between design and result of a trial in the Discussion section. Considering the example in the Discussion of a trial involving 1800 patients in each

	arm to compare 7 to 10% event rate. Such a sample size lead to many reullts concluding superiority, such as 4 vs 10%, 2 vs 12% ect…..With the latter comparison and a linear relationship between dose and response the sample size would be much lower than 16000 patients.

REVIEWER	Sunita Rehal MRC CTU at UCL, UK.
REVIEW RETURNED	16-Oct-2017

GENERAL COMMENTS	Overall, this is a nice paper that cautions researchers of potential pitfalls with biocreep for non-inferiority studies. Aside from some minor corrections listed below, the only thing lacking is some discussion/literature with what can be done about biocreep if anything? And if not, perhaps this is a call to find a methodological solution? For example, the Gladstone/Vach, paper the authors reference suggests defining a 3rd margin in addition to the FDAs suggested M1, M2 margins taking the lowest margin forward as the threshold to determine non-inferiority. The FDA (FDA's consideration of evidence from certain clinical trials) recommend consultation with agencies when choosing the standard treatment. Abstract: 1) Suggest to switch the results to match the order presented in the results section in the paper. 2) The authors quote "77.8% vs. 39.7% P<0.001", but I couldn't find this result mentioned in the main paper. Introduction: Change "trails" to "trials". Methods: 1) Suggest to exclude "prior to closing it and beginning analysis". 2) Lines 27-32 describing the inclusion criteria isn't clear. Were the cluster randomised designs etc. additional exclusions to articles that weren't included? Or have the authors mentioned a selected few types of trials that were excluded before review? Suggest to list all exclusion criteria. 3) Suggest to change "abstracted" to "extracted". 4) The authors mention a random sample was crosschecked with an author, what was the proportion of the sample checked? 5) Suggest to clearly define the CONSORT declaration. Results: 1) What was the proportion of favourable/unfavourable declaration of NI? 2) The authors have stated a rate of 58.3% vs. 82.2% but have presented a risk difference. Surely these are proportions? 3) Lines 24-25, unclear whether the authors have repeated the 58.2% vs. 82.2% result (in which case the p-value does not match) or whether this is an error. Conclusion: Line 50: the pre-specified margin of NI is important but they also need to be adequately justified.

REVIEWER	Beryl Primrose Gladstone

	Infectious Diseases, Department of Internal Medicine I, Tübingen University Hospital, Tübingen, Germany
REVIEW RETURNED	17-Oct-2017

GENERAL COMMENTS

It is a very interesting question that the authors ask regarding NI trials with reduced intensity therapies and the probability of biocreep among these specific areas. The authors have compared NI trials studying reduced intensity therapies (RIT) with those who did not study RIT.

Major comments:

1. Methodologically speaking, a first step has been made to answer the study question by describing the proportion of non-inferior or superior conclusions based on the trial results and providing the basic descriptive statistics (mean) of the outcome measures. However, to provide an empirical evidence of whether the RIT group (NI-RIT) is testing consistently less effective treatments as compared to the others, it would be necessary at least to perform a meta-analysis of the outcome measures to be able to provide the pooled effect estimate as well the distribution of the effect estimate (which would represent the distribution of the true effect as the authors state that there was no evidence for bias). The methodology can be seen in the article (Gladstone BP and Vach W. About half of the noninferiority trials tested superior treatments: a trial-register based study. 2013. Journal of Clinical Epidemiology, Volume 66 , Issue 4 , 386 – 396).

2. It is true that NI trials testing RITs could be expected to be more often inferior/non-inferior rather than superior (which the authors found out) for the same reason which the authors state, that the RITs are of a lower dose and according to the dose response relationship, they would be of lower efficacy. And it is not surprising and is at least satisfactory that about 42% (100-58%) of the NI-RITs lead to either an inferior or inconclusive compared to the 18% (100-82%) declared among NI-nRIT. This reflects the fact that the investigators studying RITs more often at inferior treatments, however the trials are sieving out the inferior treatments and retaining the actual non-inferior treatments. As mentioned above, the distribution of the treatment effects based on meta-analysis would reflect the authors concluding statements better.

3. The reason behind why NI-RITs are done seems to be usually to use the standard therapy among a subgroup of target patients with a specific characteristic who benefit from reduced intensity, for example, who either cannot tolerate the side effects or developing country patients not able to afford the costs of a long term therapy. Hence it would be interesting and of added value to study whether the benefits are more often mentioned among these trials and whether they are of a different pattern compared to the others.

4. There is evidence of regulatory authorities requiring non-inferiority of a new treatment (or a treatment declared superior in terms of efficacy) in terms of adverse effect in as compared to placebo and NI designs being used to study these. (Pocock SJ, Clayton TC, Stone GW. Challenging Issues in Clinical Trial Design: Part 4 of a 4-Part Series on Statistics for Clinical Trials. Journal of the American College of Cardiology. 2015 Dec 29;66(25):2886–98.) Hence it would be worthwhile to check whether the 6 included placebo

	controlled trials are safety trials or not. 5. Similarly, the authors state that they expect a stringent margin among the NI-RITs. The discussed issues are a matter of concern among all NI trials generally and apply to all NI trials. When these NIRITs are more often belonging to a specific area such as cardiology, neonatalogy, etc, there is a more probable danger of biocreep. It would be interesting to see whether there is a difference in any of the trial characteristics between the two groups. Metaregression could also help study the contribution of other features contributing to a difference in treatment effects if any. Table 1 could present the two groups and its characteristics. 6. The study by Gladstone et al referred to as a simulation study in the discussion is not just a simulation study based on assumptions but based on empirical data and of course assumption for the true treatment effect which is again based on empirical data. Minor comments: 1. Page 5 line 21 – it is the same as what is in line 25 and p value is different? A typo! 2. Reference to be added for the published parallel paper 3. 5th reference - typo to be corrected 4. Headings for table 1 missing and need to expand to include information on two groups

VERSION 1 – AUTHOR RESPONSE

We thank the editors and reviewers for taking the time to consider our manuscript. The reviewers have made many insightful observations and comments about the manuscript, almost all of which we have incorporated into it, much improving the revised manuscript. Below and in attached files, please find our specific responses to each reviewer's comments.

Thanks you again for this opportunity to improve our manuscript, and for your generous contributions of time in reviewing it.

Kindest regards,

Scott K Aberegg, MD, MPH on behalf of all authors.

Editorial Requirements:
- Please revise your title to state the research question and study design. This is the preferred format for the journal. We changed the title to the preferred format.
- Please revise the Strengths and Limitations section (after the abstract) to focus on the methodological strengths and limitations of your study rather than summarizing the results. The strengths and limitations section has been revised. Original points 1-3 all refer to methodological issues and do not summarize any results. Point #4 has been changed; both the original text and the revision refer clearly to methodological issues relating to interpretation – correlation and causation and ecological fallacy - and do not summarize any results.

Reviewer(s)' Comments to Author:
Note to all reviewers: A proof of the original submission with the page and line numbers generated by scholar one during original submission has not been made available to the authors. Thus, we cannot accurately refer to specific line numbers, but we are confident that we were able to locate and identify all the specific sections referred to by reviewers.

Reviewer: 1
Reviewer Name: Ben Ewald
Institution and Country: University of Newcastle, NSW, Australia
Please state any competing interests: none declared

Please leave your comments for the authors below

Details
Line 22-23 of introduction. In each described comparison the lower intensity is listed first. Unclear if this applies to androgen deprivation. Is Continuous lower intensity than intermittent? We have changed this to improve clarity.
Line 29. The a priori assumption that reduced intensity will be less efficacious is not obvious to me. There are plenty of treatments where there is a markedly diminished return for increased dose. Examples that come to mind are inhaled steroids for asthma, or statins for CVD prevention. At least some of the topics chosen for this kind of research will be ones where clinicians suspect this to be the case. For others such as the example given of ABVD without the B the assumption seems well founded, and figure 3 supports the assumption. We discuss these issues further in the discussion section. If standard (active control) dosing is at the upper end of a dose response curve that has flattened, then the reviewer is correct, there are diminishing returns for increased doses. However diminished returns are still returns, and we would say they represent “reduced effects”. We also later discuss how the highest tolerable dose is generally selected for superiority trials in order to maximize separation of the two populations and increase delta and effective study power. If the reduced dose used in the noninferiority trial had been used in the original superiority trials, effective power would be expected to be reduced, else why did the designers of the superiority trials select the higher dose that became the standard of care dose that was subsequently used as active control in the noninferiority trial?
Introduction: There are 2 questions here. One is the basic design of non inferiority trials, the other is biocreep. It would be good to identify these as separate problems, and check in the analysis if biocreep was a possibility in any of the trails reported. In the first paragraph of the introduction, we introduce, as a background, several general limitations of noninferiority trials. Subsequently, throughout the manuscript, we focus exclusively on the concept of reduced intensity therapies. The basic design issues of noninferiority trials, we think, are an important background to provide in order to show that a.) there are general problems with these trials that are already known; and b.) that we are extending this knowledge with novel empirical evidence of a theoretical problem. Theoretically, there is a possibility of biocreeep in any noninferiority trial – however this is difficult to demonstrate without testing a therapy shown to be noninferior in an actively controlled noninferiority trial to placebo in another trial and this is not generally feasible. Thus we heretofore have relied on simulation studies to show noninferiority as described in the discussion section. We posit that our study is unique because it exploits a natural experiment that allows us to infer the possible presence of biocreep in trials of reduced intensity therapies because they have results less favorable than other noninferiority trials.
“Clinicians may be advised to carefully inspect the results with an emphasis on the delta margin utilised ….” This is good advice for the reader of any non inferiority trial, but I fear it is little understood. We agree wholeheartedly with this statement, however our manuscript is not intended as

a didactic on the general interpretation of noninferiority trials. Those points were summarized in our other recent manuscript1
Figure 1: Caption should indicate that this is hypothetical. We have added a phrase to the caption to clarify this.
Overall
This is an interesting exploration of a novel facet of the methods of this increasingly important study design. Thank you. We agree that these trials merit ongoing and increased scrutiny.

Reviewer: 2
Reviewer Name: Philippe Flandre
Institution and Country: INSERM France
Please state any competing interests: None declared

Please leave your comments for the authors below

Aberegg and colleagues discussed the bio-creep phenomenon in noninferiority trials of reduced intensity therapies. Through non-inferiority clinical trials, a new therapy may be approved even if it is less effective than the previous therapy. This raises the possibility that, after a series of non-inferiority trials with each new drug being a little worse than the previous drug, an ineffective or harmful therapy may falsely be deemed efficacious. This phenomenon is known as 'bio-creep'. It is true that this phenomenon is likely to happen with a 'successful' series of trials of reduced intensity therapies. Aberegg and colleagues draw their conclusions based on the review of trials published in the five highest impact general medical journals during a 5-year recent period. The topic is interesting and should be highlighted to clinicians. However, there are few major issues with this paper and the manuscript is not carefully written.

1. The manuscript should be re-read to remove vagueness or mistake. For example, Introduction (page 3, line 21) what is TAP ? (line 22) what is ABVD ?; Results section (page 5, line 5), Figure 1 should be Figure 2; page 5 line 31 Figure 2 should be figure 3; line 24 what is RIT ?,I guess Reduced Intensity Therapies but it is not written. In the figure 1 all horizontal lines representing 95% CI cross the vertical (not dashed) line so noninferiority is not demonstrated whereas the legend of the Figure 1 and the text say the opposite. Overall, more care must be brought to the writing. In the paragraph starting page 5 line 17, the comparison 58.3% vs 82.2% p=0.002 is mentioned two times….. TPA is tissue plasminogen activator, a therapy for stroke, and ABVD is Adriamycin Bleomycin, Vinblastine, Dacarbazine. We added explanations for these abbreviations into the text. We corrected the misnumbering of the figures in the Results section. We removed RIT as an abbreviation for Reduced Intensity Therapies. In figure 1, the 95% CIs represented by horizontal lines all cross zero difference but not delta, so the correct interpretation is that noninferiority criteria were met in panels 2-6 as described in the caption and as explained in Figure 1 of reference #16. We removed the redundant sentence of the comparisons. When we did this we reran our analyses and found some minor errors in the reporting of the differences and associated p-values in the abstract and the results which we have corrected. These small errors do not change the results.
2. From the introduction it is not clear to understand that the comparison is between trials using reduced intensity therapy and other trials. I wrongly understood that the focus was on former trials. So what are the exact inclusion criteria mentioned in the beginning of the Results section? This database was generated to capture all noninferiority trials during the study period in the 5 highest impact journals that met our inclusion criteria described in the methods. This was originally done for another analysis which has since been published1. The present analysis compares the subset of those trials that compared a reduced intensity therapy as the new therapy to full intensity as active control to all other trials in the database. This is described in the first sentence of the abstract:

“To identify noninferiority trials within a cohort where the experimental therapy is the same as the active control comparator but at a reduced intensity, and determine if these noninferiority trials of reduced intensity therapies have less favorable results than other noninferiority trials in the cohort.” In addition we have added further language at the end of our introduction section to assure clarity of this key distinction.

3. Much more details of the 31 trials and the 36 comparisons should be given either in the table 2 or e-table in an appendix. We have added to table 1 caption stating that additional details of included studies can be found in reference 15. The authors leave it to the editors to determine if we should include a bibliography of all 31 trials and 36 comparisons of reduced intensity therapies. It is our opinion that this would be of little use to all but a tiny minority of readers. We reason that Table 2, listing some examples, makes it clear what we mean by trials of reduced intensity therapies, and that expanding it would lengthen the article without increasing its quality or utility for readers. Alternatively, we can make our entire database available for exploration by any interested parties.

4. As introduced by Figure 1 the bio-creep phenomenon is due to a chain of noninferiority trials. So results based on the 36 comparisons of a reduced intensity therapy are somewhat paradoxical. The authors concluded that such trials increase the risk of bio-creep but less trials of RIT conclude to noninferiority (58.3% vs 82.2%) breaking the chain of RIT trials. In fact, the rank of the trial in the chain is very important. The authors should provide the rank of the RIT trials used in their analysis because the first RTI trial (#2 in Figure 1) as a different impact in the bio-creep phenomenon than the fourth trial (#5 in Figure 1). If noninferiority is not demonstrated in the first trial there is no ‘bio-creep’ while if noninferiority is not demonstrated in the fourth trial it is likely that there is already a ‘bio-creep’. The authors should discussed this point. This is an excellent point that we failed to consider, and we have added text to the discussion to address it. However, when 58% of trials of RIT do in fact conclude noninferiority or superiority, the risk of biocreep in subsequent trials remains for the majority of trials. If it were the case that only a small minority of trials of RIT met superiority or noninferiority criteria, this proposed phenomenon, that of “breaking the chain”, would take on greater importance.

5. The analysis that provides Figure 3 has no meaning to me. I understand that the 151 absolute differences come from the 182 noninferiority comparisons mentioned page 5 line 9. Such a comparison is like comparing apples and strawberries. I guess that primary endpoints of these trials are quite different. For example, in noninferiority trials involving HIV-1 infected patients the primary endpoint is often virologic failure. In many oncology trials the primary endpoint is still mortality. That one of the reason noninferiority margins are quite different according to the primary endpoint. In Figure 3 what is the meaning of comparing a 0.5% mortality difference with a 2% virologic failure difference or with a 2% rate of adverse events….This is an inherent limitation of meta-research projects, and one reason we did not perform a formal meta-analysis, but rather a descriptive study. We reason that there is value, nonetheless, in comparing the point estimates of the results of empirical data as they relate to directionality and absolute risk differences, and there is precedent for such analyses2. We think it is fair to assume that there is no systematic difference between the choice of primary endpoints for Reduced Intensity Non-Interiority Trials vs. Non-Inferiority Trials of Novel Therapies. As such we can reasonably assume that, despite an “apples to strawberries” comparison, our analysis reveals underlying structural differences between Reduced Intensity Non-Interiority Trials and Non-Inferiority Trials of Novel Therapies.

6. Although I agree with the fact that the sample size would increase after a series of noninferiority trials there is a confusion between design and result of a trial in the Discussion section. Considering the example in the Discussion of a trial involving 1800 patients in each arm to compare 7 to 10% event rate. Such a sample size lead to many reullts concluding superiority, such as 4 vs 10%, 2 vs 12% ect…..With the latter comparison and a linear relationship between dose and response the sample size would be much lower than 16000 patients. We understand the sentiment of the

reviewer’s comment, but maintain that sample size calculations must be made a priori. If investigators seek to demonstrate half the effect size at the same power, they must still design the trial as per our example. That does not mean that they may not find a statistically significant effect smaller than the pre-specified margin (delta), but to have a priori power to do so, they must design the trial as per our example.

Reviewer: 3
Reviewer Name: Sunita Rehal
Institution and Country: MRC CTU at UCL, UK.
Please state any competing interests: None declared

Please leave your comments for the authors below

Overall, this is a nice paper that cautions researchers of potential pitfalls with biocreep for non-inferiority studies. Aside from some minor corrections listed below, the only thing lacking is some discussion/literature with what can be done about biocreep if anything? And if not, perhaps this is a call to find a methodological solution? For example, the Gladstone/Vach, paper the authors reference suggests defining a 3rd margin in addition to the FDAs suggested M1, M2 margins taking the lowest margin forward as the threshold to determine non-inferiority. The FDA (FDA's consideration of evidence from certain clinical trials) recommend consultation with agencies when choosing the standard treatment.

Abstract:
1) Suggest to switch the results to match the order presented in the results section in the paper. Thank you. We have done this.
2) The authors quote "77.8% vs. 39.7% P<0.001", but I couldn't find this result mentioned in the main paper. We thank the reviewers for pointing out some mathematical errors in our analyses which we have corrected. None of them materially changes the results.

Introduction:
Change "trails" to "trials". Thank you. We have done this

Methods:
1) Suggest to exclude "prior to closing it and beginning analysis". We have eliminated this redundancy.
2) Lines 27-32 describing the inclusion criteria isn't clear. Were the cluster randomised designs etc. additional exclusions to articles that weren't included? Or have the authors mentioned a selected few types of trials that were excluded before review? Suggest to list all exclusion criteria. First we included trials that met inclusion criteria. Next, from among those that were included, we excluded those with the listed exclusion criteria. That is, you had to meet inclusion criteria and be included before you were eligible to be excluded. Trials were also excluded if they originally appeared to meet inclusion criteria, but upon further review they did not.
3) Suggest to change "abstracted" to "extracted". Thank you. We have done this.
4) The authors mention a random sample was crosschecked with an author, what was the proportion of the sample checked? This information was added.
5) Suggest to clearly define the CONSORT declaration. We added clarification for this, and further information is available in references 17&18.

Results:
1) What was the proportion of favourable/unfavourable declaration of NI? If we leave out trials with a declaration of superiority and compare just those meeting noninferiority criteria but not superiority,

there is a small but not significant difference between RIT and non-RIT trials in the declaration of noninferiority (58 versus 65%; P=0.45). However, separating trials that showed noninferiority but not superiority is an artificial distinction because meeting noninferiority criteria is a pre-requisite for declaring superiority. Because of this and because we opine that it confuses the presentation of the results, we did not present these data.
2) The authors have stated a rate of 58.3% vs. 82.2% but have presented a risk difference. Surely these are proportions? Agree that rate is not the right word, changed to "proportion".
3) Lines 24-25, unclear whether the authors have repeated the 58.2% vs. 82.2% result (in which case the p-value does not match) or whether this is an error. We have corrected this error.

Conclusion:
Line 50: the pre-specified margin of NI is important but they also need to be adequately justified. We stated in the discussion: "Likewise, investigators designing these trials should recognize the inherent threat of bio-creep and design them with a suitably conservative margin of noninferiority." We opine that this adequately clarified this important issue.

Reviewer: 4
Reviewer Name: Beryl Primrose Gladstone
Institution and Country: Infectious Diseases, Department of Internal Medicine I, Tübingen University Hospital, Tübingen, Germany
Please state any competing interests: None declared

Please leave your comments for the authors below

It is a very interesting question that the authors ask regarding NI trials with reduced intensity therapies and the probability of biocreep among these specific areas. The authors have compared NI trials studying reduced intensity therapies (RIT) with those who did not study RIT.

Major comments:

1. Methodologically speaking, a first step has been made to answer the study question by describing the proportion of non-inferior or superior conclusions based on the trial results and providing the basic descriptive statistics (mean) of the outcome measures. However, to provide an empirical evidence of whether the RIT group (NI-RIT) is testing consistently less effective treatments as compared to the others, it would be necessary at least to perform a meta-analysis of the outcome measures to be able to provide the pooled effect estimate as well the distribution of the effect estimate (which would represent the distribution of the true effect as the authors state that there was no evidence for bias). The methodology can be seen in the article (Gladstone BP and Vach W. About half of the noninferiority trials tested superior treatments: a trial-register based study. 2013. Journal of Clinical Epidemiology, Volume 66 , Issue 4 , 386 – 396). We referenced the study by Gladstone and Vach (reference 16). Our design was not intended to be a meta-analysis, but rather a descriptive study. As such, it is preliminary, and opens the door to extension and replication by others using different methods.

2. It is true that NI trials testing RITs could be expected to be more often inferior/non-inferior rather than superior (which the authors found out) for the same reason which the authors state, that the RITs are of a lower dose and according to the dose response relationship, they would be of lower efficacy. And it is not surprising and is at least satisfactory that about 42% (100-58%) of the NI-RITs lead to either an inferior or inconclusive compared to the 18% (100-82%) declared among NI-nRIT. This reflects the fact that the investigators studying RITs more often at inferior treatments, however the trials are sieving out the inferior treatments and retaining the actual non-inferior treatments. As

mentioned above, the distribution of the treatment effects based on meta-analysis would reflect the authors concluding statements better. We agree with these statements which are similar to those made by reviewer #3, and have added comments in the discussion section to make these points clear to readers. Thank you for pointing this out.

3. The reason behind why NI-RITs are done seems to be usually to use the standard therapy among a subgroup of target patients with a specific characteristic who benefit from reduced intensity, for example, who either cannot tolerate the side effects or developing country patients not able to afford the costs of a long term therapy. Hence it would be interesting and of added value to study whether the benefits are more often mentioned among these trials and whether they are of a different pattern compared to the others. The other general manuscript that utilized this database1 (https://link.springer.com/article/10.1007%2Fs11606-017-4161-4) reported that a substantial proportion (29%) of the included trials did not make explicit mention of the justification for the noninferiority trial on the basis of a purported advantage of the new therapy. Based on your comment, we checked to determine if trials of reduced intensity therapies are more likely to report an advantage of the reduced dose. Indeed, that is the case. 86% of trials of RIT explicitly reported an advantage of the reduced dose, compared to 67% of trials not evaluating two therapies at different doses (P=0.025). We find this result interesting, but did not add it to the manuscript because it is not our intention to make the case that these trials of RIT should not be performed or are not justifiable or contrarily are better justified than other trials, but rather that our analysis suggests that their results require increased scrutiny and that they can be used to make inferences about biocreep.

4. There is evidence of regulatory authorities requiring non-inferiority of a new treatment (or a treatment declared superior in terms of efficacy) in terms of adverse effect in as compared to placebo and NI designs being used to study these. (Pocock SJ, Clayton TC, Stone GW. Challenging Issues in Clinical Trial Design: Part 4 of a 4-Part Series on Statistics for Clinical Trials. Journal of the American College of Cardiology. 2015 Dec 29;66(25):2886–98.) Hence it would be worthwhile to check whether the 6 included placebo controlled trials are safety trials or not. We excluded 14 such safety trials with a placebo comparator, as shown in Figure 2. The remaining 6 trials that compared placebo as a “new therapy” to an unproven standard of care were not safety trials, rather efficacy trials. (Incidentally, it could be argued that they should have been designed as superiority trials as we have pointed out elsewhere3.) A good example is the trial we mentioned that tested perioperative bridging anticoagulation versus none/placebo4. This and the other 5 are efficacy trials rather than safety trials, comparing an unproven standard of care as the active control to placebo as the new therapy. As mentioned, including these among the RIT trials did not change our results.

5. Similarly, the authors state that they expect a stringent margin among the NI-RITs. The discussed issues are a matter of concern among all NI trials generally and apply to all NI trials. When these NIRITs are more often belonging to a specific area such as cardiology, neonatalogy, etc, there is a more probable danger of biocreep. It would be interesting to see whether there is a difference in any of the trial characteristics between the two groups. Metaregression could also help study the contribution of other features contributing to a difference in treatment effects if any. Table 1 could present the two groups and its characteristics. The reviewer introduces another interesting question. While there appear to be more NI-RITs in certain specialties (oncology and infectious diseases – see table below), there are too few within each specialty to allow meaningful analyses with our current dataset. In addition, the classification of specialty can be somewhat arbitrary and artificial. For example, is a trial of antiviral therapy for hepatitis C virus belonging to infectious diseases or gastroenterology, or both? This introduces “researcher degrees of freedom” to the analyses, which we have taken pains to avoid. Our analysis opens the door for others to accumulate a larger database of noninferiority trials with more variables and using different methods such as meta-regression.

6. The study by Gladstone et al referred to as a simulation study in the discussion is not just a simulation study based on assumptions but based on empirical data and of course assumption for the true treatment effect which is again based on empirical data. We have added text to clarify this important point.

Minor comments:

1. Page 5 line 21 – it is the same as what is in line 25 and p value is different? A typo! Yes, a typo and corrected.
2. Reference to be added for the published parallel paper Done
3. 5th reference - typo to be corrected Done
4. Headings for table 1 missing and need to expand to include information on two groups. We have added this information to Table 1.

1. Aberegg SK, Hersh AM, Samore MH. Empirical Consequences of Current Recommendations for the Design and Interpretation of Noninferiority Trials. Journal of general internal medicine. 2017.
2. Djulbegovic B, Kumar A, Glasziou PP, et al. New treatments compared to established treatments in randomized trials. The Cochrane database of systematic reviews. 2012;10:Mr000024.
3. Hersh A, Aberegg SK. N-Terminal Pro-Brain Natriuretic Peptide Trial Design. American journal of respiratory and critical care medicine. 2017;196(4):530.
4. Douketis JD, Spyropoulos AC, Kaatz S, et al. Perioperative Bridging Anticoagulation in Patients with Atrial Fibrillation. New England Journal of Medicine. 2015;373(9):823-833.

VERSION 2 – REVIEW

REVIEWER	Ben Ewald Newcastle UNi NSW, Australia
REVIEW RETURNED	22-Nov-2017

GENERAL COMMENTS	tick

REVIEWER	Beryl Primrose Gladstone Tübingen University Hospital, Tübingen, Germany
REVIEW RETURNED	30-Nov-2017

GENERAL COMMENTS	I would suggest including 95% CI for both the proportions of 58.3% and 82.2% in the results section (page 5 line 36). I just want to explain the concern about the part included in the discussion page 24 line 9-11. The percentage of favorable result can be relatively used but does not refer to the true distribution. Figure 3 picturizes publication bias more often in RITs than non-RIT trials. So it is possible that 58% overestimates the proportion of favorable results. The estimate of 58% of favorable result may lead to biocreep, assuming that the true proportion of non-inferiority was actually lesser. The probability of having false positive results (declared non-inferior when it is actually inferior) depends not only on the delta but also on the true aprior distribution. It would be good to mention it along in the discussion.

	Table 1 – Please include the N (number of trials) for each column in the top row along with the row titles.

REVIEWER	Philippe Flandre INSERM France
REVIEW RETURNED	07-Dec-2017

GENERAL COMMENTS	Aberegg and colleagues discussed the bio-creep phenomenon in noninferiority trials of reduced intensity therapies. Through non-inferiority clinical trials, a new therapy may be approved even if it is less effective than the previous therapy. This raises the possibility that, after a series of non-inferiority trials with each new drug being a little worse than the previous drug, an ineffective or harmful therapy may falsely be deemed efficacious. This phenomenon is known as 'bio-creep'. It is true that this phenomenon is likely to happen with a 'successful' series of trials of reduced intensity therapies. Aberegg and colleagues draw their conclusions based on the review of trials published in the five highest impact general medical journals during a 5-year recent period. The topic is interesting and should be highlighted to clinicians. However, there are few major issues with this paper and the manuscript is not carefully written. 1. The manuscript should be re-read to remove vagueness or mistake. For example, Introduction (page 3, line 21) what is TAP ? (line 22) what is ABVD ?; Results section (page 5, line 5), Figure 1 should be Figure 2; page 5 line 31 Figure 2 should be figure 3; line 24 what is RIT ?,I guess Reduced Intensity Therapies but it is not written. In the figure 1 all horizontal lines representing 95% CI cross the vertical (not dashed) line so noninferiority is not demonstrated whereas the legend of the Figure 1 and the text say the opposite. Overall, more care must be brought to the writing. In the paragraph starting page 5 line 17, the comparison 58.3% vs 82.2% p=0.002 is mentioned two times….. TPA is tissue plasminogen activator, a therapy for stroke, and ABVD is Adriamycin Bleomycin, Vinblastine, Dacarbazine. We added explanations for these abbreviations into the text. We corrected the misnumbering of the figures in the Results section. We removed RIT as an abbreviation for Reduced Intensity Therapies. In figure 1, the 95% CIs represented by horizontal lines all cross zero difference but not delta, so the correct interpretation is that noninferiority criteria were met in panels 2-6 as described in the caption and as explained in Figure 1 of reference #16. We removed the redundant sentence of the comparisons. When we did this we reran our analyses and found some minor errors in the reporting of the differences and associated p-values in the abstract and the results which we have corrected. These small errors do not change the results. Therefore I understand that delta (margin) is missing from Figure 1. In the version I received the quality of the figure is poor.

2. From the introduction it is not clear to understand that the comparison is between trials using reduced intensity therapy and other trials. I wrongly understood that the focus was on former trials. So what are the exact inclusion criteria mentioned in the beginning of the Results section?

This database was generated to capture all noninferiority trials during the study period in the 5 highest impact journals that met our inclusion criteria described in the methods. This was originally done for another analysis which has since been published1. The present analysis compares the subset of those trials that compared a reduced intensity therapy as the new therapy to full intensity as active control to all other trials in the database. This is described in the first sentence of the abstract:

“To identify noninferiority trials within a cohort where the experimental therapy is the same as the active control comparator but at a reduced intensity, and determine if these noninferiority trials of reduced intensity therapies have less favorable results than other noninferiority trials in the cohort.”

In addition we have added further language at the end of our introduction section to assure clarity of this key distinction.

3. Much more details of the 31 trials and the 36 comparisons should be given either in the table 2 or e-table in an appendix.

We have added to table 1 caption stating that additional details of included studies can be found in reference 15.

The authors leave it to the editors to determine if we should include a bibliography of all 31 trials and 36 comparisons of reduced intensity therapies. It is our opinion that this would be of little use to all but a tiny minority of readers. We reason that Table 2, listing some examples, makes it clear what we mean by trials of reduced intensity therapies, and that expanding it would lengthen the article without increasing its quality or utility for readers. Alternatively, we can make our entire database available for exploration by any interested parties.

4. As introduced by Figure 1 the bio-creep phenomenon is due to a chain of noninferiority trials. So results based on the 36 comparisons of a reduced intensity therapy are somewhat paradoxical. The authors concluded that such trials increase the risk of bio-creep but less trials of RIT conclude to noninferiority (58.3% vs 82.2%) breaking the chain of RIT trials. In fact, the rank of the trial in the chain is very important. The authors should provide the rank of the RIT trials used in their analysis because the first RTI trial (#2 in Figure 1) as a different impact in the bio-creep phenomenon than the fourth trial (#5 in Figure 1). If noninferiority is not demonstrated in the first trial there is no ‘bio-creep’ while if noninferiority is not demonstrated in the fourth trial it is likely that there is already a ‘bio-creep’.

The authors should discussed this point.

This is an excellent point that we failed to consider, and we have added text to the discussion to address it.

However, when 58% of trials of RIT do in fact conclude noninferiority or superiority, the risk of biocreep in subsequent trials remains for the majority of trials. If it were the case that only a small minority of trials of RIT met superiority or noninferiority criteria, this proposed phenomenon, that of “breaking the chain”, would take on greater importance.

I did not find where you discussed that point in the Discussion.

5. The analysis that provides Figure 3 has no meaning to me. I understand that the 151 absolute differences come from the 182 noninferiority comparisons mentioned page 5 line 9. Such a comparison is like comparing apples and strawberries. I guess that primary endpoints of these trials are quite different. For example, in noninferiority trials involving HIV-1 infected patients the primary endpoint is often virologic failure. In many oncology trials the primary endpoint is still mortality. That one of the reason noninferiority margins are quite different according to the primary endpoint. In Figure 3 what is the meaning of comparing a 0.5% mortality difference with a 2% virologic failure difference or with a 2% rate of adverse events….

This is an inherent limitation of meta-research projects, and one reason we did not perform a formal meta-analysis, but rather a descriptive study. We reason that there is value, nonetheless, in comparing the point estimates of the results of empirical data as they relate to directionality and absolute risk differences, and there is precedent for such analyses2. We think it is fair to assume that there is no systematic difference between the choice of primary endpoints for Reduced Intensity Non-Interiority Trials vs. Non-Inferiority Trials of Novel Therapies. As such we can reasonably assume that, despite an “apples to strawberries” comparison, our analysis reveals underlying structural differences between Reduced Intensity Non-Interiority Trials and Non-Inferiority Trials of Novel Therapies.

6. Although I agree with the fact that the sample size would increase after a series of noninferiority trials there is a confusion between design and result of a trial in the Discussion section. Considering the example in the Discussion of a trial involving 1800 patients in each arm to compare 7 to 10% event rate. Such a sample size lead to many reullts concluding superiority, such as 4 vs 10%, 2 vs 12% ect…..With the latter comparison and a linear relationship between dose and response the sample size would be much lower than 16000 patients.

We understand the sentiment of the reviewer’s comment, but maintain that sample size calculations must be made a priori. If investigators seek to demonstrate half the effect size at the same power, they must still design the trial as per our example. That does not mean that they may not find a statistically significant effect smaller than the pre-specified margin (delta), but to have a priori power to do so, they must design the trial as per our example.

I’m not convinced by the answer. Of course, it is true that the sample size is computed a priori but the computation is based on the results of previous trials not on hypotheses of previous trials.

VERSION 2 – AUTHOR RESPONSE

Author responses are in red colored font in the Supplementary File of this response to the decision letter, which will be far easier for the reviewers and editors to view.

From: info.bmjopen@bmj.com
To: scottaberegg@gmail.com
CC: scottaberegg@gmail.com, andrewmhersh@gmail.com, matthew.samore@hsc.utah.edu
Subject: BMJ Open - Decision on Manuscript ID bmjopen-2017-019494.R1
Body: 08-Dec-2017

Dear Dr. Aberegg:

Manuscript ID bmjopen-2017-019494.R1 entitled "Reduced Effects in Noninferiority Trials of Reduced Intensity Therapies" which you submitted to BMJ Open, has been reviewed. The comments of the reviewer(s) are included at the bottom of this letter.

The reviewer(s) have recommended publication, but also suggest some minor revisions to your manuscript. Therefore, I invite you to respond to the reviewer(s)' comments and revise your manuscript.

To revise your manuscript, log into https://mc.manuscriptcentral.com/bmjopen and enter your Author Center, where you will find your manuscript title listed under "Manuscripts with Decisions." Under "Actions," click on "Create a Revision." Your manuscript number has been appended to denote a revision.

You may also click the below link to start the revision process (or continue the process if you have already started your revision) for your manuscript. If you use the below link you will not be required to login to ScholarOne Manuscripts.

*** PLEASE NOTE: This is a two-step process. After clicking on the link, you will be directed to a webpage to confirm. ***

https://mc.manuscriptcentral.com/bmjopen?URL_MASK=138bfea7aa04466faec8e2907f773a6e

You will be unable to make your revisions on the originally submitted version of the manuscript. Instead, revise your manuscript using a word processing program and save it on your computer. Please also highlight the changes to your manuscript within the document by using the track changes mode in MS Word or by using bold or colored text.

Once the revised manuscript is prepared, you can upload it and submit it through your Author Center.

When submitting your revised manuscript, you will be able to respond to the comments made by the reviewer(s) in the space provided. You can use this space to document any changes you make to the original manuscript. In order to expedite the processing of the revised manuscript, please be as specific as possible in your response to the reviewer(s).

You will receive a proof if your article is accepted, but you will be unable to make substantial changes to your manuscript, please take this opportunity to check the revised submission carefully.

IMPORTANT: Your original files are available to you when you upload your revised manuscript. Please delete any redundant files before completing the submission.

Because we are trying to facilitate timely publication of manuscripts submitted to BMJ Open, your revised manuscript should be submitted within two weeks. If it is not possible for you to submit your revision by this date, we may have to consider your paper as a new submission.

Once again, thank you for submitting your manuscript to BMJ Open and I look forward to receiving your revision.

Sincerely,
Hemali Bedi
Assistant Editor, BMJ Open
hbedi@bmj.com

Editorial Requirements:
- Your title is unacceptable. Please revise your title to state the research question and study design. We do not accept declarative titles.
We changed the title on the title page but neglected to do so on manuscript central. He have not changed it on the latter.
- Please revise the Strengths and Limitations section (after the abstract) to focus on the methodological strengths and limitations of your study only, rather than summarizing the results. Bullet point number 4 contains a study finding. Please remove this.
We have removed #4.
- We felt that some of your previous responses to the reviewers comments were unclear. Please clarify what your previous response was to reviewer 3's first comment regarding what can be done about biocreep.
OK.
- Going forward, please provide a more detailed response to the reviewers comments. Please include specific page and line numbers indicating where changes have been made. This is the standard we expect from all authors submitting revisions.
OK.

Reviewer: 1
Reviewer Name: Ben Ewald
Institution and Country: Newcastle UNi, NSW, Australia
Please state any competing interests: none

Please leave your comments for the authors below
tick

Reviewer: 2
Reviewer Name: Philippe Flandre
Institution and Country: INSERM France
Please state any competing interests: None declared

Please leave your comments for the authors below

Aberegg and colleagues discussed the bio-creep phenomenon in noninferiority trials of reduced intensity therapies.

Through non-inferiority clinical trials, a new therapy may be approved even if it is less effective than the previous therapy. This raises the possibility that, after a series of non-inferiority trials with each new drug being a little worse than the previous drug, an ineffective or harmful therapy may falsely be deemed efficacious. This phenomenon is known as ‘bio-creep’. It is true that this phenomenon is likely to happen with a ‘successful’ series of trials of reduced intensity therapies. Aberegg and colleagues draw their conclusions based on the review of trials published in the five highest impact general medical journals during a 5-year recent period. The topic is interesting and should be highlighted to clinicians. However, there are few major issues with this paper and the manuscript is not carefully written.

1. The manuscript should be re-read to remove vagueness or mistake. For example, Introduction (page 3, line 21) what is TAP ? (line 22) what is ABVD ?; Results section (page 5, line 5), Figure 1 should be Figure 2; page 5 line 31 Figure 2 should be figure 3; line 24 what is RIT ?,I guess Reduced Intensity Therapies but it is not written. In the figure 1 all horizontal lines representing 95% CI cross the vertical (not dashed) line so noninferiority is not demonstrated whereas the legend of the Figure 1 and the text say the opposite. Overall, more care must be brought to the writing. In the paragraph starting page 5 line 17, the comparison 58.3% vs 82.2% p=0.002 is mentioned two times…..

TPA is tissue plasminogen activator, a therapy for stroke, and ABVD is Adriamycin Bleomycin, Vinblastine, Dacarbazine. We added explanations for these abbreviations into the text. We corrected the misnumbering of the figures in the Results section. We removed RIT as an abbreviation for Reduced Intensity Therapies. In figure 1, the 95% CIs represented by horizontal lines all cross zero difference but not delta, so the correct interpretation is that noninferiority criteria were met in panels 2-6 as described in the caption and as explained in Figure 1 of reference #16. We removed the redundant sentence of the comparisons. When we did this we reran our analyses and found some minor errors in the reporting of the differences and associated p-values in the abstract and the results which we have corrected. These small errors do not change the results.

Therefore I understand that delta (margin) is missing from Figure 1. In the version I received the quality of
the figure is poor.

We apologize. However, we do not understand why the quality of the figure is poor. It is a 300 DPI figure as per editorial requirements, and I just double checked the properties of the file to confirm this. As regards the delta margin, the caption for figure 1 states:
"Figure 1. Diagram showing loss of presumed superiority to placebo with reduced intensity aspirin therapy in a hypothetical sequence of trials. The experimental therapy is on the left in each panel and the control is on the right; point estimates are represented as black ovals with bisecting horizontal lines representing 95% confidence intervals – point estimates on the left of the center line favor the experimental therapy and point estimates on the right favor the active control. In panels #2-6, the vertical dashed line represents the margin of noninferiority. In panel #1, aspirin 325 mg is superior to placebo control in a superiority trial. In panel #2, reduced dose aspirin at 162 mg as the experimental therapy is compared to full dose aspirin as active control. The difference favors full dose aspirin, but the reduced dose meets noninferiority criteria because the upper bound of the 95% confidence interval does not cross the noninferiority margin. The dose of aspirin is successively reduced in panels #3-5, with the reduced dose from the previous panel serving as the active control in the subsequent panel. By panel #6, the dose of active control aspirin is 20 mg, and the experimental therapy is aspirin at a dose of 0 mg (i.e., placebo) and placebo is noninferior to aspirin – a highly paradoxical result compared to panel #1 where aspirin was superior to placebo. This result obtains because in panels #2-6, reduced efficacy of the experimental therapy is concealed in the margin of noninferiority. This phenomenon has been called "bio-creep."

2. From the introduction it is not clear to understand that the comparison is between trials using reduced intensity
therapy and other trials. I wrongly understood that the focus was on former trials. So what are the exact inclusion
criteria mentioned in the beginning of the Results section?
We have made changes to clarify, in the methods, the inclusion criteria, which are now explicitly stated:
"This study used a dataset that was created for a different analysis of noninferiority trials1.. We searched MEDLINE for iterations of noninferiority (e.g., non-inferiority, noninferior)2 combined with the MEDLINE-recognized names of the five highest impact general medical journals (New England Journal of Medicine, Lancet, JAMA, British Medical Journal, Annals of Internal Medicine) to identify manuscripts reporting the results of prospective parallel group randomized controlled trials using a test of noninferiority for the primary hypothesis published between June, 2011 and October, 2016 (inclusion criteria)."

This database was generated to capture all noninferiority trials during the study period in the 5 highest impact
journals that met our inclusion criteria described in the methods. This was originally done for another analysis
which has since been published1. The present analysis compares the subset of those trials that compared a reduced
intensity therapy as the new therapy to full intensity as active control to all other trials in the database. This is
described in the first sentence of the abstract:

"To identify noninferiority trials within a cohort where the experimental therapy is the same as the active control comparator but at a reduced intensity, and determine if these noninferiority trials of reduced intensity therapies have less favorable results than other noninferiority trials in the cohort."

In addition we have added further language at the end of our introduction section to assure clarity of this key distinction.

3. Much more details of the 31 trials and the 36 comparisons should be given either in the table 2 or e-table in an appendix.

We have added an Appendix 1, to be included at the discretion of the editors, that is a complete bibliography of the 31 trials of reduced intensity therapies.

We have added to table 1 caption stating that additional details of included studies can be found in reference 15.

The authors leave it to the editors to determine if we should include a bibliography of all 31 trials and 36 comparisons of reduced intensity therapies. It is our opinion that this would be of little use to all but a tiny minority of readers. We reason that Table 2, listing some examples, makes it clear what we mean by trials of reduced intensity therapies, and that expanding it would lengthen the article without increasing its quality or utility for readers. Alternatively, we can make our entire database available for exploration by any interested parties.

4. As introduced by Figure 1 the bio-creep phenomenon is due to a chain of noninferiority trials. So results based on the 36 comparisons of a reduced intensity therapy are somewhat paradoxical. The authors concluded that such trials increase the risk of bio-creep but less trials of RIT conclude to noninferiority (58.3% vs 82.2%) breaking the chain of RIT trials. In fact, the rank of the trial in the chain is very important. The authors should provide the rank of the RIT trials used in their analysis because the first RTI trial (#2 in Figure 1) as a different impact in the bio-creep phenomenon than the fourth trial (#5 in Figure 1). If noninferiority is not demonstrated in the first trial there is no 'bio-creep' while if noninferiority is not demonstrated in the fourth trial it is likely that there is already a 'bio-creep'.

The authors should discussed this point.

Figure 1 is intended for illustrative purposes. We stated in the introduction that the process need not be iterative:

This sequence culminates in the paradoxical result in panel 6, where the dose of the experimental therapy is reduced to zero, making it a placebo which is noninferior to aspirin. In this hypothetical sequence, inferiority of reduced dose aspirin is obscured within the margin of noninferiority in panels 2-5. However, the process need not be iterative – some loss of efficacy and thus presumed superiority to placebo occurs with just one dose reduction in panel 2. This problem will be exacerbated with larger margins of noninferiority and greater reductions in therapy intensity. Though this phenomenon, called "bio-creep", could happen in any noninferiority trial, the likelihood would appear to be greater in trials of reduced intensity therapies because of fundamental dose-response considerations.

This is an excellent point that we failed to consider, and we have added text to the discussion to address it.
However, when 58% of trials of RIT do in fact conclude noninferiority or superiority, the risk of biocreep in
subsequent trials remains for the majority of trials. If it were the case that only a small minority of trials of RIT met
superiority or noninferiority criteria, this proposed phenomenon, that of “breaking the chain”, would take on greater
importance.

I did not find where you discussed that point in the Discussion.
In our first revision, we added an entire paragraph to the discussion to acknowledge Reviewer 2’s observations:
“An alternative interpretation of our results was offered by two reviewers. The reviewers noted that since noninferiority or superiority criteria were met for only 58% of trials of reduced intensity therapies, the proposed sequence of biocreep illustrated in Figure 1 was interrupted for 42% of the trials with the first noninferiority trial. That is, the noninferiority trials were effective in filtering out truly noninferior therapies. We agree that it is reassuring that many noninferiority trials of reduced intensity therapies fail to demonstrate superiority or noninferiority but note that the majority do meet noninferiority criteria. This is concerning because any declaration of noninferiority is highly sensitive to the choice of delta – with a large enough delta any therapy can be declared noninferior.”
Unfortunately, we have no way of ascertaining where our trials fit in any hypothetical sequence of trials. Figure 1 is for illustrative purposes, and is clearly labeled as a “hypothetical sequence of trials” in the caption. We included it to clarify the proposed phenomenon of biocreep because reviewers for previous submissions had failed to firmly grasp the concept and we opined that other readers will also benefit from the illustration.

5. The analysis that provides Figure 3 has no meaning to me. I understand that the 151 absolute differences come
from the 182 noninferiority comparisons mentioned page 5 line 9. Such a comparison is like comparing apples and
strawberries. I guess that primary endpoints of these trials are quite different. For example, in noninferiority trials
involving HIV-1 infected patients the primary endpoint is often virologic failure. In many oncology trials the
primary endpoint is still mortality. That one of the reason noninferiority margins are quite different according to the
primary endpoint. In Figure 3 what is the meaning of comparing a 0.5% mortality difference with a 2% virologic
failure difference or with a 2% rate of adverse events….
Since the underlying processes are stochastic, deviations from zero in any measured outcome are worthy of being considered in aggregate, as part of a random process.
There is precedent for doing this with disparate outcomes in meta-research. For example, see: Djulbegovic et al3 :
Djulbegovic B, Kumar A, Glasziou PP, et al. New treatments compared to established treatments in randomized trials. The Cochrane database of systematic reviews 2012;10:Mr000024. doi: 10.1002/14651858.MR000024.pub3 [published Online First: 2012/10/19]
I am at a loss as to how to further address this criticism, but I note that the other 3 reviewers did not take issue with Figure 3. We can remove Figure 3 if the editors wish for us to do so.

This is an inherent limitation of meta-research projects, and one reason we did not perform a formal meta-analysis, but rather a descriptive study. We reason that there is value, nonetheless, in comparing the point estimates of the results of empirical data as they relate to directionality and absolute risk differences, and there is precedent for such

analyses2. We think it is fair to assume that there is no systematic difference between the choice of primary endpoints for Reduced Intensity Non-Interiority Trials vs. Non-Inferiority Trials of Novel Therapies. As such we can reasonably assume that, despite an “apples to strawberries” comparison, our analysis reveals underlying structural differences between Reduced Intensity Non-Interiority Trials and Non-Inferiority Trials of Novel Therapies.

6. Although I agree with the fact that the sample size would increase after a series of noninferiority trials there is a confusion between design and result of a trial in the Discussion section. Considering the example in the Discussion of a trial involving 1800 patients in each arm to compare 7 to 10% event rate. Such a sample size lead to many reullts concluding superiority, such as 4 vs 10%, 2 vs 12% ect…..With the latter comparison and a linear relationship between dose and response the sample size would be much lower than 16000 patients.

We understand the sentiment of the reviewer’s comment, but maintain that sample size calculations must be made a priori. If investigators seek to demonstrate half the effect size at the same power, they must still design the trial as per our example. That does not mean that they may not find a statistically significant effect smaller than the pre-specified margin (delta), but to have a priori power to do so, they must design the trial as per our example.

I’m not convinced by the answer. Of course, it is true that the sample size is computed a priori but the computation is based on the results of previous trials not on hypotheses of previous trials.
We went back through this example, running several simulations with different assumptions for delta for the noninferiority trial. Reviewer 2 is correct as long as as a reasonable delta, smaller than the point estimate of the superiority trial, is used for the first noninferiority trial. Thus, we agree that our example is misleading and inaccurate, and we have removed it from the manuscript. Thank you for persisting with this point and forcing us to evaluate it more carefully.

Reviewer: 4
Reviewer Name: Beryl Primrose Gladstone
Institution and Country: Tübingen University Hospital, Tübingen, Germany
Please state any competing interests: None declared

Please leave your comments for the authors below

I would suggest including 95% CI for both the proportions of 58.3% and 82.2% in the results section (page 5 line 36).
We have added these data.

I just want to explain the concern about the part included in the discussion page 24 line 9-11. The percentage of favorable result can be relatively used but does not refer to the true distribution. Figure 3 picturizes publication bias more often in RITs than non-RIT trials. So it is possible that 58% overestimates the proportion of favorable results. The estimate of 58% of favorable result may lead to biocreep, assuming that the true proportion of non-inferiority was actually lesser. The probability of having false positive results (declared non-inferior when it is actually inferior) depends not only on the delta but also on the true aprior distribution. It would be good to mention it along in the discussion.
We added a sentence into the discussion to clarify this point.

Table 1 – Please include the N (number of trials) for each column in the top row along with the row titles.
We have done this.

Date Sent: 08-Dec-2017

VERSION 3 – REVIEW

REVIEWER	Beryl Primrose Gladstone University Hospital Tuebingen, Germany
REVIEW RETURNED	08-Jan-2018

GENERAL COMMENTS	No comments!