Author’s response to reviews

Title: Optimising trial designs for identifying appropriate antibiotic treatment durations

Authors:

Koen Pouwels (k.b.pouwels@gmail.com)
Mo Yin (MoYin@tropmedres.ac)
Christopher Butler (christopher.butler@phc.ox.ac.uk)
Ben Cooper (Ben@tropmedres.ac)
Sarah Wordsworth (sarah.wordsworth@dph.ox.ac.uk)
Sarah Walker (sarah.walker@ndm.ox.ac.uk)
Julie Robotham (julie.robotham@phe.gov.uk)

Version: 2 Date: 11 Apr 2019

Author’s response to reviews:

We thank the reviewers for their helpful comments which we consider have substantially improved our manuscript.

Point-by-point responses are provided below.

Reviewer #1:
1. As background, please consider delineating the difficulty in assessing treatment duration in observational studies - what are the difficulties that cannot be well circumvented.
   
   Response:
   The most obvious limitation of observational studies comparing different durations is confounding. In situations where it is feasible to perform randomised controlled trials, such as when comparing different treatment durations, one should avoid making conclusions on observational studies that are potentially affected by unmeasured confounding.

   To address this comment we added the following: ‘Because observational studies comparing different antibiotic durations are potentially confounded by unmeasured patient factors influencing need for prolonged treatment, evidence about the optimal treatment duration should come from randomised controlled trials (RCTs).’
2. Figure 1: difficult to understand. Would start with defining the graphs (diamonds, solid/ dashed lines). Are the lines hypothesized or actual event rates, given that they show non-compliance? The last sentence of the legend is very unclear. I'm not sure this figure is in its place, because it addresses compliance, which is mentioned only much further on in the manuscript. The statements (a bit repetitive) that "such a trial does not answer the more important question of 'what is the optimal antibiotic treatment duration for prostatitis?' and Such RCTs provide information about the two evaluated durations, but do not provide much information about durations not considered in the trial" do not require a figure; they are clear as presented in text. Furthermore, when you raise the problem of compliance in duration trials (Line 284-292), you do not provide a solution: "The effect of non-compliance, which is likely not completely random, can be taken into account". This is an important issue, can you provide more specific guidance on how to take into account? Can you merge an explanation with the figure, explaining better what the figure shows?

Response:
To address this comment, we changed the order of the sentences. The figure title is immediately followed by the following sentences, which clearly indicate that it reflects hypothesised event rates: ‘Diamonds show hypothesised event rates for the two randomised groups as designed. The solid and dot-dashed lines show different hypothesised duration-response curves that are compatible with those hypothesised event rates.’

We do not agree with the reviewer that the failure of the traditional two-arm RCT to identify an optimal duration in the case presented in Figure 1(b) is clear – our experience is that most will assume the longer duration is optimal in this case which may be far from the truth. We therefore think it is very important to keep this figure.

When discussing non-compliance later in the main text we now also refer to figure 1, and provide references to literature that explains how to do this, since we do not have space to discuss this issue in detail:

‘An issue encountered with all antibiotic duration trials is potential non-compliance. Non-compliance can provide a distorted picture of the efficacy of treatment durations when performing an intention-to-treat (ITT) analysis (Figure 1, top panel).’

&

‘The effect of non-compliance, which is likely not completely random, can be taken into account using instrumental variable approaches and/or g-methods as described in more detail by Berry and Hernan et al.’

3. Table 1: under "Obtain information about which durations" you write "considered and non-considered durations" for the trial designs other than the simple 2-arm design. Is this true? This wording might be confusing. All durations allocated in the trial are considered in advance. All trial designs are interventional, meaning the interventions were pre-planned, thus considered.

Response:
We agree that the wording may have been confusing. We intended to state here that these designs can model the entire duration-response curve, including estimates for durations that were not used in the trial (as explained in the text). We have changed the sentence in the table to: ‘Can model the entire duration-response curve, including estimates for durations that were not used’.
4. Power/sample size calculation is the most significant challenge for the trial designs proposed. There is no free lunch; assessing more durations and patient subgroups requires a larger sample size to avoid type II errors. Can you provide guidance? How would the stopping rules be taken into account when computing the same size?

Response:
In a standard trial, Type II error relates to missing a genuine difference between experimental and control treatments. Identifying an optimal duration out of many durations is a different type of problem, since they are intrinsically related by the duration-response curve, which will depend on many factors that vary by indication, specific outcome, population, existence of potentially relevant subgroups, type of antibiotic, etc. We acknowledge therefore that general guidance on sample size calculations is difficult to provide and an area where more work is needed – although we do not think this invalidates our points about the inefficiency of traditional designs for identifying the optimal duration. We now explicitly include the following to address this point:

‘In the recent proposal for the fixed duration design, simulations showed that a sample size of 500 patients divided into 5-7 equidistant arms was sufficient to estimate the duration-response curve within a 5% error margin in 95% of the simulations, suggesting that a trial using similar methodology is feasible in practice. Similar simulations focusing on numbers needed to estimate duration-response curves for the other designs do not exist yet. In general, using standard pairwise comparisons, the more arms included the greater the sample size – but it is not clear that such pairwise comparisons are ideal for determining optimal treatment duration.’

&

‘We have only provided theoretical considerations regarding these 4 designs, but we urge the research community to consider developing, testing and applying alternative trial designs that can identify optimal treatment durations, including sample size calculations.’

5. In line 300 you mention that it is preferable to define futility stopping criteria at the planning stage. This is not preferable, but required.

Response:
We intended to discuss here that a DMC can also decide to terminate a trial or drop specific arms when using a fixed design based on unexpected adverse events. We clarified this by changing this paragraph as follows:

‘For all designs, including fixed trial designs, continuous response monitoring for serious and unexpected adverse events or lack of efficacy of certain durations by an independent data monitoring committee (DMC) can ensure that patients are protected from being randomised to an unsafe arm. For adaptive designs, futility stopping criteria are defined at the planning stage. This can be done for both frequentist and Bayesian trials and provide statistical rules to help the DMC decide whether an arm should be dropped. The advantage of having the option to drop poorly performing arms (drop-the-loser design) is that it potentially reduces the number of patients allocated to unfavourable antibiotic durations. This is not only ethically desirable, but may also convince more patients to participate in the trial.’

6. An example of a planned trial would be appreciated, putting the suggested design into practice. You started with the example of prostatitis; it would be helpful if you provide the design of the optimized trial to assess treatment duration of prostatitis (but could be any infection), including the assessment of short term outcome, rarer, long term events and resistance development (not optional, but required in
antibiotic duration trials), with the power of the trial to detect the different outcomes. This could be provided in a supplement.

Response:
The purpose of our opinion piece is to highlight to the research community that we need to move away from trials that compare two durations towards multi-arm trial designs that estimate duration-response curves. The current draft is already at the word limit (word count: 3040; guidance indicates ~3000 words), and we do not consider that any of the material is unnecessary and hence could be deleted to make space for this. Our feeling is that providing a complete trial design, including the extensive simulations that are needed to assess potential error rates and statistical efficiency, is worth a paper on its own and hence out of the scope of this work.

Reviewer #2:
1. Overall, this manuscript lacks structure and is hard to follow. A clearer path should be stated early on to help the reader follow along.

Response:
We restructured the manuscript, removed paragraphs that made the manuscript difficult to follow or were repetitive, and added a clearer path at the end of the introduction:
‘Next, we discuss i) the main issue with conventional 2-arm trial designs; ii) how to assess the ‘optimal’ antibiotic treatment duration; iii) four alternative trial designs that can estimate much needed duration-response relationships (subsequently denoted duration-response curves); and iv) which of these designs has the most desirable properties.’

2. The manuscript is also repetitive. In spite of this, it is still unclear which clinical trial designs perform better or poorer with respect to the biases discussed. I believe the reviewers might try to reduce repetition and use the leftover space for a specific example that showcases the strengths of the multi-arm randomization design with the option to drop duration arms.

Response:
We have removed the repetition from the manuscript and used the freed up space to be more specific about the strength and weakness of the designs:

‘When using an open-label design that preferentially allocates patients to specific durations with better outcomes (RAR), clinicians will be able to determine during the trial that these durations are associated with better outcomes, thereby increasing the risk of allocation concealment (selection) bias. This knowledge can change which patients get randomised in the trial and how endpoints will subsequently be assessed. The other designs (play the winner, fixed duration and drop-the-loser design) all reduce the risk of selection bias because clinicians cannot change selection of patients based on observed changes in allocation probabilities for these designs.’

&

It is often cautioned that calendar time trends – which are common with infectious diseases – may introduce bias when using response-adaptive randomisation. However, one can take advantage of the fact that randomisation probabilities are not constantly changing with most RAR designs. A calendar-time stratified analysis, with equal randomisation probabilities within each stratum, eliminates potential time-trend bias. A larger sample size is needed with such stratified analyses, but it is important to avoid trying to gain small improvements in efficiency at the cost of introducing bias.36 While the fixed duration design is not vulnerable to time-trends due to its design, the RAR and play the winner design
require a less efficient calendar-time stratified analysis to avoid this type of bias. When considering a drop-the-loser design one should avoid comparison of patients assigned to the dropped duration with patients that were randomised to other arms after dropping the clearly inferior arm to avoid this bias. However, this may not be problematic given that there was already enough information to deem the duration clearly inferior.

‘The advantage of having the option to drop poorly performing arms (drop-the-loser design) is that it potentially reduces the number of patients allocated to unfavourable antibiotic durations. This is not only ethically desirable, but may also convince more patients to participate in the trial.’

‘Given the considerations laid out above, the fixed duration and the drop-the-loser duration design seem to be the designs that theoretically have the most potential to identify optimal antibiotic treatment durations. These designs 

i) are less vulnerable to allocation concealment bias than the RAR design;

ii) are not (fixed duration) or less (drop-the-loser) vulnerable to time-trend bias compared to the RAR or the play the winner design;

iii) are not associated with important logistical challenges that often come with adaptive trials that allow for changes in the allocation ratios (pick the winner and RAR design); and

iv) are more likely to have sufficient number of patients in each arm and/or subgroup at the end of the trial to estimate the complete duration-response curve with sufficient precision than the RAR and play the winner design, and hence enable evaluation of the potential for important differences in the optimal duration within specific subgroups;

A potential advantage of the drop-the-loser design over the fixed duration design is that the former can drop duration arms that are clearly inferior vs the standard (maximum) duration based on formal statistical analysis. This may ethically be more acceptable by reducing the number of patients allocated to inferior treatment durations.

We have only provided theoretical considerations regarding these 4 designs, but we urge the research community to consider developing, testing and applying alternative trial designs that can identify optimal treatment durations, including sample size calculations.’

3. A brief sentence or two on what happens to the subjects in dropped arms would also helpful for the reader. Also, the authors might consider citing some statistics that highlight why registry/real-world-evidence/chart review/retrospective studies are insufficient to answer the optimal duration question.

Response:
To address these comments we added the following:

‘After dropping an arm, follow-up will continue for patients assigned to this duration.’

‘Because observational studies comparing different antibiotic durations are potentially confounded by unmeasured patient factors influencing need for prolonged treatment, evidence about the optimal
treatment duration should come from randomised controlled trials (RCTs).’

4. In summary, the paper reviews very briefly some trials for possible use to ascertain optimal antibiotic treatment durations in a randomized trials and reads like an opinion piece that ends with a recommendation without good justification on statistical ground. This is the biggest weakness of the paper.

Response:
Our manuscript is intended as an opinion piece. To address this comment we have
- toned down our recommendation (see response to comment 2);
- made the advantages and disadvantages of the different designs more explicit (see response to comment 2);
- made clear that our recommendation is based on theoretical considerations and that more research into the properties of alternative designs that can estimate duration-response curves are needed, as well as actual application of such designs to better identify optimal antibiotic treatment durations (see response to comment 2).

Suggestions:
5. Pg 4, Line 42: "For many infectious conditions, the ideal antibiotic course length is unclear."
Response: We have changed this sentence accordingly.

6. Pg 4, Line 47: "We argue that alternative trial designs, which allow allocation of patients to multiple different treatment durations, are needed to better identify optimal antibiotic durations."
Response: We have changed this sentence accordingly.

7. Pg 5, Line 78: "An important challenge is that for many infectious conditions, the ideal antibiotic course length is unclear."
Response: We have changed this sentence accordingly.

8. Pg 5, Line 86: Replace "started" with "initiated"
Response: We have amended as suggested.

9. Pg 6, Line 115-116: Delete. Repetitive with the last two sentence of prior paragraph.
Response: We have deleted this sentence.

10. Pg 7, Line 123-125. It is unclear why patients who are non-compliant in the shorter duration end up receiving the standard duration. I interpret the sentence here to mean that the subjects who are non-compliant within the shorter duration will still finish their antibiotics (drawing the duration longer). Is this right? If so, wouldn't it be fair to assume that non-compliant subjects in the longer duration arm will also eventually finish their antibiotics, lengthening their regimen as well?
Response:
In our hypothetical example we assumed that patients would be more likely to return to their doctor and hence continue antibiotic if they were not fully cured (e.g. still had persisting minor symptoms which might then relapse without further antibiotic treatment) by the end of their assigned duration. In practice not all patients will require the same duration and, at the population level, the proportion of patients that are not cured will likely decrease with increasing assigned duration, as will the proportion with non-compliance, creating the dot-dashed line in the top panel of figure 1.
We tried to clarify this by adding the following: ‘The dot-dashed line in the top panel could occur if there is some non-compliance with the shorter duration as randomised, because patients who are not fully cured (e.g. still have persisting minor symptoms which might then relapse without further antibiotic treatment) at the end of their assigned duration (and hence more patients end up receiving the standard duration (28 days) despite being randomised to a shorter duration). In practice not all patients will require the same duration and, at the population level, the proportion of patients that are cured will likely decrease with increasing assigned duration, as will the proportion with non-compliance, creating the dot-dashed line in the top panel.

11. Pg 7, Line 134. Insert a comma after "long-term"
Response: We have inserted a comma.

Response: We have made the spacing between paragraphs consistent.

13. Pg 8, Line 143: using DOOR/RADAR where conventional trial designs…. where? when? while? the sentence is a little hard to follow.
Response: We have changed this sentence to make it easier to follow: ‘This unverified strong assumption could lead to demonstration of non-inferiority using DOOR/RADAR when conventional trial designs may show that shorter duration are not non-inferior.’

14. Pg 8, Line 144. replace "duration" with "durations"
Response: We have amended as suggested.

15. Pg 8, Line 148. add a comma after "approach"
Response: We have added a comma as suggested.

16. Pg 9, Line 171. are these "various prior opinions and utility functions" that were discussed in the prior paragraph (which pointed out the difficulty in multiple individuals coming to an agreement on these opinions/functions)"
Response: These are indeed the opinions and utility functions compatible with the opinions of different stake-holders (see 2 sentences above) – we have added this to clarify.

17. Pg 9, Line 178. "stopping criteria" is a term frequently used for criteria to stop a clinical trial early. Given this manuscript does discuss clinical trials, the term "stopping criteria" should be avoided when its meaning is associated with ideal antibiotic treatment duration.
Response: We have deleted this sentence to shorten the manuscript and because, when rereading the manuscript, we felt that this phrase discussed a different kind of question.

18. Pg 10, Line 190. "play-the-winner"
Response: We have now used play-the-winner throughout the manuscript.

19. Pg 10, Line 196. add a comma after "generating the data"
Response: We have added a comma.

20. Pg 10, Line 204. "compared to the standard duration based on Bayesian posterior predictive probabilities, or based on other predefined stopping criteria using a frequentist test-statistic."
Response: We have changed this sentence accordingly.
21. Pg 11, Line 208. add "of" after "Main characteristics"
Response: We have amended as suggested.

22. Pg 13, Line 245. How preferentially assigning patients to better-performing arms hampers evaluation of the complete duration-response curve is unclear. Because some arms will have insufficient number of subjects? Because of bias? Please clarify.
Response: This is due, for example, to insufficient numbers of patients in arms of shorter durations, that may nevertheless be optimal in a subgroup of patients. We have clarified this as follows:
‘However, a potential issue with adaptive designs that preferentially assign patients to better performing arms (RAR and play the winner designs), is that this may hamper proper evaluation of the complete duration-response curve due to insufficient number of patients receiving different durations.’

23. Pg 13, Line 250. add a comma after "randomization"
Response: We have added a comma after randomization.

24. Pg 13, Line 260. remove comma after "subgroup"
Response: We have removed this comma.

25. Pg 13, Line 265. Perhaps this section can be presented earlier, and then the designs can be presented, pointing out their ability to address these risk of bias issues
Instead of presenting this section earlier, we have signposted more clearly what can be found in which section at the end of the Introduction.

26. Pg 14, Line 274. Consider terming this as "selection bias" rather than "allocation concealment bias"
Response: We’ve now used added the term ‘selection bias’ because allocation concealment bias can result in selection bias. However, we have also kept allocation concealment bias in the text, because potential changes in how outcomes are assessed have nothing to do with selection bias, and because allocation concealment is one of the specific RCT biases directly evaluated in the CONSORT trial checklist, and hence considered important.

27. Pg 14, Line 279. stratified or subgroup? stratified analysis usually involves analyzing different randomization strata
Response: We apologise for lack of clarity in the original text. As explained in reference 38, calendar-time stratified analysis (resulting in constant randomisation probabilities within each stratum) can be used to eliminate time-trend bias. Subgroup analysis is something else, and typically involves splitting the patients into groups based on patient-characteristics, while calendar-time stratification is not based on patient-characteristics.

To clarify, we have amended this paragraph as follows:
‘It is often cautioned that calendar time trends – which are common with infectious diseases – may introduce bias when using response-adaptive randomisation. However, one can take advantage of the fact that randomisation probabilities are not constantly changing with most RAR designs. A calendar-time stratified analysis, with equal randomisation probabilities within each stratum, eliminates potential time-trend bias. A larger sample size is needed with such stratified analyses, but it is important to avoid trying to gain small improvements in efficiency at the cost of introducing bias.’

28. Pg 14, Line 279-280. It is unclear how this addresses calendar time trends.
Response: see response to comment 27.
29. Pg 15, Line 305. "The main reason for the increasing interest in adaptive trial designs may be that, under some circumstances,…"
Response: We have amended this sentence as suggested.

30. Pg 15, Line 312. clarify how subgroup analysis here helps prevent bias due to time trends.
Response: As explained in our response to comment 27 above, we are not referring to subgroup analysis here, but to calendar-time stratified analysis. As explained in more detail in reference 38, because randomisation probabilities are equal in each stratum, a calendar-time stratified analysis eliminates potential time-trend bias.

31. Pg 15, Line 315. Delete "would"
We have deleted the word ‘would’.