REVIEWER REPORTS

Thank you to the reviewers for taking the time to review our manuscript. Please find below our responses to each comment.

Reviewer Comments:

Reviewer 1
The authors solve the issue of partial ordering and how we can implement the trial design.

1. Recently, the use of model-assisted design, which is easier to implement, is expanding, but is it possible to apply PO to model-assisted design?

Theoretically, it should be possible to implement PO into a model-assisted design or at least produce a model-assisted approach that deals with this issue. However, to our knowledge, this has not yet been done. In terms of implementation from our perspective, it was easier to attempt to implement methodology that already existed versus having to develop new methodology and then implement that, even if ultimately that new methodology would be easier to implement. There is definitely scope to explore this as part of further research and it would be interesting to see how results from a model-assisted approach would compare.

2. The introduction of the design and the simulation settings were mixed up, making it difficult to understand the simulation settings. In particular, it was not clear to what doses were pre-defined and to what doses were unknown.

In the trial there are six dose levels (-1, 0, 1, 2a, 2b and 3). For dose level 3 there are two potential doses defined. The dose would be patients receive either 120mg daily of AZD6738 across 10 days or 80mg daily across 20 days. At the point in the trial where we reach dose level 3 a decision will be made by our independent safety committee based on the data accrued and observed in the trial as to what the specific dose of dose level 3 should be.

For the purposes of simulation this was treated as a singular dose. This assumed that the rate of toxicity would be the same at dose level 3 regardless of which specific dosing schedule was determined.

In section 2.2 of the text, we include a paragraph that explains this. "*In terms of dose level 3 only one of the doses in that tier will be investigated, it was unclear as to which dose level would be best due to a lack of historical data. The choice of dosing for this dose-level will be determined based on data observed throughout the trial. Even though dose level 3 is not yet specified in terms of modelling and simulations it was treated as singular dose. This was done as clinicians thought that it would be unlikely that we would reach these doses and that the probability of toxicity between them would be similar.* "

We added some clarifying text to the results section to specify the dose levels included in the simulations.

3. Can you compare PO-CRM and PO-TITE-CRM? If they are relatively out, how much faster would it be to add a TITE?

Due to the yearlong follow-up period we wished to observe patients for in ADePT-DDR it would be unfeasible to run a standard PO-CRM without fundamentally changing this aspect of the design. As such we did not initially consider this approach at the design stage.

Once the trial was set-up, as part of further research, we did explore how the design would perform without the TITE component. Generally speaking, the PO-CRM slightly outperformed the PO-TITE-CRM in terms of selecting the correct dose. However, that came at the cost of time. The average trial length was in the region of 11-16 years depending on the scenario. Comparatively, the PO-TITE-CRM averaged about 4-6 years in duration, which is still quite lengthy for a dose-finding study.

Perhaps more interestingly it would have been better to compare a PO-CRM design that only used a 12-week observation window. In this case, we hypothesise that the duration of the trial would be similar to that of the PO-TITE-CRM design with a yearlong follow-up but obviously, you would miss out on collecting longer-term toxicity data which would fundamentally change the questions the trial is trying to answer.

Essentially by adding the TITE component we over half the estimated duration of the study based on the simulations and scenarios that we ran.

Reviewer 2

The authors provided great contents with an application on a little-known method. Even though the partial order TITE CRM (PO TITE CRM) was proposed ten years ago by Wages et al., this method has not been applied yet and no code is available to implement it at this time.

The authors carried out extensive simulations to show how the PO TITE CRM perform with the setting of the ADePT-DDR trial.

Specific comments

1) Section 1 page 2 : Please rephrase the penultimate sentence "For example, either dose or treatment duration could be increased and even if patients receive an equal dose…"

It's strange to indicate an increase in dose and an equal dose. Perhaps keep only the treatment duration as an example. I think you want to transcribe the idea in table 1.

This sentence was trying to highlight that two doses could prescribe the same overall total dose but depending on the duration it's prescribed the toxicity rate could potentially differ which then leads to partial ordering. For example, 400mg of an IMP given over 5 days (80mg daily) could be more or less toxic than given over 20 days (20mg daily). It would depend on the treatment and disease area which would be more or less toxic but this is another way in which partial ordering may be observed.

We have rephrased the sentence so hopefully it is clearer now. "*For example, two doses could prescribe the same overall total dose but be over different treatment durations and hence have higher and lower daily doses. In this situation, it could be unclear as to whether prolonged exposure to a lower daily dose is more toxic than short exposure to a higher daily dose, which implies a partial ordering of toxicity probabilities.*"

2) Section 2.1 page 4 : You should introduce the Beta parameter in equation (1)? And discuss the choice of your model, have you tried other models like the logistic model with 1 or 2 parameters?

We only tried implementation using the power model. Other dose-toxicity models such as the one or two parameter logistic model could be implemented by replacing equation (3).

During the initial design stages of the trial other models were briefly considered. For a standard CRM the two-parameter logistic model has been shown to be better at estimating the dose-toxicity relationship compared to one parameter models. However, it's not clear if this will still be the case for a CRM in the presence of partial ordering, due to the extra complexity around the different orderings. Also, the main aim of the trial is to determine a TD25, which one parameter models have been shown to be better at estimating. In terms of selecting a one-parameter model as the original authors used the power model for both the PO-CRM and PO-TITE-CRM we felt this would be adequate for this trial and there would not be much benefit in using an alternative one-parameter model.

We have included some text to explain the Beta parameter and included some comments in the discussion regarding our choice of model.

3) Page 9: Can you check figure 2 again? I don't understand how your dotted lines are placed and the colors in the legend (green, blue and orange) don't match those on the figure.

The figure has been updated, and some text added to the manuscript to explain the dotted lines. The first dotted line is placed after 7 weeks when patients finish treatment. The second line is at 15 weeks (8 weeks post treatment) for the minimum follow-up period where patients are weighted at 60%. The third line shows when the weighting hits 80% at 19 weeks (12-weeks post treatment).

4) Section 2.3 page 8: You've made an interesting choice by estimating your weights with two linear functions, one between 8 and 12 weeks and the other between 12 and 52. Can you explain the choice of a smoother weighted function? In reality, the transition at 12 weeks is not so abrupt. Discuss the two slopes that are applied here and how certain polynomial or other functions can solve the problem or provide an alternative.

The weight function we have specified was motivated to a large extent by clinical input. It was decided, due to the nature of treatment patients were receiving, that an extend DLT period would be used. Based on clinical experience it was expected that most DLTs would occur in the first 3 months. This led to the decision of weighting patients at 80% who reached the 12-week mark without experiencing a DLT. This allowed for some room in case patients did eventually have a DLT.

The 8-week timepoint was introduced as we wanted to make dose decisions slightly earlier without having to wait for that full 12-week period. This would help speed up the conduct of the trial and the TITE methodology allowed us to implement this into the design. Discussions were then held around what weighting patients who reach the 8-week mark without DLT should be. If we were just to implement a standard linear function between the 12-week value of 80% and 52-week value of 100% and then extrapolate backwards the weight at the 8-week mark would have been 78%. This was felt to be too high as this window between 8- and 12-weeks post follow-up was felt most likely when any late-onset DLTs would occur. We settled on a weight of 60% for patients with 8 weeks of follow-up. Therefore, patients without a DLT at 8 weeks contribute less to the model ensuring that the subsequent dose recommendations would be more conservative. This was felt to be favourable over potentially overweighting patients and escalating to higher dose levels too quickly.

The reason we choose to implement these as two linear weight functions was due to its simplicity. It's arguably easier to understand how a patients weight increases over time throughout these two periods, 5% per week during the 8- and 12-week period and then 0.05% per week throughout the 12- and 52-week period. The other reason is due to the fact there is uncertainty around what the rate of increase in weight should actually be. If a smoother function were to be applied there may be certain time-points in which we are overestimating the weight that patients should be given. It may be easier to justify a different weight function if external data were available. However as this is a new combination of treatments there is limited to no data on the timing at which toxicities occur, especially over a period as long as this.

In terms of practical application, it is felt that the choice of weight function would have limited impact to actual decisions that are made. As this is a two-stage design, it may be the case that patient data has already been accrued before the model is invoked. We also plan to conduct sensitivity analyses for each dose recommendation based on how we calculate the time patients have spent in the trial. This may yield different weights for each patient and so we can see if the dose decision changes with different weights.

Overall, we feel this weight function should be adequate for the purposes of this trial. We have added some text to the discussion discussing our choice of weight functions.

5) I don't find in your paper a result on posterior probability associated with the m-parameter. The strength of the design lies in its ability to update the ordering as the data accumulates. It would be interesting to visuzalize those probabilities in addition to those on the correct choice of TD25.

The reason as to why results for m-parameter were not presented is due to comments from the original authors of the methodology discussed in their manuscript. They stated that the main aim of this design is to maintain accurate estimation of the MTD when the order of the treatments is only partially known. They also mention it is only of indirect interest to establish the correct order. It is explained that this is due to the possibility that the estimate of the MTD itself may remain unaffected by certain orders.

So, whilst the design is able to provide probabilities relating to the ordering it's not the primary focus of the design. Similarly, the main aim of the trial was to focus on selecting a TD25 not establishing which order was more likely. So, when we came to investigate the performance, we focused on evaluating selection probabilities of the correct dose and number of patients at each dose.

Nevertheless, it may be meaningful to evaluate how the design performs in terms of selecting the correct ordering and compare this to other designs that aim to establish dosing schedules.