**Reviewer 1 Major Comments**

**1. About the model, in Section 2.2, the log HR between the internal and external controls λICvEC shares the same parameters for the reference trial and the current trial. I think such assumption can be too strong and lack of flexibility. In the real data example (Table 2), the control treatments for different reference trials are difference. Even with the same treatment, the population of the subjects can be very different. So I think it is more reasonable to involve possible heterogeneity of the reference trial in the model. Or can the authors can provide any reasons why not take the heterogeneity into account.**

We agree with the reviewers comments that there will be heterogeneity between studies, and that this heterogeneity should be modelled. This is the role of the parameter σ2 introduced in equation (2). We have expanded the description of equation (2) to clarify this:

“*so that μ is a measure of the average bias in the external control arms across studies and σ2 is a measure of between study variability in the true log HR between the internal and external controls. σ2 allows for heterogeneity between the different situations studied by different studies, capturing variability due to features such as different control treatments or different study inclusion criteria*”

**2. In the real data application, it seems the control treatments are very different for the reference trials. I am curious what is the inclusion criteria of the reference trials?**

We have expanded Table 2 to include the patient populations for each of the reference studies.

**3. I think in practice, it can be difficult to find enough suitable historical trials for the current study, especially, when we require patient-level data. To estimate λˆ ICvsEC,j , it seems we need patient-level data because we should use the propensity score matching. Do the authors have any comments on this.**

We agree with the reviewer. We discuss this in a paragraph in the Discussion which we have expanded to discuss this further. The updated paragraph is:

“*A potential challenge is that there may not be a large set of suitable reference studies with patient level data, especially for rare diseases or especially novel treatments. Our simulations provide some guidance on the required number of studies and suggest that reasonable estimates can be obtained even when the number of reference studies is smaller. For instance, in cases where there were as few as 4 reference studies, estimates were unbiased and coverage probabilities were too high, implying that conclusions would be conservative. Put differently, our approach helps control type I error even when the number of reference studies is small because it adjusts for bias and results in estimates of the standard error that are conservative. Of course, increasing the number of reference studies is beneficial since it reduces the uncertainty of the estimates and increases power. An area for further research will be how wide can we cast the net for reference studies by including studies with different but related treatments or in different but related diseases. Relaxing the criteria for including studies will increase the number of reference studies available, but may be expected to increase the inter-study heterogeneity which would impact the power of treatment comparisons in the new study.*”

**4. About simulation studies, while the real data involve covariates, it seems the simulation studies do not have covariates. I think to better understand the method, it is more desirable to include covariates in the simulation.**

Whilst the real data analysis would typically involve covariates, these don’t play a part in the statistical properties of the estimates, which are based upon the well-established properties of the Cox Proportional Hazards model, and so we didn’t consider that including covariates into the simulations would add much value. Within this set up there are already a significant number of moving parts in the simulations and adding in covariates which could relate to outcome, treatment differences and study heterogeneity in a wide number of different ways could explode the number of scenarios we are investigating which has the potential to cause confusion without adding much value. As pointed out the example analysis of the advanced non-small cell lung cancer cohort of studies included covariates giving the readers an opportunity to view how the methodology behaves with covariates.

**Reviewer 1 Minor Comments**

**1. The notations used in Section 4.5 are a bit confusing.**

We thank the reviewer for their feedback. We have refined the writing of the opening two paragraphs of section 4.5, removing unnecessary references to *y* and , which we hope has made it clearer. The updated text is:

“*To check the model, we split the reference studies into training and test sets. The model was “trained" by performing the meta-analysis on the training reference studies and “tested” by predicting a treatment effect for the test study, assuming it was a single-arm trial without an internal control. Leave-one-out cross-validation was used to generalize performance, whereby reference studies were partitioned into training and test sets 14 times (with 13 of the 14 reference studies used for training and the remaining reference study used for testing), resulting in 14 separate predictions. To allow observation of the impact of the method we made both adjusted and unadjusted predictions. The adjusted predictions were given by the posterior median of λtestTRTvIC and the unadjusted predictions were the estimated log HRs from the direct comparison between the treatment and external control,ˆtestTRTvEC.*

*To assess the quality of the predictions, we computed standardised residuals for both the adjusted and unadjusted analyses by subtracting the actual log HR estimated from the trial,ˆλtestTRTvIC and dividing this difference by the standard deviation of λtestTRTvIC. QQ-plots of the standardized residuals of both the unadjusted and adjusted analyses against a theoretical normal distribution is displayed in Figure 7* ”

**2. Figure 1 illustrates the model clearly. But the colors between moderately dark and darkest nodes are too close. It can be better to make them more distinguishable.**

We thank the reviewer for the suggestion. We have made the darkest nodes in Figure 1 darker

**3. In “TRT”, why there is an additional space between T and RT throughout the manuscript.**

Thank you for pointing this out. This was due to a LaTex formatting issue which has now been resolved throughout the manuscript.

**Reviewer 2 Comments**

**1) On page 3 line 41, I don’t think the prospective single-arm trial has three arms. It only includes an oracle (latent) internal control arm. Please clarify this.**

To clarify this point we have added the following text:

“*although for the new study the internal control arm is a counterfactual which is not observed*”

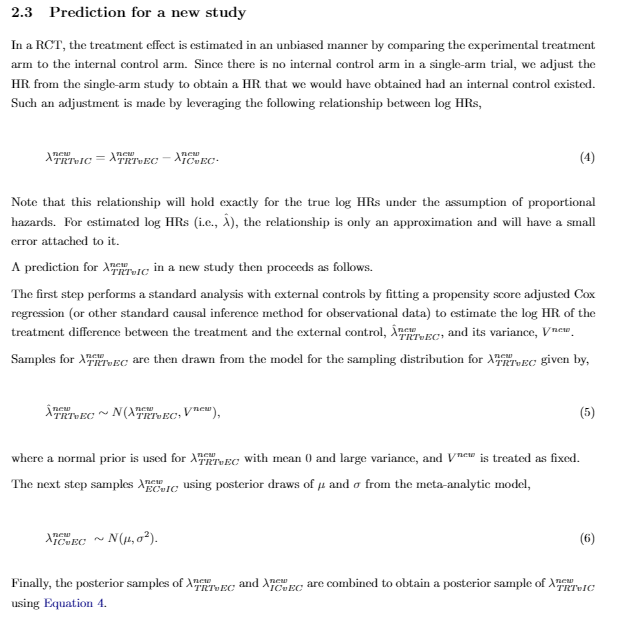
**2)      On page 5 line 44, the interpretation of the variable parameter \sigma^2 is relatively vague. Different reference studies have different control arms, this is a major issue the authors should put more clarification on. Mathematically, the model is correct, but whether such an approach is clinically meaningful should be discussed.**

We have expanded the description of *σ2* to clarify this:

“*so that μ is a measure of the average bias in the external control arms across studies and σ2 is a measure of between study variability in the true log HR between the internal and external controls. σ2 allows for heterogeneity between the different situations studied by different studies, capturing variability due to features such as different control treatments or different study inclusion criteria*”

**3)      On page 7, the description on lines 24-30 is relatively vague.**

We have updated the text to try and make the process easier to follow. Please see the new text below:



**4)      On page 7 line 33, the equation is not derived from the Bayesian perspective.**

That is correct. The standard analysis for time-to-event data within the pharmaceutical industry is a classical Cox proportional hazards model. We wanted to apply the same model, rather than changing to a Bayesian alternative, within our methodology to facilitate its acceptance and adoption as it is based around a standard, familiar and accepted analysis.

**5)      On page 19 line 32, it’s weird that the resulting type I error rate is below 0.025 based on the 95% credible interval. Theoretically, this should be around 0.05.**

This is a consequence of the somewhat confusing standard terminology used around one-sided and two-sided tests. Although a two-sided 5% significance level is typically quoted, we are generally only interested in a one-sided test (the novel treatment is an improvement on the control or standard of care), so we are only really interested in a one-sided test with type 1 error rate of 0.025.

**6)      The proposed approach should be compared to the naïve method to see the relative gain.**

The impact of the approach compared to the naïve method is made in section 4.5. We have updated the title of the section containing this to signpost it more clearly to the reader:

“*Model checking and Comparison to unadjusted comparison*”

**7)      On page 13, line 10, does the proposed method require patient-level data? If so, please clarify it at the beginning.**

We have clarified this point by adding in additional text both in the introductory section and in the methodology section.

“*To further increase comparability, if patient-level data is available then propensity score models can be developed for each reference study to match the external control to the randomised treatment arm.”*

“*We assume we have patient-level data available for all studies and for the external control arms, allowing us to use methods such as propensity scoring to account for differences in measured covariates between the study and external control populations*.”

**8)      Section 4.4, please include the investigational agents for the external control.**

The investigational and control treatments are all listed in Table 2.

**9)      The proposed approach assumes consistency in the indirect comparison. However, such a consistency assumption may not hold in many cases. This has been extensively studied in network meta-analysis. The authors may discuss the limitations of the proposed along such a direction.**

We thank the referee for pointing this out We have added the following text to address this point and give a direction for issues to consider for anyone implementing this methodology.

*“The methodology assumes that the relationship between the external control and the hypothetical internal control for a new study follows the same distribution as observed across the reference studies. Confidence that this assumption holds could be increased by confirming that the characteristics of the new study such as the patient population and the control treatment are compatible with the set of reference studies and that there is no evidence of any drift over time.”*