Author’s response to reviews

Title: Individual Patient Data Network Meta-Analysis using either Restricted Mean Survival Time Difference or Hazard Ratios: is there a difference? A case study on locoregionally advanced nasopharyngeal carcinomas

Authors:

Claire Petit (claire1.petit@gustaveroussy.fr)

Pierre Blanchard (pierre.blanchard@gustaveroussy.fr)

Jean-Pierre Pignon (jean-pierre.pignon@gustaveroussy.fr)

Béranger Lueza (berangerlueza@gmail.com)

Version: 1 Date: 18 Dec 2018

Author’s response to reviews:

SYSR-D-17-00398

Individual Patient Data Network Meta-Analysis using either Restricted Mean Survival Time Difference or Hazard Ratios: is there a difference? A case study on locoregionally advanced nasopharyngeal carcinomas.

Claire Petit, M.D; Pierre Blanchard; Jean-Pierre Pignon; Béranger Lueza

Systematic Reviews

Please include a point-by-point response within the 'Response to Reviewers' box in the submission system and highlight (with 'tracked changes'/coloured/underlines/highlighted text) all changes made when revising the manuscript. Please ensure you describe additional experiments that were carried out and include a detailed rebuttal of any criticisms or requested revisions that you disagreed with. Please also ensure that your revised manuscript conforms to the journal style, which can be found in the Submission Guidelines on the journal homepage.

Reply: First of all, the authors would like to thank the reviewers for the time dedicated to their careful review and for their comments.
Reviewer reports:

Reviewer #1:

The manuscript illustrates the impact of using restricted mean survival differences (rmstD) instead of hazard ratios (HR) in a network meta-analysis (NMA). The methodology presented is sound, the study well conducted, and the manuscript presents the first example of the use of rmstD for NMA, at least to my knowledge.

Major comments

1. Locoregional control was used as a secondary outcome, which adds the complexity of competing risks. In the analysis, distant failures as first events were treated as censored observations, which if fine to estimate HR: it resorts on computing a cause-specific hazard ratio. But the corresponding Kaplan-Meier estimator is biased for the probability of locoregional failure, because those with distant failure are considered as if they were still at similar risk of locoregional failure as patients who remained under follow-up. As a result, it is unclear what the area under the Kaplan-Meier curve estimates, as well as what lies under the definition of a RMST in that case. Extensions of RMST to competing risks exist (even in a recent BMC Medical Research Methodology paper by Calkins et al.) The authors acknowledge that competing risks were not considered, but this should at least be discussed. Likely, if there were only very few distant failures as first events, the impact of ignoring competing risks is negligible. But if not, then more appropriate analyses should be considered.

Reply: We agree that the method used to study competing risk was suboptimal, in particular in this study with 16% of the patients with distant failure. We have modified one sentence in the Methods section acknowledging that “Patients with a distant failure as a first event were censored for locoregional failure, thus not taking into account competing risks.”

We also added the following sentences in the discussion with the suggested reference at the end of the paragraph on limitation:

“For loco-regional control, we did not use competing risk model to take into account distant failure. However, in case of a non-negligible number of competing events, the corresponding Kaplan-Meier estimator is biased, and thus the RMST should be estimated under a competing
risk framework. For instance, one could use the methodology developed by Calkins et al who recently published an extension of RMST to competing risks.”

2. Two time horizon $t^*$ are used, 10 years for the primary analysis (with extrapolation if the last event occurred before 10 years) and 5 years as a sensitivity analysis. For HR analysis, however, the available follow-up was likely used. So for comparison both treatment effect measures are not exactly treated the same. For instance, HRs for follow-up restricted to 5 years could also be used, and could lead to different results, especially when hazards are not proportional. This issue could also be discussed further.

Reply: For sensitivity analysis at 5 years, we did use HR censored at 5 years as mentioned in the manuscript page 14: “one with another estimation of the rmstD at $t^*$= 5 years and compared with HR censored at 5 years”. We performed this sensitivity analysis to compare the two methods in the absence of extrapolation.

We have inserted the following sentence in the method section to better reflect this adjustment.

“Of note, in case of $t^*$ = 5 year we censored the follow-up for HR estimation at 5 years so that both treatment effect measures – rmstD and HR – are more comparable.”

3. The figure 1 is informative, but is not specific to the NMA. I would add a second panel with the p-scores for all treatments with HR and rmstD (or put the figure 1 as supplementary material and replace it by my suggestion).

Reply: Figure-1 provides a visual synthesis of the treatment effect estimated using both rmstD and HR measures. We respectfully believe, as reviewer 2, that this figure has a place in the paper and not in the supplementary material.

Concerning the p-scores, they are given in Table 3 for both rmstD($t^*$=10 years) and HR, for the three endpoints (OS, PFS, LRC). We sincerely apologized if we misunderstood the reviewer’s suggestion.

4. RmstD is not insensitive to non-proportional hazards (NPH), though it does not require such an assumption. In fact, if survival curves cross, for instance, the choice of the time horizon $t^*$
may be critical, and a treatment may offer a positive benefit in RMST at shorter follow-up times, but a negative one if $t^*$ was set at a longer time. This may be acknowledged, especially when discussing NPH for HR estimation.

Reply: Thank you for the comments. The following sentence was added in the discussion after the sentence “The most important requirement is to prespecify the time horizon(s) $t^*$ at the time of the study design in accordance to clinical interest[8].”:

“Indeed, if survival curves cross - one of the scenarios in which hazards are not proportional and HR estimation will be biased - the choice of the time horizon $t^*$ may be critical for the estimation of rmstD, as a treatment may offer a positive benefit in RMST at shorter follow-up times, but a negative one if $t^*$ was set at a longer time.“

Minor comments

1. Secondary outcome could be listed in the methods part of the Abstract, since they are (briefly) alluded to in the results.

Reply: Modification done: “secondary endpoints were progression-free survival and locoregional control” was added in the Patients and methods part of the abstract.

2. In the table 1, the term "trial comparison" is unclear. Do the author mean trial? Also, the second column should give the treatment comparison (currently last column). This would facilitate reading. I would also put the $p$-value for PH testing just after the confidence interval (CI and not IC) for the HR, since it directly related to that measure of treatment effect.

Reply: We used the term “trial comparison” and not only “trial” because of multi-arm trial: one trial is divided (QMH-95) and each comparison is individually considered.

The column “treatment comparison” was moved in second position.

The column “$p$-value test for non-proportionality” was moved after the column CI of the HR (and the inversion was done for CI).
3. "Significance change" could be changed to "change in significance". Similarly, page 11, "but changed the significance" would better be "but changed in significance".

Reply: Thank you for the advice, six modifications were done to “change in significance”.
Modification was done for “but changed in significance”.

4. The percentages given close to the curves on supplementary figure 2 are slightly misleading, except if one is really deep in the paper. Indeed, cumulative event rates could easily be mixed up with cumulative incidence, especially when the y-axis label also states "Cumulative event rate". I would suggest removing them except if this point is clarified and described more clearly than currently done.

Reply: We agree that these curves need to be clarified, but that the information provided is useful. The curves are the cumulative incidence, but the percentage on the curves the proportion of the total number of events observed at 5 years and 10 years. The title of supplementary figure 2 was changed: “Cumulative incidence curves for overall survival, progression free survival and loco-regional control events.” And in the legend of the figure, we changed the sentence “The percentages reported on the figure correspond to the percentage of events at 5 years and 10 years.” by “The percentages reported on the figure correspond to the proportion of the total number of events observed at 5 years and 10 years.” We also changed the y-axis label for “Cumulative incidence”. Finally we changed in the manuscript the sentence “The cumulative event rates at 10 years were 94.8%, 96.6% and 99.0% respectively for OS, PFS and LRC” by “The proportions of total number of events observed at 10 years were 94.8%, 96.6% and 99.0% respectively for OS, PFS and LRC”.

5. Still on this issue, page 16, it seems that Supplementary figure 2 should be referred to instead of Supplementary figure 4 in the sentence "The cumulative event rates at 10 years were 94.8%, 96.6% and 99.0% respectively for OS, PFS and LRC (Supplementary Figure 4)."

Reply: Modification done : “Supplementary Figure 2” (and not 4).
Reviewer #2:

I apologise for the delay in the submission of my review (due to illness).

Overall recommendation: Reject.

Two main reasons:

1) Limited original contribution: The article does present a better version of an existing method. This method was, however, not developed by the authors, but applied to a novel context (network-meta-analysis). Area metrics have been used as superior alternative to point-estimates for continuous time-series data such as event-related potentials for decades.

Reply: We agree that the originality of the paper is on its application of rmstD in a novel context.

2) Largely descriptive focus: To qualify as a methodology article, I think the article would need to move beyond its current descriptive focus. This would require a much more concise reporting of the current content to make space for an extensive exploration of why and when metrics differ. Otherwise I would suggest to submit the article to a different category, and as the Research category of Systematic Reviews does not quite seem appropriate, that may mean submission to a different journal.

Reply: We acknowledge that this work, though it is the first one to use the rmstD in network meta-analysis and to compare its results with the ones obtained using HR, has the limitations associated to a case study, limitations which are discussed in the paper. However, we respectfully believe that our paper meets to some extent the criteria defined by Systematic Review for Methodology paper. Please note that we would kindly approve the choice of the Research section, if deemed appropriate by the editor.
Comments

This is a well written, clear (albeit dense) article that addresses an alternative method of treatment benefit for time-to-event data (replacing a point estimate with an area estimate) important to the majority of clinical trials (e.g. licencing applications) and related health technology assessments (e.g. NICE approval in the UK). The methods employed are strong (e.g. prospectively registered protocol, clearly specified, logical choice of primary and secondary outcomes, but see comment re: choice of 10 and 5-year follow-up points). The original contribution is application context (network meta-analysis) rather than method per se (area measures are not new).

Figures and Tables (esp. Table 2 and Figure 1) provide nice, compact representation of results. I would move Supplementary Figure 2 to the main body of the publication.

Reply: We respectfully believe that Supplementary Figure 2 is a simple description of the data being used and do not reflect the context of rmstD or network meta-analysis. Even though it is useful to understand the results we would prefer to let it as supplementary data.

The article is good, but not excellent, because the focus is on *if* metrics differ (i.e. largely descriptive) with some consideration of *how* they differ, but only tangential consideration of *why* and *when* (inferential). As this is a methodology article, I would expect a detailed exploration of why and when the metrics differ (inferential), because it can reasonably be expected that any area metric will differ from point estimate. Methodologically, the key is to understand why metrics will differ and in what way under varying conditions, but the article barely touches upon these questions. The authors do make some important observations in passing (e.g. difference not solely attributable to non-proportional hazards or extrapolation), but largely relegate a detailed exploration of why and when to further research, namely simulation studies. Overall, the article seems to lean more towards estimation of clinically relevant outcomes than methodology development and it might be better to submit the article in a category reflecting this.

Reply: We agree that to answer the answer about *why* and *when* is the key question, but we felt that without simulation it is difficult to answer. The objective of this work was to point out
the potential usefulness of this (relatively) new method in NMA and to promote further research, including simulation studies.

Suggestions

One type of figure that would be essential to this article if submitted in the methodology section, but is entirely missing, are raw curves from which HR and rmstD(t*) are calculated, grouped by effect (i.e. no difference between metrics, difference in one direction (HR < rmstD(t*)), difference in other direction (HR < rmstD(t*)). I would de-emphasise significance (p values), especially given number of comparisons (suggest to consider FDR procedure to correct). I would consider relegating all but Table 2, Figure 1, and Supplementary Figure 2 to the appendix to make space for these figures if necessary.

Reply: We sincerely apologize, but because of our policy with the investigators of all the trial’s data we used, we are not able to provide raw curves for publication as individual patient data may be extracted from high quality survival curves. The complete set of curves is provided in the document “rawcurves.docx” as confidential documents for the reviewers only.

We agree that confidence interval is more important than p-value and significance. But as the two metrics studied are not directly comparable, we reported the significance as well as the confidence interval. We also agree that significance should be interpreted with caution.

We added the following sentence at the end of the paragraph on the limitations of the study: “Lastly, we focused on significance of the estimations as the two metrics studied are not directly comparable. However, to avoid too much emphasis on significance, especially given number of comparisons, confidence intervals should also be taken into account to interpret the results.”

Table 3 is a summary table of the NMA results for both HR and rmstD for all endpoints and provides p-score. We do believe that this table has its place in the main paper. Leaving the other tables and figures in the manuscript helps improve the clarity of the manuscript, as many readers would not read in detail the appendix, and we feel that this could be detrimental to the understanding of this research.

P. 7, line 20: Specify why and how a 10 and 5 year follow-up period was chosen in reference 15 (and therefore current article).
For the estimation of the rmstD(t*), we selected t*= 10 years for the primary analysis and t*= 5 years for a sensitivity analysis, as these were the two time points of clinical interest in the first update of MAC-NPC[15]. Because the median follow-up of trials was 7.4 years, the majority of the trials included in the NMA had a follow-up long enough using a t* of 10 years. When the latest event in a group of treatment occurred before t*, an extrapolation until t* was performed using the method proposed by Brown et al[16]. No trial needed extrapolation when the time horizon t* was set at 5 years. Of note, in case of t* = 5 years we censored the follow-up for HR estimation at 5 years so that both treatment effect measures – rmstD and HR – are more comparable.

Five years survival rates is the most frequent rate reported in clinical paper and in this particular case allow estimation of rmstD without extrapolation. The 10 years point was selected as long term effect is important and our data allowed estimating it. Moreover, the longer is the follow-up, higher is the probability to observe non proportional curves.

We added the following sentence at the end of the paragraph on the strengths of the study.
“Lastly, data on long term follow-up was available increasing the probability to observe non proportional hazard.”

As a very minor comment, it would be good to be consistent in decimal use. Currently, some p values are reported to 3 decimal places, most to 2, I would suggest to report all to 3 places.

Reply: We used 1 decimal places for rmstD as the result was in months to simplify the reading in the text. The results in tables had 2 decimal places so to be consistent in the decimal places used, we changed for 2 decimal places for all the results. Supplementary text was also modified with 2 decimal places.