Peer Review File

Article information: https://dx.doi.org/10.21037/jtd-21-1826.

Reviewer A

Comment: It is not mentioned that SBRT does not involve node assessment. Please mention briefly.

Reply: The issue of confounding by the degree of stage evaluation was considered extensively and is depicted in each of the tables as a domain of confounding (Hi Stage, defined in the table legends as: occult stage inaccuracy due to differences in extent of assessment). It is a complex issue – a bigger issue if p-stage for surgical patients is compared to c-stage for non-surgical, less so if c-stage is used throughout, less concern if PET is used extensively and if EBUS is used in non-surgical patients, a greater issue if the cohort includes larger tumors etc.) Thus, we think that we have extensively addressed this issue for each study in each table with an explanation in the legend.

Nevertheless, the reviewer's comment highlights that readers may miss the extent that potential confounding was considered, since it is described only briefly in the methods in the part 4 paper and many readers may not go to the part 2 paper they are referred to for more detail. In response, we have added a sentence describing the domains of potential confounding (which include discrepancies in stage classification and treatment) in the methods section of the part 4 paper.

Reviewer B

Comment 1: The authors should explain more clearly what "functional status" means in this context.

Reply 1: This is changed to functional capacity to be more clear, and in the results section it is explained that we used PFTs as a surrogate because no data on functional capacity was available.

Comment 2: How was the issue dealt with that the degree of stage evaluation is different for surgery and non-surgical therapies? Also adjuvant therapy.

Reply 2: The issue of confounding by the degree of stage evaluation was considered extensively and is depicted in each of the tables as a domain of confounding (Hi Stage, defined in the table legends as: occult stage inaccuracy due to differences in extent of assessment). This is also described in the Part 2 papers with details in the appendix (both referred to in the part 4 paper). It is a complex issue – a bigger issue if p-stage for surgical patients is compared to c-stage for non-surgical, less so if c-stage is used throughout, less concern if PET is used extensively and if EBUS is used in non-surgical patients, a greater issue if the cohort includes larger tumors etc.). Thus, we think that we have extensively addressed this issue for each study in each table with an explanation in the legend.

Similarly, the issue of discrepancies in adjuvant therapy was considered and is depicted in each of the tables as a domain of confounding (Q Treatmt, defined in the table legends as: quality of

the treatment [e.g. margin distance, adjuvant therapy]). The degree of risk of confounding depends on several factors which were taken into account when judging each study.

Nevertheless, the reviewer's comment highlights that readers may miss the extent that potential confounding was considered, since it is described only briefly in the methods and many readers may not go to the part 2 paper they are referred to for more detail. In response, we have added a sentence describing the domains of potential confounding (which include discrepancies in stage classification and treatment) in the methods section of the part 4 paper.

Comment 3: How was the issue of different time periods handled as a source of confounding.

Reply 3: The issue of discrepancies in time periods was considered and is depicted in each of the tables as a domain of confounding (Time Span, defined in the table legends as: adjustment for changes during the study period or differential use of the interventions).

Comment 4: Various stage groups were included in some studies - can you clarify is some received adj therapy.

Reply 4: The issue of confounding due to adjuvant therapy was considered extensively as noted above and considered potentially to be a greater issue in studies that included larger (higher stage) tumors, but less if the multivariate/propensity adjustment included adjuvant therapy. There is no concise way to represent in the text or a column in the tables how many patients received adjuvant therapy that also reflects that the implications are different depending on the number, the time period, whether this was adjusted for etc. We think the column depicting this domain of confounding represents the best approach.

Comment 5: It would be clinically more relevant to focus on studies including patients of at least 75 years and/or present the median age (with range) for the SBRT- and surgical group respectively – this may be a source of confounding.

Reply 5: In the section and tables on older patients we have presented the evidence available. If desired, the reader can focus only on the studies of patients >75 years of age. We have also organized the table to explore whether there is a signal of a greater or lesser disparity between interventions depending on the age cohort – a detail that is obscured by only reporting on a single age classifier (e.g. age >75).

Furthermore, each of the studies adjusted for the age (and often other demographic characteristics). We assessed and report in the tables the degree of potential residual confounding.

Comment 6: For "compromised patients" I would suggest to add a column with the number of patients having PS ≥ 2 for surgical patients and SBRT-patients respectively.

Reply 6: Adding a column would necessitate removing a different column. We had picked columns to include by assessing what data is available, relevant and most revealing. Only 1 of

the studies of compromised patients reported PS by group, and in this one the analysis was carried out after matching for PS – thus not a useful column to include.

General remarks:

Comment 7: Some of the above mentioned issues are addressed in the tables partly with colour coding, but given the complexity of the tables, the manuscript would improve by highlighting a few of these issues in the text as well.

Reply7: We have given this due consideration. However, there is not a concise way of doing this. We have devoted a great deal of effort to addressing and exposing sources of uncertainty. We have focused on RCTs and NRCs that adjusted for confounders. It is not really possible to talk about it in the text beyond what we have laid out in the methods, appendices and the tables themselves. The degree of potential residual confounding depends on aspects of each study – i.e. what was adjusted for, details of the patients and settings involved (time periods, PET, other staging studies, tumor size, stage) to mention a few. Going through this for any domains of confounding study be study is verbose and not conducive to getting an overall impression. Instead, we think interested readers are better served by looking carefully at the tables and the details of the methods. We fully acknowledge potential confounding and have built a methodologic structure that addresses it as best as possible.

Comment 8: Some of the tables could be improved by stipulating the number of patients who received SBRT and surgery separately.

Reply 8: We had gone through multiple iterations of the tables. We tried to include a lot of detail, but also to create a structure that was not overwhelming. We settled on a total number of patients as the best compromise. Citing each cohort separately would require deleting 1-2 other columns or a smaller font size. Note that, as explained in the legend, studies were required to have >50 patients per arm for inclusion, and that many of the number-of-patients entries involved matched pairs (i.e. and equal number of patients in each cohort).

Comment 9: The table of QoL: reference no 20 included 44% of patients with pulmonary metastases which is presented in the legend. Yet, it might be misleading to keep this reference since metastatic patients often will have progressive disease with new lung metastases and the decrease in QoL might be attributed to this circumstance rather than the SBRT-treatment.

Reply 9: True, and we have pointed this out with a footnote. However, there was little effect on QOL among the patients in this study (i.e. whether they had mets or NSCLC). Only dyspnea is worse at 24 months, but it is unlikely that this was related to whether the patient had NSCLC vs extrathoracic metastases.

Comment 10: Regarding the cumulative incidence on toxicity post SBRT; it should be commented that for these, generally medically compromised patients, it is difficult to certainly distinguish between treatment related toxicity and the natural course of underlying comorbidities such as COPD.

Reply 10: The text on QOL/PFTs in both the general results and compromised patients sections address what we know about the natural course of COPD and how it affects the interpretation of studies of complications.

Comment 11: The design of the tables needs some consideration since a) some references are wrongly spelled b) the headings of some of the columns are partly hidden.

Reply 11: Thank you for pointing this out. The submission system reformatted this and we didn't catch it.

Comment 12: Table 3: the arrows and colour coding are a bit difficult to understand. Does red always mean a deterioration or is it the direction of the arrows? Ref 20: dyspnea at 24 mo compared to ref 87: cogni at 1 mo. Which describes deterioration?

Reply 12: This is explained in the legend of the table and the footnote "a": "for symptoms ↑ indicates worse state (increased pain/dyspnea), ↓ indicates improvement" The color coding is also explained: red means a deterioration and green an improvement. In the case of general QOL measures the scale is such that a higher score represents an improvement; however in the symptom scales of QOL tools, a higher degree of symptoms has a negative impact on QOL, hence represented as red.

Comment 13: Result, line 185: The term "central" here needs to be clarified since ultracentral tumors are not as easily treated and carry a higher risk for high-grade toxicity.

Reply 13: It was defined in the nuances section but definition also added here.

Comment 14: Line 300 states that the HILUS-trial and SUNSET-trial treat patients with 8-18 fractions. The HILUS-trial had 8 fractions and the SUNSET-trial has three levels (8,10 and 15 fractions).

Reply 14: Error corrected.