Interobserver variability in GTV contouring in non-spine bone metastases

Carolina de la Pinta*, Raquel García LaTorre, Alberto Martínez-Lorca, Eva Fernández, Raul Hernanz, Mercedes Martín, Jose A. Domínguez, Teresa Muñóz, Elena Canales, Carmen Vallejo, Marina Alarza, Asunción Hervás, Manuel Garví, Vanesa Pino, Sonsoles Sancho

*Corresponding author
Carolina de la Pinta
Radiation Oncology Department. Ramón y Cajal University Hospital, IRYCIS, Alcalá University, 28034 Madrid, Spain.

Handling editor:
Michal Heger
Department of Pharmaceutics, Utrecht University, the Netherlands
Department of Pharmaceutics, Jiaxing University Medical College, Zhejiang, China

Review timeline:

Received: 22 June, 2022
Editorial decision: 24 July, 2022
Revision received: 20 August, 2022
Editorial decision: 20 August, 2022
Revision received: 6 September, 2022
Editorial decision: 8 September, 2022
Revision received: 9 September, 2022
Editorial decision: 9 September, 2022
Revision received: 11 September, 2022
Editorial decision, 14 September, 2022
Published online: 22 October, 2022

1st Editorial decision
24-Jul-2022

Ref.: Ms. No. JCTRes-D-22-00080
Interobserver variability in GTV contouring in non-spine bone metastases.
Journal of Clinical and Translational Research

Dear Ms De la Pinta,

Reviewers have now commented on your paper. You will see that they are advising that you revise your manuscript. If you are prepared to undertake the work required, I would be pleased to reconsider my decision.

For your guidance, reviewers' comments are appended below. Please note that, in addition to addressing the experts' points of critique, your manuscript must abide by our academic English level requirements as stipulated in the author guidelines of our website.

If you decide to revise the work, please submit a list of changes or a rebuttal against each point which is being raised when you submit the revised manuscript. Also, please ensure that the track changes function is switched on when implementing the revisions. This enables the reviewers to rapidly verify all changes made.

Your revision is due by Aug 23, 2022.

To submit a revision, go to https://www.editorialmanager.com/jctres/ and log in as an Author. You will see a menu item call Submission Needing Revision. You will find your submission record there.

Yours sincerely

Michal Heger
Editor-in-Chief
Journal of Clinical and Translational Research

Reviewers' comments:

Reviewer #1: Abstract

Page 1, Line 14: change "delineate" to "delineated"
Page 1, Lines 12 and 13: change "ten" to "10" since numerals should be used for numbers 10 and above. Please make this change wherever it applies throughout the paper.
Page 1, Line 23: It is unclear what the p-value is for. What statistical test was used and what was being compared? It seems the p-value is referencing the previous sentence and if so, it should be placed in the previous sentence. Please include what statistical test was used as well.
Page 1, Line

Introduction

Page 1, Line 38: change "Stereotactic Body Radiotherapy" tp "stereotactic body radiotherapy" as this word does not need to be capitalized.
Page 1, Line 49: change "(Computed Tomography)" to "(computed tomography)" as this word does not need to be capitalized.
Page 1, Line 51: change "(Magnetic Resonance Imaging)" to "(magnetic resonance imaging)" as this word does not need to be capitalized.
Page 1, Line 52: change "(18FluDesoxyGlucose Positron-Emission Tomography)" to "(18F-fluorodeoxyglucose positron emission tomography)" as it is misspelled and does not need to be capitalized.
Page 1, Line 55: change "delimitation" to "delineation" as it is misspelled.
Page 1, Line 57: change "(Gross Tumor Volume)" to "(gross tumor volume)" as this word does not need to be capitalized.

Material and methods

Page 2, Line 3: Specify how many of each image type. For example "Between January 2019 to December 2019, images of 7 patients with 10 bone metastases were included. In total, 10

CT, 10 MRI and 5 18FDG PET scans were included." From discussion section, it seems that there may be 5 MRI scans per patient. Please give more details of total number of scans and the breakdown numbers for each.
Page 2, paragraph 1: Specify what role, if any, the radiologists had in this study. As it was mentioned that "no oncologists received additional assistance from radiologists" the word "additional" makes it seem like there was assistance at some point beforehand.
Page 2, line 8: How experienced were the radiation oncologists? Please add average years of experience as a radiation oncologist with the minimum and maximum years.
Page 2, line 9: It is unclear what is meant by trial. Does a trial refer to a CT scan?
Page 2, paragraph 1: It is known that the choice of window/level can also affect how a observe contours objects in medical images. Since the window/level was not kept consistent, please include some details about this in the discussion about how this may impact a contouring study.

Page 2, lines 26 and 27: remove extra space between word and ® symbol in all instances. Also the ® symbol should be a superscript.
Page 2: In all instances of occurrence, specify what the "specialist" is is. Is this the radiation oncologist? If so, please use consistent words throughout to avoid confusion for reader.
Page 2: "delimitation" and "delimited" are generally not as commonly used. I would use the word "contour" or "delineate". However, I will leave this choice to the authors as it is a word preference.
Page 2, line 44: Please give more details as to what the subjective assessment survey is.

Page 2, Line 50: Contour "group" is referred to here but earlier the word "trial" was used (page 2, line 9). Could you add more clarification as to what a group or trial is. Were there repeat contours by the same radiation oncologists? Did each radiation oncologist contour each medical image scan once? Or multiple times? Or is a contour group for contours derived from a specific medial image type such as there is a contour group for the MRI-based contours and a group for the T-based contours. This is not immediately clear in the methods section. I am assuming if there are 10 CT scans, 10 MRI scans x 5 (5 different sequences), and 5 18F-FDG, then there are 10+(10x5)+5 which equals 65 medical image scans total. If each scan is contoured 3 times, then there are 65x3=195 contours total. I may be wrong, but please add more details to this section so reader knowns the total scans, the breakdown of each scan, number of contours for each scan and the total number of contours.

Page 2, Line 52: Add why the kappa statistic was used. For example "A kappa statistic was used to take into account change agreement." Also, add a reference to this sentence for the kappa statistic as it relates to interobserver reliability.
Page 2, Line 53: Add a reference to the sentence for the Landis and Koch interpretation. What was this statistic used to assess? Agreement within a group? The description in the sentence is vague as it only mentions that it "was used for assessment."
Page 2, Line 55: What differences are being referred to? I assume it is differences in GTV volumes. Please specify. Also are the differences in GTV volumes between interobservers for a given type of medical image? These details should be made more clear in the methods section.

Page 2, Line 58: "Kappa" does not need to be capitalized.

Results

Page 3, Line 1: Please list how many scans had the location of each metastases. Example: The location of bone metastases was pelvis (n=2), extremities (n=4) and calotte (n=3).
Page 3, Line 7: change "in ten metastases (p=0.25)." to "in the 10 cases, respectively (p=0.25)."
Page 3, Line 9 and 10: Unclear what "median index of agreement" as this is not mentioned in statistical analysis. Please specify or use consistent wording between what is proposed in statistical analysis section and what is explained in results section.
Page 3, Line 13: replace "respectively" with "on 18FDG PET"
Page 3, Line 17 and 18: Unclear what "median index of agreement" as this is not mentioned in statistical analysis. Please specify or use consistent wording between what is proposed in statistical analysis section and what is explained in results section.
Page 3, Line 22: change "panelists" to "observers" if this is correct. It is easier for reader when consistent words are used. Radiation oncologists and observers can be used throughout.
Page 3, Line 23: change "table 2" to "Table 2" as this should be capitalized.

Discussion

Page 3, Line 49: When discussing your results, please reference which figure this was shown so reader can quickly reference the data. For example, add "Table 1" to this following sentence, "Our study demonstrated a larger volume in MRI delineated lesions in most cases (Table 1) as was also shown in Prins' study."
Page 3, Line 54: Did all patients with MRI scans, have scans from 5 different MRI sequences? I would suggest as part of the analysis, to look at volume difference between the different MRI sequences as the premise of this paper is to understand if there is an optimal imaging test for GTV delineation. Already the literature shows that GTV delineation is more accurate than that of CT so what would be more interesting to see in this paper is if there is much differences in GTV delineations between different MRI sequences of the same patient, or if interobserver variability is more consistent in specific MRI sequences compared to others. Making a Table 4, which compares the metrics suggested in the statistical analysis for different MRI sequences across the patients would be of interest and add some new information to the literature.

Page 4, Line 16-18: In the introduction it was mentioned that in some instances 18F-FDG PET is a better tool for delineating the GTV. But in the discussion, you have pointed out that the highest interobserver variability is observed with 18F-FDG PET. Please elaborate more on what this means. These seem to be contradictory statements.
Page 4, Line 21: change "Clinical Target Volume" to "clinical target volume" as this does not need to be capitalized.

Limitations

This does not need its own section but rather should just be another paragraph in the discussion. Other limitations that should be addressed are the lack of using a consistent window level which can affect how one delineates structures in medical images.

Page 4, Line 34: "Gold" should not be capitalized.

Page 4, Line 35-36: I would remove this part of the sentence: "However, we do not believe that these shortcomings affect the main conclusion of this study that CT" and instead focus on your actual findings in terms of what they added to literature and what they supported already known findings.

Conclusions

Page 4, Line 42: remove the word "remarkable" and be more specific. The results showed that variances and standard deviations between observers in CT and MRI were minor and were also minor in CT, MRI and F18-FDG PET. If these results were minor, then it does not convince me that contouring variability is an urgent issue that needs to be addressed. Maybe reword some of the results, discussion and conclusion so that the overall story being told is more consistent.

Main suggestions:
- Using consistent terminology throughout
- The objective, results and conclusion are not in agreement. In your objective, you hypothesize that MRI and/or 18F-FDG PET will optimize the accuracy of tumor delineation, but in results you highlight that interobserver variability in MRI and 18F-FDG PET are minor. Then in the conclusion you say the interobserver variability was remarkable and that there is a need to reduce variability. If the interobserver variability is minor, then why would you need to set up a training platform to reduce this variability? Also, in your intro there is some focus on how it is not clear what the ideal MRI sequence is for imaging non-vertebral bone tumors but in your study design you only compare all of MRI with CT. It seems that you have enough data to look at interobserver variability for each MRI sequence and to have some discussion on if interobserver variability was better for a specific MRI sequence compared to another.
- Overall, the details in the paper are vague. It is unclear exactly how many medical scans there are, and how many contours total and contours per scan. It would also be wise to include information on the CT and MRI scans, such as the slice thickness and pixel spacing so that one can interpret how many voxels are in disagreement when there is a volume difference of 0.7cc. The statistical analysis lacks details. Also the statistical analysis mentions metrics such as the overlap ratio and kappa index which was never discussed in the results.
- While I have given a lot of critical feedback, I do think the authors have enough data to make a good manuscript. However, I do not think the authors have fully leveraged the data and I think the manuscript needs more details as well as clear objectives and conclusions that are supported by the data in the results/discussion.

Reviewer #2: **Summary Comments:**

- I have reviewed the article and do not feel that it is ready for publication as there are several flaws.
- In terms of scope, after reviewing the journal, I feel it does fit within the scope of the Journal of Clinical and Translational Research
- In terms of novelty, while the concept of this paper is not new (and closely mimics the ideas

behind Raman *et al.* 2018), I do feel that more data is needed in this field of study, which the authors would provide with their original article.
- However, at this point, I do not feel that the methods used support the objectives stated in the introduction, but this may also be due to the fact that the methods are very unclear on what they did. Specifically they make say, in somewhat a round about way, that they are trying to investigate the ideal imaging for delineation of non-vertebral bone tumours, specifically when evaluating inter-observer variability. They seem to wish to compare the impact of CT, MRI, and 18FDG PET on this metric.
- It is not clear from the methods how this was studied. The methods state that each set of images were studied independently (Page 3, Line 3) but then go on to state that the ROs had the fusions of MRI and FDG available to them when contouring. It is not clear how the study was actually set up. In addition, there were other issues in terms of inconsistencies in terminology and metrics used. This is only one example of clarity issues. Please see my point-by-point discussion below.
- I think it would benefit the authors to look at using additional metrics to quantify their data for easier comparison to the literature e.g., generalized conformity index. Currently they state that they used the Kappa Index but it is not clear how they modified this statistics for multiple observers.
- I also feel that basic information for reproducibility is missing from the methods e.g., the image acquisition protocol.
- There are also things included that have no relevance to the paper e.g., the mention of a subjective assessment survey of contour delineation. The conclusions stated also seem more of a future work than conclusion. I do not feel I got a sense for what the authors felt of their results and how they fit into the broader literature (i.e., do not make any strong conclusions on whether MRI or PET is helpful and should be used)

Point-by-Point comments
NOTE: this list is not comprehensive - the manuscript requires significant revisions and so correcting for grammer etc. will have to be done after that
NOTE: "L" refers to line number

- **Introduction**
- It should be more strongly stated in the introduction the impact of having good contours in SBRT type treatments (e.g., small margins = high chance of geographic miss)
- **Methods**
- L7 page 3 - since the same ROs did not contour throughout, there is additional variability added and less trust in the results.
- This is important when considering that intra-observer variability can be significant
- Why was intra-observer variability not reported
- The writers talk about using the Kappa statistic to measure inter-observer variability but then refer to using a correlation coefficient. These mean different things - it has been assumed that this was a mistake and the authors used Cohens but clarity needs to be made
- Cohen's kappa is typically used for 2 raters. There are extensions to include multiple raters including Fleiss' Kappa but it is not clear from the text how this was accounted for.
- No information was given regarding how the registration was performed, who performed it, etc. Was only 1 registration used for all ROs (it is assumed so)? This is important for repeatability
- Were the MRI and PET images taken in the same position as CT? Were they all acquired the same day?

- There is no information given on acquisition protocols for any of the image sets, as well as no information on the MRI sequences used (name, any suppression schemes, etc.)
- No indication of what threshold was used with PET which is important when discussing contouring with PET which is known to be very subjective to the window level used
- L18 page 3 "All sets of images were studied independently" - it is not clear what this means in the context of the other methods that state that ROs had all imaging information available to them when contouring.
- The only metrics reported were Cohen's kappa and volume. It would be beneficial to describe using other metrics that have been used in the literature such as Generalized conformity index (CI_gen), distance difference between center of mass etc. Using one of the studies they cite (Raman et al) would be good as a template (e.g., table 2 in Raman et al).
- A survey was mentioned in the methods but no results from it given. What was the purpose of the survey? What did it ask? At this point, extraneous information not reported on.
- With 10 cases, some of which are from the same patient, the validity of the t-test is questionable. Likely need to use non-parametric tests.
- Significant issues with what is said to be reported in the methodology and what is actually reported. Say that they report the "overlap ratio" (never defined) and the kappa index but in the results they
- 1. Never refer/report the overlap ratio
- 2. Never refer to the kappa index
- 3. Report the "correlation coefficient" and the "Index of agreement"
- I assume that there is some sort of consistency error but at this point cannot determine what they actually reported.
- **Results**
- Do not need to report both the variance and standard deviation - standard deviation is sufficient and variance only adds extraneous information
- Units should be reported for variance and standard deviation (assumed to be cc^2 and cc, respectively)
- It is sometimes unclear what is being reported - it seems that median and range is being shown in several areas but should be clarified.
- Why was the median index of agreement used? Why not mean?
- L12 page 4 - The sentence needs to be clarified - is this comparing the CT and FDG? Why does the author then go on to compare with MRI in the same paragraph. Should likely have a separate paragraph
- It is not clear what the purpose of putting in figure 1 is. The figure is small enough that they could likely include at least a few patients for readers to get a better sense of how the agreement looks (multipanel - see Raman et al reference)
- L25 page4 - What is meant that correlation coefficients were insignificant in most cases? Do they mean statistically not significant? Or just that they are low values?
- Table 1 - How can the correlation coefficient be 1 between 6 different observers for patients 5-b, 6, and essentially 1 for 5-a and 7? These seem unrealistic
- Table 2 - Why is there no correlation coefficient for patient 3?
- Table 1,2,3 - summary data should be given at the bottom in its own row.
- Figure 1 - What is the purple contour? The caption should better describe the image
- All MRI data is referred to as "MRI" - what happened to having T1 and T2 data? Why weren't these separately investigated as was hinted at in the introduction?
- **Discussion**
- First sentence should likely be an introduction point
- If previous studies showed diffusion weighted imaging was useful, why was this not

investigated in this trial?
- L49-59 Page 4 - this paragraph is difficult to understand and I am not sure what point they are trying to make
- Page 5 lines 1-15 (first paragraph on this page) - the authors show that their results differ from other studies but provide no possible reasons as to why this may be the case.
- L16-L19 page 5 - FDG paragraph needs to be expanded on what the authors feel the role of FDG is and what they plan to do in the future with it (potentially investigate better standardization of its use)
- L20 page 5 - CTVs are not to be used for compensating for interobserver variability as is implied - that is the role of the PTVs. No suggestions are given on what potential margins would be required (when margins are typically not used due to doses in SBRT being so high that the fall -off treats microscopic disease in many sites)
- **Conclusions**
- Conclusions paragraph does not actually conclude anything relevant to the study and seems to be a future work statement. Needs to be adjusted to match standard conclusions paragraphs.

Authors' response

Reviewers' comments:

Reviewer #1: Abstract

Page 1, Line 14: change "delineate" to "delineated". Changed according.
Page 1, Lines 12 and 13: change "ten" to "10" since numerals should be used for numbers 10 and above. Please make this change wherever it applies throughout the paper. Changed according.
Page 1, Line 23: It is unclear what the p-value is for. What statistical test was used and what was being compared? It seems the p-value is referencing the previous sentence and if so, it should be placed in the previous sentence. Please include what statistical test was used as well. Included in the statistics section, thank you

Introduction

Page 1, Line 38: change "Stereotactic Body Radiotherapy" tp "stereotactic body radiotherapy" as this word does not need to be capitalized. Changed according.
Page 1, Line 49: change "(Computed Tomography)" to "(computed tomography)" as this word does not need to be capitalized. Changed according.
Page 1, Line 51: change "(Magnetic Resonance Imaging)" to "(magnetic resonance imaging)" as this word does not need to be capitalized. Changed according.
Page 1, Line 52: change "(18FluDesoxyGlucose Positron-Emission Tomography)" to "(18F-fluorodeoxyglucose positron emission tomography)" as it is misspelled and does not need to be capitalized. Changed according.
Page 1, Line 55: change "delimitation" to "delineation" as it is misspelled.
Page 1, Line 57: change "(Gross Tumor Volume)" to "(gross tumor volume)" as this word does not need to be capitalized. Changed according.

Material and methods

Page 2, Line 3: Specify how many of each image type. For example "Between January 2019 to December 2019, images of 7 patients with 10 bone metastases were included. In total, 10 CT, 10 MRI and 5 18FDG PET scans were included." From discussion section, it seems that there may be 5 MRI scans per patient. Please give more details of total number of scans and the breakdown numbers for each. Changed according.
Page 2, paragraph 1: Specify what role, if any, the radiologists had in this study. As it was mentioned that "no oncologists received additional assistance from radiologists" the word "additional" makes it seem like there was assistance at some point beforehand. Changed according.
Page 2, line 8: How experienced were the radiation oncologists? Please add average years of experience as a radiation oncologist with the minimum and maximum years. Changed according.
Page 2, line 9: It is unclear what is meant by trial. Does a trial refer to a CT scan?—We refer a image study, we clarify this.
Page 2, paragraph 1: It is known that the choice of window/level can also affect how a observe contours objects in medical images. Since the window/level was not kept consistent, please include some details about this in the discussion about how this may impact a contouring study. Not included because it was not significant the panelists who changed the window, thank you.

Page 2, lines 26 and 27: remove extra space between word and ® symbol in all instances. Also the ® symbol should be a superscript. Changed according.
Page 2: In all instances of occurrence, specify what the "specialist" is is. Is this the radiation oncologist? If so, please use consistent words throughout to avoid confusion for reader. Changed according.
Page 2: "delimitation" and "delimited" are generally not as commonly used. I would use the word "contour" or "delineate". However, I will leave this choice to the authors as it is a word preference. Changed according.
Page 2, line 44: Please give more details as to what the subjective assessment survey is. Changed according.

Page 2, Line 50: Contour "group" is referred to here but earlier the word "trial" was used (page 2, line 9). Could you add more clarification as to what a group or trial is. Were there repeat contours by the same radiation oncologists? Did each radiation oncologist contour each medical image scan once? Or multiple times? Or is a contour group for contours derived from a specific medial image type such as there is a contour group for the MRI-based contours and a group for the T-based contours. This is not immediately clear in the methods section. I am assuming if there are 10 CT scans, 10 MRI scans x 5 (5 different sequences), and 5 18F-FDG, then there are 10+(10x5)+5 which equals 65 medical image scans total. If each scan is contoured 3 times, then there are 65x3=195 contours total. I may be wrong, but please add more details to this section so reader knowns the total scans, the breakdown of each scan, number of contours for each scan and the total number of contours. Changed according.

Page 2, Line 52: Add why the kappa statistic was used. For example "A kappa statistic was used to take into account change agreement." Also, add a reference to this sentence for the kappa statistic as it relates to interobserver reliability. Changed according.
Page 2, Line 53: Add a reference to the sentence for the Landis and Koch interpretation. What

was this statistic used to assess? Agreement within a group? The description in the sentence is vague as it only mentions that it "was used for assessment." Changed according.

Page 2, Line 55: What differences are being referred to? I assume it is differences in GTV volumes. Please specify. Also are the differences in GTV volumes between interobservers for a given type of medical image? These details should be made more clear in the methods section. Changed according.

Page 2, Line 58: "Kappa" does not need to be capitalized. Changed according.

Results

Page 3, Line 1: Please list how many scans had the location of each metastases. Example: The location of bone metastases was pelvis (n=2), extremities (n=4) and calotte (n=3). Changed according.

Page 3, Line 7: change "in ten metastases (p=0.25)." to "in the 10 cases, respectively (p=0.25)." Changed according.

Page 3, Line 9 and 10: Unclear what "median index of agreement" as this is not mentioned in statistical analysis. Please specify or use consistent wording between what is proposed in statistical analysis section and what is explained in results section. Clarified, thank you.

Page 3, Line 13: replace "respectively" with "on 18FDG PET". Changed according.

Page 3, Line 17 and 18: Unclear what "median index of agreement" as this is not mentioned in statistical analysis. Please specify or use consistent wording between what is proposed in statistical analysis section and what is explained in results section. Clarified, thank you.

Page 3, Line 22: change "panelists" to "observers" if this is correct. It is easier for reader when consistent words are used. Radiation oncologists and observers can be used throughout. Changed according.

Page 3, Line 23: change "table 2" to "Table 2" as this should be capitalized. Changed according.

Discussion

Page 3, Line 49: When discussing your results, please reference which figure this was shown so reader can quickly reference the data. For example, add "Table 1" to this following sentence, "Our study demonstrated a larger volume in MRI delineated lesions in most cases (Table 1) as was also shown in Prins' study." Changed according.

Page 3, Line 54: Did all patients with MRI scans, have scans from 5 different MRI sequences? I would suggest as part of the analysis, to look at volume difference between the different MRI sequences as the premise of this paper is to understand if there is an optimal imaging test for GTV delineation. Already the literature shows that GTV delineation is more accurate than that of CT so what would be more interesting to see in this paper is if there is much differences in GTV delineations between different MRI sequences of the same patient, or if interobserver variability is more consistent in specific MRI sequences compared to others. Making a Table 4, which compares the metrics suggested in the statistical analysis for different MRI sequences across the patients would be of interest and add some new information to the literature. A table specifying the MRI sequences is included.

Page 4, Line 16-18: In the introduction it was mentioned that in some instances 18F-FDG

PET is a better tool for delineating the GTV. But in the discussion, you have pointed out that the highest interobserver variability is observed with 18F-FDG PET. Please elaborate more on what this means. These seem to be contradictory statements. Clarified according.
Page 4, Line 21: change "Clinical Target Volume" to "clinical target volume" as this does not need to be capitalized. Changed according.

Limitations

This does not need its own section but rather should just be another paragraph in the discussion. Changed according. Other limitations that should be addressed are the lack of using a consistent window level which can affect how one delineates structures in medical images. Changed according.

Page 4, Line 34: "Gold" should not be capitalized. Changed according.
Page 4, Line 35-36: I would remove this part of the sentence: "However, we do not believe that these shortcomings affect the main conclusion of this study that CT" and instead focus on your actual findings in terms of what they added to literature and what they supported already known findings. Changed according.

Conclusions

Page 4, Line 42: remove the word "remarkable" and be more specific. Changed according.
The results showed that variances and standard deviations between observers in CT and MRI were minor and were also minor in CT, MRI and F18-FDG PET. If these results were minor, then it does not convince me that contouring variability is an urgent issue that needs to be addressed. Maybe reword some of the results, discussion and conclusion so that the overall story being told is more consistent. Changed according.

Main suggestions:
- Using consistent terminology throughout. Changed according.
- The objective, results and conclusion are not in agreement. In your objective, you hypothesize that MRI and/or 18F-FDG PET will optimize the accuracy of tumor delineation, but in results you highlight that interobserver variability in MRI and 18F-FDG PET are minor. Then in the conclusion you say the interobserver variability was remarkable and that there is a need to reduce variability. If the interobserver variability is minor, then why would you need to set up a training platform to reduce this variability? Also, in your intro there is some focus on how it is not clear what the ideal MRI sequence is for imaging non-vertebral bone tumors but in your study design you only compare all of MRI with CT. It seems that you have enough data to look at interobserver variability for each MRI sequence and to have some discussion on if interobserver variability was better for a specific MRI sequence compared to another. Clarified, thank you. We included table 1 with different sequences of MRI.
- Overall, the details in the paper are vague. It is unclear exactly how many medical scans there are, and how many contours total and contours per scan. It would also be wise to include information on the CT and MRI scans, such as the slice thickness and pixel spacing so that one can interpret how many voxels are in disagreement when there is a volume difference of 0.7cc. The statistical analysis lacks details. Also the statistical analysis mentions metrics such

as the overlap ratio and kappa index which was never discussed in the results.
Changed according.

- While I have given a lot of critical feedback, I do think the authors have enough data to make a good manuscript. However, I do not think the authors have fully leveraged the data and I think the manuscript needs more details as well as clear objectives and conclusions that are supported by the data in the results/discussion. Changed according, thank you.

Reviewer #2: **Summary Comments:**

- I have reviewed the article and do not feel that it is ready for publication as there are several flaws. Thank you for your review
- In terms of scope, after reviewing the journal, I feel it does fit within the scope of the Journal of Clinical and Translational Research. Thank you for your comment.
- In terms of novelty, while the concept of this paper is not new (and closely mimics the ideas behind Raman *et al.* 2018), I do feel that more data is needed in this field of study, which the authors would provide with their original article. Thank you for your comment.
- However, at this point, I do not feel that the methods used support the objectives stated in the introduction, but this may also be due to the fact that the methods are very unclear on what they did. Specifically they make say, in somewhat a round about way, that they are trying to investigate the ideal imaging for delineation of non-vertebral bone tumours, specifically when evaluating inter-observer variability. They seem to wish to compare the impact of CT, MRI, and 18FDG PET on this metric. Changed according.
- It is not clear from the methods how this was studied. The methods state that each set of images were studied independently (Page 3, Line 3) but then go on to state that the ROs had the fusions of MRI and FDG available to them when contouring. It is not clear how the study was actually set up. In addition, there were other issues in terms of inconsistencies in terminology and metrics used. This is only one example of clarity issues. Please see my point-by-point discussion below. Changed according.
- I think it would benefit the authors to look at using additional metrics to quantify their data for easier comparison to the literature e.g., generalized conformity index. Currently they state that they used the Kappa Index but it is not clear how they modified this statistics for multiple observers. Clarified according.
- I also feel that basic information for reproducibility is missing from the methods e.g., the image acquisition protocol. Clarified according.
- There are also things included that have no relevance to the paper e.g., the mention of a subjective assessment survey of contour delineation. The conclusions stated also seem more of a future work than conclusion. I do not feel I got a sense for what the authors felt of their results and how they fit into the broader literature (i.e., do not make any strong conclusions on whether MRI or PET is helpful and should be used) Changed according.

Point-by-Point comments
NOTE: this list is not comprehensive - the manuscript requires significant revisions and so correcting for grammer etc. will have to be done after that
NOTE: "L" refers to line number

- **Introduction**

- It should be more strongly stated in the introduction the impact of having good contours in SBRT type treatments (e.g., small margins = high chance of geographic miss) Changed according.
- **Methods**
- L7 page 3 - since the same ROs did not contour throughout, there is additional variability added and less trust in the results. This variability has been taken into account, thank you.
- This is important when considering that intra-observer variability can be significant. This variability has been taken into account, thank you.
- Why was intra-observer variability not reported. This is a secondary objective of the study
- The writers talk about using the Kappa statistic to measure inter-observer variability but then refer to using a correlation coefficient. These mean different things - it has been assumed that this was a mistake and the authors used Cohens but clarity needs to be made. Correct, we used cohen, corrected accordingly
- Cohen's kappa is typically used for 2 raters. There are extensions to include multiple raters including Fleiss' Kappa but it is not clear from the text how this was accounted for. Clarified in the manuscript.
- No information was given regarding how the registration was performed, who performed it, etc. Was only 1 registration used for all ROs (it is assumed so)? This is important for repeatability. Clarified in the manuscript. "Image fusion was performed with rigid registration, available in the planner, the same for all oncologist"
- Were the MRI and PET images taken in the same position as CT? Were they all acquired the same day? Clarified in the manuscript
- There is no information given on acquisition protocols for any of the image sets, as well as no information on the MRI sequences used (name, any suppression schemes, etc.) Changed according.
- No indication of what threshold was used with PET which is important when discussing contouring with PET which is known to be very subjective to the window level used. Changed according.
- L18 page 3 "All sets of images were studied independently" - it is not clear what this means in the context of the other methods that state that ROs had all imaging information available to them when contouring. Clarified in the manuscript
- The only metrics reported were Cohen's kappa and volume. It would be beneficial to describe using other metrics that have been used in the literature such as Generalized conformity index (CI_gen), distance difference between center of mass etc. Using one of the studies they cite (Raman et al) would be good as a template (e.g., table 2 in Raman et al). The most interesting and significant data were included, thank you
- A survey was mentioned in the methods but no results from it given. What was the purpose of the survey? What did it ask? At this point, extraneous information not reported on. Has finally been eliminated
- With 10 cases, some of which are from the same patient, the validity of the t-test is questionable. Likely need to use non-parametric tests. The most interesting and significant data were included, thank you
- Significant issues with what is said to be reported in the methodology and what is actually reported. Say that they report the "overlap ratio" (never defined) and the kappa index but in the results they
- 1. Never refer/report the overlap ratio
- 2. Never refer to the kappa index

- 3. Report the "correlation coefficient" and the "Index of agreement"
- I assume that there is some sort of consistency error but at this point cannot determine what they actually reported.

Changed accoding
- **Results**
- Do not need to report both the variance and standard deviation - standard deviation is sufficient and variance only adds extraneous information. Changed accoding
- Units should be reported for variance and standard deviation (assumed to be cc^2 and cc, respectively). Changed accoding
- It is sometimes unclear what is being reported - it seems that median and range is being shown in several areas but should be clarified. Changed accoding
- Why was the median index of agreement used? Why not mean? The median is generally used to return the central tendency in the case of skewed numerical distributions. The median provides the typical value even when the data set is skewed to one side or the other.
- L12 page 4 - The sentence needs to be clarified - is this comparing the CT and FDG? Why does the author then go on to compare with MRI in the same paragraph. Should likely have a separate paragraph. Changed accoding
- It is not clear what the purpose of putting in figure 1 is. The figure is small enough that they could likely include at least a few patients for readers to get a better sense of how the agreement looks (multipanel - see Raman et al reference). Deleted, we refer only to figure 2
- L25 page4 - What is meant that correlation coefficients were insignificant in most cases? Do they mean statistically not significant? Or just that they are low values? They are low values
- Table 1 - How can the correlation coefficient be 1 between 6 different observers for patients 5-b, 6, and essentially 1 for 5-a and 7? These seem unrealistic. Revised and corrected, thank you
- Table 2 - Why is there no correlation coefficient for patient 3? Revised and corrected, thank you
- Table 1,2,3 - summary data should be given at the bottom in its own row. I do not understand your comment, could you help me?
- Figure 1 - What is the purple contour? The caption should better describe the image. Deleted

- All MRI data is referred to as "MRI" - what happened to having T1 and T2 data? Why weren't these separately investigated as was hinted at in the introduction? Included in table 1
- **Discussion**
- First sentence should likely be an introduction point. Changed accoding
- If previous studies showed diffusion weighted imaging was useful, why was this not investigated in this trial? If it was done, the data are included
- L49-59 Page 4 - this paragraph is difficult to understand and I am not sure what point they are trying to make Clarified accoding
- Page 5 lines 1-15 (first paragraph on this page) - the authors show that their results differ from other studies but provide no possible reasons as to why this may be the case. Changed accoding
- L16-L19 page 5 - FDG paragraph needs to be expanded on what the authors feel the role of FDG is and what they plan to do in the future with it (potentially investigate better standardization of its use) Changed accoding
- L20 page 5 - CTVs are not to be used for compensating for interobserver variability as is

implied - that is the role of the PTVs. No suggestions are given on what potential margins would be required (when margins are typically not used due to doses in SBRT being so high that the fall -off treats microscopic disease in many sites)
Clarified accoding
- **Conclusions**
- Conclusions paragraph does not actually conclude anything relevant to the study and seems to be a future work statement. Needs to be adjusted to match standard conclusions paragraphs.
Clarified accoding

2nd Editorial decision
20-Aug-2022

Ref.: Ms. No. JCTRes-D-22-00080R1
Interobserver variability in GTV contouring in non-spine bone metastases.
Journal of Clinical and Translational Research

Dear Ms De la Pinta,

Reviewers have now commented on your paper. You will see that they are advising that you revise your manuscript. If you are prepared to undertake the work required, I would be pleased to reconsider my decision.

For your guidance, reviewers' comments are appended below.

If you decide to revise the work, please submit a list of changes or a rebuttal against each point which is being raised when you submit the revised manuscript. Also, please ensure that the track changes function is switched on when implementing the revisions. This enables the reviewers to rapidly verify all changes made.

Your revision is due by Sep 19, 2022.

To submit a revision, go to https://www.editorialmanager.com/jctres/ and log in as an Author. You will see a menu item call Submission Needing Revision. You will find your submission record there.

Yours sincerely

Michal Heger
Editor-in-Chief
Journal of Clinical and Translational Research

Reviewers' comments:

Dear authors,

Thank you for submitting a revised draft.

I have gone through the draft and concluded that the paper has been significantly improved. Thanks for the work; the paper is important and useful.

However, and even in line with tradition, there are still many linguistic inconsistencies that must be resolved before we can procced with re-review of the paper. Moreover, your point-by-point response is very brief and forces the reviewers and editor to sort out and analyze what changes were made and how they sync up with the reviewers' comments. Please be more elaborate for the points that are fundamental (so not spelling corrections) so that the reviewers know from the rebuttal how their comments were addressed.

Several individual examples of linguistic inconsistencies and errors, which should be extended towards the rest of the manuscript, include:

- "The safety and efficacy of SBRT treatment requires precise localization of the GTV, ensuring local control and reducing the volume of healthy tissue irradiated" is not proper English. The second part of the sentence should read '..., ensuring local control and limiting the irradiation of healthy tissue.'

- Please remove all registered trademark symbols from the text.

- "...the same for all oncologist." The term 'oncologist' should be plural.

- "Radiation oncologist delineate each image only one time" should read 'The radiation oncologist delineated each image only once.'

- "In PET images we used to define tumor SUV more than 5." This is not a complete sentence.

I could go on and on, but this is enough to prove my point regarding spelling and sentence structure.

Furthermore:

- Table 1 locations should be written in English. Craneo, calota are not English.

- Please ensure the the use of the same number of decimals in each value. If the SD or variance is so large that it does not justify the use of 2 decimal points, please round off to a single decimal point.

Thank you for taking care of these.

Kindest regards,

Michal Heger
Editor

Authors’ response

Reviewers' comments:

Reviewer #1: Abstract

Page 1, Line 14: change "delineate" to "delineated". Changed according.
Page 1, Lines 12 and 13: change "ten" to "10" since numerals should be used for numbers 10 and above. Please make this change wherever it applies throughout the paper. Changed according.
Page 1, Line 23: It is unclear what the p-value is for. What statistical test was used and what was being compared? It seems the p-value is referencing the previous sentence and if so, it should be placed in the previous sentence. Please include what statistical test was used as well. Included in the statistics section, thank you. “GTV volumes in cc were calculated for each contour group for each patient including minimum, maximum and median volumes, variance and standard deviation. The correlation coefficient gave information on quantifying the strength of the linear relationship between two variables in a correlation analysis. The kappa index was calculated as the ratio of the intersection of the contours delineated for a given observer with the corresponding contours of another observer and their mean. A kappa statistic including Fleiss' Kappa was used to account for multiple observers (10). The Landis and Koch interpretation was used for assessment agreement within each group of bone metastasis images. GTV volumes comparisons between groups were performed with the T-student test. Differences were considered significant when $p< 0.05$.

To establish the ideal test in each of the cases, the variability index obtained in each of the tests, the correlation coeficient and the kappa index were evaluated.”

Introduction

Page 1, Line 38: change "Stereotactic Body Radiotherapy" tp "stereotactic body radiotherapy" as this word does not need to be capitalized. Changed according.
Page 1, Line 49: change "(Computed Tomography)" to "(computed tomography)" as this word does not need to be capitalized. Changed according.
Page 1, Line 51: change "(Magnetic Resonance Imaging)" to "(magnetic resonance imaging)" as this word does not need to be capitalized. Changed according.
Page 1, Line 52: change "(18FluDesoxyGlucose Positron-Emission Tomography)" to "(18F-fluorodeoxyglucose positron emission tomography)" as it is misspelled and does not need to be capitalized. Changed according.
Page 1, Line 55: change "delimitation" to "delineation" as it is misspelled.
Page 1, Line 57: change "(Gross Tumor Volume)" to "(gross tumor volume)" as this word does not need to be capitalized. Changed according.

Material and methods

Page 2, Line 3: Specify how many of each image type. For example "Between January 2019 to December 2019, images of 7 patients with 10 bone metastases were included. In total, 10 CT, 10 MRI and 5 18FDG PET scans were included." From discussion section, it seems that there may be 5 MRI scans per patient. Please give more details of total number of scans and the breakdown numbers for each. Changed according. “In total, 10 CT, 10 MRI and 5 18FDG PET scans were included”

Page 2, paragraph 1: Specify what role, if any, the radiologists had in this study. As it was mentioned that "no oncologists received additional assistance from radiologists" the word

"additional" makes it seem like there was assistance at some point beforehand. Changed according, deleted.
Page 2, line 8: How experienced were the radiation oncologists? Please add average years of experience as a radiation oncologist with the minimum and maximum years. Changed according.
Page 2, line 9: It is unclear what is meant by trial. Does a trial refer to a CT scan?—We refer a image study, we clarify this. “Delineation was performed by 6 experienced radiation oncologists (mean 8.6 years (5-13 years)”
Page 2, paragraph 1: It is known that the choice of window/level can also affect how a observe contours objects in medical images. Since the window/level was not kept consistent, please include some details about this in the discussion about how this may impact a contouring study. Not included because it was not significant the panelists who changed the window, thank you.

Page 2, lines 26 and 27: remove extra space between word and ® symbol in all instances. Also the ® symbol should be a superscript. Changed according.
Page 2: In all instances of occurrence, specify what the "specialist" is is. Is this the radiation oncologist? If so, please use consistent words throughout to avoid confusion for reader. Changed according. “Radiation oncologist”.

Page 2: "delimitation" and "delimited" are generally not as commonly used. I would use the word "contour" or "delineate". However, I will leave this choice to the authors as it is a word preference. Changed according.

Page 2, line 44: Please give more details as to what the subjective assessment survey is. Changed according. We deleted this information.

Page 2, Line 50: Contour "group" is referred to here but earlier the word "trial" was used (page 2, line 9). Could you add more clarification as to what a group or trial is. Were there repeat contours by the same radiation oncologists? Did each radiation oncologist contour each medical image scan once? Or multiple times? Or is a contour group for contours derived from a specific medial image type such as there is a contour group for the MRI-based contours and a group for the T-based contours. This is not immediately clear in the methods section. I am assuming if there are 10 CT scans, 10 MRI scans x 5 (5 different sequences), and 5 18F-FDG, then there are 10+(10x5)+5 which equals 65 medical image scans total. If each scan is contoured 3 times, then there are 65x3=195 contours total. I may be wrong, but please add more details to this section so reader knowns the total scans, the breakdown of each scan, number of contours for each scan and the total number of contours. Changed according. “The radiation oncologist delineated each image only once”

Page 2, Line 52: Add why the kappa statistic was used. For example "A kappa statistic was used to take into account change agreement." Also, add a reference to this sentence for the kappa statistic as it relates to interobserver reliability. Changed according. “The kappa index was calculated as the ratio of the intersection of the contours delineated for a given observer with the corresponding contours of another observer and their mean. A kappa statistic including Fleiss' Kappa was used to account for multiple observers (10). The Landis and Koch interpretation was used for assessment agreement within each group of bone metastasis images”
Page 2, Line 53: Add a reference to the sentence for the Landis and Koch interpretation. What

was this statistic used to assess? Agreement within a group? The description in the sentence is vague as it only mentions that it "was used for assessment." Changed according. “The kappa index was calculated as the ratio of the intersection of the contours delineated for a given observer with the corresponding contours of another observer and their mean. A kappa statistic including Fleiss' Kappa was used to account for multiple observers (10). The Landis and Koch interpretation was used for assessment agreement within each group of bone metastasis images” Page 2, Line 55: What differences are being referred to? I assume it is differences in GTV volumes. Please specify. Also are the differences in GTV volumes between interobservers for a given type of medical image? These details should be made more clear in the methods section. Changed according. “GTV volumes in cc were calculated for each contour group for each patient including minimum, maximum and median volumes, variance and standard deviation.” Page 2, Line 58: "Kappa" does not need to be capitalized. Changed according.

Results

Page 3, Line 1: Please list how many scans had the location of each metastases. Example: The location of bone metastases was pelvis (n=2), extremities (n=4) and calotte (n=3). Changed according. “In total, 10 CT, 10 MRI and 5 18FDG PET scans were included” Page 3, Line 7: change "in ten metastases (p=0.25)." to "in the 10 cases, respectively (p=0.25)." Changed according. Page 3, Line 9 and 10: Unclear what "median index of agreement" as this is not mentioned in statistical analysis. Please specify or use consistent wording between what is proposed in statistical analysis section and what is explained in results section. Clarified, thank you. “GTV volumes in cc were calculated for each contour group for each patient including minimum, maximum and median volumes, variance and standard deviation. The correlation coefficient gave information on quantifying the strength of the linear relationship between two variables in a correlation analysis. The kappa index was calculated as the ratio of the intersection of the contours delineated for a given observer with the corresponding contours of another observer and their mean. A kappa statistic including Fleiss' Kappa was used to account for multiple observers (10). The Landis and Koch interpretation was used for assessment agreement within each group of bone metastasis images. GTV volumes comparisons between groups were performed with the T-student test. Differences were considered significant when p< 0.05.

To establish the ideal test in each of the cases, the variability index obtained in each of the tests, the correlation coeficient and the kappa index were evaluated.”

Page 3, Line 13: replace "respectively" with "on 18FDG PET". Changed according. Page 3, Line 17 and 18: Unclear what "median index of agreement" as this is not mentioned in statistical analysis. Please specify or use consistent wording between what is proposed in statistical analysis section and what is explained in results section. Clarified, thank you. We deleted this information. Page 3, Line 22: change "panelists" to "observers" if this is correct. It is easier for reader when consistent words are used. Radiation oncologists and observers can be used throughout. Changed according. Page 3, Line 23: change "table 2" to "Table 2" as this should be capitalized. Changed according.

Discussion

Page 3, Line 49: When discussing your results, please reference which figure this was shown so reader can quickly reference the data. For example, add "Table 1" to this following sentence, "Our study demonstrated a larger volume in MRI delineated lesions in most cases (Table 1) as was also shown in Prins' study." Changed according.
Page 3, Line 54: Did all patients with MRI scans, have scans from 5 different MRI sequences? I would suggest as part of the analysis, to look at volume difference between the different MRI sequences as the premise of this paper is to understand if there is an optimal imaging test for GTV delineation. Already the literature shows that GTV delineation is more accurate than that of CT so what would be more interesting to see in this paper is if there is much differences in GTV delineations between different MRI sequences of the same patient, or if interobserver variability is more consistent in specific MRI sequences compared to others. Making a Table 4, which compares the metrics suggested in the statistical analysis for different MRI sequences across the patients would be of interest and add some new information to the literature. A table specifying the MRI sequences is included.
Page 4, Line 16-18: In the introduction it was mentioned that in some instances 18F-FDG PET is a better tool for delineating the GTV. But in the discussion, you have pointed out that the highest interobserver variability is observed with 18F-FDG PET. Please elaborate more on what this means. These seem to be contradictory statements. Clarified according. “Some studies have studied the role of 18FDG PET in bone lesion delineation (17). The threshold SUV level and accuracy of this technique for lesion delineation is not defined. In our study it was the test with the highest interobserver variability and lowest correlation coefficient probably due to these problems”
Page 4, Line 21: change "Clinical Target Volume" to "clinical target volume" as this does not need to be capitalized. Changed according.

Limitations

This does not need its own section but rather should just be another paragraph in the discussion. Changed according. Other limitations that should be addressed are the lack of using a consistent window level which can affect how one delineates structures in medical images. Changed according. “The study has several limitations. The number of patients and the absence of a gold standard showing the true extent of the tumor. Another limitation is the use of different MRI sequences and the possibility of changing the window”

Page 4, Line 34: "Gold" should not be capitalized. Changed according.
Page 4, Line 35-36: I would remove this part of the sentence: "However, we do not believe that these shortcomings affect the main conclusion of this study that CT" and instead focus on your actual findings in terms of what they added to literature and what they supported already known findings. Changed according.

Conclusions

Page 4, Line 42: remove the word "remarkable" and be more specific. Changed according. The results showed that variances and standard deviations between observers in CT and MRI were minor and were also minor in CT, MRI and F18-FDG PET. If these results were minor, then it does not convince me that contouring variability is an urgent issue that needs to be addressed. Maybe reword some of the results, discussion and conclusion so that the overall story being told is more consistent. Changed according. “In our study, interobserver variability in non-spine bone metastases was higher in CT than in MRI and PET. To reduce this variability, MRI and PET are tests that can help. However, radiation oncologists are not trained in delineation on these tests, so, although less, some variability was also observed. For this reason, the authors have designed an action plan using a training platform for radiation oncologists. There is a need for standardization and protocols with contour guidelines for these situations, which will become increasingly frequent in our clinical practice”

Main suggestions:
- Using consistent terminology throughout. Changed according.
- The objective, results and conclusion are not in agreement. In your objective, you hypothesize that MRI and/or 18F-FDG PET will optimize the accuracy of tumor delineation, but in results you highlight that interobserver variability in MRI and 18F-FDG PET are minor. Then in the conclusion you say the interobserver variability was remarkable and that there is a need to reduce variability. If the interobserver variability is minor, then why would you need to set up a training platform to reduce this variability? Also, in your intro there is some focus on how it is not clear what the ideal MRI sequence is for imaging non-vertebral bone tumors but in your study design you only compare all of MRI with CT. It seems that you have enough data to look at interobserver variability for each MRI sequence and to have some discussion on if interobserver variability was better for a specific MRI sequence compared to another. Clarified, thank you. We included table 1 with different sequences of MRI.
- Overall, the details in the paper are vague. It is unclear exactly how many medical scans there are, and how many contours total and contours per scan. It would also be wise to include information on the CT and MRI scans, such as the slice thickness and pixel spacing so that one can interpret how many voxels are in disagreement when there is a volume difference of 0.7cc. The statistical analysis lacks details. Also the statistical analysis mentions metrics such as the overlap ratio and kappa index which was never discussed in the results. Changed according. “Images were acquired by a Toshiba Aquileon multislice helical CT (64 slices), with a Philips Achieva 1.5 Tesla MRI and Siemens mct biograph 18FDG PET. In PET images we used to define tumor SUV more than 5. CT, MRI and PET slices were performed every 3 mm. MRI sequences included T1, T2, and diffusion. For the study of contours in MRI, different sequences were included using the one that showed the least variability and compared with CT and PET.“
- While I have given a lot of critical feedback, I do think the authors have enough data to make a good manuscript. However, I do not think the authors have fully leveraged the data and I think the manuscript needs more details as well as clear objectives and conclusions that are supported by the data in the results/discussion. Changed according, thank you.

Reviewer #2: **Summary Comments:**
- I have reviewed the article and do not feel that it is ready for publication as there are several flaws. Thank you for your review

- In terms of scope, after reviewing the journal, I feel it does fit within the scope of the Journal of Clinical and Translational Research. Thank you for your comment.

- In terms of novelty, while the concept of this paper is not new (and closely mimics the ideas behind Raman *et al.* 2018), I do feel that more data is needed in this field of study, which the authors would provide with their original article. Thank you for your comment.
- However, at this point, I do not feel that the methods used support the objectives stated in the introduction, but this may also be due to the fact that the methods are very unclear on what they did. Specifically they make say, in somewhat a round about way, that they are trying to investigate the ideal imaging for delineation of non-vertebral bone tumours, specifically when evaluating inter-observer variability. They seem to wish to compare the impact of CT, MRI, and 18FDG PET on this metric. Changed according.
- It is not clear from the methods how this was studied. The methods state that each set of images were studied independently (Page 3, Line 3) but then go on to state that the ROs had the fusions of MRI and FDG available to them when contouring. It is not clear how the study was actually set up. In addition, there were other issues in terms of inconsistencies in terminology and metrics used. This is only one example of clarity issues. Please see my point-by-point discussion below. Changed according.
- I think it would benefit the authors to look at using additional metrics to quantify their data for easier comparison to the literature e.g., generalized conformity index. Currently they state that they used the Kappa Index but it is not clear how they modified this statistics for multiple observers. Clarified according.
- I also feel that basic information for reproducibility is missing from the methods e.g., the image acquisition protocol. Clarified according. “Images were acquired by a Toshiba Aquileon multislice helical CT (64 slices), with a Philips Achieva 1.5 Tesla MRI and Siemens mct biograph 18FDG PET. In PET images we used to define tumor SUV more than 5. CT, MRI and PET slices were performed every 3 mm. MRI sequences included T1, T2, and diffusion. For the study of contours in MRI, different sequences were included using the one that showed the least variability and compared with CT and PET.“

- There are also things included that have no relevance to the paper e.g., the mention of a subjective assessment survey of contour delineation. The conclusions stated also seem more of a future work than conclusion. I do not feel I got a sense for what the authors felt of their results and how they fit into the broader literature (i.e., do not make any strong conclusions on whether MRI or PET is helpful and should be used) Changed according. Conclusions: “In our study, interobserver variability in non-spine bone metastases was higher in CT than in MRI and PET. To reduce this variability, MRI and PET are tests that can help. However, radiation oncologists are not trained in delineation on these tests, so, although less, some variability was also observed. For this reason, the authors have designed an action plan using a training platform for radiation oncologists. There is a need for standardization and protocols with contour guidelines for these situations, which will become increasingly frequent in our clinical practice”

Point-by-Point comments
NOTE: this list is not comprehensive - the manuscript requires significant revisions and so correcting for grammer etc. will have to be done after that

NOTE: "L" refers to line number

- **Introduction**

- It should be more strongly stated in the introduction the impact of having good contours in SBRT type treatments (e.g., small margins = high chance of geographic miss) Changed according.
- **Methods**

- L7 page 3 - since the same ROs did not contour throughout, there is additional variability added and less trust in the results. This variability has been taken into account, thank you.
- This is important when considering that intra-observer variability can be significant. This variability has been taken into account, thank you.
- Why was intra-observer variability not reported. This is a secondary objective of the study
- The writers talk about using the Kappa statistic to measure inter-observer variability but then refer to using a correlation coefficient. These mean different things - it has been assumed that this was a mistake and the authors used Cohens but clarity needs to be made. Correct, we used cohen, corrected accordingly
- Cohen's kappa is typically used for 2 raters. There are extensions to include multiple raters including Fleiss' Kappa but it is not clear from the text how this was accounted for. Clarified in the manuscript. “”The kappa index was calculated as the ratio of the intersection of the contours delineated for a given observer with the corresponding contours of another observer and their mean. A kappa statistic including Fleiss' Kappa was used to account for multiple observers (10). The Landis and Koch interpretation was used for assessment agreement within each group of bone metastasis images.”

- No information was given regarding how the registration was performed, who performed it, etc. Was only 1 registration used for all ROs (it is assumed so)? This is important for repeatability. Clarified in the manuscript. “Image fusion was performed with rigid registration, available in the planner, the same for all oncologist”

- Were the MRI and PET images taken in the same position as CT? Were they all acquired the same day? Clarified in the manuscript. “in different positions”

- There is no information given on acquisition protocols for any of the image sets, as well as no information on the MRI sequences used (name, any suppression schemes, etc.) Changed according. Table 1.

- No indication of what threshold was used with PET which is important when discussing contouring with PET which is known to be very subjective to the window level used. Changed according. In PET images we used to define tumor SUV more than 5.
- L18 page 3 "All sets of images were studied independently" - it is not clear what this means in the context of the other methods that state that ROs had all imaging information available to them when contouring. Clarified.

- The only metrics reported were Cohen's kappa and volume. It would be beneficial to describe using other metrics that have been used in the literature such as Generalized conformity index (CI_gen), distance difference between center of mass etc. Using one of the studies they cite (Raman et al) would be good as a template (e.g., table 2 in Raman et al). The most interesting and significant data were included, thank you

- A survey was mentioned in the methods but no results from it given. What was the purpose of the survey? What did it ask? At this point, extraneous information not reported on. Has finally been eliminated

- With 10 cases, some of which are from the same patient, the validity of the t-test is questionable. Likely need to use non-parametric tests. The most interesting and significant data were included, thank you

- Significant issues with what is said to be reported in the methodology and what is actually reported. Say that they report the "overlap ratio" (never defined) and the kappa index but in the results they
- 1. Never refer/report the overlap ratio
- 2. Never refer to the kappa index
- 3. Report the "correlation coefficient" and the "Index of agreement"
- I assume that there is some sort of consistency error but at this point cannot determine what they actually reported.

Changed according

- **Results**
- Do not need to report both the variance and standard deviation - standard deviation is sufficient and variance only adds extraneous information. Changed accoding
- Units should be reported for variance and standard deviation (assumed to be cc^2 and cc, respectively). Changed accoding, cc
- It is sometimes unclear what is being reported - it seems that median and range is being shown in several areas but should be clarified. Changed accoding
- Why was the median index of agreement used? Why not mean? The median is generally used to return the central tendency in the case of skewed numerical distributions. The median provides the typical value even when the data set is skewed to one side or the other.
- L12 page 4 - The sentence needs to be clarified - is this comparing the CT and FDG? Why does the author then go on to compare with MRI in the same paragraph. Should likely have a separate paragraph. Changed accoding
- It is not clear what the purpose of putting in figure 1 is. The figure is small enough that they could likely include at least a few patients for readers to get a better sense of how the agreement looks (multipanel - see Raman et al reference). Deleted, we refer only to figure 2
- L25 page4 - What is meant that correlation coefficients were insignificant in most cases? Do they mean statistically not significant? Or just that they are low values? They are low values
- Table 1 - How can the correlation coefficient be 1 between 6 different observers for patients 5-b, 6, and essentially 1 for 5-a and 7? These seem unrealistic. Revised and corrected, thank you
- Table 2 - Why is there no correlation coefficient for patient 3? Revised and corrected, thank you
- Table 1,2,3 - summary data should be given at the bottom in its own row. I do not understand your comment, could you help me?
- Figure 1 - What is the purple contour? The caption should better describe the image. Deleted

- All MRI data is referred to as "MRI" - what happened to having T1 and T2 data? Why weren't these separately investigated as was hinted at in the introduction? Included in table 1

- **Discussion**
- First sentence should likely be an introduction point. Changed according
- If previous studies showed diffusion weighted imaging was useful, why was this not investigated in this trial? If it was done, the data are included
- L49-59 Page 4 - this paragraph is difficult to understand and I am not sure what point they are trying to make Clarified according
- Page 5 lines 1-15 (first paragraph on this page) - the authors show that their results differ from other studies but provide no possible reasons as to why this may be the case. Changed according
- L16-L19 page 5 - FDG paragraph needs to be expanded on what the authors feel the role of FDG is and what they plan to do in the future with it (potentially investigate better standardization of its use) Changed according
- L20 page 5 - CTVs are not to be used for compensating for interobserver variability as is implied - that is the role of the PTVs. No suggestions are given on what potential margins would be required (when margins are typically not used due to doses in SBRT being so high that the fall -off treats microscopic disease in many sites) Clarified according
- **Conclusions**
- Conclusions paragraph does not actually conclude anything relevant to the study and seems to be a future work statement. Needs to be adjusted to match standard conclusions paragraphs. Clarified according. “In our study, interobserver variability in non-spine bone metastases was higher in CT than in MRI and PET. To reduce this variability, MRI and PET are tests that can help. However, radiation oncologists are not trained in delineation on these tests, so, although less, some variability was also observed. For this reason, the authors have designed an action plan using a training platform for radiation oncologists. There is a need for standardization and protocols with contour guidelines for these situations, which will become increasingly frequent in our clinical practice”

3rd Editorial decision
08-Sep-2022

Ref.: Ms. No. JCTRes-D-22-00080R2
Interobserver variability in GTV contouring in non-spine bone metastases.
Journal of Clinical and Translational Research

Dear author(s),

Reviewers have submitted their critical appraisal of your paper. The reviewers' comments are appended below. Based on their comments and evaluation by the editorial board, your work was FOUND SUITABLE FOR PUBLICATION AFTER MINOR REVISION.

If you decide to revise the work, please itemize the reviewers' comments and provide a point-by-point response to every comment. An exemplary rebuttal letter can be found on at http://www.jctres.com/en/author-guidelines/ under "Manuscript preparation." Also, please use the track changes function in the original document so that the reviewers can easily verify your responses.

Your revision is due by Oct 08, 2022.

To submit a revision, go to https://www.editorialmanager.com/jctres/ and log in as an Author.

You will see a menu item call Submission Needing Revision. You will find your submission record there.

Yours sincerely,

Michal Heger
Editor-in-Chief
Journal of Clinical and Translational Research

Reviewers' comments:

Dear authors,

Thank you for submitting a revised draft and for incorporating all the reviewers' comments.

The manuscript has now passed though peer review and is suitable for publication after linguistic revision.

This should come at no surprise, as there is always a linguistic battle between us, and this time it is not different.

There are grammar, spelling, and even syntax errors that really must be corrected before I can accept the paper.

Please perform your final task very seriously, and preferably engage a native speaker (co-author, contract service, or JCTR editor).

Thanks and good luck,

Michal Heger
Editor

4th Editorial decision
09-Sep-2022

Ref.: Ms. No. JCTRes-D-22-00080R3
Interobserver variability in the contouring of non-spine bone metastases.
Journal of Clinical and Translational Research

Dear author(s),

Reviewers have submitted their critical appraisal of your paper. The reviewers' comments are appended below. Based on their comments and evaluation by the editorial board, your work was FOUND SUITABLE FOR PUBLICATION AFTER MINOR REVISION.

If you decide to revise the work, please itemize the reviewers' comments and provide a point-by-point response to every comment. An exemplary rebuttal letter can be found on at http://www.jctres.com/en/author-guidelines/ under "Manuscript preparation." Also, please use the track changes function in the original document so that the reviewers can easily verify your responses.

Your revision is due by Oct 09, 2022.

To submit a revision, go to https://www.editorialmanager.com/jctres/ and log in as an Author. You will see a menu item call Submission Needing Revision. You will find your submission record there.

Yours sincerely,

Michal Heger
Editor-in-Chief
Journal of Clinical and Translational Research

Reviewers' comments:

Still replete with linguistic errors and inconsistencies. To cite a few, in the abstract alone:

- Gross Tumor Volume should not be capitalized;
- SBRT was not written out in full, but only used once. So please write out in full but do not abbreviate;
- "MRI and/or 18FDG PET" should be written without an article ("the");
- will optimize should be optimizes as this should be present tense given that the techniques are already being used;
- compared to instead of with when the comparison concerns a difference;
- We evaluated (past tense since the research was performed in the past) - should be extended to other sections as well;
- what is a center panel?
- why is IRB plural? In the Methods section you only refer to a single ethics committee;

and the list goes on

We will stay in this cycle until the manuscript is in pristine shape linguistically, I am afraid.

Thanks!

Michal Heger
Editor

5th Editorial decision
14-Sep-2022

Ref.: Ms. No. JCTRes-D-22-00080R4
Interobserver variability in the contouring of non-spine bone metastases.
Journal of Clinical and Translational Research

Dear authors,

I am pleased to inform you that your manuscript has been accepted for publication in the Journal of Clinical and Translational Research. Please see my comments below - THESE ARE IMPORTANT.

You will receive the proofs of your article shortly, which we kindly ask you to thoroughly review for any errors.

Thank you for submitting your work to JCTR.

Kindest regards,

Michal Heger
Editor-in-Chief
Journal of Clinical and Translational Research

Comments from the editor:

Since you were incapable of properly writing a manuscript, I did it for you.

Please ensure I interpreted everything correctly after you receive the proofs.