Response to the reviewers

We appreciate the time spent by the editors and reviewers in assessing our manuscript. Before addressing the issues raised by the reviewers, we would like to point out that we discovered a bug in the proton image processing within a week of the initial submission and changed the results and text accordingly. Also, a couple co-authors provided recommendations after the initial submission which explains certain aesthetic differences (e.g., the inversion of the red and green labelings in some of the figures). Please see below a point-by-point to the issues raised by the reviewers. Please note that textual changes in the manuscript are in blue font.

Reviewer 1

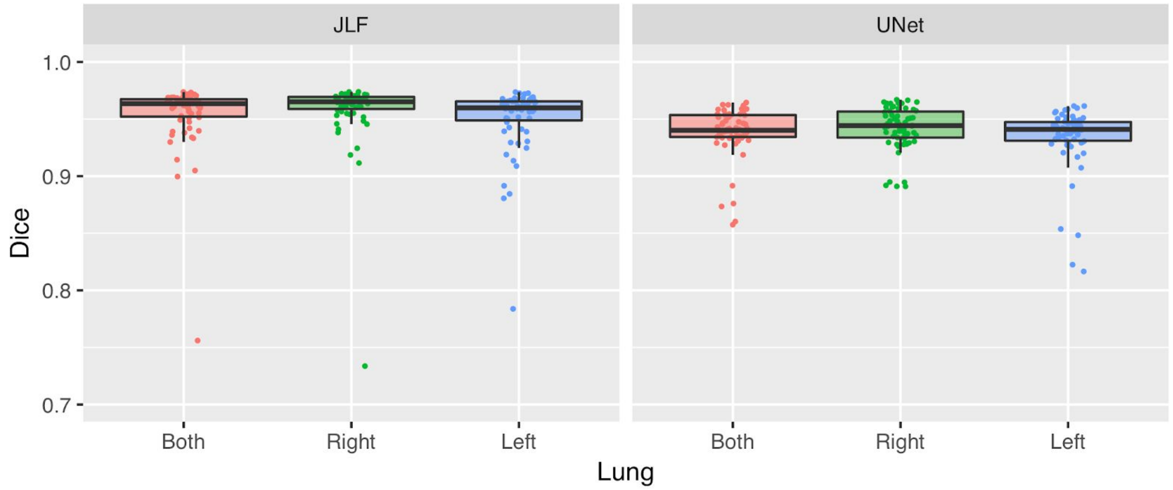
The authors proposed an automatic segmentation pipeline to accelerate VDP measurements using hyperpolarized gas imaging, improving upon the segmentation speed of the previously published JLF method. This improvement could potentially advance the clinical translation of this technique by streamlining the entire process of quantitative post-analysis. The paper is well written overall. In addition, the algorithm developed by the authors is based on ANTs software, the use of which is consistently expanding within the field. I strongly recommend this paper for publication. Nevertheless, I have several comments in below.

> *We appreciate the assessment of our work from the Reviewer and hope the responses below adequately respond to the Reviewer's concerns.*

1. The new deep learning pipeline reduced the accuracy of segmentation compared to JLF. An ~87% Dice coefficient is not very good given the relative ease of segmenting proton MRI images. Did the algorithm take advantage of the dramatic boundary intensity contrast? Can the author show an imperfect segmentation using this AI algorithm? What could be potential reasons for this imperfection, and what are some potential solutions?

   As mentioned above, we discovered a bug in the proton MRI processing. We invite the reviewer to reassess based on these new results (reproduced below).

   > Accuracy for CNN-based proton lung MRI segmentation (in terms of Dice overlap) was left lung: 0.93 ± 0.03, right lung: 0.94 ± 0.02, and whole lung: 0.94 ± 0.02. Although slightly less accurate than our previously reported joint label fusion (JLF) approach (left lung: 0.95 ± 0.02, right lung: 0.96 ± 0.01, whole lung: 0.96 ± 0.01), processing time is < 1 second per subject for the proposed approach versus ~30 minutes per subject using JLF.



2. Can the authors discuss the potential translation of this algorithm to other imaging modalities such as CT, which typically offer higher resolution? This algorithm's potential to speed up the segmentation of CT images could ultimately be more beneficial than its application to MRI. On the other hand, CT images in the case of various pathologies display a similar intensity to that of the chest wall.

   We added the following sentence to the Discussion section:

   > Additionally, as the U-net architecture is application-agnostic, investigators can apply the contributions discussed in this work to their own data, such as lung CT.

   The Reviewer raises a couple interesting points concerning the application to lung CT. We know of a couple of other groups who are looking at this application. CT certainly offers higher resolution but that implies larger image sizes which presents certain practical difficulties. Nevertheless, the motivation for this line of research is obvious but additional insight and discussion is beyond the scope of this work.

3. As the authors mention, this technique is significant for data augmentation. Would it be possible and/or legal for the authors to share their atlas and images?

   Not mentioning the availability of the proton lung MRI and masks was an oversight on our part which we have remedied in the current manuscript. We added the following sentence to the **Processing specifics** *subsection in* **MATERIALS AND METHODS***:*

   > These data are publicly available for download at
   > https://doi.org/10.6084/m9.figshare.4964915.v1.

   Could the ANTs network somehow optimize the atlas by using other users' data if they agree to share?

Yes. We added the following sentence to the Discussion section:

> More immediate benefits could result from augmenting the limited, single-site data set used in this work to include data contributed from other groups which could translate into more robust models.

Reviewer 2

This manuscript presents a novel and useful framework built into an open-source software package for automated segmentation of proton lung MRI and automated labeling of ventilation defect in hyperpolarized gas MRI. This work has potential to influence research involved in structural and functional imaging, as well as potential for clinical use in image processing. Deep learning is a hot topic for medical image processing, and a common challenge in this context is limited availability of annotated training data. In this work, the data augmentation approach used to generate additional training data from existing training data is unique and innovative, avoiding the problem of spatially distorting images used to learn shape-sensitive features. The benefit of the deep learning approach is demonstrated with comparable accuracy to the previous rule-based approach yet almost 1000-fold reduction in computation time. However the overall manuscript writing style does not appropriately delineate content into different sections, and the novel data augmentation strategy is not described in sufficient detail in the methodology.

We appreciate the assessment of our work from the Reviewer and hope the responses below adequately respond to the Reviewer's concerns.

Major comments

1. Much of the methods section contains background information and qualitative comparison of methodologies, yet not enough detail description of actual methodology. non-methods-related subject matter should be relegated to the introduction and/or discussion. In general, keep background/motivation/rationale in the introduction. Methods should contain only description of what you did. Qualitative and quantitative comparisons should be kept in discussion. The overall quality and focus of this manuscript would be improved by moving, combining, or removing misplaced portions of the text to conform to these general content guidelines. This issue pertains particularly to the methodology of the deep learning approach:

page 8, line 34-48

Done.

page 11, line 13-27

We disagree with the Reviewer with respect to this particular section and choose to leave it as is.

page 13, line 20-38

After re-reading this section, we believe it best to simply remove it entirely.

Similar to above comment, description of methodology does not belong in results section. in particular, descriptions of statistical tests used and comparisons made should be indicated in methods.

page 16, line 37-53

We partially agree with the Reviewer and have removed the top portion of this section as it is redundant with what was written at the end of the **MATERIALS AND METHODS** *section.*

page 17, line 40-57

Done.

page 18, line 9-13

We disagree with the Reviewer with respect to this particular section and choose to leave it as is.

2. The manuscript would be improved by additional detail for one of the key contributions of this paper, namely the data augmentation approach.

Done. Please see changes in Figure 4 and the subsection **Template-based data augmentation**.

- page 11, line 37: "propagated to the space of every other image" -- it may be helpful to more rigorously and clearly specify this process, e.g. something along the lines of: applying the forward template transformation of one training image followed by the inverse template transformation of another, resulting in a unique combination of the first image transformed into the space of the second.

Done. Please see changes in Figure 4 and the subsection **Template-based data augmentation**.

- Figure 2, consider replacing the zigzag bidirectional arrows with two parallel unidirectional arrows, represent the separate forward and inverse transformations. These could be color-coded (or solid vs. dashed), and used in the legend to explicitly describe the process used to map from one image space to another through the template.

> *The symbology in the manuscript is consistent with other ANTs publications so we choose to keep this as is. However, to clarify usage, we have explicitly specified it's meaning in the caption of Figure 4.*

- It may also be more intuitive for some readers to present a few simple symbols and equations relating this process in the text.

> *Done. Please see changes in Figure 4 and the subsection* **Template-based data augmentation***.*

- A detailed description of the data augmentation pipeline w.r.t. structural vs. functional images

> *Done. Please see changes in Figure 4 and the subsection* **Template-based data augmentation***.*

3. Are the same transformations applied when augmenting the proton vs. ventilation image from a single subject? if not, ventilation defects may alter perceived shape and distort registration to a template. if i understand correctly, it seems as though the template transformations should be constructed using structural images only, and the resulting transformations applied to both structural and functional images. this would also clarify the issue of how you augment the left-right masks for use as the additional channel for training the ventilation neural network. the process for data augmentation of this additional channel is not described in the methodology. an explicit step-by-step description of this augmentation pipeline might alleviate this concern. i understand that there is a major difference between 3D structural images and 2D functional images, but this only further warrants explanation of how this ancillary channel from the structural image CNN was used as input for the functional image CNN. do you identify a single slice from the structural image mask output by the first CNN to pass as input to the second CNN? if so, how is that decision made?

> *Please see changes in Figure 4 and the subsections* **Template-based data augmentation** *and* **Processing specifics***. In addition, here are specific answers to a couple of the Reviewer's comments with selections from the current text:*

- Are the same transformations applied when augmenting the proton vs. ventilation image from a single subject?

  Since U-net model generation is completely separate for the proton and ventilation data, template-based data augmentation is also isolated between the two protocols.

- if not, ventilation defects may alter perceived shape and distort registration to a template.

> In contrast, the ventilation template is created directly from the training ventilation images using a more elastic-type transform resulting in the grayscale template in the center of the right panel of Figure 4. An alternative strategy for ventilation template creation could have employed relabeling the ventilation masks to a single value and aligning the lung boundaries, somewhat similar to the proton protocol. This approach would be necessary for severe lung disease where major portions of the lung are absent in the ventilation image.

The Reviewer should also keep in mind that precise alignment with the template is not crucial, and in our estimation, would not negatively affect the results if misalignments were to occur. In the case of possible distortion, it is important to remember that the same transformation (correct or incorrect) will be applied to any ancillary (channel) images as well as the corresponding segmentation so the intensity information learned during model training will not be incorrect. In addition, we have not found evidence that the U-net model generation is learning any sort of ventilation defect shape:

> We built a 2-D U-net model for the ventilation images as the functional image segmentation does not take advantage of obvious anatomical factors.

This also explains why 2-D models were sufficient for the good performance shown in Figure 6. Contrast this with the structural protocol where we

> built a 3-D U-net model to take advantage of the characteristic 3-D shape of the lungs.

4. It would strengthen the merit and justification for the proposed approach to see some sort of result or quantitative analysis supporting the hypothesis that the proposed data augmentation strategy is beneficial for applications in medical imaging, as opposed to training with conventional data augmentation or without any augmentation.

We added the following footnote in the results section:

> Although the need for data augmentation techniques is well-established within the deep learning research community, we performed a smaller 2-D experiment to illustrate the potential gain using template-based data augmentation. Training data consisted of 50 coronal proton lung MRI with lung segmentations. Data and scripts are available in the companion repository to the ANTsRNet package, called ANTsRNetExamples, which contains examples for available architectures. Accuracy, in terms of Dice overlap, achieved with augmentation was left lung: 0.94 ± 0.02, right lung: 0.92± 0.04, and whole lung: 0.93 ± 0.03. Accuracy achieved without augmentation was left lung: 0.88 ± 0.13, right lung: 0.83 ± 0.21, and whole lung: 0.86 ± 0.16.

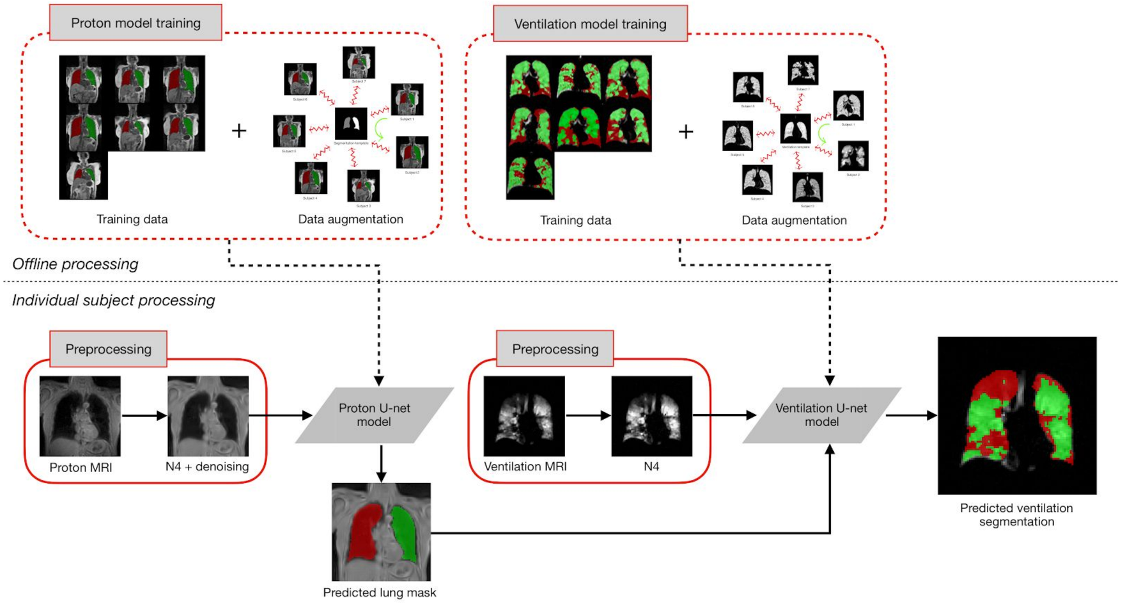
Minor comments

page 3, line 60: "common use case scenarios in which proton images are used for quantifying corresponding ventilation images". it would help to be more specific here, e.g., "proton images are used to identify regions of interest in corresponding ventilation images, which typically contain no discernible boundaries for anatomic structures". otherwise the need/motivation may not be clear to some readers.

Done.

figure 1: This figure provides a helpful guide through the processing pipeline, but there are a few issues that could be addressed: Visual representation of images for data augmentation do not match training data (red/green vs grayscale). Also, number of subjects does not match: training data set shows N=6 subjects but data augmentation shows N=7. Finally, if the ventilation model is trained using mask input, then there should be some indication of that in this figure. in addition to major comment above, this calls into question whether the registration/warping is different for the structural vs. functional data during augmentation.

We have altered Figure 1



and Figure 4 and added to the subsection **Template-based data augmentation**.

page 7, line 29: "Unlike other methods which rely solely on intensity distributions which discards spatial information, our technique..." -- this sentence is syntactically awkward. Consider replacement by "Unlike other methods that rely solely on intensity distributions, thereby discarding spatial information, our technique..."

Done.

page 13, line 10: "(as well as more focused techniques)" - such as? can you provide an example?

Sure. The sentence in question currently reads:

> Basic image operations such as classification, object identification, segmentation, as well as more focused techniques, such as predictive image registration [42], have significant potential for facilitating basic medical research.

Figure 8. Legend mentions specific effects of partial voluming, perhaps an arrow or other visual indicator may help the reader quickly identify this ROI.

Done.   We also specified this in the caption.

page 3, line 32: change "has" to "have"

Done.

page 4, line 9: change "requirements"

Done.

page 5, line 14: helium and xenon are not proper nouns and should not be capitalized

Done.

page 5, line 24: change "studies was" to "studies were"

Done.

page 5, line 26: change "either ... and" to "either ... or"

Done.

page 5, line 26: helium and xenon are not proper nouns and should not be capitalized

Done.

page 7, line 23: change "based an" to "based on"

Done.

page 10, line 47: "sandwiching" is informal and idiomatic, consider alternatives

Done.

page 11, line 27: "within the range within" -- consider replacement by "within the typical range exhibited by"

Done.

page 13, line 6: change "has" to "have"

Done.

page 13, line 28: insert comma: "implementations, leading"

Done.

page 15, line 8: spell out "repository"

Done.

page 19, line 38: "shown in 7" to "shown in figure 7"

Done.

page 19, line 46: "approaches are" to "approaches is"

Done.

page 19, line 49-53: awkward phrasing.

Done.