*We appreciate the time spent by the editors and reviewers in assessing our manuscript.*

*Please see below for a point-by-point response to the issues raised. Textual changes in the manuscript are denoted in blue font.*

**Editor:**

Thanks for submitting to Magn Reson Med. Your manuscript was reviewed by two experts in the field. There is support for eventual publication of the manuscript, but only after important revisions are made to the paper. We invite you to respond with a revised manuscript that seeks to address the points made. Please note that you are already beyond the word limit for a Full Paper. We would appreciate you making compensatory reductions for any further word count that is required to address the reviewer critiques, or by moving suitable material to Supporting Information. I also note that several of your figures are quite dense and may not reproduce well in the final paper. I suggest that you consider whether any of these figures can be simplified and made more legible. Please be aware that you can use Supporting Information (see Author Instructions) to show fuller data and figures that may not be appropriate for the core document. In addition to the reviewers' comments, please address the following formatting issue:

-Upload your figures separately in TIFF or EPS format.

*X Done.*

**Deputy editor:**

The reviewers found merit in the work judging by the generally high scores, which place the work above the threshold required for publication. The importance and timelines of the work was particularly noted. Nevertheless, the quality of the work will need to be improved before consideration for publication. Specifically, the important take-aways of the work are clouded by a lack of clarity, substantially weakening the work.

*After rereading the submitted manuscript in the context of the Reviewers’ comments, we made significant changes to the text to clarify the overall message. Please see below for clarifications in response to specific Reviewer items.*

**Reviewer 1**

This paper presents an important step forward in the increasing widespread use of HP gas MRI through reducing the time needed to analyse the data generated by this technique. Additionally, the authors provide a comparison with the current state of the art techniques employed in the field. Most admirably the authors have applied their method in the well-known ANTs R/python implementations to allow for reproduction of their results and comparison with other sites implementations of deep learning methods. I feel the authors have carried out a study which meets the purpose of the paper and have only minor comments on the work presented in this paper which are enumerated below:

*We appreciate the positive assessment of our work, particularly the recognition of open-source availability. Please allow us to clarify, though, that computational time considerations (assuming “reducing the time needed to analyse the data” refers to computational time for the different algorithms) are not an integral part of this work. In fact, although the time to run each algorithm for single image processing is different (considering typical image sizes), none of the processing times would, in our assessment, be considered prohibitive.*

Data/analysis comments:

1. While the rationale of demonstrating how the number of images in a reference set affects the output in linear binning is logical, in a real-world setting reference datasets of n<5 would not be used (one would hope!) and thus I suggest removing those analyses from Figure 6

*The intention of Figure 6 is not to demonstrate “real-world” feasibility of any particular reference set size for performing studies. In fact, although the Reviewer suggests a cut-off of n = 5, our measures show that even sizes of n ∈ {8,9} produce a concerning amount of cluster definitional variance (Figure 6, bottom) which would call into question “real-world” study utility. As such, we prefer to keep the entire range as nothing is lost by including all of n ∈ {1,...,10}. In addition, we see two advantages for inclusion: 1) the convergence trends are more easily visualized and 2) it implicitly points to another potentially problematic issue with the reference distribution component associated with the linear binning algorithm. Although related to the results in the subsection* **Effect of reference image set selection***, we omitted explicitly discussing this particular issue due to space constraints and wanting to avoid focusing excessively on any single histogram-based algorithm. ~~However, in the case of Reviewer interest, we provide the following elaboration:~~*

*In linear binning, clustering defines clinically relevant ventilation categories. Those clustering definitions are determined directly from the parameters of a fitted Gaussian distribution of the combined set of all normal cohort voxels. It is these Gaussian parameters (i.e., the mean and standard deviation) that are exploited within the optimization domain as meaningful histogram features. However, as shown in Figure 6 (reference set n = 1), the amount of variance in both the mean and standard deviation for individual subjects is relatively large. Due to this wide variance for these defining (in terms of normal vs. non-normal) parameters within the individual members of the healthy cohort, it would seem that rather than these Gaussian parameters being clinically meaningful, this algorithmic component of a Gaussian-based reference distribution is quite possibly an unrealized post hoc, pseudo-scientific rationalization for what amounts to a simple manifestation of the central limit theorem.[[1]](#footnote-0)*

*Further, we find this use of a reference distribution to be a particularly “brittle” (i.e., non-generalizable) component of the linear binning algorithm (as demonstrated with the UVa data). In addition, as mentioned in the manuscript, the published values (based on 10 normal subjects from a single site) of the reference distribution is N(0.52, 0.18), which define cluster thresholds in the normalized space [0,1] as {0, 0.16, 0.34, 0.52, 0.70, 0.88, 1.0}. This just happens to almost precisely define the boundaries if one simply performed a linear subdivision of [0, 1] into six equally spaced regions, i.e. 1 / 6 \* {0, 1, 2, 3, 4, 5, 6} = {0, 0.17, 0.33, 0.5, 0.67, 0.83, 1.0}. In the spirit of Occam’s razor, the latter would be a much more robust, less assumption-laden approach to defining cluster definitions for the linear binning algorithm. We note that the authors of the linear binning algorithm did explore ventilation intensity distributions which were non-Gaussian. However, instead of revising (or, at least, revisiting) the Gaussian assumptions associated with the original algorithm, a Box-Cox transform (with the clinically unjustified transformation parameter, lambda) is used to force the data to conform to these seemingly unvalidated distribution requirements.*

2. Given the data is publicly available the authors may not have seen fit to include SNR details, however, given it plays such a key role in the analysis of HP gas MRI I would suggest the inclusion of some form of SNR reporting (of original images and distorted images – e.g. median (IQR))

*This is a great idea. We added the following description to the Methods section:*

**Data simulations**

Both datasets were transformed by adding Gaussian noise, nonlinear histogram- based intensity warping, and their combination. The peak signal-to-noise ratio (PSNR) is defined as

*PSNR=20⋅log10⁡(max⁡(Ioriginal )) - 10⋅log10⁡(mse(Ioriginal,Isimulated))* (1)

where mse denotes the mean-squared error between the simulated image and the corresponding original image. The median PSNR values for the simulated UVa dataset are noise: 20.7dB, nonlinearities: 29.9dB, and noise and nonlinearities: 19.6dB. Analogous values for the Dataverse dataset are noise: 19.8dB, nonlinearities: 26.6dB, and noise and nonlinearities: 19.4dB.

3. From a personal point of view I feel, that whilst the matching of the segmentations is important, what is more important is having reliable volume calculations – particularly as VDP is so commonly used in this field, as stated by the authors themselves, and thus would be interested to see the variation introduced in ventilation defect volume rather than simply cluster overlap

*The Reviewer raises a very important point. We agree that VDP is of particular interest to the field in clinically-based quantification of functional lung images. In fact, we would expect it to be one of the most widely used measures reported in the literature derived from the segmentation algorithms discussed in this paper (and the only one that we mention in the manuscript).*

*It should be noted we do use VDP in our measurement bias evaluation located in subsection* **Diagnostic prediction** *where the prediction models employ VDP-based measures (cf Equation (2)). This shows that certain histogram-based algorithms perform just as well (if not slightly better) than El Bicho for these specific diagnostic scenarios. As we write in the original submission:*

In the absence of ground truth, this type of evaluation does provide evidence that all these algorithms produce measurements which are clinically relevant although, it should be noted, that this is a very coarse assessment strategy given the global measures used (i.e., cluster volume percentage) and the general clinical categories employed. In fact, even spirometry measures can be used to achieve highly accurate diagnostic predictions with machine learning techniques (51).

*However, similar to the lossy relationship between the image and its corresponding intensity histogram, purely volumetric-based measures (such as VDP difference) are lossy distillations of the segmentation overlap information and, as such, smooth over potentially important algorithmic differences. In fact, we believe it is precisely the nearly exclusive use of VDP and its corresponding coarse quantification properties (in terms of segmentation assessment) which has caused these algorithmic issues to go unnoticed, at least formally, for so long. As we write in the Introduction:*

Ultimately, we are not claiming that these algorithms are erroneous, per se. Much of the relevant research has been limited to quantifying differences with respect to ventilation versus non-ventilation in various clinical categories and these algorithms have demonstrated the capacity for advancing such research *through the use of clinically useful measures such as ventilation defect percentage (VDP)*. Furthermore, as the sample segmentations in Figure 3 illustrate, when considered qualitatively, each segmentation algorithm appears to produce reasonable segmentations even though the voxelwise differences are significant (as are the corresponding histograms). However, the aforementioned artefact issues influence quantitation in terms of core scientific measurement principles such as precision (e.g., reproducibility and repeatability (11,33)) and bias which are obscured in isolated considerations but become increasingly significant with multi-site (34) and large-scale studies. In addition, generally speaking, refinements in measuring capabilities correlate with scientific advancement so as acquisition and analysis methodologies improve, so should the level of sophistication and performance of the underlying measurement tools.

*with the emphasis denoting added text to explicitly address the Reviewer’s point. We also added the following paragraph to the* **Discussion**

As illustrated in Figure 2, measures based on the human visual system seem to quantify what is understood intuitively; that image-based information is much more robust than its corresponding histogram-based information in the presence of image transformations, such as common MR artefacts. This observation is not intended to imply that the histogram-based approaches are useless in performing research. In fact, ventilation defect percentage is perhaps the most widely used clinical measurement reported in the literature and it is easily quantified from the image histogram. Thus, even relatively simple histogram-only segmentation algorithms will provide some utility which was observed in the measurement bias experiments employing a variant of ventilation defect percentage to predict diagnostic accuracy. However, similar to the lossy relationship between the image and its corresponding histogram, such volumetric-based measures are lossy distillations of the segmentation information and might obscure important algorithmic characteristics and relative differences as well as discard potentially useful spatial information which is why additional experimentation explored measurement precision in the presence of MR artefacts.

*It should be noted that a good Dice coefficient value is indicative of good segmentation label overlap whereas volumetric differences afford no such guarantees. Or, to use the words of the Reviewer --- if one has good “matching of segmentations,” then one will have “reliable volume calculations.” The converse is not necessarily true. For an evaluation investigating segmentation performance, variation for a less precise measure would not be as informative. And given that we are evaluating segmentation algorithm performance and need to prioritize evaluation measures based, at least partly, on journal-based space constraints (as mentioned by the Editor above), we choose to use the Dice overlap measure for our precision-based evaluation and invite the interested Reviewer to use the publicly available resources associated with this manuscript to explore other measures of interest. In fact, while compiling the Dice values for all of our simulations, we also compiled VDP values which are available as .csv files in the Histograms GitHub repo. For example, for the UVa simulated data the .csv files containing the VDP values are available online at:*

* *Noise and nonlinear intensity variation -*

[*https://github.com/ntustison/Histograms/blob/main/Analysis/Data/SimulationExperiments\_NoiseAndNonlinearities/varianceStudy.csv*](https://github.com/ntustison/Histograms/blob/main/Analysis/Data/SimulationExperiments_NoiseAndNonlinearities/varianceStudy.csv)

* *Noise only -*

[*https://github.com/ntustison/Histograms/blob/main/Analysis/Data/SimulationExperiments\_Noise/varianceStudy.csv*](https://github.com/ntustison/Histograms/blob/main/Analysis/Data/SimulationExperiments_NoiseAndNonlinearities/varianceStudy.csv)

* *Nonlinear intensity variation only -*

[*https://github.com/ntustison/Histograms/blob/main/Analysis/Data/SimulationExperiments\_Nonlinearities/varianceStudy.csv*](https://github.com/ntustison/Histograms/blob/main/Analysis/Data/SimulationExperiments_NoiseAndNonlinearities/varianceStudy.csv)

*Similar .csv files are available for the “No N4” version of the UVa data experimental simulations as well as the Harvard Dataverse data.*

4. Whilst spatial overlap and volume differences are important considerations, recent work has shown that distance measures (e.g. Hausdorff/Euclidean) may provide additional understanding in the performance of segmentation methods and thus I would suggest the inclusion of one of these measures

*We are aware of certain contexts in which distance-based measures might augment segmentation performance and offer additional evaluation assessment possibilities (e.g., white matter hyperintensity segmentation in the brain[[2]](#footnote-1)). However, such contexts are highly specific and we do not see how the particulars of those scenarios translate to similar context-specific advantages in the functional lung imaging domain. Also, related to the discussion of Item 3, the Hausdorff average distance is not a reliable indicator of VDP (unlike the Dice coefficient).*

*Therefore, given space limitations, we choose to keep the current Dice-based evaluation. Nevertheless, we would be more than willing to reconsider if the Reviewer has a citation specific to hyperpolarized gas imaging in which distance-based evaluation measures are employed (and provide an advantage or complement existing evaluation measures) and/or has further reasoning as to what additional insight could be provided over the current Dice-based evaluation. Otherwise, we invite the Reviewer and other interested readers to explore the use of such evaluation measures through the previously mentioned resources. To facilitate this exploration, we recently added Hausdorff distance calculation to both the ANTsR and ANTsPy packages (see ANTsR::hausdorffDistance and ANTsPy::hausdorff\_distance).*

Minor comments:

1. There are a number of footnotes which, in the opinion of this reviewer are unnecessary – e.g. footnotes 1,4 and 5. Additionally, Footnote 2 could be included in parentheses in the main text as could footnote 3.

*The footnotes have been removed. In addition to the suggestion by the Reviewer, we noticed that the Style Guide states that “[t]he use of footnotes is discouraged.”[[3]](#footnote-2)*

2. page 3, line 10 - Early appears twice

*Fixed.*

3. page 5, line 9 - .. missing “that”? - “algorithms, that the intensity thresholds between class labels…”

*Fixed.*

4. Page 11, line 7 missing seconds or s after <10

*Fixed.*

5. Page 12, lines 23-24 – “a header-modified version of these data” appears twice

*Fixed.*

6. Page 14, line 31 – suggest changing “So, in summary,…” to “ To summarise…”

*Done.*

7. Figure 7 – some visible text underneath the “Cluster” label

*Fixed.*

8. Figure 8 – “Cluster” label partially obscures y axis values

*Fixed.*

9. Page 22 line 19, suggest semicolon after “intuitively”

*Fixed.*

10. Page 22, line 24 – suggest changing “Figure” to “Figures”

*Done.*

**Reviewer 2**

In this work, the authors rigorously compare various histogram- and image-based segmentation methods for hyperpolarized gas ventilation images. The authors demonstrate several of the failings of histogram-based techniques and provide a roadmap for future development. However, the manuscript was at times hard to follow due to somewhat disjointed organization, and the results were not fully discussed. Moreover, several stylistic requirements of Magnetic Resonance in Medicine, most notably citation style, were not followed.

*We appreciate the generally positive assessment of our work. We have made several changes based on the Reviewer’s comments including a complete revision of the Discussion section.*

Major Comments

1. In text citation and bibliography are formatted incorrectly. MRM requires the citations be listed in the order in which they appear, not alphabetical. Also, in-text citation should be done with superscripts.

*Done. The Style Guide[[4]](#footnote-3) specifies that “[r]eferences should use the Citation-Sequence format,” which has been corrected in the current manuscript, and also indicates “citing references in the text by a number in parentheses” which has also been followed in the current manuscript.*

2. Ultimately, five segmentation strategies (linear binning, Kmeans, Fuzzy c-means, Atropos, and El Bicho) are compared. The introduction briefly alludes to these, but they are not explicitly defined, or even all referenced by name, until the Methods section. The reader would benefit from a short description of each of these methods in the introduction.

*The following is included in the* **Introduction** *to further describe the methods contextualized within the primary message of this work (i.e., histogram-based vs. image-based considerations in pulmonary image segmentation):*

Linear binning is a simplified type of MR intensity standardization in which images from healthy controls are normalized to the range [0, 1] and then used to calculate the cluster intensity boundary values, based on an aggregated estimate of the parameters of a single Gaussian fit. Subject images to be segmented are then rescaled to this reference histogram (i.e., a global affine 1-D transform). This mapping results in alignment of the cluster boundaries such that corresponding labels are assumed to have similar clinical interpretation. Variants of the well-known k-means algorithm constitute an algorithmic approach with additional flexibility over linear binning as it employs prior knowledge in the form of a generic clustering desideratum (i.e., minimizing within-cluster intensity variance) for optimizing a type of nonlinear MR intensity standardization. However, as with binary thresholding, both linear binning and k-means completely discard spatial context in optimizing voxelwise cluster membership.

Additional sophistication incorporating spatial considerations is found in the fuzzy spatial c-means (26) and Gaussian mixture-modeling (GMM) with a Markov random field (MRF) prior algorithms. The former, similar to k-means, optimizes over the within-class sample variance but includes a per-sample membership weighting (27) whereas the latter is optimized via the expectation-maximization (EM) algorithm (28). These algorithms have the advantage, in contrast to histogram-only algorithms, that the intensity thresholds between class labels are softened which demonstrates some relative robustness to certain imaging distortions, such as noise. However, all these algorithms are flawed in the inherent assumption that meaningful structure is found, and can be adequately characterized, within the associated image histogram in order to optimize a multi-class labeling.

Related, was there a reason you did not include binary thresholding in your comparisons? I don't think that it needs to be included, but it is introduced with the other methods and then never discussed again.

*We agree with the Reviewer that binary thresholding does not need to be included but acknowledge that our reasoning for exclusion should have been mentioned. We corrected this as follows (in* **Methods: Algorithmic implementation***):*

In support of the discussion in the Introduction, we performed various experiments to compare the algorithms mentioned previously, viz. linear binning (7), hierarchical k-means (9), fuzzy spatial c-means (13), GMM-MRF (specifically, ANTs-based Atropos tailored for functional lung imaging) (14), and a trained CNN with roots in our earlier work (40), which we have dubbed “El Bicho.” Note that we consider the binary thresholding variants to be simplified versions of linear binning and, therefore, omit them from consideration in this work.

3. Lack of clarity regarding how determination of the thoracic cavity mask. Several of the methods that you compare (e.g. linear binning, kmeans) require a mask of the thoracic cavity, as the histogram-based methodology, as you point out, is agnostic to the image. However, you provide no details regarding how this thoracic cavity mask is generated.

*Although tangential to the point of the manuscript (i.e., the criticisms would still apply given any reasonable definition of a thoracic cavity mask), we added the following to the* **Methods: Algorithmic implementation** *section:*

Lung masks for the UVa data were created using segmentation functionality described in (40) and inspected/edited by one of the co-authors (M. H.). The lung masks for the Harvard Dataverse 129Xe data are publicly available with the online image repository (31).

4. Much of the Results section is dedicated to describing methods. eg. pages 14, 15, 16(lines 21-53), and 19 (lines 47-56) are almost exclusively dedicated to describing methods. This makes the paper hard to follow. These should be moved to the Methods section.

*We agree with the Reviewer and also note that, in addition to making the paper hard to follow, it also distracts from our main intent which is to generalize to histogram-based vs. image-based optimization segmentation considerations. We significantly revised the Discussion section to better contextualize the results:*

**Discussion**

Over the past decade, multiple algorithms have been proposed for the segmentation of hyperpolarized gas images into clinically based functional categories. These algorithms are optimized using the histogram information primarily (with many using it exclusively) much to the detriment of algorithmic robustness and segmentation quality. This is due to the simple fact that these approaches discard, or do not optimally leverage, a vital piece of information essential for accurate quantitative image interpretation---the spatial relationships between voxel intensities. While simplifying the underlying complexity of the segmentation problem, these algorithms are deficient in leveraging the general modelling principle of incorporating all available prior information to any solution method. In fact, this is a fundamental implication of the “No Free Lunch Theorem” (54)—algorithmic performance hinges on available prior information.

As illustrated in Figure 2, measures based on the human visual system seem to quantify what is understood intuitively; that image-based information is much more robust than its corresponding histogram-based information in the presence of image transformations, such as common MR artefacts. This observation is not intended to imply that the histogram-based approaches are useless in performing research. In fact, ventilation defect percentage is perhaps the most widely used clinical measurement reported in the literature and it is easily quantified from the image histogram. Thus, even relatively simple histogram-only segmentation algorithms will provide some utility which was observed in the measurement bias experiments employing a variant of ventilation defect percentage to predict diagnostic accuracy. However, similar to the lossy relationship between the image and its corresponding histogram, such volumetric-based measures are lossy distillations of the segmentation information and might obscure important algorithmic characteristics and relative differences as well as discard potentially useful spatial information which is why additional experimentation explored measurement precision in the presence of MR artefacts.

Common MR artefacts of noise and intensity nonlinearities can produce quantifiable differences in the segmentation results and the degree of deviation (i.e., lack of measurement precision) largely corresponds to the algorithmic choice of optimization domain, i.e., image-based vs. histogram-based, with those algorithms leveraging the former providing improved segmentation repeatability. Notably, El Bicho generally yields the best segmentation overlap measures over the specified clusters and MR artefacts most likely due to optimization of the governing network weights over hierarchical image features found in the training set as opposed to strictly relative intensities and/or more simplistic neighborhood intensity information. In addition, this network demonstrates site acquisition generalizability as these performance gains are also seen in the Harvard Dataverse dataset.

In addition to motivating a renewed assessment of current algorithmic approaches to pulmonary hyperpolarized gas segmentation, there are other avenues for further research. El Bicho was developed in parallel with the writing of this manuscript merely to showcase the incredible potential that deep learning can have in the field of hyperpolarized gas imaging (as well as to update our earlier work (40)). We certainly recognize and expect that alternative deep learning strategies (e.g., hyperparameter choice, training data selection, data augmentation, etc.) would provide comparable and even superior performance to what was presented with El Bicho. However, that is precisely our motivation for presenting this work—deep learning, generally, presents a much better alternative than histogram approaches as network training directly takes place in the image (i.e., spatial) domain and not in a transformed space where key information has been discarded. Just as important, deep learning provides other avenues for research exploration and development. For example, given the relatively lower resolution of the acquisition image, exploration of the effects of deep learning-based super-resolution might prove worthy of application- specific investigation (55). Also, with the same network software libraries, high-performing classification networks can be constructed and trained which might yield novel insights regarding image-based characterization of disease. One additional modification that we did not explore in this work, but is extremely important, is the confound caused by multi-site data which has yet to be explored in-depth. With neural networks, such confounds can be handled as part of the training process or as an explicit network modification. Either would be important to consider for future work.

Admittedly, this work was limited in its exploration of MR artefacts. Noise variation was limited to a zero-mean Gaussian distribution and nonlinear intensity variation was explored strictly through smoothly varying histogram deformation. Inclusion of other noise models (e.g., shot, salt-and-pepper) might further characterize algorithmic differences and provide additional realistic data augmentation strategies. Specific to nonlinear intensity variation, a recent addition to the ANTsX ecosystem allows for the possible simulation of bias fields which would also expand data augmentation and, significantly, in the spirit of algorithmic parsimony, could potentially remove the dependency of N4 bias correction as an unnecessary preprocessing step.

Finally, although ventilation defect percentage has proven to be a compelling quantity for clinical studies, the results from the diagnostic prediction evaluation and the previous discussion implies that this popular measure does not fully leverage the spatial information of the segmentation information from any of these algorithms. Perhaps the results of this work, in addition to pointing to the need for rethinking algorithm innovation direction, also point to possibly investigating differentiating spatial patterns within the images as evidence of disease and/or growth and correlations with non-imaging data using sophisticated voxel-scale statistical techniques which intrinsically leverage spatial information (e.g., similarity-driven multivariate linear reconstruction (52,53)).

5. The first half of the Discussion section (page 21-line 17 of page 22) seems as though it would be more appropriate in the introduction. You do not discuss any of your results in these lines, just provide general information about the segmentation methods.

*Please see our previous response to Item 4.*

6. You provide very little discussion of your results in the Discussion section: Experiments showcased in figure 4, figure 5, and figure 6 are not discussed at all. Also related to figure 4 - you state "All four algorithms perform significantly better than a random classifier", but provide no other discussion of the results displayed in figure 4. However, there are clear differences in the ROC curves among the segmentation techniques - Linear Binning and El Bicho clearly have the best overall performance (by this metric), and some discussion of this would be helpful.

*We would also consider the text following “All four algorithms perform significantly better than a random classifier” to be relevant in discussing the results of Figure 4:*

In the absence of ground truth, this type of evaluation does provide evidence that all these algorithms produce measurements which are clinically relevant although, it should be noted, that this is a very coarse assessment strategy given the global measures used (i.e., cluster volume percentage) and the general clinical categories employed. *This complicates attempts at additional inferences concerning voxelwise bias performance with this type of evaluation strategy.* In fact, even spirometry measures can be used to achieve highly accurate diagnostic predictions with machine learning techniques (51).

*where the italicized portion was added to relay to the reader that, despite the obvious differences in ROC curves for this particular scenario, we do not believe that further conclusions are warranted, particularly as it relates to voxelwise segmentation accuracy. As mentioned to Reviewer 1, VDP (as used in this diagnostic prediction scenario) is a lossy distillation of the segmentation information. We discuss this further in the* **Discussion** *section:*

As illustrated in Figure 2, measures based on the human visual system seem to quantify what is understood intuitively; that image-based information is much more robust than its corresponding histogram-based information in the presence of image transformations, such as common MR artefacts. This observation is not intended to imply that the histogram-based approaches are useless in performing research. In fact, ventilation defect percentage is perhaps the most widely used clinical measurement reported in the literature and it is easily quantified from the image histogram. Thus, even relatively simple histogram-only segmentation algorithms will provide some utility which was observed in the measurement bias experiments employing a variant of ventilation defect percentage to predict diagnostic accuracy. However, similar to the lossy relationship between the image and its corresponding histogram, such volumetric-based measures are lossy distillations of the segmentation information and might obscure important algorithmic characteristics and relative differences as well as discard potentially useful spatial information which is why additional experimentation explored measurement precision in the presence of MR artefacts.

Minor Comments:

1. Most of the figures were displayed prior to their first reference in the text, which made the paper hard to follow.

*Done. As requested by the editors, the figures were taken out of the main body and uploaded separately.*

2. the word "Early" is repeated twice in the first sentence of the introduction.

*Fixed.*

3. Page 5, lines 7-12, starting with "These algorithms have the advantage", seems to have a typo that makes this sentence hard to understand.

*Fixed.*

4. Figure 1- What are the numbers listed on the left size - what does Max SD stand for? Greater description of the intensity warping and what each panel shows n the caption would be useful.

*Fixed. As these numbers are not relevant to the discussion in the Introduction, we simply relabeled them as “Simulation 1,” “Simulation 2,” and “Simulation 3.” We also revised the caption to read as follows:*

Figure 1: Illustration of the effect of MR nonlinear intensity warping on the histogram structure using a representative sampling of the simulations used in the experiments in this work. By simulating these types of nonlinear intensity changes, we can visualize the effects on both the image and the corresponding intensity histograms and investigate the effects on salient outcome measures. These simulated intensity mappings, although relatively small and difficult to distinguish in the image domain, can have an algorithmically consequential effect on the histogram structure.

5. Figure 3- One of the useful features of some of the algorithms, such as linear binning, is that their published descriptions include a color scale that is intuitive - red is low, green is normal, blue is high. While such color scaling is non-essential for the current work, using intuitive colors for segmentation maps would make images more clear and improve the likelihood of eventual clinical application.

*A clinically intuitive color scheme would definitely be beneficial to the field (obviously assuming widespread acceptance). We are aware of the various preferences of the different groups in the field and would be very supportive of working towards a consensus. The aesthetic choices associated with the segmentations illustrated in this manuscript were made outside of any such considerations (and with absolutely no intent to promote a particular scheme to the community). Specifically, the first author, being a long-time ITK/ANTs developer and ITK-snap user, simply employed the ITK-SNAP default color scheme. This facilitates reproducibility for anyone interested in viewing the results while also avoiding having to make difficult decisions in reconciling the different number of clusters used by each algorithm, each with possibly different clinical meaning.*

6. On page 12, you state "none of the evaluations use these categorical definitions in a cross-algorithmic fashion", but in figures 7 and 8, you compare the cluster volumes across the various algorithms. Moreover, on page 19 going into page 20, you state "This allowed us to compare between algorithms". Can you clarify this apparent contradiction?

*In reference to “evaluations us[ing] these categorical definitions in a cross-algorithmic fashion”, we were specifically referring to the use of STAPLE[[5]](#footnote-4) or related approaches to bootstrap consensus-based labelings as ground-truth approximations. We alluded to this in the previous version of the* **Results** *section:*

In addition to the fact that the number of ventilation clusters is not consistent across algorithms, it is not clear that the ventilation categories across algorithms have identical clinical definition. This prevents application of various frameworks accommodating the lack of ground-truth for segmentation performance analysis (e.g., (48)) to these data.

*However, to avoid any confusion, we removed the text in question.*

7. Figures 7 and 8 - In the Tukey HSD plots, some of the lines are solid, and some are dotted, but there is no legend or caption to explain what this means. Additionally, these figures show a lot of data and are, at first glance a little hard to understand. The caption would benefit from a little more description. It may be useful to highlight some of the comparisons that show large positive values to break up the plot and make it easier to read.

*We separated these two figures into four separate figures and added explanatory text to both the captions and within the manuscript. The Tukey HSD plots were lengthened and further edited to better group the results according to the specific pairwise comparison. The caption included additional text to explain the correspondence between statistical significance and line type. We also added alluvial diagrams to more easily visualize the relative performance relationships between the algorithms quantified with Tukey’s testing. For example, Figure 8 is now as follows:*

|  |
| --- |
|  |
| Figure 8: University of Virginia image cohort. (Left) Results from Tukey’s test following one-way ANOVA to compare the resulting overlaps between algorithms (cf Figure 7). Higher positive values indicate increased robustness to simulated image distortions. A solid line indicates statistical significance at the 0.05 level whereas the dashed line indicates no statistically significant difference. (Right) To further visualize the Tukey results, a simplified alluvial diagram is used to provide connections illustrating relative performance between algorithms where the algorithms listed on the left have improved performance relative to their connected algorithms on the right with the width of the connection being proportional to difference in performance. |

8. In the Discussion, page 21, you state, "with its questionable assumption of Gaussianity". In reference 43, though, the use of non-Gaussian reference distributions is explore

*We removed the text in question based on this and other issues brought up by the reviewers/editors. We are aware of the explorations in the cited reference in which the Box-Cox transform is used to convert a non-Gaussian reference distribution to one more Gaussian-like in order to apply the standard linear binning algorithm. It is this ultimate requirement of a Gaussian distribution for linear binning which we (or, at least the first, author) find questionable.*

9. No true conclusions are provide at the end of the manuscript. A summary of the main takeaways would be a useful addition.

*Interestingly enough, the working title of the paper was “Histograms should not be used to segment functional lung MRI,” which would constitute the main takeaway. This verbiage was softened in multiple places within the manuscript. Specifically, please see the revised* **Discussion** *section mentioned in response to Item 4. Several other secondary takeaways are discussed in this section as well.*

1. *An analogy can be drawn with the brain imaging community where population templates (reference distributions of spatially-defined voxel intensities) are constructed by combining brain images from normal individuals which have a close approximation to the representative template. During template generation, clinical features (e.g., ventricles, cortical regions) are aligned using nonlinear registration. However, if one were to simply perform voxelwise addition without nonlinear alignment, one would just end up with a bright blob with little or no discernible features.* [↑](#footnote-ref-0)
2. [*https://pubmed.ncbi.nlm.nih.gov/33080507/*](https://pubmed.ncbi.nlm.nih.gov/33080507/) *and* [*https://pubmed.ncbi.nlm.nih.gov/30908194/*](https://pubmed.ncbi.nlm.nih.gov/30908194/)*, respectively.* [↑](#footnote-ref-1)
3. *https://onlinelibrary.wiley.com/pb-assets/assets/15222594/Style%20Guide%2012.2.19-1575389345850.pdf* [↑](#footnote-ref-2)
4. *https://onlinelibrary.wiley.com/pb-assets/assets/15222594/Style%20Guide%2012.2.19-1575389345850.pdf* [↑](#footnote-ref-3)
5. *https://pubmed.ncbi.nlm.nih.gov/15250643/* [↑](#footnote-ref-4)