We are grateful for all the reviewers for their constructive feedback, which will undoubtedly improve the quality of the manuscript. As four reviewers acknowledge, we present a novel segmentation approach to learning robustly in the presence of large expert disagreement in annotations, and demonstrate its utility in a range of datasets from medical imaging where such problem is particularly pertinent. Our work is the first instance of an end-to-end supervised segmentation method that simultaneously estimates the reliability of individual annotators and the true segmentation distribution from noisy observations alone. We will open-source the code and label simulator upon publication.

The main criticisms from the reviews are summarised as:

(1) **R1.3.1:** Relevance of STAPLE and its variants as baselines is unclear

(2) **R1.3.2:** The overall goal needs clarification: is it segmentation accuracy or the utility of modeling the reliability of annotators in downstream tasks?

(3) **R2.3.1:** The method and theoretical contributions are very close to those from [28].

(4) **R2.3.2:** Need to include the *"global"* confusion matrix (CM) based model in the performance comparison.

(5) **R3.3.1:** Assuming annotations of different pixels are independent might limit the model performance.

(6) **R3.3.2:** The LIDC-IDRI dataset has little inter-reader variability, and may be more credible to test on a dataset with higher variability e.g., Gleason'19.

(7) **R4.3.1:** All annotations are simulated, and the method should be evaluated on a dataset with real noisy annotations.

Reg. (1), STAPLE is the most prevalent label fusion framework used in the curation of training datasets for medical image segmentation (including many public benchmarks), where the initial labels are typically noisy. The advancement made in our work can also be viewed as a translation of the STAPLE framework to the supervised learning setting. Thus it is necessary to know how much improvement the learning based method achieved compared to such label fusion methods.

Reg. (2), our primary goal is to learn to segment unseen images accurately when the annotations in the training data are very noisy. We attain this by modelling explicitly the reliability of the individual annotators, and as R1 highlights, such information can be potentially used in downstream applications e.g., education and active label acquisition. However, this remains future work, and to clarify, we have removed the last sentence of the abstract and expand on the future outlook in discussion.

Reg. (3), while the work [28] builds on their conceptual framework, extending to segmentation demands substantial technical differences. Firstly, to capture the correlations involved in image segmentation, the method yields the estimate of annotator reliability in every pixel, which is crucial in capturing the complex spatial variations in annotators' characteristics, and is absent in [28] which only addressed image classifications. Secondly, we modelled the reliability of annotators as a function of the input image in a stark contrast with [28] that merely estimated the population average as additional parameters — this enables estimating the annotator's reliability on a per example basis. Lastly, the theoretical justification for the trace regularisation was extended from the "global" confusion matrix scenario to the "local" setting—this is not trivial and we could only prove a weaker (but sufficient) statement that the relevant column of the CMs is recoverable but not the entire matrices. We will state these differences more clearly in intro and section 3.

Reg. (4), we agree with R2 that it is critical to include such a baseline to assess the benefits of modelling CMs as a function of the image—thank you for the suggestion. We have implemented this and included its performance in all 8 quantitative comparisons. While performing better than mean/mode label baselines, the table below shows the global CM model performs consistently worse than our model based on the image-dependent pixel-wise CMs with differences indicated in red:

Dataset	DICE (dense labels)	CM estimation, mse (dense labels)	DICE score (1 label per image)	CM estimation, mse (1 label per image)
MNIST	79.21 ± 0.41 (-3.71)	0.1132 ± 0.0028 $(+0.0239)$	59.01 ± 0.65 (-17.47)	0.1953 ± 0.0041 $(+0.0624)$
MS	61.58 ± 0.59 (-5.97)	0.1449 ± 0.0051 $(+0.0638)$	40.32 ± 0.68 (-16.11)	0.1974 ± 0.0063 $(+0.0432)$
BraTS	47.33 ± 0.28 (-6.14)	0.1673 ± 0.1021 $(+0.0488)$	41.76 ± 0.71 (-4.45)	0.2419 ± 0.0829 $(+0.0843)$
LIDC-IDRI	70.94 ± 0.19 (-3.18)	0.1386 ± 0.0052 $(+0.0935)$	63.25 ± 0.66 (-4.87)	0.1382 ± 0.0175 $(+0.0795)$

Reg. (5), we agree with R3 that independence assumption between pixels in annotation might limit the performance. We note, however, that the annotations are assumed to be only conditionally independent between pixels given the input image, and thus the model can still capture some correlations in the output segmentation labels that are explained in the input image. We additionally note that such independence assumption is typically made in most of the deep learning based segmentation methods, and thus a posprocessing method such as Gaussian CRF is commonly used to capture the missed correlations. We believe the same problem applies to the annotation modelling—we note this limitation and mention such correlation modelling as future work in the discussion.

Reg. (6), the LIDC-IDRI dataset contains some cases with very high inter-reader variability, and we will clarify in section 4. For example, Fig.14 in the supplementary material shows that many examples have low consensus levels in the range of 30%-50% mIoU between annotators. The third row in Fig.7 also shows one such example where Annotator 4 completely misses the abnormality. We believe that our initial promising results motivate us to test the method on an even more challenging dataset such as Gleason'19 in the future. However, we do note that Gleason'19 does not have curated ground-truth and resort to a label fusion method to create "gold standard" labels. An inspection with our clinical collaborators has revealed that many of such "gold standard" are not realistic, implying that an extra care is needed.

Reg. (7), we have in fact evaluated our method with real annotations—the LIDC-IDRI dataset contains annotations per input from 4 different radiologists. Unlike the MNIST, MS and Brats datasets with synthetic noisy labels, we use LIDC-IDRI to evaluate the utility of our work in the presence of real-world noisy labels. We will clarify in Section 4.