Dear Dr Peiro Lo Monaco,

Thank you for your email and the opportunity to revise our paper on "***Liver Shape Analysis using Statistical Parametric Maps at Population Scale***". We also appreciate your comments, and the suggestions offered by the reviewers have been very useful.

I have included the Reviewer's comments and our responses immediately after this letter. All modifications in the manuscript have been highlighted in yellow.

Yours sincerely,

Marjola Thanaj

Reviewer #1

The authors have mostly addressed my concerns regarding the predictive potential for S2S distances towards identifying disease state by adding a matched case control analysis. However, in Table 4, the authors should also report sensitivity/specificity in addition to the F1 score so that predictive potential can be evaluated under consideration of disease prevalence. Furthermore, the F1 score depends on class balance and may generalize poorly across cohorts. Additionally, p-values from a statistical test such as the Delong test should be added to support the claim that S2S distances improve AUC in prediction of liver disease, since the overlapping confidence intervals in Figure S7 do not necessarily imply a statistically significant difference.

Response:

- **We appreciate your feedback and have made the necessary amendments to address your concerns. In Table 4, we have included accuracy, sensitivity and specificity alongside the F1 score to provide a more comprehensive evaluation of predictive potential, taking into account disease prevalence. We have also performed a statistical test, specifically the Delong test, and reported the corresponding p-values to support our claim that S2S distances improve the AUC in the prediction of liver disease. We thank you for your valuable input.**

However, in my opinion, the authors have not responsively addressed concerns regarding quality control and segmentation errors. In their response, the authors cite a paper previously published by their group claiming that quality control was therein thoroughly conducted. However, this reviewer notes that the paper cited (Liu et al., ELife 2021) does not report any details of liver segmentation performance or quality control thereof. While the authors claim that experientially, "we have confidence in [our pipeline's] ability to produce accurate segmentations," this cannot be taken as scientific evidence towards its reliability. It should be noted that both reviewers requested the authors to elucidate the reliability of S2S distances considering sensitivity to errors and variabilities associated with (i) segmentation [R1+R2], and (ii) registration [R1]. This reviewer notes that while additions were made to the manuscript to clarify the quality assurance procedure, no attempt was made to rebut, evaluate, or discuss the effects of segmentation and registration errors on the reliability of the method. This reviewer believes that these sources of segmentation and registration errors are nonnegligible and are likely to corrupt the reliability of S2S distances, especially when attempting to make predictive inferences at the patient-specific level. I repeat that given the modest magnitudes of the beta coefficients, the model is likely quite sensitive to segmentation and registration errors. The authors must make an honest attempt to address these concerns.

This said, the authors did effectively assuage my other concerns outlined in the previous review.

Response:

- **Thank you for your thoughtful comments and concerns regarding the quality control and segmentation and registration errors in our study.**
- **Regarding the quality control discussed in a previous publication by Liu et al., eLife 2021, the approach included a rigorous iterative process involving visual inspections of extreme volumes for each distinct organ/structure and spot checks of hundreds of random subjects. The training data were continuously enriched to include problematic cases, and this procedure was repeated until the results no longer displayed outliers for extreme subjects or any of the random spot checks. The performance metrics for the liver segmentations can be found in the supplementary material from our previous publication, including metrics such as Dice similarity, Jaccard index, negative predictive value, precision, recall, and specificity for the Dixon liver segmentation Dixon Liver (validation set): Dice Similarity 0.932, Jaccard 0.873, Negative Predictive Value 1.000, Precision 0.936, Recall 0.930, Specificity 1.000). These metrics, in addition to the visual quality control, demonstrate the robustness and reliability of our segmentation procedure. Hence, this rigorous quality control process ensures that the effects of any residual segmentation errors are minimised. These details were not included in the current manuscript considering them already comprehensively described in our previous paper (Liu et al., eLife 2021). Readers interested in this area of the research are now pointed towards this data in the manuscript.**
- **Regarding possible registration errors and the potential impact on the reliability of S2S distances. It is important to note that when employing the mass univariate regression framework with S2S distances as the response variable, the analysis extends across a substantial cohort. In this context, a statistically significant regression coefficient can represent a much smaller unit of change when compared to what might be apparent at an individual level, as the model is based on an average effect across a large cohort. Hence, we believe that any potential impact of segmentation inaccuracies on the S2S distances would be minimal.**
- **We believe that our quality control procedures and the robustness of our analysis framework, which operates at the cohort level, provide confidence in the reliability of our results.**

Reviewer #2

All review comments have been adequately addressed by the authors. The addition of the predictive analysis also adds a critical dimension to the manuscript in addressing questions of the practical clinical utility of the S2S distances. Although this was preliminary and the differences in the volume and S2S models were not statistically significant, it still shows potential use as an additional biomarker and should be followed up in future studies along with longitudinal analyses.

Response:

- **Thank you for your positive feedback. We appreciate your insightful comments and agree with the predictive analysis. We fully intend to explore this further in future studies, including longitudinal analyses, to better understand the role of S2S distances as an additional biomarker. Your feedback is encouraging and will guide our future research in this direction.**