Title: Liver Shape Analysis using Statistical Parametric Maps at Population Scale

Version: 4 **Date:** 06 Sep 2023

Reviewer's report:

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.

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 non-negligible 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.