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

Version: 2 **Date:** 17 Apr 2023

Reviewer's report:

In this paper, the authors apply a registration-based shape comparison model to a very large liver imaging cohort from the UK Biobank. Under a guiding premise that variations in liver shape could potentially inform risk factors and etiology of liver disease, the authors use regression analysis to show that patient-specific deviations in liver shape from a template anatomical atlas are associated with certain clinical findings including type 2 diabetes, fatty liver, AST/ALT ratio, fibrosis score, etc. While the scope of the authors' work should be commended and the clinical methodology seems sound, there are several limitations and gaps in clarity that this reviewer believes are not adequately recognized.

Major Critiques:

The authors only present a purely descriptive regression analysis. It is a major shortcoming that no exploration of predictive capacity is performed. How well could these variations in liver shape be used as imaging biomarkers to quantitatively inform disease state? For example, could a patient be matched against the template to predict their disease status? Are the effect sizes of the regression model large enough to translate prognostically?

It is not clear to the reviewer what is the actual clinical utility of this study. What additional value can be demonstrated with the authors' shape-driven approach? Can it be shown that basic volumetry of liver segmentations fails to produce similarly significant regression coefficients to clinical variables?

Besides assessment of the final S2S distances, what quality assurance measures were put in place to screen and control deep learning segmentation errors? State-of-the-art deep learning methods for the liver are typically associated with maximum segmentation errors still on the order of several voxels. With expected segmentation errors of a few millimeters, the true variances surrounding the beta estimates are likely larger than the regression model would estimate. Furthermore, given the modest magnitudes of the beta coefficients, the model is likely quite sensitive to segmentation and registration errors.

From this reviewer's understanding, p-values provided by the TFCE procedure at each vertex do not imply spatially localized statistical significance. A significant p-value at a vertex only indicates that there exists at least one cluster threshold (out of all possible cluster thresholds) where the vertex belongs to a significant cluster. Therefore, significance area may be a misleading measure of the spatial prevalence of the effect, and the contour boundaries in Figure 1 would similarly only be approximately representative. It is not clear why S2S significance *area* should serve as the endpoint for characterizing the shape effect.

Other Comments:

The introduction should more clearly stress the contributions of the current manuscript compared to the already published prior work (e.g. ref. 26, 27 of manuscript).

How was the S2S cutoff for the manual quality assurance step (L181) determined? Since it is not practical to manually review over 33,000 segmentations and registrations, a discussion towards the nature of and the way segmentation and registration errors within the unverified cases may affect the analysis is warranted.

L172: “… each vertex is anatomically accurate and consistent across all subjects”. This cannot be guaranteed due to the use of label consistency similarity for registration. Label consistency will minimize distance between the boundaries of the template and subject segmentations, but will not produce accurate correspondence in the tangential surface directions. For example, a vertex at the falciform ligament of the template mesh will not necessarily correspond to the falciform ligament of the subject after registration.

It is unclear whether ANTs is also used for the template-to-subject registrations or if another more time-efficient deformable image registration algorithm was used.

What is the rationale for investigating each of the interaction terms in regression Model 2? What additional information can be drawn from including these, especially considering the wavering significance of the interaction terms outlined in L433-437?

The paragraphs starting on L412 and L445 of the discussion could better summarize the key regional variations seen in liver shape associated with each of the significant clinical variables. To this reviewer, the paper falls short in clearly demonstrating how the nuances of regional shape changes of the liver might overcome current limitations to improve the ability to define or track disease state.