Dear Prof. Jack H. Noble,

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 comments and our responses immediately after this letter. All modifications in the manuscript have been highlighted in yellow.

Yours sincerely,

Marjola Thanaj

Reviewer #1

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:

1) 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?

Response:

- **We appreciate your valuable feedback and your suggestion to explore the predictive capacity of our model. We agree that this is an important aspect to consider and we have now included it in our analysis.**
- **We have conducted additional analyses to evaluate the predictive power of our model and investigated whether the emerging 3D liver mesh-derived phenotype can add to the prediction of disease outcomes. Specifically, we identified participants with liver disease and T2D that were diagnosed after the imaging visit and identified a control-cohort without any reported conditions, matched with every case by age, gender and BMI and designed a case-control study for each disease population. Due to having a large number of S2S values for small population groups, we first calculated the sparse PCA and extracted 40 principal components modes of the S2S distances for each disease case-control group, as they were sufficient to describe over 90% of the S2S distances in both cohorts**

(which were sufficient to capture 90% of the cumulative variation for each cohort). We then performed two logistic regression models where in the first model (the volume model), the disease outcome was regressed on age, gender, ethnicity, BMI, WHR, AST/ALT, FIB-4 index, liver volume, PDFF and iron concentration and in the second model (the S2S model), we included all the covariates from the volume model, adding 40 principal component scores of the S2S values for each disease group. We have included this analysis in our revised manuscript. Thank you again for your insightful comments, which we believe will greatly enhance the scientific value of our work.

2) 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?

Response:

- **Thank you for your valuable questions. In response to your first comment, we agree that exploring the utility of liver shape variations as imaging biomarkers demonstrates the potential clinical value of our approach. We have conducted additional analyses to assess the predictive performance of our model using our emerging 3D liver mesh-derived phenotype and compare it to other standard approaches, such as volume of liver segmentations. We demonstrate that the model using the principal component scores of the S2S distances improved the prediction of liver disease; however, there was no improvement in T2D compared to the model with liver volume. We show that our methods may provide additional information in monitoring disease and potentially predicting outcomes.**
- **Regarding your second point, by utilising the statistical parametric maps (SPMs), the main aim is not to directly compare the results with the volume measurements, but rather to investigate the potential value that can be derived from the high-dimensional representation of the liver shape in relation to relevant clinical variables. This approach allows us to assess if there are spatial patterns within the liver that demonstrate the associations with clinical variables. While basic volumetric measurements remain valuable in certain clinical contexts, our study aims to demonstrate the potential utility and added insights provided by considering liver shape. By identifying spatial patterns and statistically significant regions in the SPMs, we gain a deeper insight into the relationship between liver morphology and relevant clinical variables. We have included these observations in our revised manuscript. Thank you for bringing up these important points and helping us to improve our work.**

3) 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.

Response:

- **Thank you for your comments regarding the quality assurance measures. In our previous study (Liu Y, Basty N, Whitcher B, Bell JD, Sorokin EP, van Bruggen N, et al. Genetic architecture of 11 organ traits derived from abdominal MRI using deep learning. Elife. 2021;10), we conducted rigorous quality control for our segmentations and reported those procedures and our quantitative metrics.**
- **Regarding the comment on the segmentation errors, our pipeline has been extensively evaluated and we have confidence in its ability to produce accurate segmentations. We further employed a rigorous quality control process to minimise the effects of any residual segmentation errors. We have provided more details on the quality control in our revised manuscript.**
- **With respect to the regression model, we have conducted a rigorous hypothesis testing procedure with permutation testing and adjusted for relevant variables with correction for multiple comparisons while considering the spatial dependence to characterise the statistical output. The regression models effectively summarise the impacts of variables across a large sample size of over 33,000 participants and are resilient to minor levels of noise.**

4) 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.

Response:

- **We thank the reviewer for their thoughtful comments regarding the interpretation of p-values provided by the TFCE procedure and the use of significance area as an endpoint for characterising the shape effect. A previous study (*Smith, Stephen M. and Thomas E. Nichols. "Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference." NeuroImage 44 (2009): 83-98.*), reports that the TFCE approach aims to enhance areas of signal that exhibit some spatial contiguity without relying on hard-threshold-based clustering. The authors also report that the TFCE method keeps the benefit of cluster-based thresholding while minimising the issues occurring from hand-tuning and pre-smoothing of the signal providing better sensitivity and stability than other cluster-based methods. The authors have also compared the TFCE procedure with a cluster-based thresholding in brain images from schizophrenic patients compared matched control adolescents and found that TFCE finds major areas in Heschl's gyrus / parietal operculum and in the supplementary motor area, whereas, cluster-based thresholding does not find the smaller areas, does not find supplementary motor area, and the areas that are found have lower statistical significance than the equivalent peaks found by TFCE. Not only is TFCE more sensitive than cluster-based thresholding in this example, but it also retains information about relative significance of effect within the reported areas of**

group difference and provides information regarding local maxima in the final significance map, which is not possible with cluster-based thresholding. The authors of the aforementioned publication, also report that the region in the data that causes a voxel to reach that level of significance, is obtained by defining the smallest area of the local neighbourhood that contributes to the TFCE score at that voxel, resulting in statistical significance. This area can be therefore interpreted as the region in the original data that contributes to the result at the voxels of interest.

- **Hence, we chose to use significance area as a widely known term (*de Marvao A et al., Precursors of Hypertensive Heart Phenotype Develop in Healthy Adults: A High-Resolution 3D MRI Study, JACC: Cardiovascular Imaging, 2015, 1260-1269, Biffi C, de Marvao A, Attard MI, Dawes TJW, Whiffin N, Bai W, et al. Three-dimensional cardiovascular imaging-genetics: a mass univariate framework. Bioinformatics. 2018;34:97–103*), to demonstrate the overall spatial distribution of the effect across the liver surface. We have provided additional details on the interpretation of the significance area in our revised manuscript.**

Other Comments:

1) 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.

Response:

- **Thank you for your comments. The current manuscript contributions to the field are threefold. Firstly, we investigate the impact of the population size and the robustness of the construction of the liver template. Secondly, we extended the use of SPMs of the liver images from a large population dataset of over 33,000 individuals and investigated the impact of relevant anthropometric, phenotypic, and clinical factors on the regional geometry of the liver. Lastly, we extracted shape features derived from the 3D mesh-derived phenotype by dimensionality reduction and evaluated whether these shape features were better predictors of disease outcomes than the conventional volumetric measurements. We have emphasised these contributions in our revised manuscript.**
- **Regarding the S2S cutoff for manual quality assurance, we agree that this should have been more clearly explained in the manuscript. We simply report S2S summary statistics after applying quality control. We have provided more details on quality control in our revised manuscript.**
- **Regarding the potential effects of segmentation and registration errors within the unverified cases, we acknowledge that this is an important consideration. While it is not practical to manually review all cases, we have applied a quality control process and excluded cases that did not meet our criteria. We have addressed these points in our revised manuscript.**

2) 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.

Response:

- **We thank you for your insightful comments. By using label consistency, we aim to achieve approximate associations with the same anatomical locations. It is a valuable metric that provides a more comprehensive evaluation of the correspondence between template and subject shapes (*A. F. Frangi, D. Rueckert, J. A. Schnabel and W. J. Niessen, "Automatic construction of multiple-object three-dimensional statistical shape models: application to cardiac modeling," in IEEE Transactions on Medical Imaging, vol. 21, no. 9, pp. 1151-1166, Sept. 2002, doi: 10.1109/TMI.2002.804426*). However we acknowledge that the term "anatomically accurate" may not be entirely precise in this context. Hence, we will make the necessary correction in the wording to reflect that the correspondence between vertices is approximately accurate rather than claiming anatomical accuracy. We appreciate your keen observation and ensuring the clarity and accuracy of our statements is essential to us.**

3) 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.

Response:

- **Thank you for your comments. We have used ANTs for the template construction and IRTK for the template-to-subject registrations as it has provided more suitable functionalities for the template-to-subject registration. We have provided further clarification in our revised manuscript.**

4) 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?

Response:

- **Thank you for your feedback. As this is an exploratory study, our aim was to investigate a wide variety of statistical hypotheses. Interaction terms using disease status provide a way to investigate the potential for different linear relationships between disease-based subsets of the UKBB participants; e.g., does the S2S distance in participants with type-2 diabetes differ with age to those participants without type-2 diabetes and similarly for liver PDFF. While some of the interaction terms have not yielded statistical significance, this is still informative and we believe that including them provides a more comprehensive understanding of the relationships between variables in the model and informs future research endeavours.**

- **Additionally, we believe that the lack of statistical significance in the interaction with liver disease, compared with the interactions that involve T2D, provides insight into potential morphological variations that are disease specific. Overall, we deemed it is valuable to include all interaction terms for a more comprehensive analysis.**

5) 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.

Response:

- **Thank you for your feedback. We agree that it would be useful to summarise the key regional variations associated with each of the significant clinical variables. We have revised the manuscript accordingly to make it more clear.**

- **Regarding the ability of regional shape changes to improve the ability to define or track disease state, we acknowledge that this is an important area for further research. Our study provides evidence for regional differences in liver shape associated with disease, but further work is needed to determine how these changes can be utilised to improve diagnosis or monitoring of disease progression. We plan to explore this topic in future research and have included a discussion of potential applications in our manuscript revisions.**

Reviewer #2

The manuscript presents a novel morphometric image analysis method for the analysis of liver shape and size, and presents the associations between shape variations from a population template with several anthropometric variables, liver IDPs and disease. The associations are investigated on a population of over 33k participants — including participants with diagnosed liver disease — which is a commendable sample number for statistical conclusions, and the authors appropriately discuss the findings in the context of previous literature. Ample information is provided about the data and the statistical methods used, and the writing is easy to comprehend, demonstrating a high level of clarity and coherence.

The significance of the research question in quantifying spatial variations in liver shape is clearly explained. The statistical analysis of the data is sound, and the use of 3D SPM visualizations is a valuable tool to localize areas of morphometric difference.

However, the paper could be revised to address potential limitations:

Major changes:

1) The analysis uses previously published segmentation methods. The sensitivity of the S2S distance to the segmentation error must be clarified. Additionally, in the selection of templates, differences less than 8 mm were deemed to be not important — the authors should clarify how

much this error might be compounded by the segmentation error and whether that can cause some differences to be impactful.

Response:

- **We appreciate the reviewer's feedback. Regarding the sensitivity of the S2S distances on the segmentation errors, we acknowledge that these errors can affect the accuracy of the S2S distance however, in our previous study (*Liu Y, Basty N, Whitcher B, Bell JD, Sorokin EP, van Bruggen N, et al. Genetic architecture of 11 organ traits derived from abdominal MRI using deep learning. Elife. 2021;10*), we have conducted quality control for our segmentations and we reported procedures and our quantitative metrics. Our pipeline has been extensively evaluated and we have confidence in its ability to produce accurate segmentations.**
- **Regarding the selection of templates, we understand the importance of considering small differences that may affect the S2S distances. In our revised manuscript, we have provided further clarification on this matter. Specifically, we have quantified the differences between the 20-participant and 200-participant templates, as well as between the 100-participant and 200-participant templates. For example, the median absolute distances between the 20-participant and 200-participant templates was found to be -1.1 (IQR: 3.2) mm, indicating relatively small variations. We have also mentioned that the distances between the 100-subject and 200-subject templates were even smaller (median: -0.4 (IQR: 1.8) mm).**

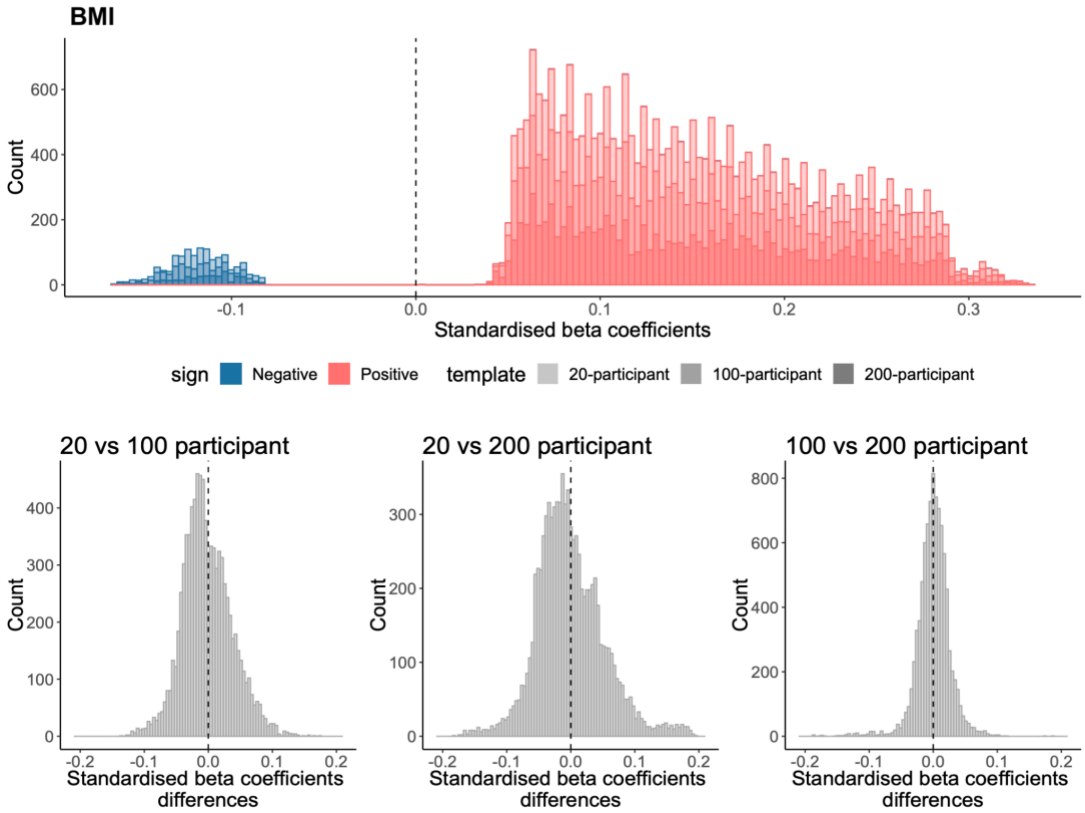
2) The statistical methods to compare the 3 templates and the selection of one needs to be more rigorous. On top of considering the total significance areas — which do not seem to have a spatial component — the authors should add an overlap metric on the 3D meshes to quantitatively demonstrate the difference (or lack thereof) of the statistically significant areas between templates.

Response:

- **Thank you for your valuable feedback. We appreciate your suggestion to quantitatively demonstrate the differences of the statistically significant areas between templates. In our analysis, we have visually presented the 3D SPMs, with the TFCE corrected p-values, of BMI and WHR with the S2S distance on the 500-participants cohort (Supplementary Fig. S3) as well as the cohort with liver disease (Supplementary Fig. S4). Additionally, we have provided tables showing the significance areas of their associations across the three templates (Table 1 and Table 2). By combining visualisation and presentation, we have conducted a comprehensive qualitative and quantitative analysis. Our results demonstrate that the distribution of the corrected p-values were consistent across all three different templates and that there was no apparent difference in the areas of association between BMI and WHR with S2S distances across the three templates. We have amended our manuscript accordingly to make it more clear.**

- **Furthermore, to ensure a comprehensive comparison between templates, we have taken several steps. We first have aligned each template to the same space,**

which allows for a consistent and comparable analysis. To identify the position differences between templates, we have utilised the closest points approach. Regarding the significance areas, we agree that it is important to consider not only the total significance areas but also the spatial component. We have taken this into account and have analysed the beta coefficients within the significance areas for the associations between BMI and the S2S distances on the 500-participants cohort. By finding the closest points between the templates, we were able to determine the spatial variations within the significance areas (covering a significance area ~ 50% of the liver for all three templates) and compute the differences of the beta coefficients across different participant groups (e.g., 20 vs 100, 20 vs 200, and 100 vs 200 participants, see plots below). Note that the significance areas of the three templates overlapped on 8,757 vertices representing over 80% the significance areas for each template. By examining the differences within the significance areas, we observe small differences of the beta coefficients within the statistically significant areas between the templates (-0.006 (0.05) (median (IQR)), for 20 vs 100 participants; -0.007 (0.07), for 20 vs 200 participants; -0.0009 (0.03), for 100 vs 200 participants). Thank you again for your insightful comments, which we believe will greatly enhance the scientific value of our work.



3) Additionally, the choice of subjects for each of the 3 templates is not clear. Since the selection of the template is a significant step towards quantifying the variations in liver shape, and this selection is done based on the number of samples, presumably the specific subjects are not relevant — therefore, for statistically sound conclusions, the template construction process should be repeated multiple times (at least 5) at each N (20, 100 and 200) with different samples from the population, and the average metrics should be taken from each to

compare. This can help avoid sampling bias and to add robustness to the S2S measurements compared to the template.

Response:

- **We thank the reviewer for their thoughtful comments regarding the choice of subjects for each of the three templates. We agree that the choice of subjects for each template should be made more clear. Taking into account the reviewer's suggestion, we have addressed this by constructing 5 templates, each constructed from different samples drawn from the population of 20 participants each. The Dice coefficients of the template images for the 20-participant template experiment consistently demonstrates a high level of overlap across the distinct cohorts (see table below). We have included a table in our supplementary material showcasing these results and we have incorporated a statement in the results section to acknowledge this.**
- **It is important to note that for constructing templates using larger cohort sizes (i.e., 100 or 200 participants), it is expected that the variability will be reduced due to the averaging effect. Based on these findings, we are confident in the robustness and consistency of our template construction process.**

Samples	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
Sample 1					
Sample 2	0.924				
Sample 3	0.942	0.947			
Sample 4	0.933	0.94	0.951		
Sample 5	0.917	0.939	0.941	0.938	

The following recommendations are in the realm of major changes; however they are not binding since they can reasonably be out of scope:

1) The methods can be strengthened even further by repeating this process across multiple N values beyond 200 as well and plotting metrics of difference against sample size. While this might be out of scope for this manuscript, it is an important step to be able to select an optimal sample size.

Response:

- **Thank you for the suggestion. We agree that exploring the effect of different sample sizes on the template selection process is an important step and could provide valuable insights. However, this might be beyond the scope of the current manuscript. We plan to explore this in future work and will consider your suggestion when designing our future studies.**

2) The experimental results compare shape variations across the population; however, it lacks a longitudinal component. Especially given how the variations in liver shape are non-uniform and not fully elucidated, the correlation of patient specific changes in liver shape to disease

progression over time cannot be fully concluded in this study. This should be addressed in the discussions and next steps as necessary follow up studies before this method can be used in tracking disease progression in specific patients.

Response:

- **Thank you for your comment. We completely agree that a longitudinal study is needed to investigate the correlation between patient-specific changes in liver shape and disease progression over time. We agree that this is an important aspect to consider and we have included it in our analysis using a small number of longitudinal data.**
- **We have conducted additional analyses to evaluate the predictive power of our model and investigated whether the emerging 3D liver mesh-derived phenotype can add to prediction of disease outcomes. We have included these results in our revised manuscript and we have addressed this limitation in our discussion and highlighted the need for future longitudinal studies to validate the use of liver shape analysis in tracking disease progression in specific patients.**

Additionally, the following minor changes are recommended:

Minor changes:

1) Please list the default parameters for the template construction for full reproducibility

Response:

- **Thank you for your suggestion. We have incorporated this into our revised manuscript.**

2) Line 181: Please clarify which steps constituted the “manual quality control on the S2S values”.

Response:

- **Thank you for your comment. We have provided more details on the quality control in our revised manuscript.**

3) Line 396: Please add brief clarification on how the methods described in the paper can be used as a “…powerful adjunct tool in clinical trials…”

Response:

- **Thank you for your comment. We apologise for any confusion caused by the statement regarding the methods described in the paper as a "powerful adjunct tool in clinical trials". Upon reconsideration, we acknowledge that this statement may not be adequately supported by the content of the paper. We will remove this sentence to ensure clarity and accuracy in our discussion of the methods**

4) Figure S2: Please change the x-axis labels to “20 vs 200 participants” and “100 vs 200 participants” to clarify that the plots are based on differential metrics. Overall, this is a high-quality paper that contributes to the field of novel biomarkers for tracking liver disease.

Response:

- **Thank you for your suggestion and positive evaluation of the paper. We have incorporated this into our revised manuscript.**