We thank the reviewers for their insightful comments and agree with all points related to clarity of the terminology, notation, and image quality. We would correct all these points for any camera ready copy of the manuscript.

R2 is correct in suggesting that the ultimate goal of machine learning for healthcare should be explainable models. However, interpretability and explainability need not be mutually exclusive. For complex neurological and psychiatric disorders, where the causes are unclear and symptoms are highly heterogeneous, development of explainable models may be non-trivial, arguably requiring prior hypotheses of the types of variation expected. It is here where interpretable models such as ICAM can help. At the same time, ICAM may be positioned alongside other radiological support tools, such as Mckinney Nature 2016, which are already showing potential for clinic use. Indeed, the results from this paper, have led to new clinical collaborations for the development of tools for pre-surgical planning of epilepsy.

Nevertheless, as R3 suggests (points 1&2), the anatomical validity of the counter-factual augmentation must be validated for ICAM to have clinical potential. Accordingly we ran two experiments 1) we applied ICAM on Alzheimer's (unseen) images acquired at multiple time points for the same subject and compared the outputs. Fig. 1a shows that ICAM generates very similar FA maps for all images (despite them being independently acquired and processed) suggesting the method is reproducible, consistent and that anatomy is preserved. Further evidence is provided by Fig. 1b, which shows that repeat runs of ICAM on Biobank data generate very similar maps despite taking different samples from the latent space, producing low variance (≤ 0.0003) across $\times 10$ experiments. However, we also note that we would seek clinical verification of these findings for any camera ready copy. And, in response to R2 related points, we swap the attribute (class) latent space that is 3D by design, to allow class relevant spatial information to be encoded (e.g. brain atrophy), whereas subject-specific brain anatomy is encoded by the 3D content (class-irrelevant) latent space.

We find R2's request for reporting image generation quality reasonable; although we stress that the objectives of this model was generation of disease maps (for which we show clear improvements) rather than image generation. Nevertheless, the Fréchet Inception Distance (FID) score (which measures the similarity between two datasets) indicates that VA-GAN outperforms ICAM (with respective scores 14.01 and 38.05; lower is better). A better result for VA-GAN is to be expected as VA-GAN is a U-Net style network, with high level skip connections, whereas ICAM receives much more downsampled features that support the learning of a meaningful latent space for improved disease map generation.

We agree with R1 that more thorough details of the training process should go in the supplement. We also plan to upload a cleaned and commented version of the code to Github with examples (all data sources are open source and available). We also appreciate R1 literature suggestions and request for more benchmarking. Accordingly we performed guided backpropagation (GB) and guided Grad-CAM (G-CAM) on the ADNI data set and found that, while GB offers improved performance (over other baselines) of 0.541 ± 0.05 Normalised Cross-Correlation (NCC) (-) and 0.532 ± 0.05 NCC(+), it is still worse than VA-GAN and ICAM (see Table 3 in paper). G-CAM does not perform as well, with scores of 0.244 ± 0.05 NCC(-) and 0.339 ± 0.07 NCC(+). In addition, in response to R3 point 1b we applied feature attribution methods (including GB, integrated gradients (IG) and occlusion (OC)) to the encoder network of ICAM E^c. We report GB of 0.296 ± 0.06 NCC(-) and 0.301 ± 0.04 NCC(+), IG of 0.269 ± 0.05 NCC(-) and 0.289 ± 0.05 NCC(+), and OC of 0.235 ± 0.07 NCC(-) and 0.310 ± 0.05 NCC(+). These scores are not as good as the feature attribution (FA) maps generated by ICAM's generator network. As discussed in the paper, these baselines methods suffer from being low-resolution and, by design, ignore phenotypically variable features. By contrast ICAM is designed to generate high resolution disease maps sensitive to all areas of pathology.

Finally, R4 correctly highlights that the age distribution of the MCI and AD training groups should be reported. The age average and standard deviation is 74.95 ± 8.1, and 72.26 ± 7.9, for AD and MCI subjects, respectively. In response to R2's point on testing on a simulated dataset with ground-truth, we did do this for a simulated lesion HCP dataset (see Table 2 in the paper, and Figs. A.3 & B.2 in the supplement). However, we agree with R2 &R3 (point 3) that benchmarking on a natural image dataset (where the ground truth is easier to visually verify) would be informative. Therefore in Fig. 1c we show that ICAM works well (but could be further improved with hyperparameter optimisation) on the Yosemite dataset (as used in the DRIT paper), and will be added to the supplement.

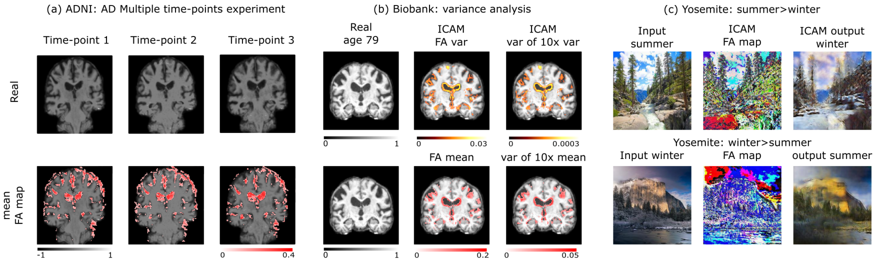

Figure 1: Additional experiments.