Identifying Causal Components in Medical Image Data for Disease Outcomes

Zhaonan Qu

March 18, 2021

1 Project Motivation

The growing availability of medical imaging data coupled with powerful computer vision and machine learning methods have accelerated medical research and practice considerably in recent years. For example, prediction models trained on labeled imaging data have been used to help doctors diagnose diseases, sometimes simultaneously uncovering interesting and previously unknown relationships between exposures and disease outcomes. However, not all such correlations correspond to actual biological pathways from exposures to disease outcomes. As a result, spurious correlations threaten to undermine the robustness of prediction models, and make it dangerous to rely purely on supervised learning for critical applications such as disease diagnosis. Figure 1 depicts a concrete example of this type of challenge. A image classification model trained using pictures of camels with desert backgrounds and pictures of cows with pasture backgrounds may learn to falsely associate the color of the background with the type of animal, which results in incorrect predictions for pictures where the backgrounds are switched. It is therefore important to ensure that a predictive model is using causally relevant information when making predictions, since doing so guards against distributional shifts in data, and also helps improve the interpretability of the model.

In this project, we are interested in discovering and quantifying causally relevant information from medical images. Instead of supervised learning, we take an unsupervised approach and start from unlabeled medical image data. We extract salient features (components) from images, and then assess the causal relevance of these image components. Compared to a supervised approach, we are not restricted to particular disease outcomes (labels) when extracting features, but on the flip side, the extracted information may be difficult to interpret and associate with particular disease outcomes. A complementary supervised approach would start from particular disease outcomes and train a predictive model using labeled images, then assess the causal content of important features that the predictive model has identified. This approach may also be interesting to investigate in future work.

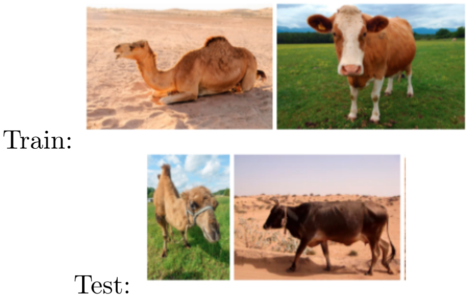

Figure 1: Distributional shifts in train and test sets can result in failure of prediction models that use spurious correlations to make predictions.

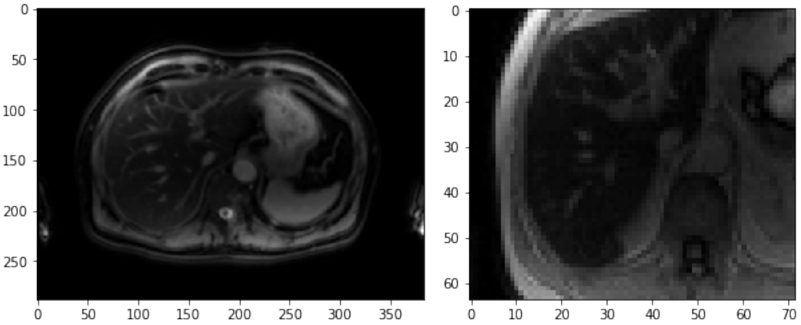

Figure 2: Left, MRI imaging of the abdomen; right, cropped image focusing on the liver region.

In any case, the ultimate goal for this broad research agenda is to uncover interesting and causally relevant features from imaging data for disease outcome prediction and diagnosis, that may not be discernable with visual inspection of the images, even by doctors and trained professionals.

2 Data and Methodology

We use the liver MRI imaging data from UK Biobank, which importantly also contains genome sequence data for each subject. To prepare images for subsequent analysis, we crop the abdomen MRI image in the region corresoponding to the liver, and resize and standardize images. See figure 2 for an example of the abdomen MRI image and cropped liver region.

Our unsupervised approach starts with extracting salient components from unlabeled image data. Ideally, to ensure that spatial information in images is captured, convolutional architecture-based methods should be used. For example, we can train an autoencoder and use the encoder to compress images into vectors. In this project, we take a simpler approach and use principal component analysis performed on vector representations of images to extract information from these images. This approach is easy to implement, but may result in significant loss in signal. A better approach is to combine the two, performing PCA on top of encoder-compressed image content.

Next, to assess the causal relevance of extracted image features, we take advantage of the availability of genetic information and use Mendelian randomization to infer causal links between image components and disease outcomes. Figure 3 gives a concise summary of the statistical reasoning behind Mendelian randomization. On a high level, variations in genetic information of individuals in a population result in variantions in disease outcomes through biological pathways. For a particular exposure-outcome pair we are interested in, if there are genetic variants (SNPs) for which the only such channel of influence to the outcome is mediated through the exposure, then we can infer the causal effect of that exposure on the disease outcome by comparing the association between such genetic variants and the exposure and between genetic variants and the outcome. In our project, exposures are the extracted features from images that we obtain from unsupervised learning, while in a supervised approach, they will be features identified from supervised learning to have predictive power for the labeled disease outcomes.

In our project, the first stage association study between SNPs and each principal component is done through GWAS, while the second stage association study between SNPs and outcomes, and the subsequent Mendelian randomization analysis are automated using MR Base. With summary statistics returned by GWAS, we threshold on p-values less than 10^{-5} in order to select SNPs that are significantly associated with principal components, and use the selected SNPs as instruments in MR analysis. For disease outcomes, we selected liver-related diseases as well as diabetes that are available on MR Base. Notably, the second stage associations with disease outcomes are not obtained from the UK Biobank dataset, but this is not a problem for Mendelian randomization. For details on the execution of the methodology, see the Github repository.[1]

3 Results and Discussion

Our preliminary results are mixed. On one hand, GWAS association analysis between genetic information and principal components for subjects in the UK Biobank reveal promising SNPs that could be used as instruments for Mendelian randomization analysis. See for example the Manhattan plots corresponding to the first and seventh principal components of liver images. We identify significant associations with the genes SLC17A1, SLC 17A2, and SLC 17A3, as well as FTO. These genes have been known to be causally linked to outcomes such as diabetes and obesity.

On the other hand, Mendelian randomization analysis using the selected SNPs and disease outcomes did not yield causal estimates that are uniformllly

[1] https://github.com/zhaonanq/Causal-Image-Analysis

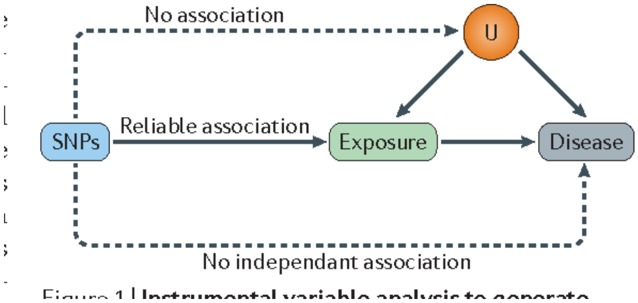

Figure 3: Main assumptions underlying Mendelian randomization. Figure taken from Holmes et al. 2017.

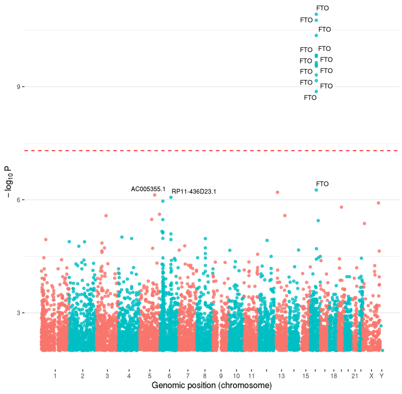

Figure 4: Manhattan plot from GWAS analysis with the first principal component of liver image data.

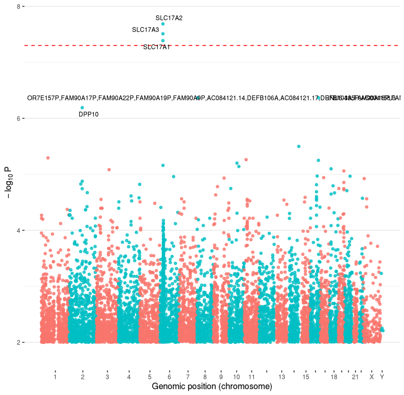

Figure 5: Manhattan plot from GWAS analysis with the seventh principal component of liver image data.

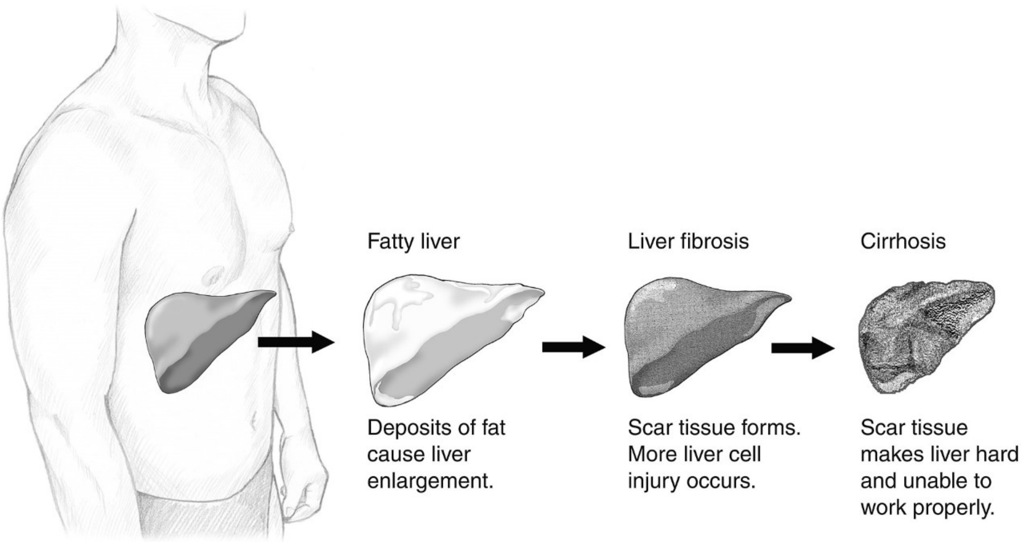
Fatty liver
Liver fibrosis
Cirrhosis
Deposits of fat cause liver enlargement.
Scar tissue forms. More liver cell injury occurs.
Scar tissue makes liver hard and unable to work properly.

Figure 6: Fatty liver is a possible precursor to fibrosis and chirrhosis of liver.

significant across a variety of different MR methods. This may be the result of weak signal from PCA, which led to an insufficient number of SNPs selected as instruments. We tried to enrich the instruments by running GWAS with principal components from pancreas images, but did not find significant associations. Subsequent attempts using VAE to extract components did not yield significantly better results. This is in contrast to a parallel study using fundus (eye-ball) images of the retina from the UK Biobank, which did yield significant causal estimates from many selected SNPs. These suggest that in order to extract more information from liver images, we may need to use better segmentation techniques. However, this may be difficult without properly labeled samples.

However, there are some interesting findings that warrant further investigations. For example, we found that the SNP rs1421085 on FTO yielded a non-zero MR estimate for PC1 on Fibrosis and Chirrhosis of liver. See Figure 6. While fatty liver is known to be a precursor of fibrosis and chirrhosis of liver, FTO is not specifically associated with fatty liver, but with general susceptibility to obesity, so this non-zero estimate from FTO could be because of FTO's association with fatty liver through the general fat content of the abdomen area, which is captured by the first principal component. Because of pleiotropy and unobserved confounding, we should not read too much into individual MR estimates. For example, a zero estimate (as seen in Figure 7) using some SNPs could be the result of unobserved confounding. Similarly, we would need many non-zero MR estimates from different SNPs in order to conclude with confidence that there is causal link between a PC and an outcome. In the absence of a significant number of SNPs that can be used as instruments for MR analysis, we

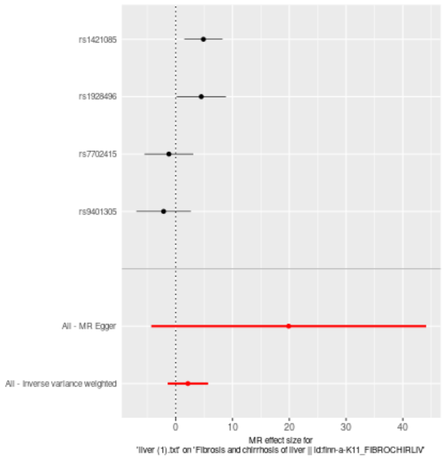

Figure 7: Individual MR estimates using different SNPs of the first principal component on fibrosis and chirrhosis of the liver.

may still be able to conclude that estimates from certain SNPs are likely to be causal. For example, if a SNP has a significant causal estimate for a number of correlated disease outcomes or across different studies, then the SNP may be a valid instrument for a particular disease outcome. This is supported by similar findings from the fundus dataset, where SNP rs1129038 on HERC2 and other SNPs on OCA2 consistently yield significant MR estimates across different but related outcomes, such as hair color and skin color.

Going forward, there are many directions to explore and refine the general research agenda. For example, as discussed before, to select relevant components from images, we can instead pursue a supervised approach, as doing so would guarantee that the components selected by the trained model have predictive power for a particular disease outcome, and we may use MR to assess the causal content of these selected features. Moreover, we can examine more disease outcomes that may be relevant to the SNPs and exposures we find. For example, OCA2 found from the fundus dataset is related to skin cancer. Better segmentation of the liver images would be ideal, but without labeled data, it is unclear if any gain can be achieved. Lastly, instead of thresholding on p-values to select SNPs as instruments, we can consider other methods such as LASSO and other sparsity-inducing methods.

Acknolwedgements. I'd like to thank Professor Manual Rivas, Professor Samuel Yang, and Dr. Yongchan Kwon for continued guidance and interesting discussions.