Representation Learning based Target Discovery from UKBB MRI data

Sivaramakrishnan Sankarapandian * Calico Life Sciences sivark@calicolabs.com

Ramprakash Srinivasan * Calico Life Sciences rams@calicolabs.com

Matt Sooknah Calico Life Sciences mds@calicolabs.com

Elena Sorokin Calico Life Sciences sorokin@calicolabs.com

Jun Xu Calico Life Sciences junxu@calicolabs.com

Abstract

Medical imaging technologies such as MRI and CT scans offer valuable insights into a person's biological condition. Phenotypes derived from these images are essential for the discovery of novel drug targets. Traditional Genome-Wide Association Studies (GWAS) on imaging derived phenotypes (IDPs) require laborious manual feature annotation, extraction of disease-related phenotypes, and subsequent analysis of their associations with genetic variations. This approach has two main limitations: (1) manual voxel-level annotations are time consuming and subjective, particularly for intricate features; (2) these annotations are often limited to a handful of human-definable features, overlooking the wealth of information present in the scans. To address these limitations, we propose an alternative approach to derive phenotypes, which we term embedding-derived phenotypes (EDPs). Our approach consists of two steps. First, we train a self-supervised representation learning model to transform scans into latent embeddings, eliminating the need for manual annotations. Second, we convert these embeddings into disease-relevant phenotypes, preserving the information that may be lost in manually derived phenotypes. Although there are numerous self-supervised representation learning methods, it is not straightforward to transform the embeddings from these models into disease-relevant phenotypes. We present two simple methods that leverage binary labels like ICD-10 codes and demonstrate that the proposed methods identify more biologically meaningful genetic associations compared to using ICD-10 codes alone as binary traits or manually derived phenotypes.

1 Introduction

Non-invasive imaging modalities such as computed tomography (CT) and magnetic resonance imaging (MRI) can provide valuable information about a person's biological state. To investigate a specific disease or trait, researchers often define phenotypes related to the conditions and conduct genome-wide association studies (GWAS)[17, 26] against these phenotypes to identify the genetic architecture of the disease. For instance, in studying aneurysm, the cross-sectional area of the aorta may be measured [15]. However, manually defining phenotypes (also known as hand-crafted phenotypes) has several drawbacks. Firstly, expert manual annotations are costly, and for fine-grained annotations, there is low inter-annotator agreement [2, 11, 25, 13]. e.g., in annotating multiple sclerosis on the brain, seven experts reported an inter-expert agreement ranging from 0.66 to 0.76 of the median Dice score with the consensus [5]. Secondly, imaging-derived phenotypes are often

*These authors contributed equally to this paper.

AI for New Drug Modalities at NeurIPS 2024.

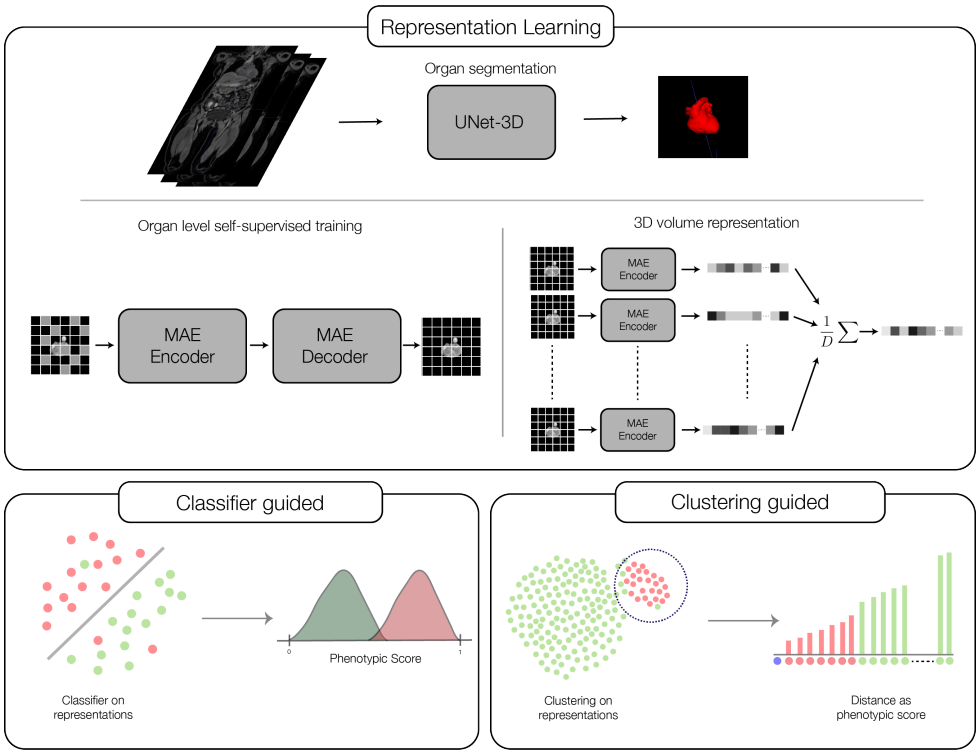

Figure 1: Two step workflow to extract embedding derived phenotypes (EDPs). First, train a self-supervised MAE on 2D slices of masked organ voxels then convert organ level embeddings into phenotype scores using ICD-10 codes

functions of only the segmentations, excluding potentially useful information present in voxels. For example, when studying atherosclerosis in aorta, deriving only the cross-sectional area of the aorta without considering the raw voxels would omit crucial information such as the presence, shape and distribution of fibrotic plaques or calcification, which are essential for understanding the disease architecture.

To overcome the limitations of manually derived IDPs, we propose using embedding-derived phenotypes (EDPs). EDPs are derived from latent embeddings of scans rather than from segmentations. The EDP derivation process involves two steps (Fig. 1). In the first step, we build a self-supervised representation learning model that compresses the information present in the scans into latent embeddings without requiring expensive annotations. The aim is to extract all critical information for deriving disease-relevant phenotypes encoded in this embedding. In the second step, we convert the embedding into a phenotype that can be used in downstream GWAS. Previous work [30] hypothesized that individual dimensions of the embeddings capture a single (disentangled) trait of the input data and can be treated as phenotypes. However, it has been shown [18] that such disentangled latents are nonidentifiable without additional supervision, and a reconstruction objective alone could not make the embeddings disentangled. Instead, we propose two simple methods: *classifier-guided* and *clustering-guided*, to convert embeddings into disease-relevant phenotypes with the help of binary labels such as ICD-10 codes. These approaches can be similarly applied to the study of other traits of interest in deep biobanks, such as drug usage, proteomic marker levels, and aging.

The classifier-guided EDP assumes that controls and cases are linearly separable in the embedding space and builds a linear projection to arrive at a phenotypic score. Although the linearity assumption may not be true for all diseases of interest, empirically, we found the score to work well in identifying biologically relevant genetic variants associated with a handful of diseases studied in this work. However, the cluster-guided EDP utilizes the strategy that scans with similar phenotypes are closer in the embedding space and produce a distance-based score. We show across multiple organs and diseases, these two methods identify more biologically relevant associations compared to just using ICD-10 codes as binary phenotypes or manually derived phenotypes.

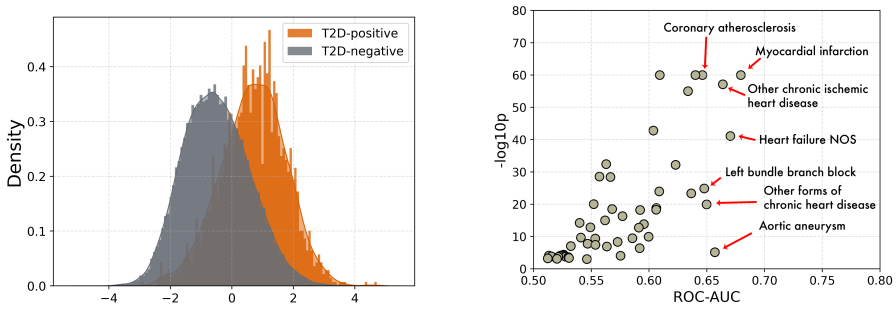

Figure 2: (**left**) The distribution of classifier guided phenotypic scores for type 2 diabetes positive (orange) and negative (gray) participants using embeddings from the liver. Higher scores indicate higher likelihood of disease. (**right**) Plots AUC score and p-value for diseases that were most predictive from classifier using embeddings from heart. Diseases of the cardiac system are labelled in green.

2 Representation Learning

Consider a data set, $\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^N$ with $\boldsymbol{x}_i \in \mathcal{R}^{H\times W\times D}$ of 3D volumes such as MRI or CT scans. We would like to derive disease-relevant quantitative phenotypes from $\mathcal{D}$. For constructing manual IDPs, segmentation models are typically trained to predict features of interest using a mask of the same dimension as the input data (ie) $\boldsymbol{y}_i \in \{0,1\}^{H\times W\times D}$. For example, $\boldsymbol{y}_i$ could be the segmentation of the aorta from MRI scans. Then a hypothesis-based summary statistic of segmentation is derived (e.g., one hypothesis could be quantifying the cross-sectional area (CSA) of the aorta, which is important for understanding atherosclerosis, so one measures CSA of the aorta in a particular slice $f_{CSA}(\hat{\boldsymbol{y}}_{i,d}) = \sum_{H,W} \hat{y}_{i,h,w,d}$. Now $f_{CSA}(\hat{\boldsymbol{y}}_{i,d})$ is treated as a quantitative phenotype of atherosclerosis for downstream GWAS.

Our alternative approach first uses self-supervised learning to convert 3D scans into vectors (i.e.) $\boldsymbol{z}_i = \text{NN}(\boldsymbol{x}_i)$ where $\text{NN} : \mathbb{R}^{H\times W\times D} \to \mathbb{R}^m$. Then we convert these latent vectors into scalar phenotypes using either classifier-guided or clustering-guided method approach (Sec. 3). Motivated by the recent success of Masked AutoEncoders (MAEs) [10], we use them to convert raw voxels into latent representations. We extract slice-level representations from a 2D MAE and use a global average pooling layer across the slices to obtain an embedding for the 3D volume. In addition, we also train Video MAE [27] on the 3D subvolumes and use a global average pooling layer at inference. Finally, we also compare against an imagenet pretrained Vit-b. These models are trained on specific organs/substructures segmented from 3D scans. We find that models trained on substructures yield more biologically relevant associations than general-purpose models trained on the whole image. Sec. The appendix describes experiments that compare the different models; however, we leave the extensive ablation of the model architecture and self-supervised learning algorithms for future work.

3 Converting latent vectors into phenotypic scores

We aim to map the latent vectors $\{\boldsymbol{z}_i\}_{i=1}^N$ from the representation learning model to scalar phenotypes associated with specific diseases. To ensure that projected phenotypes are pertinent to these diseases, we utilize binary labels that categorize scans into various disease-related groups. In this study, we utilize ICD-10 codes as our binary labels, indicating the presence or absence of each disease; however, any binary labels, such as medication indicators, could be employed.

3.1 Classifier guided scoring

We consider a single ICD-10 label $\mathcal{S} = \{s_1, ..., s_N\}$, where $s_i \in \{0, 1\}$ indicates the presence or absence of a disease. Our goal is to learn a scoring $\mathcal{T} = \{t_1, ..., t_N\}$ that correlates with the disease of interest. In classifier-guided scoring, we make the simplifying assumption that the positives and negatives indicated by $\mathcal{S}$ are linearly separable. This assumption has been empirically successful in

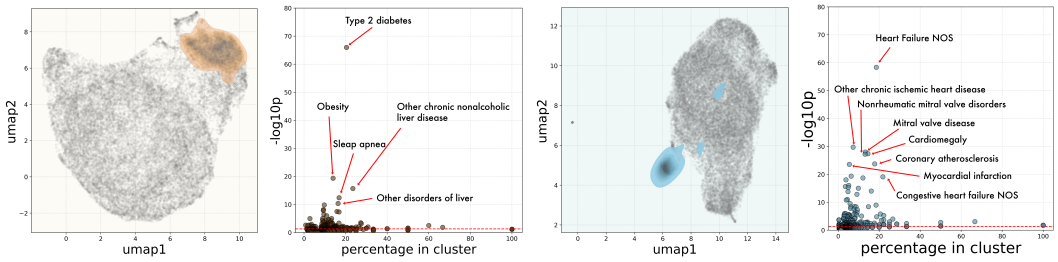

Figure 3: (**left to right**) UMAP of liver embeddings where each dot is an individual and the overlayed brown density plot shows a specific cluster identified through Leiden clustering. *percentage-pval* plot showing the enrichment of diseases in the brown cluster. UMAP of heart embeddings where each dot is an individual and the overlayed blue density plot shows a specific cluster. *percentage-pval* plot showing the enrichment of diseases in the blue cluster. Red dotted line represents p-value = $5e^{-2}$.

identifying downstream biologically relevant genetic variants for certain diseases (see Sec. 4), but it may not hold for all diseases of interest.

To identify ICD-10 codes where the linearity assumption holds, we use the ROC-AUC for a linear classifier and p-values from a nonparametric Wilcoxon rank-sum test on $\mathcal{T}_p = \{t_1, ..., t_{N_p}\}$ for positives and $\mathcal{T}_n = \{t_1, ..., t_{N_n}\}$ for negatives on a held out validation set. If the ROC-AUC is small, the linear classifier cannot fit the data well, suggesting that the embeddings do not capture the features needed for linear separation. Additionally, if the p values are large, the distributions of $\mathcal{T}_p$ and $\mathcal{T}_n$ are not statistically different, indicating that they are not linearly separable. To maximize separation between the classes after linear projection, we use Fisher's linear discriminant criterion as follows:

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} \sum_{i=0}^{N} \mathcal{L}(s_i, \boldsymbol{w}^T \boldsymbol{z}_i) + \lambda ||\boldsymbol{w}||_2 \quad \text{where } \mathcal{L}(\boldsymbol{w}) = \frac{m_p - m_n}{s_p^2 + s_n^2}$$

m_p & m_n denote projections of positive and negative class means respectively (i.e.) $m_p = \boldsymbol{w}^T \frac{1}{N_p} \sum_{p=1}^{N_p} \boldsymbol{z}_p$ and $m_n = \boldsymbol{w}^T \frac{1}{N_n} \sum_{n=1}^{N_n} \boldsymbol{z}_n$. Similarly, s_p & s_n denote within class variances of the projected data respectively (i.e.) $s_p^2 = \sum_{p=1}^{N_p} (s_p - m_p)^2$ and $s_n^2 = \sum_{n=1}^{N_n} (s_n - m_n)^2$. We treat $t_i = \boldsymbol{w}^{*T} \boldsymbol{z}_i$ as our classifier guided phenotypic score for downstream analysis.

3.2 Clustering guided scoring

Classifier-guided scoring implicitly assumes that all positive cases of a disease manifest similarly in the input scans. However, some diseases are complex and exhibit heterogeneous manifestations that cannot be captured by binary labels used to convert latent embeddings into phenotypic scores. In addition, the positives for each disease could be in different stages of progression. For instance, patients who are positive for atherosclerosis could have just fibrotic plaques building in their vessel wall or could have late stage calcified portions. We aim to discern the inherent heterogeneity of diseases using embeddings in this clustering-guided approach. One simplifying assumption we make is that, patients with each manifestation are closer together in the embedding space compared to negatives for a disease. We therefore propose to find different *islands* of positives - closest group of mostly positive individuals-interleaved by negatives in the embedding space corresponding to different manifestations through the following optimization problem,

$$i^*, K^* = \arg\min_{i,K} \sum_{j \in \mathcal{N}_K(i)}^{K} w_j d(\boldsymbol{z_i}, \boldsymbol{z_j}) - \lambda K \qquad \text{where} \quad w_j = \begin{cases} 1, & \text{if } s_j = 1. \\ \gamma, & \text{otherwise.} \end{cases}$$

Intuitively, we wanted to find a patient i^* whose largest K nearest neighbors $\mathcal{N}_{K^*}(i^*)$ are mostly positives, γ controls the number of negatives we are willing to include in the nearest neighbor set.

λ controls the size of the nearest neighbor set. $\mathcal{N}_{K^*}(i^*)$ is the group who are mostly positives and are homogeneous in their manifestations. But, solving this discrete optimization problem could get computationally expensive for larger values of i and K while also finding optimal values for γ & λ. So, we opted for a simpler strategy that directly optimizes the objective function, in that we follow a two-step framework. **Step-1:** We cluster the embeddings into pre-specified number of $\mathcal{Q}$ clusters. This enables the grouping of subjects into distinct clusters based on their underlying imaging characteristics potentially separating the individuals that have the same ICD-10 code but with different manifestations into different clusters. **Step-2:** We map the ICD-10 annotations to each of the clusters, now each patient gets a cluster label and ICD-10 binary label. Therefore, clustering partitions the positives into different clusters and we treat each positive cluster as an approximation to $\mathcal{N}_{K^*}(i^*)$.

Once the clusters of positives are identified, to get to a phenotypic score, we propose to compute the distance from the mean embedding of positives within each cluster with the rest of the dataset. Concretely, consider a single ICD-10 label as before $\mathcal{S} = \{s_1, ..., s_N\}$ where $s_i \in \{0, 1\}$ denote the presence or absence of a disease / trait. After clustering, every embedding gets a cluster membership assignment $m_{i,q}$ where $m_{i,q} = 1$ when $\boldsymbol{z}_i$ belongs to cluster q and 0 otherwise. For each cluster q, we take the embeddings that are positive for both ICD-10 code and cluster label (i.e.) $\mathcal{Z}_{q^*} = \{\boldsymbol{z}_i | m_{i,q^*} = 1, s_i = 1\}$ and compute the mean embedding $\bar{\boldsymbol{z}} = \sum_{\mathcal{Z}_{q^*}} \boldsymbol{z}_j$. Then we compute the distance with respect to all datapoints $t_i = d(\bar{\boldsymbol{z}}, \boldsymbol{z_i})$ and treat that as the phenotypic score. This process yields a score that quantifies the semantic similarity of all embeddings to the average appearance of a specific manifestation of positive cases.

4 Experiments

We use 45,714 abdominal MRI & T1 brain scans from UK BioBank (UKBB) with abdominal MRIs measuring fat and water content for all of our experiments along with lifestyle / health information such as ICD-10 codes and genotypes. The details of the microarray and imputation of variants can be found here [17]. We preprocess and extract organ-level segmentation using the protocol developed in [17] and study each organ independently. To train representation models, we obtain the organ-level segmentation mask and mask out all voxels outside of the predicted mask, then extract the smallest fitting cuboid around the masked voxels and use them as $\boldsymbol{x}_i$. We then train an MAE on 2D slices from $\boldsymbol{x}_i$'s with an additional `[cls]` token to the input sequence. We use a global average pooling layer on the `[cls]` tokens from all 2D slices in the input $\boldsymbol{x}_i$, which yields a 384-dimensional embedding for every organ of every individual. For training the Video MAE, we use subvolumes of 16 slices from $\boldsymbol{x}_i$, and similarly use global average pooling on `[cls]` tokens. (See appendix for architectural and hyperparameter details). For all of our GWAS analysis, we regress out age, age^2, self-reported sex, BMI, genotyping array, imaging center, and first 10 PCs of the genotype, from the final phenotypes. We use *regenie* GWAS [20] package and to map rsID to gene, we use the highest variant-to-gene (v2g) score gene from OpenTargets [8] for every associated variant.

In some cases, the number of positive cases for an ICD-10 code is significantly lower than the negatives. For example, only a small fraction of individuals in the whole cohort are positive for Splenomegaly. To create a robust linear classifier, we apply SMOTE [4] with an oversampling ratio set to balance the class proportions. For each organ and ICD-10 code, we train a linear model using 10-fold cross-validation and select the best model to compute the final score, which is then used for ROC-AUC and p-value calculations. A sample histogram illustrating the classifier-guided scores for type 2 diabetes code ICD-10 is shown in Fig. [2] (left), with liver representations color-coded with binary labels. To identify the ICD-10 codes with the highest linear separability, we create *AUC-pval* plots for each organ and select the codes that fell into the upper right region, indicating a high ROC-AUC and low p-value. An example of this is presented in Fig. [2] (right).

In clustering-guided scoring, we take inspiration from single-cell RNA sequencing literature, where communities of cells are determined by gene counts. We use Leiden clustering with resolution=1, n_neighbors=10, and cosine similarity as the distance metric to compute communities of individuals from representations. We also found Leiden clustering working well empirically in identifying homogeneous populations of ICD-10 codes compared to k-means or agglomerative clustering. Since most ICD-10 codes have positives in every cluster, we need a method to identify ICD-10 codes that are statistically enriched in each cluster. We achieve this by running a Fisher exact test to determine if the proportion of positives within a cluster is statistically different from the positives outside

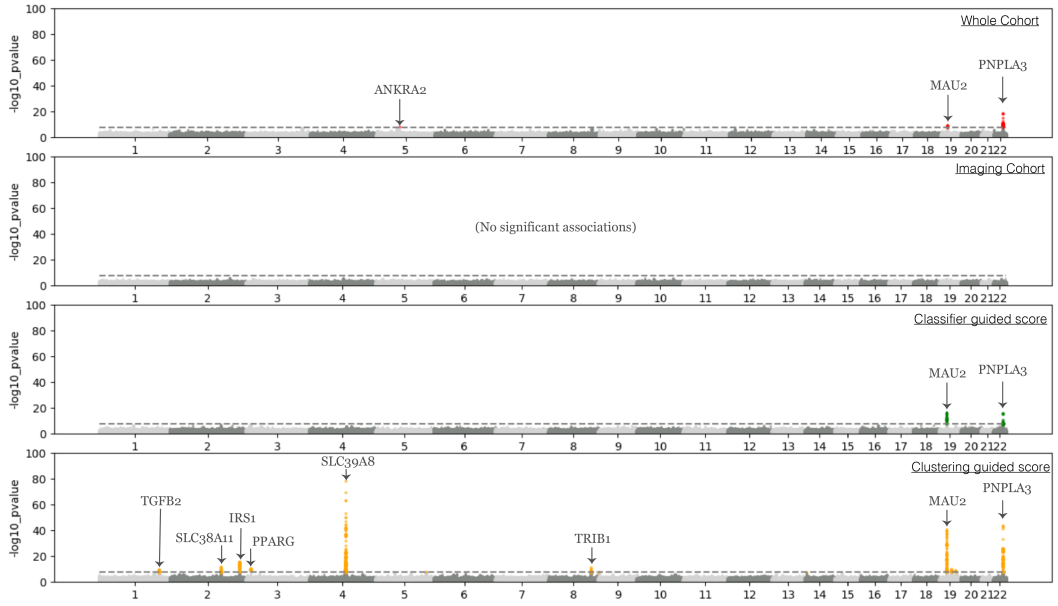

Figure 4: **(Top to bottom)** GWAS summary statistics for *Other-chronic-nonalcoholic-liver-disease* on whole UKBB cohort (N=408,961) using ICD-10 code as binary trait, imaging cohort (N=45,714) using ICD-10 code as binary trait, imaging cohort with classified-guided quantitative score and imaging cohort with clustering-guided quantitative score

the cluster. To select ICD-10 codes enriched in each cluster, we create *percentage-pval* plots. In these plots, we examine the percentage of each ICD-10 code in a cluster and its associated p-value from the Fisher exact test pertaining to that cluster. Examples of *percentage-pval* plots for liver embeddings & heart embeddings are shown in Fig. [3]). As illustrated, the brown cluster in the left panel has a statistically significant number of "chronic nonalcoholic liver disease" positive individuals among others. Similarly, the blue cluster in the right panel is enriched for heart conditions such as "heart failure" and "mitral valve disease". Once an ICD-10 code is selected, as detailed in 3.2, we intersect the individuals who are positive for that ICD-10 code and have the same cluster label. We then take the mean of their embeddings and compute the (1 - cosine similarity) between that mean embedding and everyone else. Fig. [4] presents an example GWAS for the brown cluster, Other-chronic-nonalcoholic-liver-disease phenotypic score from liver embeddings.

4.1 GWAS

To validate phenotypic scores calculated from classifier-guided and cluster-guided methods capture biologically relevant information, we conduct genome-wide association studies (GWAS) using the scores as quantitative traits for each ICD-10 codes and benchmark the number of gene loci that pass the Bonferroni correction threshold, p-value $> 5e^{-8}$.

We compare the proposed methods with two baselines - 1) genetic variants associated with the ICD-10 code of interest as a binary trait only on the *imaging cohort* (N=45,714), 2) genetic variants associated with the ICD-10 code of interest in the *whole UKBB cohort* (N = 408,961: $\sim$ 10 times more than the imaging cohort). As seen from columns 4 to 6 in Table. 1, both classifier and clustering guided phenotypes yielded strictly more genetic hits compared to baseline-1 on the imaging cohort. Fig. [4] shows an example of a GWAS summary for *other-chronic-nonalcoholic-liver-disease*. As seen, both MAU2 and PNPLA3 which were identified in the whole cohort (and independenly found in many studies [23, 29]), were recovered by both classifier and clustering guided methods. We also emphasize that with the same sized cohort running GWAS with ICD-10 label as a binary trait without using the representations did not yield any statistically significant associations. In addition, clustering guided scoring identified more hits that also holds biological relevance(See 5 for more details) which were missing in the whole cohort which is 10x size of the imaging cohort. This trend continues for other diseases as well, for instance, when the proposed methods were run on representations of

Disease	Organ	Whole cohort	Imaging cohort	Classifier guided	Clustering guided	Classifier ∩ WC	Clustering ∩ WC
Chronic NALD	Liver	7	0	5	38	5	6
Heart Failure	Heart	0	0	1	8	0	0
Type 2 Diabetes	Kidney	118	1	8	29	0	1
Splenomegaly	Spleen	0	0	1	53	0	0
Osteoporosis	Thigh Bones	13	0	16	7	2	0

Table 1: Number of GWAS hits that pass Bonferroni threshold. Classifier guided and Clustering guided are run on the image cohort. Classifier ∩ WC & Clustering ∩ WC represent number of genes that overlap between classifier guided scoring, clustering guided scoring with the whole cohort respectively.

spleen and for the ICD-10 *splenomegaly*, GWAS on the whole cohort yielded no hits, while both classifier and clustering guided yielded more genetic associations. We also note that the classifier guided strategy tends to generate fewer associations than clustering, as the classifier strictly projects to a subspace that separates out disease positives and negatives, while clustering may incorporate more features in the phenotype, that can further be interpreted using PheWAS (see appendix).

4.2 Comparison with hand-crafted image derived phenotypes

To evaluate the proposed EDPs against manually derived phenotypes, we compare the number of GWAS hits generated from organ volume data from [17] with those obtained using our two proposed methods. [17] developed segmentation models for each organ and further ran downstream GWAS using the volume of the segmented organ as manually derived IDPs. As shown in Table 2, EDPs yield more genetic associations than using organ volume alone, regardless of whether MAE or Video MAE architectures are employed. This suggests that these scans contain additional information that can enhance our understanding of the genetic architecture of diseases, beyond simply measuring volume.

Method	Liver	Kidneys	Lungs	Heart
MAE	343	232	229	296
Video MAE	297	288	192	311
Volume	12	7	11	18

Table 2: Number of GWAS associations, comparing EDPs (combining classifier and clustering guided strategies.) vs volume based IDPs from [17] on organs in abdominal MRI scans.

To demonstrate that our representations encompass manually derived phenotypes, we train a linear regressor to predict organ volumes. Table 4 shows that we achieve nearly perfect correlation with ground truth volumes, indicating that embeddings effectively capture these phenotypes. Additionally, since the likelihood of many diseases increases with age, we compare the performance of a linear regressor in predicting chronological age using both our representations and organ volumes. As shown in Table 3, there is a significant gap in age prediction in terms of the mean absolute error between the representations (both MAE and Video MAE) and the volumes, further highlighting that our representations capture more biological information than manually derived traits.

4.3 Flexibility to remove information

In target discovery, genetic associations with the most promising therapeutic potential may not have the most significant associations, e.g., the most significant associations from retinal fundus images code for eye color [1], which is not interesting from a therepeutic standpoint. This inherently makes interpreting GWAS for target discovery challenging, where the ability to *remove* factors that dominate associations is very useful. Since we construct phenotypes from scan embeddings, we can adjust for known factors simply by regressing these variables out. For example, if $\{v_i\}_{i=1}^N$ denotes the factor that one would like to remove, continuous or discrete, we could do $\boldsymbol{z}_{i(\text{adjusted})} = \boldsymbol{z}_i - \hat{\beta} v_i$, where $\hat{\beta}$ is the estimated regression coefficient between the factor v_i and the embedding $\boldsymbol{z}_i$.

Method	Liver	Kidneys	Lungs	Heart	Bones	Combined Organs
MAE	4.2	4.1	3.7	3.9	3.8	2.9
Video MAE	4.1	4.1	3.9	3.8	4.0	2.8
Volume	18.3	16.1	13.7	19.2	12.4	9.7

Table 3: Mean Absolute Error (in yrs) between ground truth and predicted chronological age from a linear model on different organ level representations & IDPs. It can be observed that organ volume is not a good predictor of chronological age, while EDPs when combined across organs are comparable to the horovath biological ageing clock [12]

Metric	Liver	Kidneys	Lungs	Heart
Pearson Correlation	0.95	0.96	0.97	0.96
Mean Absolute Error %	1.2	1.3	0.79	1.6

Table 4: Metrics after fitting a linear model to predict volume from organ level representations from MAE. It can be seen that volume can be reconstructed with high fidelity.

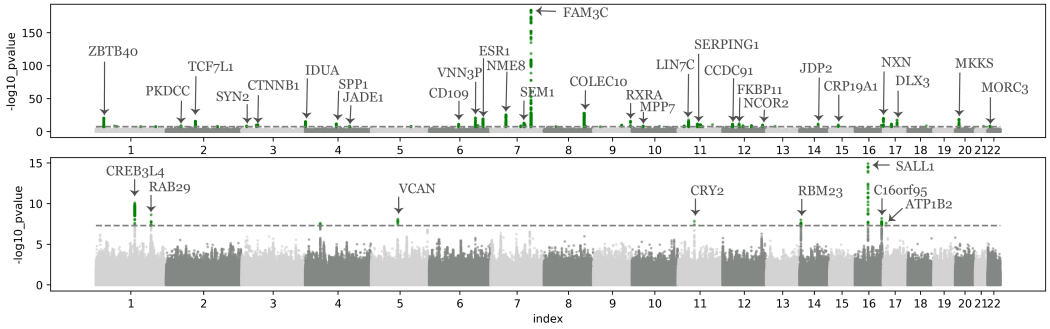

Figure 5: (**Top**) Manhattan plot of GWAS associations of clustering-guided phenotype for *mental disorders*. The results capture a lot of variation associated with bone mineral density (BMD), that are not of interest. PC 1 captures a lot of this variance. (**Bottom**) Manhattan plot of GWAS associations for the same phenotype after adjusting for PC 1, leading to more novel associations that don't code for bone mineral density.

As an example, we notice that the first principal component (PC) of image embeddings from brain T1 MRIs reflects variations in Bone Mineral Density (BMD), influencing clustering-guided scoring. This is illustrated by numerous genetic associations (e.g., ZBTB40, COLEC10, FAM3C) linked to BMD [9], as seen in Fig. 5 (top) for the ICD-10 code of *mental disorders*. To eliminate the influence of BMD associations, we performed clustering-guided scoring while regressing out the first PC, resulting in the GWAS shown in Fig. 5 (bottom). This analysis successfully removed all associations related to BMD, revealing CREB3L4 and RBM23 as potentially novel targets for further investigation.

4.4 Few-shot phenotypes

In clustering-guided scoring, we compute phenotypes using a distance metric between an average-looking positive individual (after clustering) and the rest of the dataset. This means we technically only need one representative positive individual to arrive at a phenotype. To demonstrate that clustering-guided scoring yields biologically relevant phenotypes with fewer ICD-10 labels, we consider a scenario where we have image embeddings for the entire population but labels for only a subset of the positives. It's worth noting that this scenario is commonly encountered in real-world settings. For instance, when using other labels like medication usage, one might have access to MRI

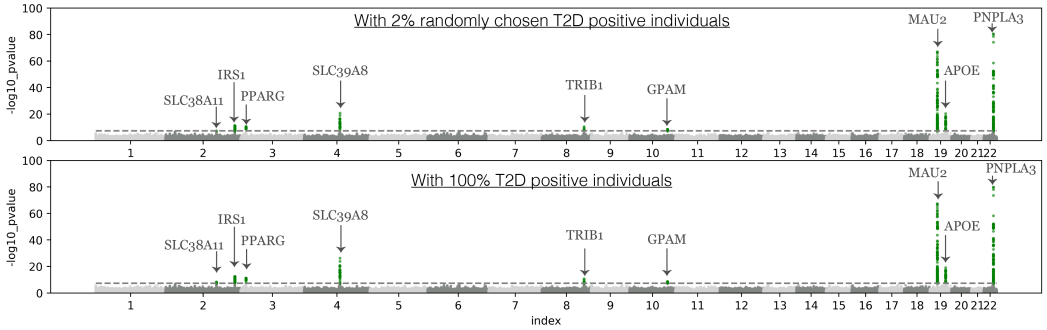

Figure 6: **(Top)** Manhattan plot for *type-2-diabetes* clustering guided phenotypic score using all liver embeedings but with only 2% of the positive labels, **(Bottom)** Manhattan plot for *type-2-diabetes* clustering guided phenotypic score from liver embeddings using all the labels and embeddings

scans, but medication usage might be sparse or not diligently captured in the patient management system.

Fig. [6] shows the GWAS ran using clustering guided phenotypic score for the ICD-10 code of type 2 diabetes from liver embeddings. As shown, the genetic associations found through all the ICD-10 labels (bottom) and only 2% of the positive labels (top) - randomly sampled - are the same with minor variations in p-values. This suggests that clustering guided EDPs could be used in scenarios where very few labels are available.

5 Discussion

Here, we describe a framework that leverages representation learning based embeddings guided by ICD-10 code annotations to derive phenotypes that are directly applicable for identifying disease-relevant genetic variants. Table [1] demonstrates the utility of our method in extracting genetic associations related to diseases, producing a large number of associations that **cannot** be derived directly from the analysis of ICD-10 code annotations as a binary trait. Notably, these associations are derived from a sample size that is only 10% the whole UK biobank imaging cohort, where directly using the ICD-10 code annotations on this cohort yield almost no associations. Thus, our method may potentially be leveraged to work with dataset sizes that are considered underpowered. Table [2] illustrates the ability of our method to discover significantly more disease relevant genetic associations than traditional IDPs like organ volumes.

The genetic associations we identify often capture known disease biology. As a case study, the analysis of chronic NAFLD in liver yielded genes that include PNPLA3 which is a known modulator of triglycerides in hepatocytes [3]. SLC39A8, MAU2, PNPLA3 are associated with liver fat and other metabolic traits [23]. PPARγ is a known enhancer for genes coding for lipid and glucose metabolism [28]. Our analysis of osteoporosis in thigh bones yield a number of genetic associations (TNFRSF11, RUNX2, ZBTB40) which are implicated in osteogenesis, osteoblast and bone mineral density modulation [16, 22, 6]. We also observe pleiotropic genes that are implicated in multiple diseases and organs, like SLC39A8, WNR4, CCDC91. Most exciting are novel associations, like CASP9 being consistently implicated in kidney disease.

Since clustering guided scoring works by computing similarity with a single embedding, we ablate by randomly keeping the labels for 2% of the *Type 2 diabetes* positives in the imaging cohort and we observe the hits were nearly identical to 100% of positives suggesting our work could potentially be used in studying the genetic architecture of rare diseases (see supplementary material).

Recent works that utilize image embeddings for genetic discovery focus on utilizing the features of the embedding (or linear combination of features using PCA) independently as phenotypes in association studies [24, 14]. Our methodology is the first to our knowledge to holistically incorporate the entire image embedding in deriving disease-relevant phenotypes. Our methodology is general purpose, it can be extended to utilize **any** binary trait to guide image embeddings.

6 Acknowledgement

This study was carried out using UK Biobank Application number 44584.

References

[1] Ahadi, S., Wilson, K.A., Babenko, B., McLean, C.Y., Bryant, D., Pritchard, O., Kumar, A., Carrera, E.M., Lamy, R., Stewart, J.M., et al.: Longitudinal fundus imaging and its genome-wide association analysis provide evidence for a human retinal aging clock. Elife **12**, e82364 (2023)

[2] Association, A.H., et al.: Icd-10-cm field testing project: Report on findings: Perceptions, ideas and recommendations from coding professionals across the nation. ICD-10-CM Field Testing Project: Report on Findings: Perceptions, Ideas and Recommendations from Coding Professionals Across the Nation/AHIMA, American Health Information Management Association (2003)

[3] Bruschi, F.V., Tardelli, M., Claudel, T., Trauner, M.: Pnpla3 expression and its impact on the liver: current perspectives. Hepatic medicine: evidence and research pp. 55–66 (2017)

[4] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority oversampling technique. Journal of artificial intelligence research **16**, 321–357 (2002)

[5] Commowick, O., Istace, A., Kain, M., Laurent, B., Leray, F., Simon, M., Pop, S.C., Girard, P., Ameli, R., Ferré, J.C., et al.: Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. Scientific reports **8**(1), 13650 (2018)

[6] Doolittle, M.L., Calabrese, G.M., Mesner, L.D., Godfrey, D.A., Maynard, R.D., Ackert-Bicknell, C.L., Farber, C.R.: Genetic analysis of osteoblast activity identifies zbtb40 as a regulator of osteoblast activity and bone mass. PLoS Genetics **16**(6), e1008805 (2020)

[7] Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

[8] Ghoussaini, M., Mountjoy, E., Carmona, M., Peat, G., Schmidt, E.M., Hercules, A., Fumis, L., Miranda, A., Carvalho-Silva, D., Buniello, A., et al.: Open targets genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. Nucleic acids research **49**(D1), D1311–D1320 (2021)

[9] Guo, B., Wang, C., Zhu, Y., Liu, Z., Long, H., Ruan, Z., Lin, Z., Fan, Z., Li, Y., Zhao, S.: Causal associations of brain structure with bone mineral density: a large-scale genetic correlation study. Bone Research **11**(1), 37 (2023)

[10] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)

[11] Horsky, J., Drucker, E.A., Ramelson, H.Z.: Accuracy and completeness of clinical coding using icd-10 for ambulatory visits. In: AMIA annual symposium proceedings. vol. 2017, p. 912. American Medical Informatics Association (2017)

[12] Horvath, S.: Dna methylation age of human tissues and cell types. Genome biology **14**, 1–20 (2013)

[13] Joo, J., Hobbs, B.D., Cho, M.H., Himes, B.E.: Trait insights gained by comparing genome-wide association study results using different chronic obstructive pulmonary disease definitions. AMIA Summits on Translational Science Proceedings **2020**, 278 (2020)

[14] Kirchler, M., Konigorski, S., Norden, M., Meltendorf, C., Kloft, M., Schurmann, C., Lippert, C.: transfergwas: Gwas of images using deep transfer learning. Bioinformatics **38**(14), 3621–3628 (2022)

[15] Klarin, D., Devineni, P., Sendamarai, A.K., Angueira, A.R., Graham, S.E., Shen, Y.H., Levin, M.G., Pirruccello, J.P., Surakka, I., Karnam, P.R., et al.: Genome-wide association study of thoracic aortic aneurysm and dissection in the million veteran program. Nature Genetics pp. 1–10 (2023)

[16] Komori, T.: Roles of runx2 in skeletal development. RUNX Proteins in development and cancer pp. 83–93 (2017)

[17] Liu, Y., Basty, N., Whitcher, B., Bell, J.D., Sorokin, E.P., van Bruggen, N., Thomas, E.L., Cule, M.: Genetic architecture of 11 organ traits derived from abdominal mri using deep learning. Elife **10**, e65554 (2021)

[18] Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., Bachem, O.: Challenging common assumptions in the unsupervised learning of disentangled representations. In: international conference on machine learning. pp. 4114–4124. PMLR (2019)

[19] Lu, M.Y., Chen, B., Williamson, D.F., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L.P., Gerber, G., et al.: A visual-language foundation model for computational pathology. Nature Medicine **30**(3), 863–874 (2024)

[20] Mbatchou, J., Barnard, L., Backman, J., Marcketta, A., Kosmicki, J.A., Ziyatdinov, A., Benner, C., O'Dushlaine, C., Barber, M., Boutkov, B., et al.: Computationally efficient whole-genome regression for quantitative and binary traits. Nature genetics **53**(7), 1097–1103 (2021)

[21] Millard, L.A., Davies, N.M., Gaunt, T.R., Davey Smith, G., Tilling, K.: Software application profile: Phesant: a tool for performing automated phenome scans in uk biobank (2018)

[22] Odgren, P.R., Kim, N., MacKay, C.A., Mason-Savas, A., Choi, Y., Marks, Jr, S.C.: The role of rankl (trance/tnfsf11), a tumor necrosis factor family member, in skeletal development: effects of gene knockout and transgenic rescue. Connective Tissue Research **44**(1), 264–271 (2003)

[23] Parisinos, C.A., Wilman, H.R., Thomas, E.L., Kelly, M., Nicholls, R.C., McGonigle, J., Neubauer, S., Hingorani, A.D., Patel, R.S., Hemingway, H., et al.: Genome-wide and mendelian randomisation studies of liver mri yield insights into the pathogenesis of steatohepatitis. Journal of hepatology **73**(2), 241–251 (2020)

[24] Patel, K., Xie, Z., Yuan, H., Islam, S.M.S., Zhang, W., Gottlieb, A., Chen, H., Giancardo, L., Knaack, A., Fletcher, E., et al.: New phenotype discovery method by unsupervised deep representation learning empowers genetic association studies of brain imaging. medRxiv pp. 2022–12 (2022)

[25] Stausberg, J., Lehmann, N., Kaczmarek, D., Stein, M.: Reliability of diagnoses coding with icd-10. International journal of medical informatics **77**(1), 50–57 (2008)

[26] Tcheandjieu, C., Xiao, K., Tejeda, H., Lynch, J.A., Ruotsalainen, S., Bellomo, T., Palnati, M., Judy, R., Klarin, D., Kember, R.L., et al.: High heritability of ascending aortic diameter and trans-ancestry prediction of thoracic aortic disease. Nature Genetics **54**(6), 772–782 (2022)

[27] Tong, Z., Song, Y., Wang, J., Wang, L.: Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. Advances in neural information processing systems **35**, 10078–10093 (2022)

[28] Tyagi, S., Gupta, P., Saini, A.S., Kaushal, C., Sharma, S.: The peroxisome proliferator-activated receptor: A family of nuclear receptors role in various diseases. Journal of advanced pharmaceutical technology & research **2**(4), 236 (2011)

[29] Wilman, H.R., Kelly, M., Garratt, S., Matthews, P.M., Milanesi, M., Herlihy, A., Gyngell, M., Neubauer, S., Bell, J.D., Banerjee, R., et al.: Characterisation of liver fat in the uk biobank cohort. PloS one **12**(2), e0172921 (2017)

[30] Yun, T., Cosentino, J., Behsaz, B., McCaw, Z.R., Hill, D., Luben, R., Lai, D., Bates, J., Yang, H., Schwantes-An, T.H., et al.: Unsupervised representation learning on high-dimensional clinical data improves genomic discovery and prediction. Nature Genetics pp. 1–10 (2024)

A Appendix / supplemental material

A.1 Image Embedding ML Models

We experiment with a few different strategies to extract image embeddings from UKBB MRI images. All of our representation models use the ViT-b [7] architecture. We use the image representation of the `[cls]` token of the penultimate layer. If the model is designed for images, we run a forward pass across 2D slices and use a global average pooling layer over the embedding from all slices. We use a tight bounding box around the organ of interest and extract embeddings in this field of view. For 3D models (VideoMAE), we use a 3D subvolume with 16 slices (this design is to ease memory requirement of the representation models). When the organ covers more than 16 slices, we use global average pooling on multiple subvolumes. The subvolumes are 0-padded to be of fixed size, (patch size and image resolution in 6)

As a baseline 2D embedding model, we use the ViT-b model trained in the supervised classification of ImageNet categories. This is effectively used as a foundation model, with the same model weights applied for all organs, with the same model architecture. We also train an MAE model for each organ, with patch size and image sizes modified to suit the organ shape, all other hyperparameters retaining the Vit-b default. The MAE models that we trained on all the slices inside the organ from 47000 scans of the abdomen and 41000 of the brain. This results in $\mathcal{O}$(10 million) training samples. All models were trained for 30 epochs through the training set. We observe that ViT-b is approximately the optimal model size for extracting the maximum signal from these MRI scans, and it has been observed by other foundation model methods in the medical imaging field [19]. For VideoMAE, the 16 slice subvolumes have yielded the best results. We leave extensive ablation of design choices to future work.

Organ	Patch Size (MAE)	Patch Size (VideoMAE)	2D Resolution (MAE)	2D Resolution (Video MAE)
Liver	8	7	144x144	144x144
Kidneys	6	6	48x48	36x36
Lungs	12	12	180x180	156x156
Heart	8	7	144x144	144x144

Table 5: Patch size and of ViT-b tokens and image resolution. The sizes are changed to optimize for the organ shapes, smaller organs have a smaller patch size.

We observe that the self-supervised training loss tends to continue to decrease, we chose to stop our training at 30 epochs, it is likely that training for longer would yield improvements to the image embeddings, and we leave ablating the impact of the training procedure to future work. Further, we observe that training a single model across organs yields empirically less informative associations than training models for each organ/substructure. We hypothesize that this is caused by the lack of functional specialization in a vast number of pixels in the imaging volume that aren't a part of the organs of interest, leading the model to not learn concepts that are functionally specialized. This leads to models learning the main results of adiposity from the fat channel. This is another aspect of training scalable representations of radiological images that needs to be explored further.

A.2 Interpreting associations: Phenome wide scan

The image embedding derived phenotypes are designed to be used in large cohort GWAS to detect novel associations. However, in comparison to hand-crafted phenotypes, these measures are more abstract, e.g., a cluster is likely to capture variance related to a variety of traits. To further interpret the information captured by the phenotypes, we conducted a Phenome-Wide Association Study (PheWAS) to examine associations between embedding-derived phenotypes and a wide range of phenotypic traits acquired from the biobank. PheWAS is an essential tool of genetic epidemiology that enables the discovery of potential pleiotropic effects and links to other diseases and traits, our ML derived phenotypes integrate readily with this framework. We use the protocol prescribed by PHESANT [21] for comprehensive phenome-wide scans in UK Biobank, covering 683 phecodes and 234 other traits. We corrected for a set of covariates that included age, sex, BMI, imaging center, and date.

Model	Num. Associations
ViT-S	142
ViT-B	297
ViT-L	207

Table 6: Number of GWAS associations in liver from various embedding model sizes.

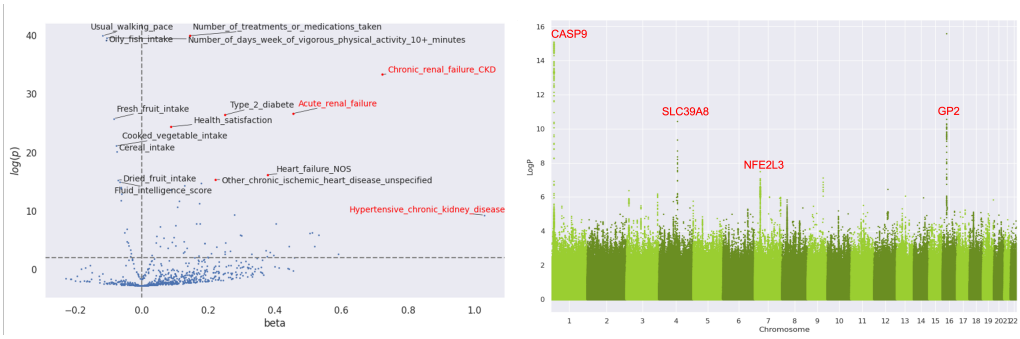

Figure 7: (**left**) Volcano plot ($-log(p)$ against effect size β) derived from PheWAS for classifier-guided phenotype derived from kidneys using *renal disease*. Red text denotes kidney related diseases that are expected from the phenotype. The other associations capture potentially new information. (**right**) Manhattan plot of GWAS associations for the same phenotype, including potentially novel association with CASP9.

Figure Fig.[7] illustrates the results of a PheWAS scan for the classifier-guided phenotype that codes for the disease *renal failure CKD* extracted from embeddings of the kidneys. We observe significant associations with a large effect size for renal failure. We also find a significant covariance with hypertensive chronic kidney disease (CKD), type 2 diabetes, and heart failure. Furthermore, it can be observed that subjects who have a low score for this phenotype (with negative values of β) are likely to have a better exercise regimen, fruit and fish intake. GWAS for this phenotype identifies loci that include CASP9, SLC39A8 and GP2. CASP9 is a novel association for CKD and needs further exploration as a target. The phenome-wide scan shed light on associated traits that can be used in a variety of downstream applications in drug discovery, e.g., to derive cellular phenotypes, adjust for covariates, identify appropriate indications, construct biomarkers, etc.