Integrating large scale genetic and clinical information to predict cases of heart failure

Author information
Kuan-Han H. Wu, MS[1], Nicholas J. Douville, MD, PhD[1,2,3], Xianshi Yu, PhD[4], Michael R. Mathis, MD[1,2,3], Sarah E. Graham, PhD[5], Global Biobank Meta-analysis Initiative (GBMI), Ida Surakka, PhD[5], Whitney E. Hornsby, PhD[5], Cristen J. Willer, PhD[1,5,6*], and Xu Shi, PhD[4*]

Affiliations
[1] Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan, USA
[2] Department of Anesthesiology, Michigan Medicine, Ann Arbor, Michigan, USA
[3] Institute of Healthcare Policy & Innovation, University of Michigan, Ann Arbor, Michigan, USA
[4] Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, USA
[5] Department of Internal Medicine, Michigan Medicine, Ann Arbor, Michigan, USA
[6] Department of Human Genetics, University of Michigan, Ann Arbor, Michigan, USA
[*] Senior author

Correspondence
Xu Shi, PhD
Correspondence: shixu@umich.edu

ABSTRACT

Background

Heart failure is a major cause of death globally and earlier initiation of treatment could mitigate disease progression. Multiple efforts have been made using genome-wide association studies (GWAS) or electronic health records (EHR) to identify individuals at high risk of heart failure (HF). However, integrating both sources using novel natural language processing (NLP) techniques and large scale global genetic predictors into heart failure prediction models has not been evaluated.

Objectives

The study aimed to improve the accuracy of HF prediction by integrating GWAS- and EHR-derived risk scores.

Methods

We previously performed the largest HF GWAS to date within the Global Biobank Meta-analysis Initiative, which includes 974,174 samples (51,274 cases; 5%) from 9 biobanks across the world, to create a polygenic risk score (PRS). Next, to extract information from the Michigan Medicine high-dimensional EHR (N=61,849 subjects), we treated diagnosis codes as 'words' and applied NLP on the data. NLP was used to learn code co-occurrence patterns and extract 350 latent phenotypes (low-dimensional features) representing 29,346 EHR codes. Next, we regressed HF on the latent phenotypes in an independent cohort and the coefficients were used as the weights to calculate a clinical risk score (ClinRS). Model performances were compared between baseline (age and sex) model and three models with risk scores added: 1) PRS, 2) ClinRS, and 3) PRS+ClinRS, using 10-fold cross validated Area Under the Receiver Operating Characteristic Curve (AUC).

Results

Our results show that PRS and ClinRS are each, separately, able to predict HF outcomes significantly better than the baseline model, up to eight years prior to HF diagnosis. Higher AUC (95% CI) were observed in the PRS model (0.76 [0.74-0.78]) and ClinRS model (0.77 [0.74-0.79]), compared to the baseline model (0.71 [0.68-0.73]). Moreover, by including both PRS and ClinRS in the model, we achieved superior performance in predicting HF up to ten years prior to HF diagnosis (AUC: 0.79 [0.77-0.82]), 2-3 years earlier than using either single risk predictor alone.

Conclusions

We demonstrate the additive power of integrating GWAS- and EHR-derived risk scores to predict HF cases prior to diagnosis. Clinical application of this approach may allow identification of patients with higher susceptibility to HF and enable preventive therapies to be initiated at an earlier stage.

INTRODUCTION

Heart failure affects an estimated 64 million patients worldwide with a growing burden anticipated as the population ages[1,2]. Echocardiographic screenings in the general population have revealed that as many as half of individuals living with heart failure may be undiagnosed, preventing earlier access to mortality-reducing treatments[3,4]. Applying risk prediction tools enables earlier identification of diseases, thereby shifting the trajectory of disease progression towards prevention. Additionally, a better understanding of which risk factors play the largest role in the development of heart failure could shed insight into the mechanisms of disease progression and guide therapeutic management, either generally or on a per-individual basis. We sought to evaluate the predictive accuracy of a modern risk assessment tool that incorporates diverse clinical and genetic data compared to genetic or clinical prediction models alone[4–6].

Clinical prediction tools for cardiovascular disease (CVD) such as Framingham and atherosclerotic cardiovascular disease (ASCVD) risk score (also the Pooled Cohort Equation [PCE]) have been widely applied and updated over time to include a variety of demographic, laboratory, hemodynamic, and medical details[7–11]. Researchers have established risk scores to predict the risk of developing heart failure[12]. However, due to the heterogeneous nature of heart failure, it is difficult to fully capture the risk using clinical data alone as these scores fail to leverage genetic data, which accounts for a portion of the unexplained risk[13–15]. Novel risk scores incorporating diverse clinical data integration with well-powered genetic data are needed to more precisely account for heart failure risk.

Genome-wide polygenic risk scores (PRS) estimate an individual's genetic risk using millions of genetic polymorphisms validated across hundreds of thousands of patients[16,17]. Multiple studies have shown that using a PRS – a weighted sum of genetic effects on certain

diseases or traits across the human genome – can enhance disease prediction and further improve early prevention[6,18]. Multiple efforts have been made to summarize genetic and clinical information to identify high risk patients, but integrating high-dimensional genome-wide association study (GWAS) and electronic health record (EHR) in heart failure prediction models has not been previously evaluated[19–21].

We explore approaches to enhance the prediction of future heart failure events leveraging both genetic and clinical data. Our study integrates recent insight on the genetic underpinning of heart failure with a novel EHR-based clinical scoring system, referred to as the clinical risk score (ClinRS), to predict future heart failure. The polygenic risk scoring was powered by the largest-to-date heart failure GWAS[22] and the clinical risk assessment used Natural Language Processing (NLP) to capture co-occurrence patterns of medical events within the structured EHR data. From the proposed approaches above, we summarized 907,272 genetic variants into a PRS and 29,346 medical diagnosis codes into a ClinRS. We hypothesized that the additive power of integrating PRS and ClinRS would result in the most powerful heart failure prediction model.

METHODS

To generate the most powered genetic predictor, we meta-analyzed multiple biobank datasets within the Global Biobank Meta-analysis Initiative (GBMI) consortium to generate a heart failure GWAS[22,23]. The GBMI consortium aims to enhance GWAS power and improve disease risk prediction via collaborative efforts through biobanks across the world and making all GWAS summary statistics open-access for researchers. The case count of the heart failure GWAS from GBMI is the largest-to-date and the PRS generated from GBMI meta-analysis GWAS is expected to have higher accuracy in predicting future heart failure events.

To extract clinical information from EHR, we developed novel machine learning methods that efficiently summarized large scale structured EHR data into a heart failure ClinRS. We treated medical diagnosis codes (i.e., International Classification of Diseases [ICD] code) as 'words' in human language and adapted NLP methods to capture the co-occurrence pattern between codes in the high-dimensional medical records. The co-occurred relationship among codes was later used to extract independent information and converted into low-dimensional numeric vectors resembling the context and semantics of medical events. The University of Michigan's Institutional Review Board approved these protocols (HUM00128472 and HUM00143523).

Michigan Medicine EHR system and biobank

Three cohorts of Michigan Medicine (MM) patients were used in this study: 1) Primary Care Provider cohort (MM-PCP; N=61,849), 2) Heart Failure cohort (MM-HF; N=53,272), and 3) Michigan Genomics Initiative cohort (MM-MGI; N=60,215) (Supplementary Figure 1). Individuals in all three cohorts underwent at least one surgical procedure within the MM

healthcare system. The data were recorded between 2000 to 2022 in the Michigan Medicine EHR system, which includes both ICD-9 and ICD-10 diagnosis codes.

Inclusion criteria for MM-PCP cohort include i) patients with primary care providers within Michigan Medicine, ii) had received an anesthetic, iii) most recent visit was in 2018 or later, and iv) had 5 or more years of medical encounter history (difference between last and first encounter year greater or equal to five) within Michigan Medicine. Exclusion criteria for this cohort include i) patients recruited in Michigan Genomics Initiative and ii) patients predefined in the Heart Failure cohort to ensure no sample overlap with datasets used to validate the clinical predictor.

The MM-HF cohort was defined by a previously validated heart failure phenotyping algorithm[4]. The phenotyping algorithm incorporated ICD diagnosis codes, medication history, cardiac imaging, and clinical notes (free text) to assign the disease outcome for each individual. Clinical expert adjudication was performed on 279 individuals to serve as the gold-standard label for algorithm validation.

The Michigan Genomics Initiative (MGI) is an EHR-linked biobank hosted at the University of Michigan with genotype data linked to EHR information to facilitate biomedical research. With both genetic and clinical data for all individuals in MM-MGI, we are able to validate the prediction models using genetic and/or clinical information. The MM-MGI cohort used in this study is from data freeze 4 (release date: July 2021)[24].

The study cohorts were subset to individuals who self-reported as European American in MM-HF and MM-MGI cohorts, to avoid having reduced performance of genetic predictors in non-white ancestries thereby biasing the model evaluation towards favoring clinical predictors.

The proportion of European American individuals in MM-HF and MM-MGI cohorts is 90% and 86%, respectively.

We refer to MM-PCP cohort as code embedding derivation set, MM-HF cohort excluding individuals in MM-MGI cohort as ClinRS weights derivation set, and the interaction of MM-MGI and MM-HF cohort as model validation set (Supplementary Figure 1). First, the code embedding derivation set was used to learn EHR code patterns and build medical code embeddings for downstream analysis. Patients with a rich medical history and active records within the system were included for code co-occurrence pattern learning in the code embedding derivation set. Next, the labels curated in the MM-HF cohort served as the outcome in the ClinRS weights derivation set to obtain the weights to calculate ClinRS for heart failure cases prediction. The ClinRS weights derivation set included 7,120 individuals from MM-HF which excluded individuals in MM-MGI. Last, the model validation set (independent from ClinRS weights derivation set) was used to validate the prediction ability of PRS and ClinRS. The model validation set consisted of 20,279 participants, representing the intersection of individuals from both MM-MGI and MM-HF cohorts. All patients in the model validation set had a phenotyping algorithm assigned label for heart failure outcome, were fully genotyped to calculate PRS, and with EHR data to generate ClinRS (Supplementary Figure 1).

Polygenic Risk Score (PRS)

The polygenic risk score was calculated using the heart failure GWAS from the Global Biobank Meta-analysis Initiative. GBMI is a global collaboration network of 23 biobanks, across 4 continents and with more than 2.2 million participants (as of April 2022)[23]. The summary statistics from nine of the GBMI heart failure contributing cohorts (BioMe, BioVU, Estonian

Biobank, FinnGen, HUNT, Lifelines, Partners Biobank, UCLA Precision Health BioBank, and UK Biobank) were meta-analyzed resulting in 974,174 individuals of European ancestry in the combined GWAS. These nine biobanks contributed a total of 51,274 heart failure cases and 922,900 healthy controls, defined by phecode 428.2 (heart failure, not otherwise specified)[25,26]. The GBMI heart failure study has the highest heart failure case number in a published GWAS study to date and more advanced genotyping imputation reference panels were used in the participating cohorts. The advancement in GBMI heart failure GWAS improved the statistical power to more precisely identify the genetic risk associated with the outcome[27,28]. In this study, we used the GBMI European-ancestry meta-analysis GWAS to generate a heart failure PRS, which is the current best performing heart failure PRS for European American individuals.

The weights used to create PRS were calculated with PRS-CS[29], using European individuals from the 1000 Genome and UK Biobank combined cohort as the LD reference panel[30,31]. The meta-analyzed heart failure GWAS summary statistics from GBMI used in this study excluded the MGI cohort, which is independent from the validation set used in the analysis to compare the effect contribution between genetic and clinical information for predicting heart failure. Possible population substructure was controlled by regressing the raw PRS on the top 10 principal components (PC) derived from the patient's genotype file. The resulting residuals were inverse normalized to transform the final PRS score into a standard normal distribution.

Clinical Risk Score (ClinRS)

To extract information from high-dimensional EHR data, we developed a novel clinical risk score, ClinRS, to summarize a patient’s longitudinal medical records into one single risk score via NLP techniques. The overall procedure is as follows. First, we treated 29,346 EHR

diagnosis ICD codes as ‘words’ and concatenated all codes documented in a patient’s whole medical history into an ‘article’ using the MM-PCP cohort. After we created the article from all patients, we applied an adapted NLP technique to obtain numeric vector representations that captured the semantic meaning and context of medical codes[32–34]. These vector representations were subsequently validated to be clinically meaningful, in the sense that it captured the concept of each code and showed high concordance with expert manually curated phenotypic grouping labels. We refer to these representations as medical code embeddings in the remainder of this manuscript.

We leveraged the medical code embeddings to generate patient-level latent phenotypes according to a patient’s code utilization, using MM-HF cohort. Next, the latent phenotypes were used to predict disease outcome and the model coefficients were utilized as weights (effect sizes) for calculation of the ClinRS. Finally, a ClinRS was created, which is a linear combination of i) coefficients learned from the ClinRS weights derivation set (MM-HF, excluding MM-MGI) and ii) patients’ latent phenotypes in the model validation set (intersection of MM-MGI and MM-HF) (Supplementary Figure 1). With these steps, we successfully reduced the data dimension from 29,346 unique ICD codes to 350 latent phenotypes, then to a single risk score. See below sections for details of latent phenotypes and ClinRS curation.

Extraction of medical code embeddings using NLP

The first step to summarizing the EHR data using NLP was to convert a patient's EHR medical codes from all healthcare encounters to paragraphs, then concatenate the patient’s paragraphs of medical codes to create an article. After converting EHR data to an article, we were able to derive the co-occurrence patterns of each pair of medical codes. We extracted the semantic meaning of each code into numeric vector representations (medical code embeddings)

that contain clinically meaningful information. See curating medical code embedding section in the supplementary materials for details on the NLP approach to generate vector representation of medical codes.

Evaluation of NLP derived medical code embeddings and parameter tuning

The algorithm for obtaining the medical code embeddings as described above has two tuning parameters including the time window *t* and embedding dimension *d* (i.e., the number of features/ elements in a code embedding). The principle used in parameter tuning is to optimize the clinical meaningfulness of the medical code embedding. The code embeddings should capture similarity of the codes and thus be able to identify whether two specific codes describe the same overall medical concept (i.e., grouping of ICD codes).

To select the optimal time window *t* and embedding dimension *d*, we developed a set of true labels for ICD code grouping using an expert curated ontology named phenow-wide association study code (phecode)[25]. Next, we evaluated whether code pairs that are mapped to the same phecode have larger cosine similarity (i.e. the cosine value of the angle between the corresponding medical code embedding vector pairs) than randomly selected pairs. The cosine similarity is a distance metric measuring how close the two codes are alike in terms of their concepts and meanings. It ranges from -1 to 1, with high cosine values representing that the selected pair of two codes have more similar semantic meaning and utilization context. These evaluations aid in the search for the most 'clinically meaningful' yet efficient version of medical code embedding with the smallest necessary dimension.

In this analysis, phecodes are rolled up to the integer level[26]. For example, ICD-9 code 428.2 (systolic heart failure) and 428.3 (diastolic heart failure) are mapped to the phecode 428.3 (heart failure with reduced EF) and 428.4 (heart failure with preserved EF), respectively. After

rolling phecode 428.3 and 428.4 up to 428 as an integer, these two ICD-9 codes (i.e., 428.2 and 428.3) belong to the same phecode group (i.e., 428). Moreover, both ICD-9 and ICD-10 codes can be mapped to the same phecode. For example, ICD-9 code 428.1 (left heart failure) and ICD-10 code I50.1 (left ventricular failure) are both mapped to phecode 428.2 (heart failure) and further rolled up to the integer 428.

To search for the most clinically meaningful medical code embeddings, we performed a classification task using phecode label and cosine scores. The classification label was the binary indicator of whether the two codes shared the same phecode. The classification score was the cosine distance score calculated between vector representations for two codes. This classification task showed whether a pair of codes mapped to the same phecode have higher cosine similarity (similar semantic representations). The classification results were evaluated using Area Under the Receiver Operating Characteristics (AUC). To distinguish the AUC used in the subsequent evaluation of the heart failure prediction model, we refer to the AUC aiding grid search for optimal NLP derived medical code embeddings based on existing clinical concept ontology as concept-AUC. Concept-AUC is used throughout the remainder of this article for evaluating whether the medical code embeddings derived from NLP is clinically meaningful, in the sense that it can aid identifying whether arbitrary pairs of codes are describing the same concept or belonging to the same general group. The time window *t* and embedding dimension *d* combination that achieves the highest concept-AUC was selected, the corresponding code embeddings were generated accordingly.

In the grid search for time windows *t* and embedding dimension *d*, cosine similarity for 430,579,185 pairs of codes among 29,346 unique codes were calculated for each time window and embedding dimension combination. Ten *t* time windows (1, 2, 7, 10, 14, 20, 30, 40, 50, and

60 days) and twelve *d* embedding dimensions (10, 30, 50, 100, 150, 200, 250, 300, 350, 400, 450, and 500) were evaluated. This results in a total of 120 concept-AUC calculated to evaluate the concept derived from NLP in EHR data that are clinically applicable.

Calculation of patient-level latent phenotypes

To create latent phenotypes for each patient, we used the medical code embeddings derived from the MM-PCP cohort curated from the previous step, and applied this information to the diagnosis codes documented in medical records of patients in the MM-HF cohort. Specifically, we summed up medical code embeddings corresponding to all codes present within a patient's medical record. These latent phenotypes summarize the information of a patient's medical diagnosis history. See details for creating patient-level latent phenotypes in supplementary material.

Time point specific latent phenotypes

We sought to evaluate how far in advance we could predict heart failure and avoid label leakage. The rationale of avoiding label leakage is to not use the information not existing in the prediction period to predict outcome, which could lead to overestimating the model performance. For example, we would like to avoid using the disease treatment or procedure information that is only available after disease diagnosis. To do this, we removed all ICD codes a year prior to the heart failure diagnosis date and then calculated the latent phenotypes. We repeated this procedure by excluding all ICD codes two years prior, in intervals of one year up to ten years prior to disease diagnosis. A total of ten sets of latent phenotypes using different time point cutoffs to remove the medical history were generated. Patients with no medical history recorded within the healthcare system prior to the cutoff time point were removed from the analysis. See Supplementary Table 1 for sample size in each time point.

Supervised training for ClinRS using LASSO

To summarize the multi-dimensional patient-level latent phenotypes into a single risk score, we applied the Least Absolute Shrinkage and Selection Operator (LASSO) for feature selection with 10-fold validation for shrinkage parameter tuning[35]. The LASSO leverages the L1 penalty on the regression coefficients to eliminate non-important variables, avoid overfitting, and achieve better prediction. Next, the coefficients yielded from the LASSO model were used as weights (effect sizes) to calculate a weighted sum of patients' clinical risk. In the ClinRS weights derivation set (individuals in MM-HF excluding MM-MGI), the patients' latent phenotypes were calculated using EHR records one year prior to heart failure diagnosis (Supplementary Figure 1). The heart failure outcome was regressed on 350 latent phenotypes and adjusted for age, sex, and healthcare utilization using logistic regression with L1 regularization. Three patient characteristics known to be predictive of the outcome (age, sex, and healthcare utilization) were forced in the model with no shrinkage. Patients' healthcare utilizations were summarized by the number of months of encounters recorded in the EHR.

Calculate ClinRS for patients in model validation set

To validate the prediction accuracy of ClinRS, we applied the ClinRS weights obtained from the ClinRS weights derivation set to an independent model validation set to summarize the entire EHR diagnosis records into one score (Supplementary Figure 1). The score was further used in the heart failure prediction model to predict patients disease outcome in the future. For each participant in the model validation set, ten ClinRS were calculated using time point specific latent phenotypes from one year up to ten years prior to disease diagnosis. Next, we performed inverse normalization to convert the ClinRS score into standard normal distribution.

Statistical analysis

We conducted analyses within cohorts of 20,279 individuals in the model validation set (intersection of MM-MGI and MM-HF) with at least 1 year of medical history prior to a heart failure diagnosis in the Michigan Medicine health system. Ten different datasets with time point cutoffs, one year apart from one year to ten years prior to disease diagnosis, were applied to the analysis. Individuals with no medical history prior to the time point cutoff were removed from the respective year specific analysis. Sample size in each time point-specific dataset decreased from one year to ten years prior to disease diagnosis, ranging from 20,279 (576 cases) to 10,391 (332 cases) participants, respectively (Supplementary Table 1).

We fit four logistic regression models to predict whether patients have heart failure and further evaluated the accuracy among models with different risk predictor(s) for all ten time points, one year apart from one year to ten years prior to disease diagnosis. The baseline model included patients' demographic information (age at diagnosis and sex), and three additional models with the risk score added: i) PRS, ii) ClinRS, and iii) PRS+ClinRS were created to compare the improvement in model accuracy from the baseline model. In the PRS and PRS+ClinRS models, the top ten PCs derived from patients' genotype data were adjusted to account for the population structure. Model performances were compared using 10-fold cross validated AUC. The analysis was performed using European ancestry samples only.

Sensitivity analysis removing circulatory system diagnosis codes

Additional analyses on ClinRS validity were conducted to examine the robustness of the co-occurrence patterns captured by the unsupervised NLP algorithm. We created a ClinRS without circulatory system information (ClinRS-NoCirc) by excluding ICD diagnosis codes

belonging to ICD-9 Seventh Chapter (390-459) and ICD-10 Chapter IX (I00-I99): Diseases of the Circulatory System. The ClinRS without circulatory system was further used in model prediction to evaluate the ability of the proposed method to predict disease outcome (heart failure) without directly associated diagnosis information (circulatory system diagnosis codes). We excluded 1,340 circulatory system diagnosis codes (459 from ICD-9 and 881 from ICD-10) and used the rest of the 28,006 codes to create patient-level latent phenotypes, and applied the newly derived latent phenotypes with ClinRS weights derived previously to generate ClinRS-NoCirc. We demonstrated that using pre-trained co-occurrence patterns from an independent dataset could be useful for disease prediction and the co-occurrence patterns aided capturing disease risks through indirect associations.

RESULTS

In this paper, we utilized three independent datasets (Supplementary Figure 1) at Michigan Medicine to achieve two main goals in this study: 1) obtain medical code embeddings using NLP in EHR data and 2) improve heart failure prediction using PRS and ClinRS. First, we used MM-PCP cohort with a total of 61,849 individuals and 159,273,800 ICD diagnosis codes recorded from 2000 to 2022 to learn the medical code co-occurrence patterns and to extract medical code embeddings representing the clinical meaning of each code. The medical code embeddings trained from MM-PCP were validated using phecodes to demonstrate that vector representations derived from unsupervised NLP method contextually are clustered in similar ways compared to expert manually curated code grouping (Supplementary Figure 2).

Next, we built two risk scores, PRS and ClinRS, in the model validation set (intersection of MM-MGI and MM-HF) to predict future heart failure cases. The PRS was calculated using heart failure GWAS summary statistics, meta-analyzed from nine biobanks in GBMI (independent from Michigan Medicine)[22]. We chose the European ancestry GWAS summary statistics (51,274 cases and 922,900 controls) as the base of our PRS due to its superior performance in the European ancestry individuals in the original publication. The ClinRS calculation required two steps: i) create patient-level latent phenotypes and ii) derive weights (effect sizes) to calculate ClinRS. We generated medical code embeddings for 29,346 medical codes from MM-PCP, and then used the medical code embeddings to create 350 latent phenotypes for each patient in MM-HF. To derive weights for the ClinRS, we regressed heart failure outcome on latent phenotypes in ClinRS weights derivation set (MM-HF, excluding MM-MGI) and extracted the effect sizes as ClinRS weights. The ClinRS weights derivation set had a heart failure incidence of 330 out of 7,120 patients (4.6%) whereas in the model validation set

we observed 576 (2.8%) heart failure cases out of 20,279 patients (Supplementary Figure 1). From these summary statistics, a total of 907,272 genetic variants were integrated into a polygenic risk score. In analogy, 29,346 medical diagnosis codes were integrated into a clinical risk score. Below we summarize our findings.

NLP extracted medical code embeddings are clinically meaningful

First, we validated whether the medical code embeddings generated in MM-PCP cohort were clinically meaningful and that NLP could capture the information hidden in the complex EHR dataset. We used the cosine distance between a pair of codes to classify whether a code pair shared the same phecode (i.e., have similar clinical concept) and calculated the concept-AUC. The results here served as a proof of concept of whether the medical code embeddings derived from the MM-PCP cohort is suitable to be used for generating a ClinRS.

We discovered two main findings: 1) smaller time window size *t* and 2) inclusion of more features *d* in a code embedding yielded higher accuracy on identifying code pairs in the same phecode group. Supplementary Figure 2 showed that holding constant embedding dimension *d* while varying time window size *t*, the highest concept-AUC was consistently found from co-occurrence matrices constructed based on codes that appeared on the same day (within 1 day). The accuracy attenuated linearly when the window size increased. For example, concept-AUC calculated from embedding dimension of 350 was the highest for codes co-occurred on the same day (1 day) with concept-AUC of 0.78, decreased to 0.76 for codes co-occurred within 1 week (7 days), and dropped to the lowest of 0.73 for codes that co-occurred within 2 months (60 days). These results indicated that diagnosis codes recorded on the same day provided the most information about code relationships. One possible explanation could be that diagnostic codes

were often all billed on the same day, likely the last day of the hospitalization. Additionally, by increasing the time window of codes considered for co-occurrence, it could also potentially introduce noise (e.g., diagnosis code not related to the same medical event) and lower the ability to construct meaningful semantic vector representations. Next, we evaluated the concept-AUC variation across different numbers of features *d* in a code embedding. In general, the higher the embedding dimension *d*, the higher the concept-AUC was observed. The concept-AUC plateaued with up to embedding dimensions of 300 to 500, depending on the time-window. This finding is similar to previous reports[36–40]. The optimal embedding dimension found in this study using Michigan Medicine EHR data was $d = 350$ (Supplementary Figure 2).

The medical code embeddings generated from time window $t = 1$ day with embedding dimension $d = 350$ yielded a concept-AUC of 0.78 (Supplementary Figure 2). This result supports that the medical code embeddings derived via unsupervised learning were clinically meaningful validated by expert manually curated phenotypic grouping. The medical code embeddings corresponding to the above chosen tuning parameters were further used to calculate patient-level latent phenotype in this analysis.

In addition to numerically evaluating the semantic resemblance of vector representations using concept-AUCs, we further assessed the semantic relationship graphically using a heatmap of the cosine similarity scores (Supplementary Figure 3). In this study, we used ICD-9 Second Chapter (140-239): Neoplasms as an example to discern how the similarity patterns were formulated among each cancer code. Cancer codes were selected to demonstrate the similarity patterns of code pairs due to its distinct organ system specific sub-chapter within the cancer codes. For example, codes from cancer of digestive organs (ICD: 150-159) and cancer of

respiratory organs (ICD: 160-165) are both cancer codes, but for different organs and were expected to have different patterns and concepts.

As anticipated, we observed that the same ICD-9 diagnosis codes and/or nearby codes (off-diagonal line in Supplementary Figure 3) had higher cosine values between their embeddings, indicated by the darker color on the off-diagonal line and the band surrounding it. Furthermore, clear distinctions crossing different sub-chapters were found. These results suggest the contextual representations were clinically meaningful since related types of cancers from the same organ system had more similar context and patterns of co-occurred comorbidities, treatments, or procedures. Conversely, lower cosine scores were found in code pairs between different sub-chapters of cancer diagnosis ICD codes.

PRS and ClinRS each predict heart failure cases up to eight years in advance

We evaluated the prediction ability of using genetic and clinical information, separately, to identify heart failure patients in the future. We used 10-fold cross validated AUC to assess how well each risk score predicted the event of heart failure at ten different time points prior to the disease diagnosis date. Ten different time points used were one year apart from one year to ten years prior to disease diagnosis; for simplicity, we refer to the ten cutoffs prior to disease diagnosis as ten time points.

We summarized the AUCs of ten time points from different models (baseline, PRS, ClinRS, and PRS+ClinRS model) in Figure 1. In this study, we found that PRS and ClinRS each and separately were able to predict heart failure outcomes significantly better than the baseline model (age and sex only), up to eight years prior to heart failure diagnosis. Results from one year prior to the diagnosis, significantly higher AUC was observed in the PRS model (AUC: 0.76

[95% CI: 0.74-0.83]) and ClinRS model (AUC: 0.85 [0.83-0.87]), compared to the baseline model with AUC of 0.70 (0.68-0.72). See Supplementary Table 1 for the specific AUC values across all ten time points and four different models. As expected, we observed that the benefit of ClinRS prediction was attenuated by censoring EHR data with increasing time thresholds prior to the event, and the model accuracy decreased when sample size is smaller due to earlier censoring. Nevertheless, better performance in both PRS and ClinRS models were continuously observed in the analysis until eight years prior to the disease diagnosis. For example, in a cohort with at least eight years or more of medical history within Michigan Medicine, the PRS and ClinRS models yielded an AUC of 0.76 (0.74-0.78) and 0.77 (0.74-0.79), respectively, significantly higher compared to baseline model with AUC of 0.71 (0.68-0.73).

In models given data from nine years prior to disease diagnosis, no significant difference was observed among PRS (AUC: 0.77 [0.74-0.79]), ClinRS (AUC: 0.77 [0.74-0.79]), and baseline (AUC: 0.72 [0.69-0.75]) models. Lack of significant difference between PRS (does not change over time because genetics is fixed at conception), ClinRS, and the baseline model from such a limited dataset from nine years before the event could potentially be due to the smaller sample size. The sample size for the ten-year censored data was 51% of that for the one-year censored data. In addition, EHR data nine- and ten-year prior to disease diagnosis provided insufficient information for complex prediction tasks.

Integrating PRS and ClinRS enhances heart failure prediction

In addition to evaluating the risk score separately, we further studied the additive power of including both risk scores together in the heart failure prediction model. Consistently across all ten time points, the highest accuracy was found in the PRS+ClinRS model. See Figure 1 and

Supplementary Table 1. Significantly higher AUC was continuously found in the PRS+ClinRS model even at ten years prior to disease diagnosis with an AUC of 0.79 (95% CI: 0.77-0.82), compared to baseline model (AUC: 0.72 [0.69-0.75]). Compared to the single risk predictor models predicted heart failure eight years prior to disease diagnosis, the model including both predictors predicted disease two years earlier than using either single risk predictor alone.

As expected, we observed that the prediction accuracy of the PRS+ClinRS model outperformed single risk score models throughout the entire one to ten years time horizons. In Supplementary Figure 4, we showed that by using both clinical and genetic risk scores to predict which individuals have high risk of future heart failure, the combined score discovered the highest proportion (28%) of individuals who had heart failure.

Sensitivity analysis on removing circulatory system diagnosis code

To examine the robustness of ClinRS and to eliminate the concerns of overfitting, we conducted a sensitivity analysis by removing all circulatory system diagnosis codes to create ClinRS-NoCirc. In Supplementary Figure 5, we presented the model performances of using ClinRS-NoCirc as the clinical risk predictor and compared to the ClinRS model. The results were largely similar with and without removing circulatory system codes, which demonstrated that we successfully built a risk score that leveraged the high-dimensional EHR records and apprehended underlying patterns to reveal disease associations. Specifically, the models using ClinRS-NoCirc to predict future heart failure events yielded significantly higher accuracy than baseline models, up to six years in advance of disease diagnosis. We observed an AUC of 0.77 (0.75-0.80) from ClinRS-NoCirc model at six years prior to disease diagnosis, which was significantly higher than baseline model at six years in advance of heart failure diagnosis (AUC:

0.72 [0.69-0.74]) (see Supplementary Figure 5 and Supplementary Table 1). Although the results derived from ClinRS-NoCirc could not predict the outcome as many years in advance as the ClinRS model, the additive power of integrating genetic and clinical information in disease risk prediction remains evident through ClinRS-NoCirc. By including both PRS and ClinRS-NoCirc in the heart failure prediction model, we were still able to distinguish patients with high risk of heart failure a decade in advance of the disease diagnosis. The model with PRS and ClinRS-NoCirs predictors showed a significantly higher AUC of 0.78 (0.76-0.81) at ten years prior to heart failure diagnosis, compared to the baseline model with AUC of 0.72 (0.69-0.75).

ClinRS insights

We dissected the composition of ClinRS for heart failure prediction and further studied the risk and protective factors associated with disease outcome in Supplementary Figure 6. In Supplementary Figure 6, we showed the ClinRS weights of risk and protective factors contributing to the heart failure outcome. The diagnoses prioritized in the ClinRS score can generally be classified by 1) organ system (cardiac versus non-cardiac) and 2) etiology (potential causal mechanism, associated comorbidity, or unclear link). As expected, 7 out of the top 10 risk factor for heart failure in ClinRS were cardiac diagnoses, exhibiting potential causal mechanisms; for example, ICD codes associated with acute myocardial infarction[a] (Supplemental Table 2). Additional potential cardiac-causal diagnoses including: i) stenosis, mitral and aortic valves (ICD: 396.0), ii) acute myocarditis (ICD: 422.0), and iii) defect, acquired cardiac septal (ICD: 429.71) were highly prioritized by the ClinRS algorithm. Also, ClinRS

[a] 410.91 = Acute myocardial infarction of unspecified site, 410.21 = ST elevation (STEMI) myocardial infarction involving other coronary artery of inferior wall, 410.41 = Acute myocardial infarction, of other inferior wall, 410.01 = Acute myocardial infarction, anterolateral wall, initial, 410.51 = Acute myocardial infarction, lateral wall, initial, 410.71 = Acute myocardial infarction, subendocardial, initial, and 410.61= True posterior wall infarction, initial

incorporates many associated-cardiac diagnoses including i) malfunction, cardiac pacemaker (ICD: 996.01) and ii) mechanical complication of automatic implantable cardiac defibrillator (ICD: 996.04). These codes are likely to co-occur in patients with heart failure, but may have limited utility in predicting new or previously undiagnosed cases - although it is noteworthy that all diagnoses included in ClinRS were documented prior to the heart failure diagnosis. Diagnoses identified by ClinRS including: i) Marfan syndrome (ICD: 759.82, 754.82)[41], ii) alcohol abuse (ICD: 303.01, 790.3, 980.0)[42], and iii) viral infection (ICD: 74.8)[43] may reflect non-cardiac, causal mechanisms of heart failure pathogenesis. Notably, non-cardiac diagnoses, unclear link with a protective effective against heart failure in the ClinRS score included a cluster of pregnancy-related conditions (ICD: 765.14, 765.25, 656.43, 678, etc) and another cluster of ophthalmologic diagnoses (ICD: 371.03, 370.03, 370.63, 374.23, 370.35, etc). No causal or mechanistic relationship should be inferred -- instead this correlation likely results from the lower-risk baseline population (childbearing females) for pregnancy related-conditions and more focused, clinical ophthalmologic assessment being less likely to diagnose heart failure, for the ophthalmologic-conditions.

DISCUSSION

This study sought to improve the accuracy of heart failure prediction by integrating high-dimensional genetic data with clinical information to advance heart failure prevention initiatives. Genetic risk was summarized by a PRS, calculated from the largest-to-date heart failure GWAS[22], and clinical risk was summarized by a ClinRS, a novel EHR-based risk score. The combined PRS and ClinRS score prediction model identified patients with a high risk of heart failure a decade in advance of the disease diagnosis (Figure 1 and Supplementary Table 1). Specifically, the PRS+ClinRS prediction model showed a significantly higher AUC at ten years prior to heart failure diagnosis with AUC of 0.79 (0.77-0.82) compared to the baseline model with AUC of 0.72 (0.69-0.74). In contrast, models with a single risk score alone can only identify heart failure cases eight years in advance, by integrating genetic and clinical information we are able to identify heart failure cases two years earlier. These findings reveal the power of integrating PRS and ClinRS to enhance disease prediction and the potential to inform heart failure prevention efforts. More broadly, this study highlights the methods and opportunity to curate ClinRS for other complex diseases and integrate with PRS to improve disease prediction accuracy.

Advances in comprehensively utilizing longitudinal and high-dimensional EHR data

The critical challenges of incorporating EHR data are its high dimensionality and longitudinal nature. We successfully developed a risk score summarizing the clinical information despite the complexity of EHR data and validated its utility in an independent dataset from an EHR-linked biobank cohort. This study treated structured EHR diagnosis codes as human language and converted the diagnosis code into paragraphs. This enabled learning the coding

patterns for patient records with any dimensionality and longitudinal history. By focusing on co-occurrence patterns of medical codes within a specified time window, we were able to utilize data from all individuals regardless of the length of healthcare utilization. Patients with only one visit to decades of medical history within the healthcare system all contributed to the medical code embedding construction. In addition, by applying NLP to transform codes to medical code embeddings, we successfully reduced the high-dimensional EHR dataset into low-dimensional features. The results present an avenue to incorporate other domains of structured EHR datasets, such as medical procedures and laboratory tests, to create a clinical risk score that could more comprehensively capture the risk of having the disease.

An integrated model (PRS + ClinRS) enables improved prediction of heart failure

We previously developed a heart failure GWAS with the largest number of cases to date to build heart failure risk prediction models[22]. We successfully reduced high-dimensional GWAS into a single predictor – PRS. Furthermore, we implemented adapted NLP techniques to capture latent phenotypes in EHR data and summarized it into a new predictor – ClinRS. Analysis results showed that both risk scores were significantly better predictors of heart failure compared to baseline demographic information alone. Additionally, adding both PRS and ClinRS together into prediction models yielded superior accuracy for predicting future heart failure outcomes. This result demonstrated the additive predictive power of leveraging genetic and clinical information in risk prediction.

In alignment with our findings, Mujwara et al., used CAD-PRS to reclassify high genetic risk patients from patients in the borderline or intermediate of PCE clinical risk pool[20]. Findings showed that using the combined PCE and CAD-PRS approach risk screening methods to initiate

early preventive treatment per 10,000 individuals screened could potentially avert 50 ASCVD events over 10 years and lead to substantial cost saving per averted event. It is promising that we have the potential to achieve more accurate prediction by using PRS and ClinRS together into prediction models. Such strategies could then inform guidelines for patient care to aid in earlier initiation of prevention treatment.

Medical code embeddings filled in missing information/ incomplete EHR history

We strengthened the evidence that leveraging genetic and clinical information improves precision health by performing a sensitivity analysis with all circulatory system diagnosis codes removed. Even though clinical information from the EHR system was partially missing, we were still able to reach high prediction accuracy one decade prior to disease diagnosis by incorporating a genetic risk score in the model (Supplementary Figure 5 and Supplementary Table 1). This analysis also indicated the potential benefit for patients with short medical history within the same healthcare system, missing information and/or unrecorded diagnosis would be able to reveal from the incomplete health records using pre-trained medical code embedding[44].

Study limitations

Heart failure is known to have separate subtypes caused by different mechanisms or genetic risk factors, with distinct treatments and phenotypic symptoms[22]. In the future, ClinRS for heart failure subtypes needs to be further validated in cohorts with larger sample sizes. Moreover, the curation of ClinRS and utilization of integrating genetic and clinical information for disease risk prediction needs to be benchmarked in other complex diseases.

Despite the high fidelity, validated clinical outcome assessed across a relatively long surveillance window in a large population, the retrospective study design imposes some intrinsic limitations. While including full diagnostic codes from the EHR, potential selection bias in both timing and medical specialty, may limit clinical relevance and applicability. Furthermore, the retrospective nature of the study has inherent limitations including the possibility of yet unidentified confounding variables.

This study solely utilized the diagnosis information derived from EHR data, however, leveraging other domains of structured and unstructured EHR data (e.g., procedure, medication, clinical notes, etc.) to assist disease prediction is needed to fully understand the additive power of integrating genetic and clinical data.

Furthermore, the limitation of any EHR-based study also includes the low transferability across different healthcare systems due to the heterogeneity of EHR data. Methodology in language models could potentially be borrowed to improve transferability of medical code embeddings and the derived latent phenotypes. Applying transfer learning techniques could also produce a more generalizable ClinRS to be applied across different healthcare systems.

Conclusion

In conclusion, the amalgamation of GWAS- and EHR-derived risk scores predicted heart failure cases 10-years prior to diagnosis. These findings highlight how application of natural language processing to complex datasets such as medical records and incorporating genetic information may enhance the identification of patients with a higher susceptibility to heart failure. Application of this approach at scale may enable physicians to introduce preventive therapies at a much earlier stage, which may prevent the onset of overt heart failure.

Acknowledgements

We would like to express our gratitude to all contributors to GBMI and the biobank participants who provided their data for biomedical research. The authors acknowledge the participants, recruitment teams and project managers of the Global Biobank Meta-analysis Initiative for providing data aggregation, management, and distribution services in support of the research reported in this publication (particularly Sinéad Chapman and Bethany Klunder).

This work was supported by the National Institutes of Health grants R35-HL135824 (CJW) and R01-GM139926 (XS).

Declaration of interests

CJW's spouse works at Regeneron pharmaceuticals. SEG works at Regeneron pharmaceuticals. MRM received grant funding support from the National Heart, Lung, and Blood Institute, Grant K01HL141701, Bethesda, MD.

FIGURES AND TABLES

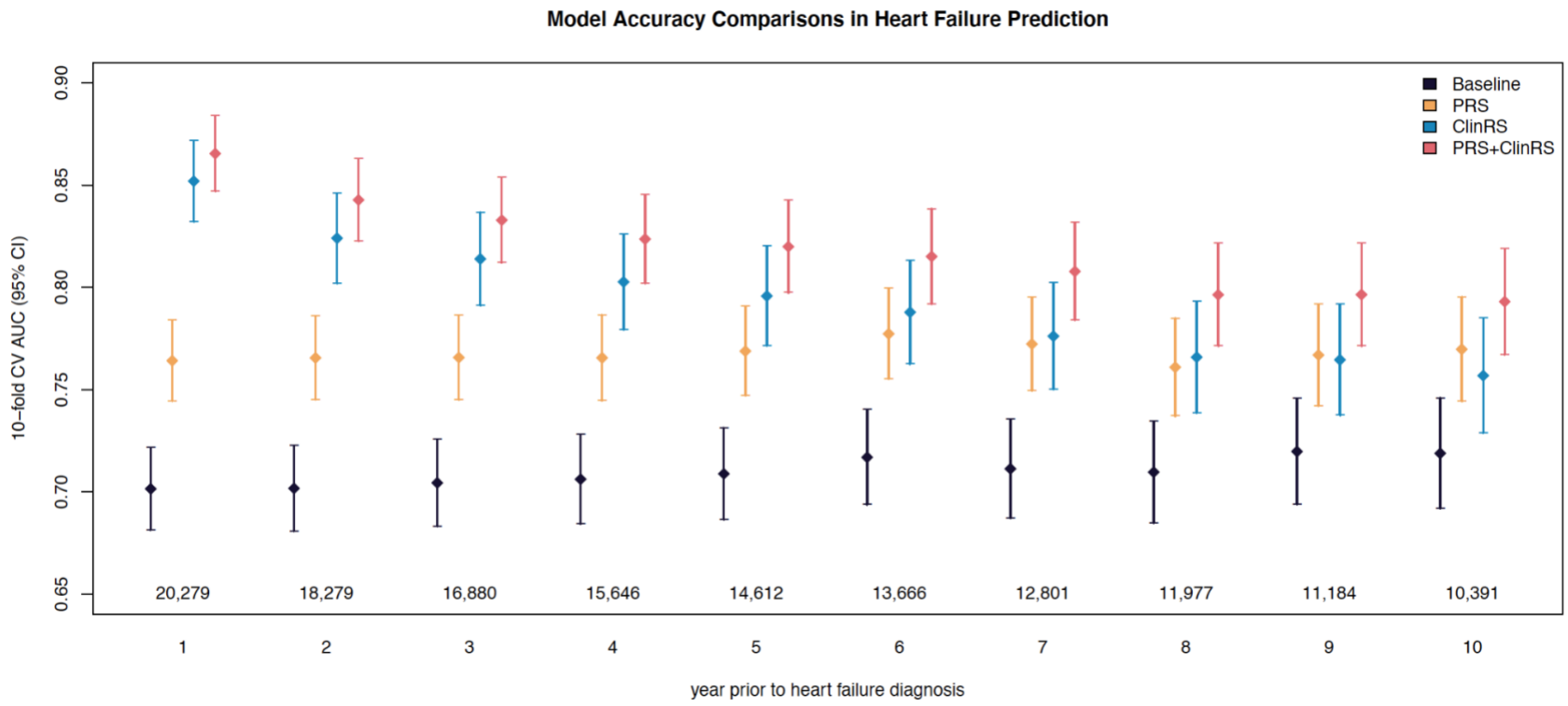


Figure 1. Forest plot comparing models' accuracy of predicting heart failure at one to ten years prior to disease diagnosis. Four models were compared with each time point: baseline (age and sex), PRS (polygeneic risk score), ClinRS (clinical risk score), and PRS+ClinRS. Numbers at the bottom of the plot indicate the sample size for each time point. Results showed that PRS and ClinRS, separately, can predict heart failure outcomes eight years in advance, and adding both risk predictors in the model can predict disease ten years in advance.

SUPPLEMENTARY MATERIAL

Curating medical code embedding

The medical code embeddings were created by learning vector representations of ICD codes based on their co-occurrence patterns in the EHR, which were obtained through adapted NLP method[32]. More specifically, the medical code embeddings were extracted by performing truncated singular value decomposition (SVD) on the shifted positive pointwise mutual information (SPPMI) matrix, which is derived from codes' co-occurrence matrix. The pipeline we developed to extract medical code embedding was based on Hong *et al*.[33] and it is publicly available at https://github.com/The-Shi-Lab/CodeEmbedding.

Co-occurrence Matrix

A co-occurrence matrix is defined with a selected time window *t*, within which the co-occurrence instances of codes are counted. Since there are 29,346 codes, the dimension of this matrix is 29,346-by-29,346, with each entry counting the number of co-occurrence instances in the EHR between the corresponding pair of codes. By this definition, the co-occurrence matrix is a symmetric matrix. Assuming that the selected time window *t* is co-occurred within 7 days, for each code (which we denote by C) and each patient, we first identify the dates when the code was assigned to the patient. Then, for each of these identified dates, we scan the EHR of the patient within the day and the following 6 days; each code assignment found is counted as an instance of co-occurrence with code C. In such a fashion, the co-occurrence matrix is obtained by aggregating the co-occurrence instances over all patients and all codes.

Calculation of medical code embedding

The medical code embeddings were obtained through dimension reduction of the SPPMI matrix, which is derived from the co-occurrence matrix, which we denote by CC. Specifically, the SPPMI matrix share the size of CC which is 29,346-by-29,346 and for each code pair C_1, C_2,

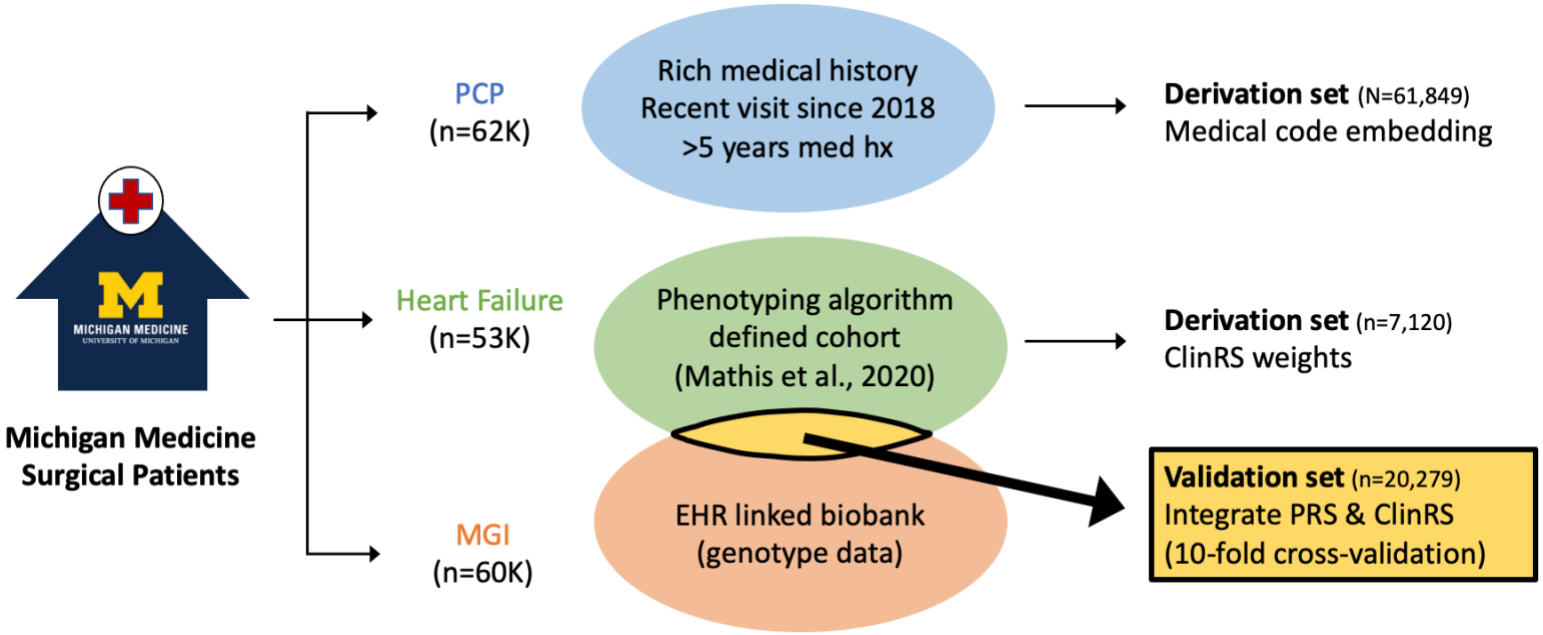
$$SPPMI(C_1, C_2) = max\{log\frac{CC(C_1,C_2)}{CC(C_1,\cdot)CC(C_2,\cdot)} - log(k),0\}$$

where $CC(C_1,\cdot)$ represents the row sum of CC on the row corresponding to C_1. The tuning parameter, negative sample k was set to 10 based on results shown in previous studies[34,40,45]. Given a SPPMI matrix and a desired semantic vector representation (SEV) dimension d, the SEVs are obtained through the truncated singular value decomposition of the SPPMI matrix, which we denote by $U_d diag(\sigma_1, \cdots, \sigma_d){U_d}^T$, where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d$ are the d largest singular values of the SPPMI matrix. Specifically, the d SEVs are the columns of $U_d diag(\sqrt{\sigma_1}, \cdots, \sqrt{\sigma_d})$, which are all vectors with 29,346 entries (one for each ICD code).

<u>Creating patient-level latent phenotypes</u>

To create latent phenotype using EHR data for individuals, we took the product of the patient-level EHR record *D*, a dataset recorded whether patients had the diagnosis code in the past, and code embedding *C*, a semantic vector representation of the EHR codes. *D* is a n by p matrix, where n is the number of patients and p is the number of unique diagnosis codes. *C* is a p by k matrix, where p is the number of unique diagnosis codes and k is the embedding dimension selected from the code embedding curation step. The final product of *D* and *C* will be the patient-level latent phenotypes with dimension of n by k. See Supplementary Figure 7 for illustration

SUPPLEMENTARY TABLES and FIGURES

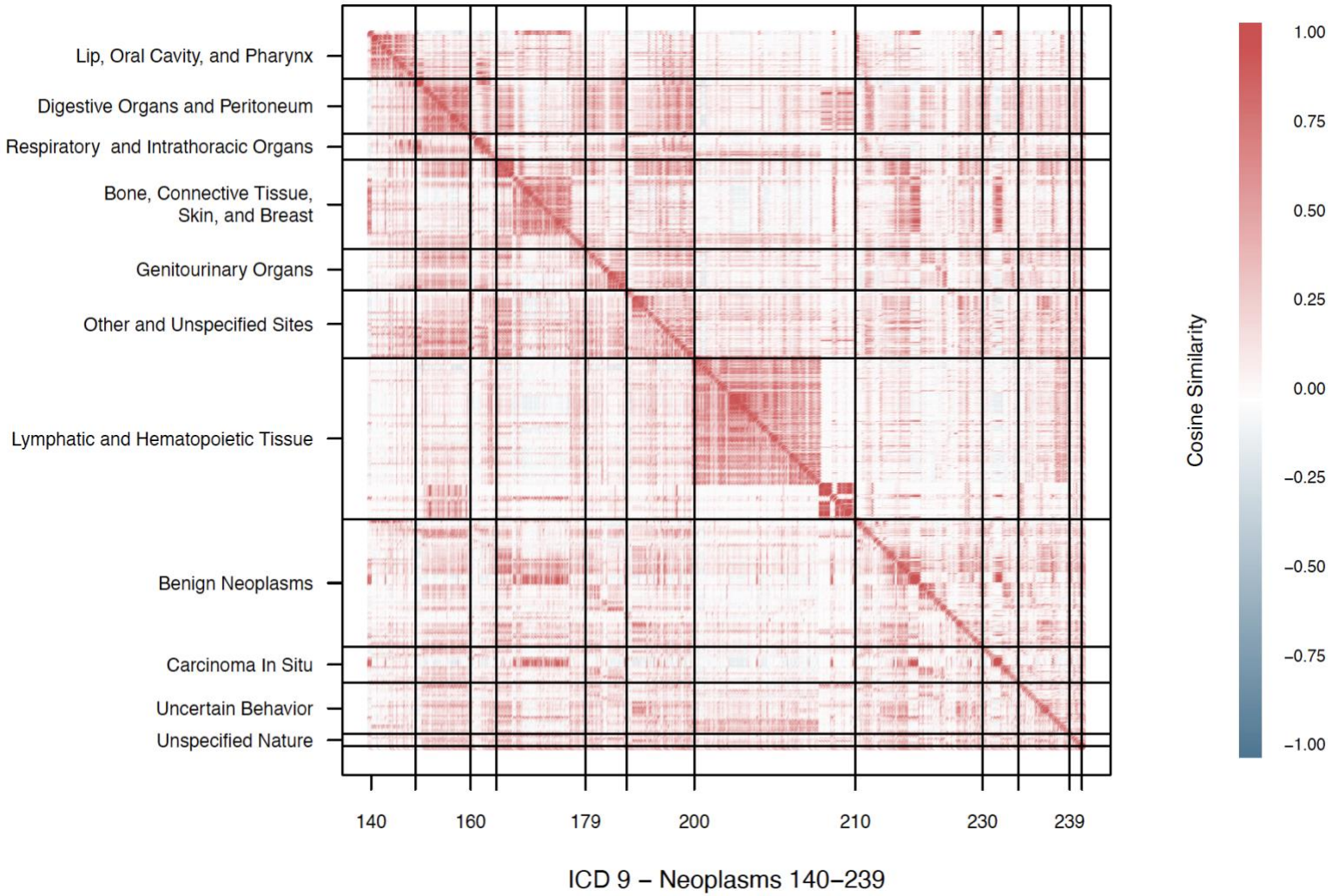


Supplementary Figure 1. Three cohorts within Michigan Medicine (MM) were used in this analysis: i) Primary Care Provider (MM-PCP), ii) Heart Failure (MM-HF), and iii) Michigan Genomics Initiative (MM-MGI). MM-PCP cohort with 61,849 individuals was used to build medical code embeddings. Subset of MM-HF (N=7,120), participants of European descent and not in MM-MGI, was used to derive the weights (effect sizes) of clinical risk score (ClinRS). Subset of MM-MGI (N=20,279), patients fully genotyped and disease outcome was predefined using Mathis et al. phenotyping algorithm[4] in MM-HF, was used to validate heart failure prediction accuracy using polygenic risk score and clinical risk score.

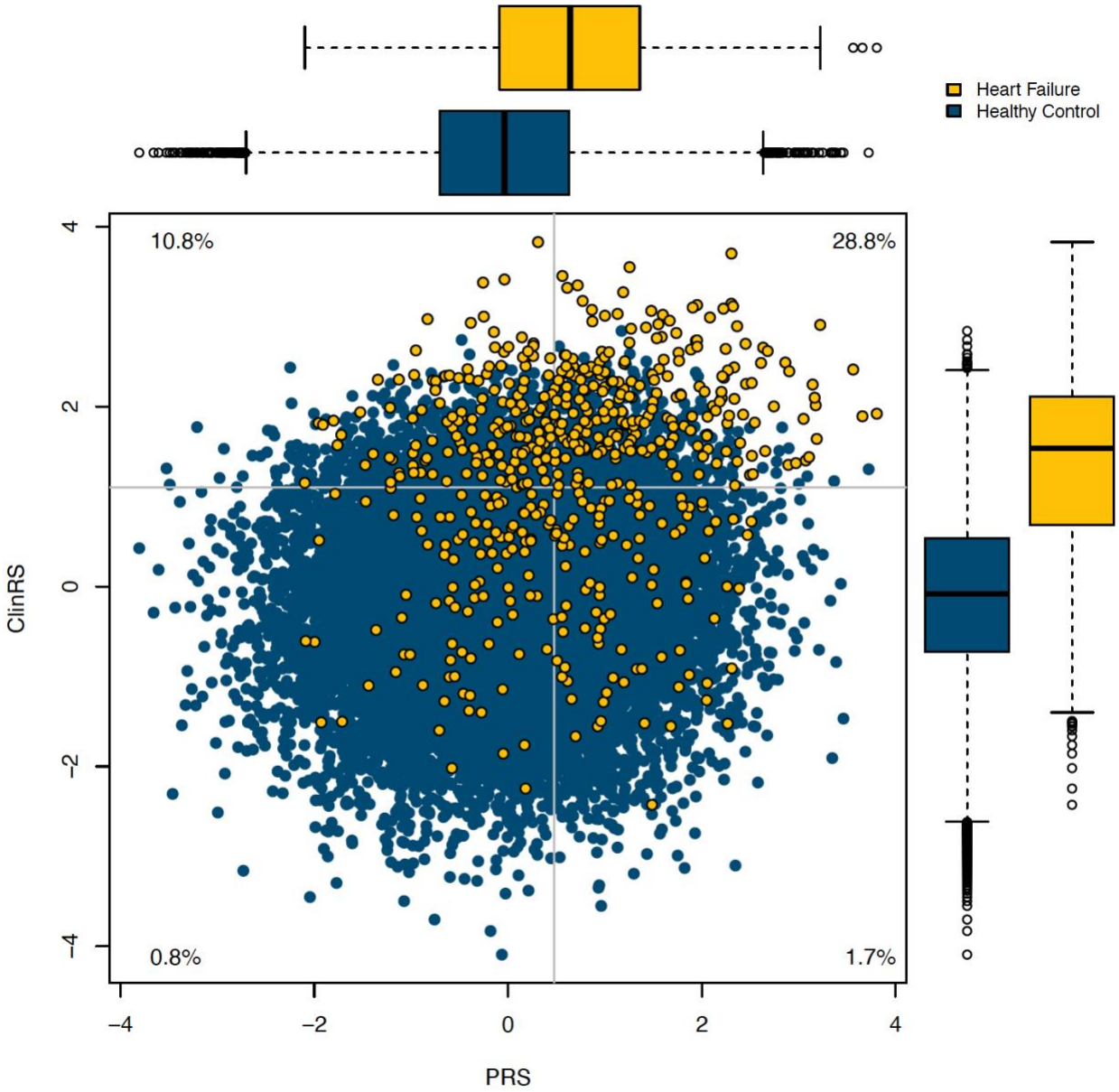
Time Window (co–occur within)	10	30	50	100	150	200	250	300	350	400	450	500
1	0.6274	0.6733	0.6999	0.7294	0.7474	0.7712	0.7789	0.7830	0.7835	0.7814	0.7824	0.7802
2	0.6180	0.6686	0.6927	0.7202	0.7425	0.7636	0.7684	0.7725	0.7734	0.7749	0.7751	0.7747
7	0.6011	0.6644	0.6793	0.7038	0.7363	0.7493	0.7556	0.7583	0.7602	0.7622	0.7647	0.7651
10	0.5951	0.6623	0.6756	0.6992	0.7332	0.7456	0.7508	0.7563	0.7561	0.7573	0.7607	0.7600
14	0.5930	0.6609	0.6733	0.6956	0.7273	0.7380	0.7447	0.7525	0.7529	0.7527	0.7553	0.7571
20	0.5888	0.6562	0.6712	0.6940	0.7228	0.7371	0.7413	0.7468	0.7491	0.7486	0.7512	0.7529
30	0.5855	0.6531	0.6721	0.6871	0.7208	0.7299	0.7363	0.7409	0.7429	0.7427	0.7464	0.7464
40	0.5861	0.6526	0.6771	0.6858	0.7176	0.7275	0.7332	0.7380	0.7394	0.7390	0.7435	0.7431
50	0.5895	0.6504	0.6743	0.6820	0.7177	0.7253	0.7310	0.7339	0.7374	0.7356	0.7396	0.7397
60	0.5908	0.6499	0.6699	0.6798	0.7151	0.7240	0.7289	0.7324	0.7347	0.7334	0.7370	0.7371

Embedding Dimension

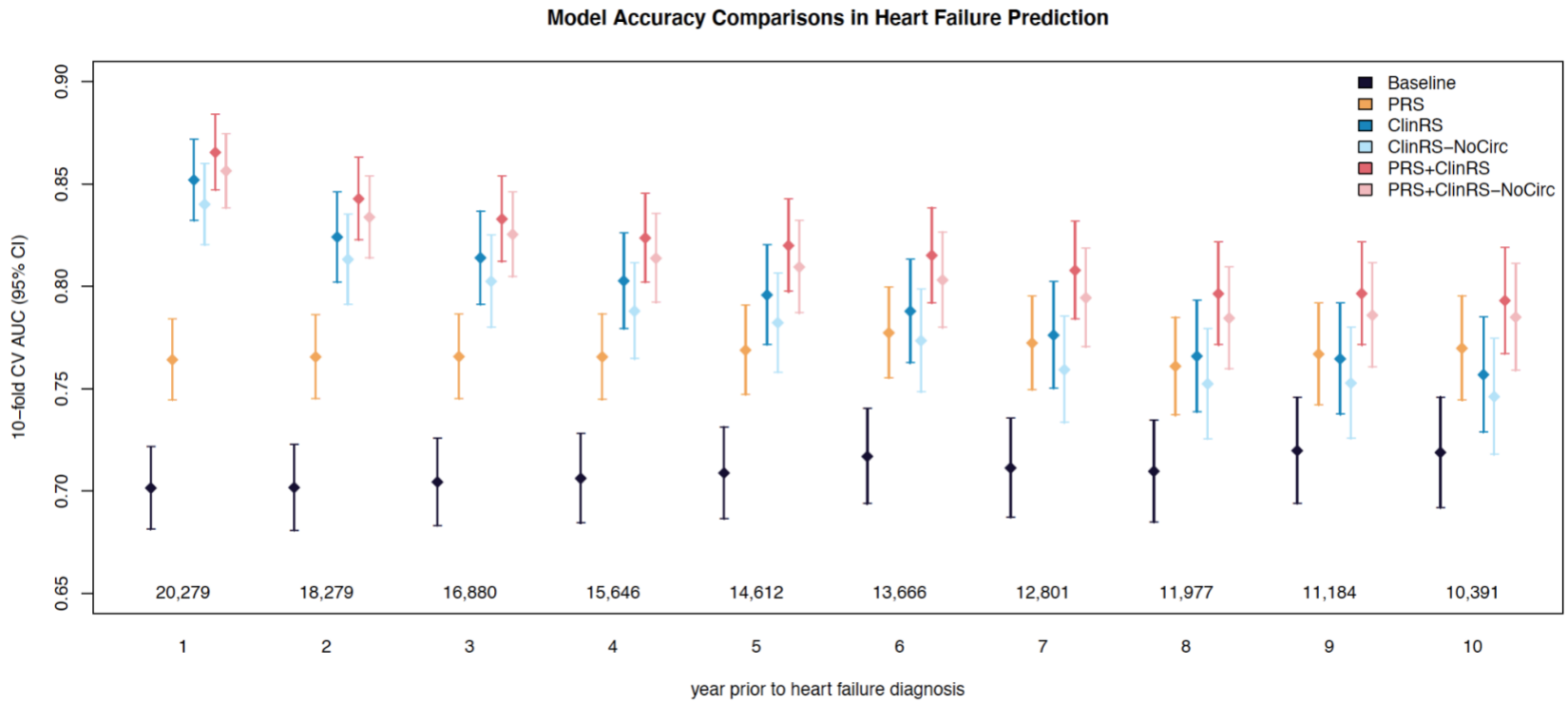
Supplementary Figure 2. Heatmap of concept-AUC across medical code embeddings derived from using 10 time windows and 12 embedding dimensions to summarize a medical code. Concept Area Under the Receiver Operating Characteristics (concept-AUC) summarized how well medical code embeddings generated from the adapted natural language (NLP) processing method capture the clinical meaning of each code. Medical code embedding built on code oc-occurred within 1 day with embedding dimension of 350 yielded the highest concept-AUC.



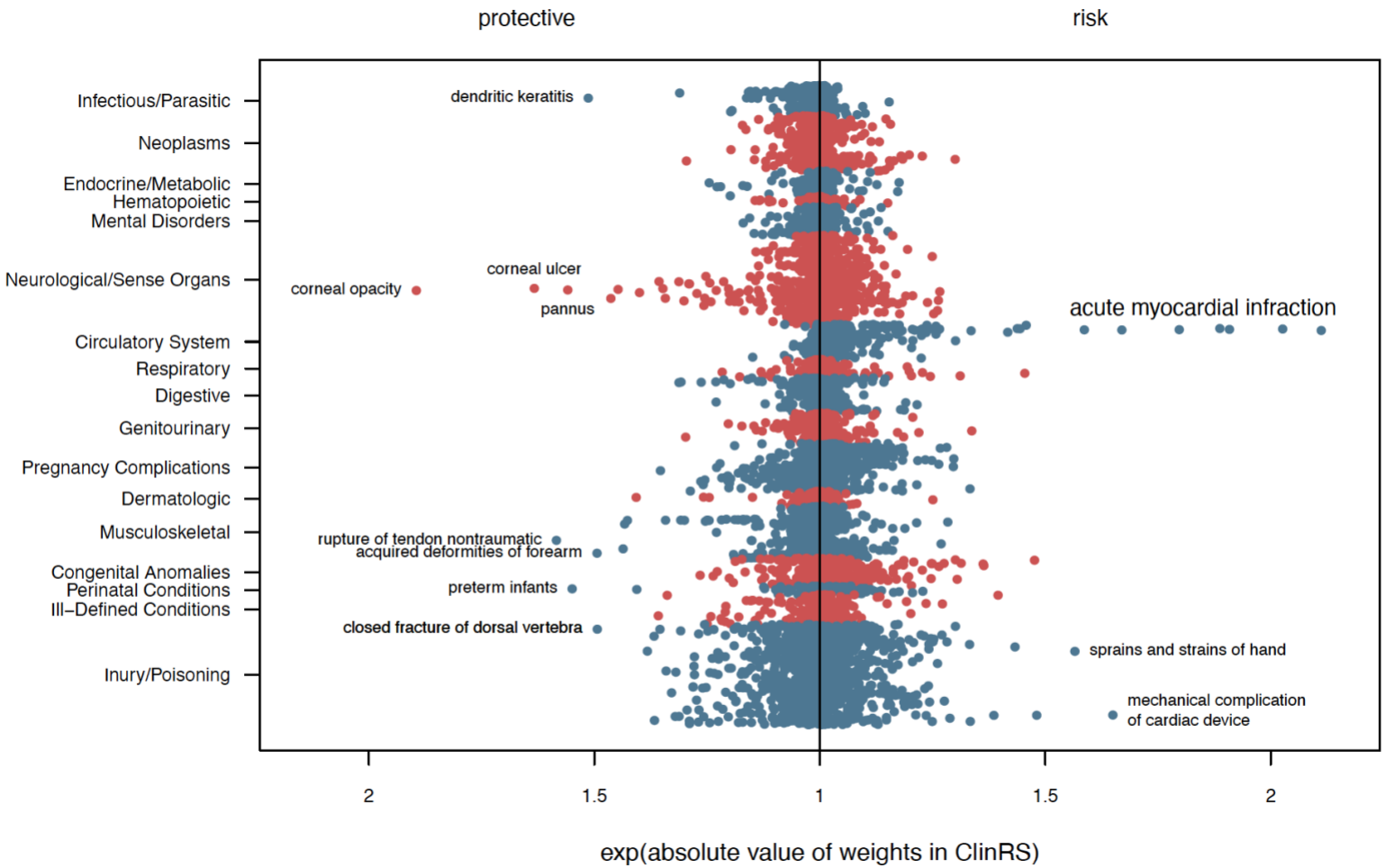
Supplementary Figure 3. Heatmap of cosine similarity score between a pair of codes within ICD-9 140 to 239 (Neoplasms) and sorted by its order. Every dot in this plot represents a pair of codes and its cosine similarity score, with the darker the red representing the closer the distance (more similar) between these 2 codes.



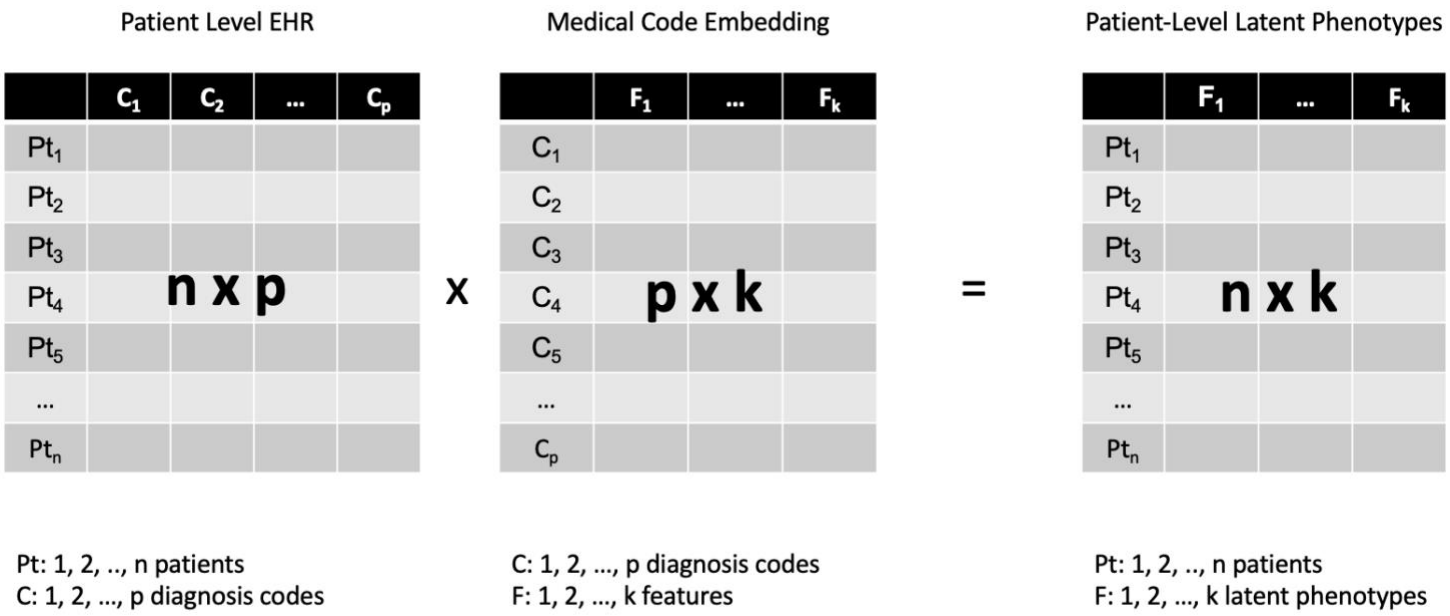
Supplementary Figure 4. Scatter plot and boxplot of patients' polygenic risk score (PRS) and clinical risk score (ClinRS) at one year prior to heart failure diagnosis, colored by disease status. Dotted gray lines indicate the cutoff of high and low risk of corresponding risk predictors. Percentage in each quadrant indicates the percentage of heart failure cases among patients classified in the corresponding risk group.



Supplementary Figure 5. Forest plot comparing models accuracy of predicting heart failure at one to ten years prior to disease diagnosis in the sensitivity analysis. Six models were compared with each time point: baseline (age and sex), PRS (polygeneic risk score), ClinRS (clinical risk score), ClinRS-NoCirc, PRS+ClinRS, and PRS+ClinRS-NoCirc. ClinRS-NoCirc was calculated by removing circulatory system diagnosis code in patients' medical records to validate the validity of ClinRS generated using the adapted natural language processing method. Numbers at the bottom of the plot indicate the sample size for each time point. Results showed that ClinRS-NoCirc can predict heart failure outcomes six years in advance, shorter than using ClinRS as a predictor. Adding both PRS and ClinRS-NoCirc in the model, the model accuracy is comparable to PRS+ClinRS model, which predicts disease ten years in advance.



Supplementary Figure 6. Manhattan plot of clinical risk score (ClinRS) weights for each ICD-9 diagnosis code by disease class. X-axis indicates the exponential of the absolute weights in ClinRS. The left panel showed the weights of the protective (negative weights; decreased risk) factor and the right panel showed the weights of the risk (positive weights; increased risk) factor.



Supplementary Figure 7. Illustration of creating latent phenotype from individual level electronic health records.

Supplementary Table 1. Sample size of heart failure cases and controls included in analysis for one to ten years prior to disease diagnosis.

year	Sample Size		10-fold Cross-Validated AUC					
	cases	controls	baseline	PRS	ClinRS	PRS+ClinRS	ClinRS-noCirc	PRS+ClinRS-noCirc
1	576	19,703	0.70 (0.68-0.72)	0.76 (0.74-0.78)	0.85 (0.83-0.87)	0.87 (0.85-0.88)	0.84 (0.82-0.8599)	0.86 (0.84-0.87)
2	539	17,758	0.70 (0.68-0.72)	0.77 (0.75-0.79)	0.82 (0.80-0.85)	0.84 (0.82-0.86)	0.81 (0.79-0.84)	0.83 (0.81-0.85)
3	515	16,365	0.70 (0.68-0.73)	0.77 (0.75-0.79)	0.81 (0.79-0.84)	0.83 (0.81-0.85)	0.80 (0.78-0.83)	0.83 (0.80-0.85)
4	494	15,152	0.70 (0.68-0.73)	0.77 (0.74-0.79)	0.80 (0.78-0.83)	0.82 (0.80-0.85)	0.79 (0.76-0.81)	0.81 (0.79-0.84)
5	459	14,153	0.71 (0.69-0.73)	0.77 (0.75-0.79)	0.80 (0.77-0.82)	0.82 (0.80-0.84)	0.78 (0.76-0.81)	0.81 (0.79-0.83)
6	427	13,239	0.72 (0.69-0.74)	0.78 (0.76-0.90)	0.79 (0.76-0.81)	0.82 (0.79-0.84)	0.77 (0.75-0.80)	0.80 (0.78-0.83)
7	407	12,394	0.71 (0.69-0.74)	0.77 (0.75-0.80)	0.78 (0.75-0.80)	0.81 (0.78-0.83)	0.76 (0.73-0.79)	0.79 (0.77-0.82)
8	376	11,601	0.71 (0.68-0.73)	0.76 (0.74-0.78)	0.77 (0.74-0.79)	0.80 (0.77-0.82)	0.75 (0.73-0.78)	0.78 (0.76-0.81)
9	353	10,831	0.72 (0.69-0.75)	0.77 (0.74-0.79)	0.76 (0.74-0.79)	0.80 (0.77-0.82)	0.75 (0.73-0.78)	0.79 (0.76-0.81)
10	332	10,059	0.72 (0.69-0.75)	0.77 (0.74-0.80)	0.76 (0.73-0.78)	0.79 (0.77-0.82)	0.75 (0.72-0.77)	0.78 (0.76-0.81)

Ten-fold cross-validated Area Under the Receiver Operating Characteristics (AUC) of six models predicting heart failure outcome across 10 time points. Model performances were calculated for baseline (age and sex) model and 5 models with risk score(s) added: i) polygenic risk score (PRS), ii) clinical risk score (ClinRS), iii) PRS+ClinRS, iv) clinical risk score calculated without circulatory system diagnosis code (ClinRS-NoCirc), and v) PRS+ClinRS-NoCirc.

Supplementary Table 2. Top 20 protective and risk factors yielded from clinical risk score (ClinRS).

Protective Factors			Risk Factors		
ICD code	**ClinRS weight**	**Diagnosis**	**ICD code**	**ClinRS weight**	**Diagnosis**
371.03	-0.6391	Opacity, central cornea	410.91	0.7476	AMI NOS, initial
370.03	-0.4905	Ulcer, central corneal	410.21	0.7063	AMI, inferolateral wall, initial
727.63	-0.4600	Rupture, hand/wrist extensor tendon	410.41	0.6463	AMI, inferior wall, initial
370.63	-0.4441	Vascularization, deep corneal	410.01	0.6400	AMI, anterolateral wall, initial
765.14	-0.4378	Preterm infant NEC, 1000-1249 gram	410.51	0.5861	AMI, lateral wall, initial
54.42	-0.4144	Herpes simplex dendritic keratitis	410.71	0.5127	AMI, subendocardial, initial
736.09	-0.4015	Deformity, acquired, forearm NEC	996.01	0.5007	Malfunction, cardiac pacemaker
806.25	-0.4011	Fx T7-T12 clsd w/spinal cd inj NOS	410.61	0.4616	True posterior wall, initial
374.23	-0.3808	Lagophthalmos, cicatricial	842.19	0.4483	Sprain/strain, hand NEC
370.35	-0.3697	Keratoconjunctivitis, neurotrophic	996.04	0.3927	Complications d/t AICD
732.7	-0.3621	Osteochondritis dissecans	743.37	0.3894	Ectopic lens, congenital
718.84	-0.3596	Drngmnt, oth joint NEC, hand	396	0.3768	Stenosis, mitral and aortic valves
717.89	-0.3554	Disruption, internal, knee NEC	512.8	0.3746	Pneumothorax, spontaneous NEC
695.14	-0.3416	SJS toxic epidermal necrolysis synd	410.11	0.3674	AMI, anterior wall, initial
765.25	-0.3407	Gestation completed 29-30 weeks	410.02	0.3631	AMI, anterolateral wall, subsequent
371.61	-0.3364	Keratoconus, stable	835.03	0.3596	Dsloc, anterior hip NEC, closed
842.12	-0.3239	Sprain/strain, metacarpophalangeal	414.2	0.3483	Chrn total occlusion coronary arter
813.54	-0.313	Fx lower radius w/ulna, open	780.32	0.3331	Symp, convulsions, febrile complex
997.4	-0.3124	Complications, digestive system	996.09	0.3264	Malfunction, cardiac dev/graft NEC
793.1	-0.3062	AbFnd, rdlog, lung field	746.86	0.3102	Block, heart, congenital

REFERENCES

1. Groenewegen, A., Rutten, F. H., Mosterd, A. & Hoes, A. W. Epidemiology of heart failure. *Eur J Heart Fail* **22**, 1342–1356 (2020).
2. Savarese, G. & Lund, L. H. Global Public Health Burden of Heart Failure. *Card Fail Rev* **3**, 7–11 (2017).
3. van Riet, E. E. S. *et al.* Epidemiology of heart failure: the prevalence of heart failure and ventricular dysfunction in older adults over time. A systematic review. *Eur J Heart Fail* **18**, 242–252 (2016).
4. Mathis, M. R. *et al.* Early Detection of Heart Failure With Reduced Ejection Fraction Using Perioperative Data Among Noncardiac Surgical Patients: A Machine-Learning Approach. *Anesth Analg* **130**, 1188–1200 (2020).
5. Povysil, G. *et al.* Assessing the Role of Rare Genetic Variation in Patients With Heart Failure. *JAMA Cardiol* **6**, 379–386 (2021).
6. Douville, N. J. *et al.* Use of a Polygenic Risk Score Improves Prediction of Myocardial Injury After Non-Cardiac Surgery. *Circ Genom Precis Med* **13**, e002817 (2020).
7. Kannel, W. B., McGee, D. & Gordon, T. A general cardiovascular risk profile: the Framingham Study. *Am J Cardiol* **38**, 46–51 (1976).
8. Lauer, M. S., Anderson, K. M., Kannel, W. B. & Levy, D. The impact of obesity on left ventricular mass and geometry. The Framingham Heart Study. *JAMA* **266**, 231–236 (1991).
9. Wilson, P. W. *et al.* Prediction of coronary heart disease using risk factor categories. *Circulation* **97**, 1837–1847 (1998).
10. Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults. Executive Summary of The Third Report of The National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, And Treatment of High Blood Cholesterol In Adults (Adult Treatment Panel III). *JAMA* **285**, 2486–2497 (2001).
11. Goff, D. C. *et al.* 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Circulation* **129**, S49-73 (2014).
12. Agarwal, S. K. *et al.* Prediction of incident heart failure in general practice: the Atherosclerosis Risk in Communities (ARIC) Study. *Circ Heart Fail* **5**, 422–429 (2012).
13. Lambert, S. A., Abraham, G. & Inouye, M. Towards clinical utility of polygenic risk scores. *Hum Mol Genet* **28**, R133–R142 (2019).
14. Santos, R. D. Screening and management of familial hypercholesterolemia. *Curr Opin Cardiol* **34**, 526–530 (2019).
15. Nordestgaard, B. G. Triglyceride-Rich Lipoproteins and Atherosclerotic Cardiovascular Disease: New Insights From Epidemiology, Genetics, and Biology. *Circ Res* **118**, 547–563 (2016).
16. Warren, M. The approach to predictive medicine that is taking genomics research by storm. *Nature* **562**, 181–183 (2018).
17. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet* **50**, 1219–1224 (2018).
18. Polygenic Risk Score Task Force of the International Common Disease Alliance. Responsible use of polygenic risk scores in the clinic: potential benefits, risks and gaps. *Nat Med* **27**, 1876–1884 (2021).

19. Surakka, I. *et al.* Sex-specific survival bias and interaction modeling in coronary artery disease risk prediction. 2021.06.23.21259247 Preprint at https://doi.org/10.1101/2021.06.23.21259247 (2021).
20. Mujwara, D. *et al.* Integrating a Polygenic Risk Score for Coronary Artery Disease as a Risk-Enhancing Factor in the Pooled Cohort Equation: A Cost-Effectiveness Analysis Study. *J Am Heart Assoc* **11**, e025236 (2022).
21. Sinha, A. *et al.* Risk-Based Approach for the Prediction and Prevention of Heart Failure. *Circ Heart Fail* **14**, e007761 (2021).
22. Wu, K.-H. H. *et al. Polygenic risk score from a multi-ancestry GWAS uncovers susceptibility of heart failure*. http://medrxiv.org/lookup/doi/10.1101/2021.12.06.21267389 (2021) doi:10.1101/2021.12.06.21267389.
23. Zhou, W. *et al. Global Biobank Meta-analysis Initiative: powering genetic discovery across human diseases*. http://medrxiv.org/lookup/doi/10.1101/2021.11.19.21266436 (2021) doi:10.1101/2021.11.19.21266436.
24. Zawistowski, M. *et al. The Michigan Genomics Initiative: a biobank linking genotypes and electronic clinical records in Michigan Medicine patients*. http://medrxiv.org/lookup/doi/10.1101/2021.12.15.21267864 (2021) doi:10.1101/2021.12.15.21267864.
25. Denny, J. C. *et al.* PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* **26**, 1205–1210 (2010).
26. Wei, W.-Q. *et al.* Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. *PLoS One* **12**, e0175508 (2017).
27. Graham, S. E. *et al.* The power of genetic diversity in genome-wide association studies of lipids. *Nature* **600**, 675–679 (2021).
28. Hong, E. P. & Park, J. W. Sample Size and Statistical Power Calculation in Genetic Association Studies. *Genomics Inform* **10**, 117 (2012).
29. Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A. & Smoller, J. W. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat Commun* **10**, 1776 (2019).
30. The 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
31. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
32. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. Distributed Representations of Words and Phrases and their Compositionality. Preprint at http://arxiv.org/abs/1310.4546 (2013).
33. Hong, C. *et al.* Clinical knowledge extraction via sparse embedding regression (KESER) with multi-center large scale electronic health record data. *NPJ Digit Med* **4**, 151 (2021).
34. Levy, O. & Goldberg, Y. Neural Word Embedding as Implicit Matrix Factorization. in *Advances in Neural Information Processing Systems* vol. 27 (Curran Associates, Inc., 2014).
35. Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58**, 267–288 (1996).
36. Choi, Y., Chiu, C. Y.-I. & Sontag, D. Learning Low-Dimensional Representations of Medical Concepts. *AMIA Jt Summits Transl Sci Proc* **2016**, 41–50 (2016).
37. De Vine, L., Zuccon, G., Koopman, B., Sitbon, L. & Bruza, P. Medical Semantic Similarity with a Neural Language Model. in *Proceedings of the 23rd ACM International Conference*

on Conference on Information and Knowledge Management 1819–1822 (Association for Computing Machinery, 2014). doi:10.1145/2661829.2661974.
38. Finlayson, S. G., LePendu, P. & Shah, N. H. Building the graph of medicine from millions of clinical narratives. *Sci Data* **1**, 140032 (2014).
39. Levy, O., Goldberg, Y. & Dagan, I. Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the Association for Computational Linguistics* **3**, 211–225 (2015).
40. Beam, A. L. *et al.* Clinical Concept Embeddings Learned from Massive Sources of Multimodal Medical Data. *Pac Symp Biocomput* **25**, 295–306 (2020).
41. Stuart, A. G. & Williams, A. Marfan's syndrome and the heart. *Arch Dis Child* **92**, 351–356 (2007).
42. Djoussé, L. & Gaziano, J. M. Alcohol consumption and heart failure: a systematic review. *Curr Atheroscler Rep* **10**, 117–120 (2008).
43. Tam, P. E. Coxsackievirus myocarditis: interplay between virus and host in the pathogenesis of heart disease. *Viral Immunol* **19**, 133–146 (2006).
44. Yu, S. *et al.* Enabling phenotypic big data with PheNorm. *J Am Med Inform Assoc* **25**, 54–60 (2018).
45. Goldberg, Y. & Levy, O. word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. Preprint at https://doi.org/10.48550/arXiv.1402.3722 (2014).