Developing ICD Code Embeddings Across Two Institutions

Dharani Krishnan Senthil Kumar[1], VG Vinod Vydiswaran[1,2], Xu Shi[3]

1. School of Information, University of Michigan
2. Department of Learning Health Sciences, Medical School, University of Michigan
3. Department of Biostatistics, University of Michigan

ABSTRACT:

Each healthcare institutions have numerous patients with a variety of different disease conditions. Some of these diseases might be common and spread across multiple locations and demographics while some might be situated or originated in certain places. Interoperability of patient data between different healthcare organizations will help in improving the quality of care provided to the patients. The patients' records, consisting of their notes, diagnoses, etc., contain numerous ICD codes in them. Mapping these through embedding created can help in understanding and utilizing the data for further studies. This work involves two intuitions namely the University of Texas Medical Branch, and Michigan Medicine where embeddings are generated for the ICD codes in both of their patient cohorts.

INTRODUCTION:

In the current era of digital information and technological advancement, the volume of data available is tremendous. In healthcare, patient data is available across different mediums such as Electronic Health Records (EHR), Patient Portals, Personal Health Records (PHI), Payer's database, etc. The type of data ranges from physician's notes, laboratory test results, personal information, family history, medical history, medications, and so on. This abundant data provides the foundation of evidence-based medicine and delivering value-based, patient-centric care. Patient data is also a fundamental resource in analyzing trends, patterns and developing diagnostic Machine Learning predictive models.

After 2014, when meaningful usage of EHRs was adopted, each patient is the source of 80 MB of data per year. It is predicted that by 2025, the compound annual growth rate of healthcare data will touch 36%. Without the Global Covid-19 Pandemic, 2020 predictions show about 2314 exabytes of data were expected to be produced (Culbertson, 2021). The majority of the data is available from the patient's care provider and the hospital. With HIPAA and other health data privacy policies existing, the data is stored securely and not exchanged or shared with other healthcare organizations. This is essential to ensure that patients' personal health data isn't leaked, stolen, or misused. On the other hand, it also prevents tapping into the abundant potential of using the data in developing retrospective healthcare solutions.

With big data being the primary ingredient in developing and implementing Healthcare Artificial Intelligence Solutions, methods to utilize the patient's data while also eliminating the possibility of leaking personal details must be employed. Cyber-encryptions, and de-identification techniques to mask personal records mentioned according to HIPAA are a few possibilities. These steps enable interoperability or sharing of data. One hospital or healthcare organization would not have sufficient data to create robust models which can perform for the majority of extreme cases. Even if the model developed with data from a specific location does work accurately, the same solution can't be adopted in another location as the demographics might be different. Thus, interoperability between different healthcare institutions in different regions is essential. One example in recent times includes the publishing of data from Wuhan medical centers on Covid-19 symptoms in patients. As this data was made public, it helped health systems across the globe to make prior preparations and have an advantage in dealing with the scenario.

Interoperability is defined as the ability of different information systems, devices, and applications to access, exchange, integrate and cooperatively use data in a coordinated manner, within and across organizational, regional, and national boundaries, to provide timely and seamless portability of information and optimize the health of individuals and populations globally (Epalm, 2021). In the healthcare domain, interoperability will enable sharing of data securely by means of dedicated health data exchange architecture, safety, and privacy standards. The exchange of data can be across different care organizations in the same community, other health networks, involved stakeholders, and the individual patient as well.

There are 4 levels or layers of interoperability namely Foundational, Structural, Semantic, and Organizational. The first level or the Foundational level establishes the interconnectivity which will enable the ability of one information system to communicate (share or receive) data with another system. The next level, Structural, lays the format, and syntax of the exchange data. It defines the standard structure in which the data can be transferred and interpreted on both sides. The penultimate level is the Semantic level in the complex layer where real-time transfer of data exchange is conducted between multiple systems for interpretation and subsequent utilization. It provides the standard for encryption and models for the data elements. The final level is the Organizational level in which policy, legal governance, and other organizational factors are established to facilitate smooth communication and usage of the data across multiple entities. This helps foster trust and builds a seamless, streamlined workflow (Lambert, 2023). Another critical piece is the user consent, in this case, the patient whose data would be shared with or without masking their personal information.

Moving forward, Heath Information Exchange (HIE) would provide seamless electronic transfer of hospital data with other clinical and healthcare organizations. The final stage of the Health Information Exchange is to facilitate access and usage of patient clinical data to deliver safe, timely, efficient, effective, affordable, and equitable patient-centered care (Epalm, 2021). Another spectrum is the usage of this data by public health authorities to assist in the analysis of the health of populations and by policymakers in drafting healthcare policies for the nation or state.

The potential of healthcare information interoperability is limitless. On top of using the information in developing data-oriented technological solutions, the data can help in efficient value-based and evidence-based care. It can also assist in increasing awareness in patients, improving productivity while decreasing the cost, and burnout of healthcare professionals. Another critical advantage would be the continuum of quality care despite patients visiting different hospitals. Financial incentives and reimbursement will also benefit from this interoperability.

Currently, the Center for Medicare and Medicaid Services is taking a tremendous effort in increasing healthcare data interoperability between patients, care providers, and community hospitals to increase support for patient care and reduce administrative burden. In 2020 May, CMS

introduced its Interoperability & Patient Access Final Rule policy which aims to enable the interoperability of healthcare for all involved stakeholders in the system (Interoperability and the Connected Health Care System | CMS).

BRAIN INJURY ASSOCIATED FATIGUE AND ALTERED COGNITION:

There are more than 2.5 million people who experience Traumatic Brain Injury (TBI) yearly with approximately 50,000 ending in death and more than 80,000 suffering permanent disability. TBI or Craniocerebral Trauma is caused by a sudden injury, such as a sports injury or vehicular accident, to the head leading to damage or trauma to the brain. It is commonly categorized as mild, moderate, or severe based on the level of trauma. Around 90% of the patients fall under the middle category (Brain Trauma Foundation - Frequently Asked Questions (FAQ) — Brain Trauma Foundation). The generic symptoms include confusion or "fuzzy or foggy brain", dysfunctions in vision and speech, short-term memory loss, and day-to-day organizational and concentration difficulties.

It was found that, in a subclass of TBI patients, there is pituitary dysfunction and abnormal growth hormone secretion. This further leads to profound fatigue and reduced cognitive aptitude in fields of memory, processing capability, and execution of tasks preventing the patients from conducting their normal life daily. This clinical condition is termed Brain Injury Associated Fatigue and Altered Cognition or BIAFAC in short. This post-TBI syndrome discovery was made by Dr. Randall Urban, University of Texas Medical Branch, and his team (Traumatic Brain Injury Impairs Hormone Production, Disrupting Sleep, Cognition, Memory, 2021).

It was found that TBI triggers a reduction in growth hormone secretion and that most TBI patients' health improves after growth hormone replacement treatment. The approach of recombinant growth hormone replacement treatment has significantly improved the conditions of patients with BIAFAC (Wright et al., 2020). Signs of improvements in fatigue and cognition were seen after 3 and 4-5 months after treatment, respectively. This is more of a treatment for the symptoms than the condition as it was observed that any pause in the treatment causes the return of the symptoms.

Initially, the study was conducted on just 18 patients. With the exact causative factor of BIAFAC being ambiguous and the patients being currently limited to around 120 only, the current data is insufficient to derive the exact reasons. This work aims to identify and implement code embedding in both the Michigan Genomics Initiative (MGI), which is a data repository of over 80,000 patients, their medical notes, readings, genomic data, etc., and the University of Texas Medical Branch (UTMB) which has the BIAFAC patients. As of now, BIAFAC doesn't have a specified official ICD code and thus, the patients at Michigan Medicine aren't diagnosed with BIAFAC. But some patients have TBI out of which some of them have the symptoms of BIAFAC while the rest don't.

MICHIGAN GENOMICS INITIATIVE:

The Michigan Genomics Initiative (MGI) is a combined research effort of care providers, patients, and researchers at Michigan Medicine to collect and maintain patients' medical records from the health system's EHR along with their genetic information for research purposes. The data is collected by means of voluntary participants on the side of patients who are consenting to share their medical records, data, history, and their DNA information. This is a protected data repository where access is granted only to the University of Michigan researchers who have undergone training and received IRB approval. As of this year, there are close to 100K MGI participants and all of them are patients who have received care in Michigan Medicine. Out of these patients, about 71K patients' genotypes and polygenic scores are available.

As of the end of the year 2022, the gender distribution of 71K patients in the repository is approximately 47% males to 53% females. The common median age is 60 years while the median for males is 62 years and for females 57 years. The distribution in terms of self-reported ethnicity is 86% Caucasian, 6.5% African American, 2.7% Asian, 0.5% American Indian or Native Alaskan, and the remaining unknown (Michigan Genomics Initiative | University of Michigan Precision Health).

The data utilized in this work is sourced from the MGI repository. There are multiple datasets employed for this study. There are datasets for patients' demographics, diagnoses, lab

results, procedures, and clinical notes. The patient demographics dataset consists of 63,455 instances. These patients are filtered from the complete MGI based on the occurrences or mention of Traumatic Brain Injury, Fatigue, Cognitive Impairment, and other symptoms of BIAFAC. The diagnoses dataset of the patients has around 60338966. The lab dataset of the patients has around 78925407. The notes dataset of the patients has around 2307522. The procedures dataset of the patients has around 24302807. The following tables are data dictionaries for the different datasets utilized for developing the code embeddings.

S.No	Column Name	Description	Data Type
1	MRN	Medical Record Number	numeric
2	PATIENTID	Unique ID Number of the Patient	alphanumeric str
3	LIVING_STATUS	Patient's Mortality	str
4	CURR_AGE_OR_AGE_DEATH	Age of the patient	numeric
5	SEX	Gender of the patient	str
6	PAYOR_NAME	Insurance Company Name	str
7	BENEFIT_PLAN_NAME	Insurance Plan Name	str
8	RACE_1	Race of the Patient	str
9	ETHNICITY	Ethnicity of the Patient	str

Table 1: Data Dictionary of Patients Dataset (MGI)

S.No	Column Name	Description	Data Type
1	MRN	Medical Record Number	numeric
2	PATIENTID	Unique ID Number of the Patient	alphanumeric str
3	DX_DATE	Date of Diagnosis	datetime value
4	DX_CODE	Diagnosis Code	float
5	DX_NAME	Name of the Diagnosis	str
6	DX_SOURCE		str

Table 2: Data Dictionary of Diagnoses Dataset (MGI)

S.No	Column Name	Description	Data Type
1	MRN	Medical Record Number	numeric
2	PATIENTID	Unique ID Number of the Patient	alphanumeric str
3	ORDER_TIME	Time of the Lab Order	datetime value
4	CPT_CODE	Current Procedural Terminology Codes	alphanumeric str
5	DESCRIPTION	Brief about the CPT Code	str
6	ORDER_DISPLAY_NAME	Name Displayed for the order	str
7	COLLECTION_TIME	Sample Collection Time	datetime value
8	RESULT_TIME	Time of result	datetime value
9	RESULT_NAME	Result of the test	str
10	BASE_NAME	Main Element of the test	str
11	RESULT_VALUE	Value of the base element	numeric
12	UNITS	Unit of the numeric value	str

Table 3: Data Dictionary of Labs Dataset (MGI)

S.No	Column Name	Description	Data Type
1	MRN	Medical Record Number	numeric
2	PATIENTID	Unique ID Number of the Patient	alphanumeric str
3	PX_DATE	Date of Procedure	datetime value
4	PX_CODE	Procedure Code	numeric
5	PX_NAME	Name of the Procedure	str
6	PX_TYPE	Type of Procedure	alphanumeric str

Table 4: Data Dictionary of Procedures Dataset (MGI)

S.No	Column Name	Description	Data Type
1	MRN	Medical Record Number	numeric
2	PATIENTID	Unique ID Number of the Patient	alphanumeric str
3	PAT_ENC_CSN_ID		numeric
4	NOTE_DATE	Date the note was created	datetime value

5	NOTE_TYPE	Type of Note	str
6	NOTE	Content of Procedure	str

Table 5: Data Dictionary of Notes Dataset (MGI)

METHODOLOGY:

PRE-PROCESSING:

The datasets contain numerous additional data columns which aren't required for this analysis. Thus, initial cleaning of the data is required to convert it into the desired structure. The diagnoses, procedures, and labs dataset were considered for the embedding process as each of them have Diagnostic and/or Procedural codes in them. For all three datasets, the MRN, Code, and Date columns. Each patient has their own MRN and Patient ID and it is unique for them. There are no cross instances of MRN and Patient ID. To maintain anonymity and to protect the personal information of the patients, their MRN must be coded numerically. Similarly, the Codes are also numerically coded to enable the determination of the Co-occurrence matrix in the next step. The Code to Code-ID mapping is preserved to use during the validation step. For the next step, the day of occurrence of the code has to be recorded. This is relative to the patients and their code instances. This data column known as 'numDays' was calculated by setting a standard date and subtracting the standard date from the date of Code occurrence. The standard date was determined by a day less than the minimum date of Code occurrence. Cleaned metadata consists of the new patient ID, new Code ID, and numDays. Finally, there was a total of 60338965 instances in the metadata.

Patient ID	numDays	ICD Code
7362	3263	1282
7362	3263	1282
7362	3263	4453
7362	3263	11088
7362	3263	11088
7362	3263	369

Table 6: Sample of Cleaned Metadata from MGI

CO-OCCURRENCE MATRIX:

Generally, the co-occurrence matrix or gray level co-occurrence distribution is defined over an image to be the distribution of co-occurring values of the pixel. This is commonly employed in texture analysis of medical images such as MRI and CT scans. In Natural Language Processing, a co-occurrence matrix can be used for the processing of words in a corpus. Fundamentally, it establishes the relationship between different words, phrases, sentences, etc. in the text collection. Each row and column in the co-occurrence matrix represent a unique element of the corpus. Each value denotes the number of occurrences of the two elements together. This will help in understanding the association between the elements (Zhao et al., 2017).

In this scenario, the co-occurrence matrix is built where the rows and columns represent the codes present in the data. This matrix will display the codes which occur with each other and the frequency of their occurrences of the codes within each patient. These occurrences are categorized into time windows such as 0, 6, 13, 29, and 59. The window length of 0 represents the co-occurring codes within a day, followed by a week, a fortnight, a month, and finally two months, respectively. This matrix not just forms the basis for text analysis, but also helps in representing the elements as embeddings. Any row or column from the co-occurrence matrix can be taken and turned into a word embedding representation. The high dimensionality of the matrix might most likely pose problems during this process but here are dimensionality reduction methods that can be utilized. A few options would be Principal Component Analysis (PCA), Singular Value Decomposition (SVD), Linear Discriminant Analysis (LDA), etc.

Each instance in the co-occurrence table consists of the two codes, their number of co-occurrences, and the window under it. It is then converted to a sparse matrix representation format for each window length. The co-occurrence sparse matrix of the larger window is the summation of all the elements within the smaller windows. For example, a window size of 13, will include co-occurrence values for the code pairs from windows 0 and 6 as well.

Code1	Code2	Count	Window
1282	1282	21243	0

1282	4453	1573	0
1282	11088	5998	0
369	1282	7430	0

Table 7: Sample of Co-Occurrence Data

POINTWISE MUTUAL INFORMATION:

Pointwise Mutual Information (PMI) is a common measure of association in statistics which in Natural Language Processing can be used in determining related words. It compares the probability of 2 words occurring together to the probability of the occurrence of words if they were independent. The PMI helps in weighing the association by obtaining the co-occurrence of the 2 words in a text corpus that a priori expected to appear in it by chance. For instance, "Machine Learning" is a word that has a certain meaning while the words separately, "Machine" and "Learning", have a different meaning altogether. On the other hand, "Great Britain" is a meaningful usage of the word "Great", but it can be used in front of other State or Country names.

Mathematically, PMI can be defined as,

$$PMI(x, y) = log_2 \frac{P(x, y)}{P(x)\,P(y)}$$

When there are two words, x, and y. If they are independent entities, the joint probability is the product of each word's individual probability (Damani, 2013). Thus, the PMI would be equal to the log of 1 which is 0. Which would mean that there is no association or specific meaning between the 2 words. PMI will help determine whether words have a high joint probability with each other while not having a high probability of occurrence if those words are separately considered. Thus, these 2 words could be called a pair having significant and relevant meanings. There exist variations to PMI like Positive PMI (PPMI) and Normalized Pointwise Mutual Information (NPMI) (Levy & Goldberg, 2014). The former is obtained by making the negative values of PMI 0 while the latter is attained by normalizing the values in between the range of -1 to 1 where -1 is for words never occurring together, 0 for independence, and 1 for complete co-occurrence.

SHIFTED POSITIVE POINTWISE MUTUAL INFORMATION:

Shifted PPMI is a word-context co-occurrence PMI matrix where each value is calculated as $maximum(0, PMI(w, c) - \log(k))$, where k is the number of negative samples considered. In other words, the SPPMI is obtained by subtracting the log(k) from the PMI matrix to obtain mostly non-negative values in the resultant matrix (Levy & Goldberg, 2014).

SINGULAR VALUE DECOMPOSITION:

To improve generalization and faster computational efficiency, the sparse representation can be reduced to a low-dimensional dense matrix. Embedding like word2vec or GloVe uses a gradient descent approach for factorization. Here, Singular Value Decomposition (SVD) is employed. The Singular Value Decomposition of a matrix, say A, is the factorization of A into the product of 3 matrices. The matrices U and V contain orthonormal columns or in other words, the orthogonal matrix in nature while matrix D is diagonal with positive real values. It can be expressed by the formula as follows,

$$A = UDV^T$$

The columns of U are the left singular vectors, and the transpose of V has rows that are the right singular vectors. Finally, the D matrix, having the same dimensions as A, has diagonal singular values. The SVD represents an expansion of the original matrix in a coordinate system where the covariance matrix is diagonal (Klema & Laub, 1980). Obtaining the SVD for A involves calculating the eigen values and eigen vectors of AA^T and A^TA. The eigen vectors of the former form the columns of matrix U while the latter's eigen vectors form the columns of matrix V. The last matrix D's values are obtained from the square roots of eigen values from both matrix multiplications. The singular values are real numbers and correspond to the diagonal values of the matrix D and are arranged in descending order (Alter et al., 2000).

The SVD has varied applications from computing the pseudoinverse of a matrix and solving homogenous linear equations to least squares and low-rank matrix approximations. There

are variations of SVD that are more commonly employed than the complete version mainly due to the computation limit and storage space. A few popular derivatives are Thin SVD, Truncated SVD, and Compact SVD.

The Spectral Dimensionality Reduction over the Shifted PPMI Matrix (SPPMI-SVD) aims to apply singular value decomposition on the SPPMI matrix. It is a factorization-based continuous dense distributional word model approach. After that, the dense singular vectors are taken as the word embeddings for the word vectors passed in the SPPMI matrix (Levy & Goldberg, 2014).

A fast approximation of SVD is the Augmented, Implicitly Restarted Lanczos Bidiagonalization Algorithm (IRLBA). This algorithm was developed by James Baglama and Lothar Reichel. This is an effective and memory-efficient approach for computing truncated singular value decomposition for highly sparse and dense matrices. The IRLBA is a partial SVD approach that finds some approximate singular values and their singular vectors for the matrix. While the SVD method provides a complete set of singular values and corresponding vectors which can be more than the provided set of singular vectors, IRLBA on the other hand, gives an estimated number of singular values equal to or less than the maximum number of set singular vectors (Baglama & Reichel, 2005).

RESULTS:

Genome-wide association studies (GWAS) found that if single-nucleotide variants occur at various locations across the genome, they can be associated with a specific trait, and it was considered a phenotype (Bastarache, 2021). Similarly, Phenome-wide association studies (PheWAS) adapted the idea of GWAS to search or analyze for phenotypes associated with specific single-nucleotide variants across numerous phenotypes, also known as phenome. Based on this approach, mappings of ICD 9 and 10 diagnoses codes in EHR to Phecodes were developed and made open source for research (Wu et al., 2019).

The Phecode data consists of the ICD code and a phecode value which is combined to group similar ICD codes into batches. These ICD groups were taken as the testing standard to see

if the embedding generated for the code pairs were significant. Each pair in the phecode data was considered and if their phecode value is the same, they are associated pairs and if their values are not the same, the code pair don't have an association with each other. The cosine similarity was calculated on the embedding to obtain the predicted association between the codes. The two sets of pairs constructed from the phecodes, and the embeddings were evaluated to find the AUC_ROC curve value. The value was evaluated with multiple lengths of dimensions from 10 to a total of 500 dimensions. The table below shows the ROV value for the different chosen dimensions of the embeddings.

S.No	ICD Code	Phecode	Rounded Phecode
1.	S62.024D	804.0	804
2.	T84.498D	858.0	858
3.	728.4	728.2	782
4.	F10.959	317.0	317
5.	716.92	716.9	716

Table 8: Sample of ICD Phecodes

Dimension	AUC_ROC Value for Different Windows				
	1 Day	7 Days	14 Days	30 Days	60 Days
10	0.5539521888797039	0.5593082156865559	0.5638041919156447	0.5656628929218283	0.565616950773861
50	0.5788491931829053	0.5773175180430431	0.5765834129578798	0.5786445588013006	0.5819482392095202
100	0.5845325851126077	0.5884192686213173	0.5910691907417622	0.5861006526016854	0.5821475078243514
300	0.5873197703165446	0.5938041822872928	0.5946707546203469	0.5932845422637602	0.5920295477722952
500	0.584734145168001	0.5895691992910649	0.5917322381776561	0.5917723287092848	0.5914700338716454

Table 9: ROC Values for Code Embeddings from MGI

Compared to the MGI repository the BIAFAC patients in UTMB are minimal. There are only about 120 patients and their corresponding records over the years of treatment. There was a total of 38764 records for a total of 119 patients who had BIAFAC conditions. The different columns in the table were,

S.No	Column Name	Description	Data Type
1	PATIENT_NUM	Unique Patient ID Number	numeric

2	START_DATE	Date the code was encountered	datetime value
3	CONCEPT_CD	ICD Code	alphanumeric

Table 10: Data Dictionary of Patients' Diagnoses (BIAFAC)

After cleaning the data to format, it in the same arrangement as the MGI metadata, it was observed that there were instances having dates in the late 1985s. Most likely, these dates were autogenerated since the original values were missing from the dataset. Since the study was started in 2004, all the instances having dates before that year were dropped as these date differences can offset the values in the co-occurrence matrix. After the embedding was obtained and cosine similarity was calculated with different dimensions of the embeddings to compute the ROC_AUC value with the phecodes as the gold standard.

Dimension	AUC_ROC Value for Different Windows				
	1 Day	7 Days	14 Days	30 Days	60 Days
10	0.532578117073145	0.5433824139599822	0.5389521435508348	0.5346649884286183	0.5283015097409528
50	0.5642949849603682	0.5706068521540714	0.565486408910506	0.5671847106003527	0.5609190895510984
100	0.5621229146473165	0.5660693410292849	0.5637295800511546	0.5644461709067351	0.5625437118952165
300	0.5600050590690202	0.5599955873350535	0.5608023122327654	0.5618713907545914	0.5577950683313043
500	0.5554382500552563	0.5573348074466209	0.5581526545614182	0.5570981037102096	0.5548329028809644

Table 11: ROC Values for Code Embeddings from BIAFAC

DISCUSSION AND LIMITATIONS:

The ROC value for both institutions is around 0.60. MGI's results are closer to the value while the BIAFAC cohort's result is slightly lesser. The phecode data contains 98,549 ICD 9 and ICD 10 codes combined. But the number of ICD codes in the MGI and BIAFAC data is less; the former has about 33513 codes while the latter has only around 1769 ICD 10 codes for the smallest window size. Thus, the number of codes the embeddings generated and evaluated is reduced and limited to the number of codes present in the main data. The embeddings were generated with 2000 iterations for the IRLBA algorithm. Increasing the number of iterations to 3000, yielded the same range of ROC. Utilizing the Sparse SVD implementation instead of the IRLBA also didn't return any significant change in the ROC result. Thus, changing the type of SVD or increasing the iterations didn't improve the accuracy of the embeddings generated for the ICD codes.

The ROC results of MGI for the different lengths of embeddings yield different values. The smallest dimension considered was 10 and the ROC value was the least across all the windows. There is an increase in the values as the number of dimensions increases. The embedding dimensions of 300 have the highest results followed by 500 which is only slightly less than the previous. Also, upon looking at the ROC results, an increase in the ROC value can be seen as the window length increases. This can be attributed to the increase in the number of codes available as the window expands. All these results have remained stable across the different dimensions and windows as the results were of the same range while testing for subsets of the code embeddings.

The ROC values are not consistent for the smallest dimensions across all windows for the BIAFAC data. After that, there is an increase that can be seen but beyond 300, there is a decrease in the value. Unlike MGI, 100 seems to be the stable and accurate dimension for the BIAFAC data.

The phecode representation is available for ICD 9 and ICD 10 diagnoses codes. There are other medical codes like procedural codes, CPT, etc. The embedding for these codes can be generated but the accuracy can't be determined with the phecodes. The codes are also spread across numerous days. The codes that co-occur beyond two months weren't included in this analysis. In the future analysis, the cosine similarity for the codes of the same window from both institutions can be compared to compute the closest, similar ICD codes.

There are a total of 1750 ICD Codes that are matching with each other in both institutions. The similarity of the embeddings for the same ICD codes on both sides were computed and the 1743 codes out of the 1750, have a similarity of over 0.90. The minimum similarity value is around 0.35, 25% quartile value, median, 75% quartile value, and the maximum value are all around 0.99. This shows that majority of the ICD codes' embedding are highly similar to each other with a similarity score of around 0.99. The table below shows that the 7 ICD codes that have the least similarity with each other on both MGI and BIAFAC side.

BIAFAC Code	Similarity
K62.89	0.358652435
M70.61	0.653337027
T83.89XA	0.730062731
L81.1	0.793964768
K51.90	0.859340931
N80.9	0.875781127
K56.60	0.892278585

Table 12: Codes having least similarity value

While initial similarity was calculated for matching codes, the embeddings for the ICD Codes in BIAFAC data was compared against each embedding in the MGI cohort, to determine the closet ICD code which has the most similarity.

BIAFAC Code	Top 3 Similar Codes					
	MGI Code	Similarity	MGI Code	Similarity	MGI Code	Similarity
K62.89	M75.100	0.36136	H66.90	0.36028	H35.433	0.36027
M70.61	864.01	0.66012	924.3	0.65552	M00.00	0.65443
T83.89XA	Z99.81	0.73092	R13.12	0.73086	A56.11	0.73079
L81.1	780.57	0.79482	V58.61	0.79477	Z12.31	0.79472
K51.90	N30.10	0.99992	R19.8	0.99992	T45.1X5A	0.99992
N80.9	T68.XXXD	0.99992	560.89	0.99992	S12.500D	0.99991
K56.60	Z94.1	0.89286	A56.11	0.89282	B01.9	0.89269

Table 13: Top 3 Similar MGI Codes to BIAFAC Codes

Furthermore, in future work, the groups of similar ICD codes can be determined by comparing the similarity between each embedding within institution, say MGI, and finding the set of codes that are similar to each other. The same approach can be performed on the other side, BIAFAC, to establish the sets. These sets can now be associated with each other to determine if the all the ICD codes in the set are matching or if there are any difference in the codes.

CONCLUSION:

The potential of interoperability between healthcare organizations in improving evidence-based and increasing the quality of care for patients is massive. Majorly in diseases that originate in certain places, rare diseases, etc. sharing of data and subsequent interpretations of it can help in a higher probability of successful treatment. BIAFAC is a similar condition, found in a single organization, but can be present in different demographics as well. Making use of this data in other locations gives the advantage of easier analysis and access to treatment options. The embedding for the ICD codes was determined based on the co-occurrence of these codes in different time windows in both the MGI data and the BIAFAC data. This will help in finding similar codes across the various locations to enable understanding of the shared health data.

REFERENCES:

Alter, O., Brown, P. O., & Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(18), 10101–10106. https://doi.org/10.1073/pnas.97.18.10101

Baglama, J., & Reichel, L. (2005). Augmented Implicitly Restarted Lanczos Bidiagonalization Methods. *SIAM Journal on Scientific Computing*, *27*(1), 19–42. https://doi.org/10.1137/04060593x

Bastarache, L. (2021). Using Phecodes for Research with the Electronic Health Record: From PheWAS to PheRS. *Annual Review of Biomedical Data Science*, *4*(1), 1–19. https://doi.org/10.1146/annurev-biodatasci-122320-112352

Brain Trauma Foundation - Frequently Asked Questions (FAQ) — Brain Trauma Foundation. (n.d.). Brain Trauma Foundation. https://braintrauma.org/info/faq

Culbertson, N. (2021, August 6). The Skyrocketing Volume Of Healthcare Data Makes Privacy Imperative. *Forbes*. https://www.forbes.com/sites/forbestechcouncil/2021/08/06/the-skyrocketing-volume-of-healthcare-data-makes-privacy-imperative/?sh=7c0913496555

Damani, O. P. (2013). Improving Pointwise Mutual Information (PMI) by Incorporating Significant Co-occurrence. In *Conference on Computational Natural Language Learning* (pp. 20–28). https://arxiv.org/pdf/1307.0596

Epalm. (2021). Interoperability in Healthcare. *HIMSS*. https://www.himss.org/resources/interoperability-healthcare

Interoperability and the Connected Health Care System | CMS. (n.d.). https://www.cms.gov/blog/interoperability-and-connected-health-care-system

Klema, V., & Laub, A. J. (1980). The singular value decomposition: Its computation and some applications. *IEEE Transactions on Automatic Control*, *25*(2), 164–176. https://doi.org/10.1109/tac.1980.1102314

Lambert, D. (2023, February 8). *What is Interoperability in Healthcare and its Benefits?* Continuum. https://www.carecloud.com/continuum/what-is-interoperability/

Levy, O., & Goldberg, Y. (2014). Neural Word Embedding as Implicit Matrix Factorization. In *Neural Information Processing Systems* (Vol. 27, pp. 2177–2185). http://papers.nips.cc/paper/5477-neural-word-embedding-as-implicit-matrix-factorization.pdf

Michigan Genomics Initiative | University of Michigan Precision Health. (n.d.). https://precisionhealth.umich.edu/our-research/michigangenomics/

Shifted PPMI Matrix - GM-RKB. (n.d.). http://www.gabormelli.com/RKB/Shifted_PPMI_Matrix

Traumatic brain injury impairs hormone production, disrupting sleep, cognition, memory. (2021, July 11). UTMB News. https://www.utmb.edu/news/article/utmb-news/2021/07/11/traumatic-brain-injury-impairs-hormone-production-disrupting-sleep-cognition-memory

Wright, T. J., Urban, R. J., Durham, W. B., Dillon, E. L., Randolph, K. M., Danesi, C. P., Gilkison, C. R., Karmonik, C., Zgaljardic, D. J., Masel, B. E., Bishop, J., Pyles, R. B., Seidler, R. D., Hierholzer, A. H., & Sheffield-Moore, M. (2020). Growth Hormone Alters Brain Morphometry, Connectivity, and Behavior in Subjects with Fatigue after Mild Traumatic Brain Injury. *Journal of Neurotrauma*, *37*(8), 1052–1066. https://doi.org/10.1089/neu.2019.6690

Wu, P., Gifford, A., Meng, X., Li, X., Campbell, H., Varley, T., Zhao, J., Carroll, R., Bastarache, L., Denny, J. C., Theodoratou, E., & Wei, W. (2019). Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development and Initial Evaluation. *JMIR Medical Informatics*, *7*(4), e14325. https://doi.org/10.2196/14325

Zhao, Z., Liu, T., Li, S., Li, B., & Du, X. (2017). Ngram2vec: Learning Improved Word Representations from Ngram Co-occurrence Statistics. In *Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/d17-1023