Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/mbe

MBE, 19(7): 6795–6813.
DOI: 10.3934/mbe.2022320
Received: 30 November 2021
Revised: 08 April 2022
Accepted: 24 April 2022
Published: 05 May 2022

Research article

Population scale latent space cohort matching for the improved use and exploration of observational trial data

Rachel Gologorsky [1,†], Sulaiman S. Somani [2,†], Sean N. Neifert [3], Aly A. Valliani [1], Katherine E. Link [1], Viola J. Chen [4], Anthony B. Costa [5] and Eric K. Oermann [3,6,*]

[1] Department of Medicine, Icahn School of Medicine, New York, NY 10028, USA
[2] Department of Medicine, Stanford University School of Medicine, Stanford, CA 94305, USA
[3] Department of Neurosurgery, NYU Grossman School of Medicine, New York, NY 10016, USA
[4] Oncology Early development, Merck & Co., Inc, Kenilworth, NJ 07033, USA
[5] NVIDIA, Santa Clara, CA 95051, USA
[6] Department of Radiology, NYU Grossman School of Medicine, New York, NY 10016, USA

† The authors contributed equally to this work.

* **Correspondence:** Email: Eric.Oermann@nyulangone.org.

Abstract: A significant amount of clinical research is observational by nature and derived from medical records, clinical trials, and large-scale registries. While there is no substitute for randomized, controlled experimentation, such experiments or trials are often costly, time consuming, and even ethically or practically impossible to execute. Combining classical regression and structural equation modeling with matching techniques can leverage the value of observational data. Nevertheless, identifying variables of greatest interest in high-dimensional data is frequently challenging, even with application of classical dimensionality reduction and/or propensity scoring techniques. Here, we demonstrate that projecting high-dimensional medical data onto a lower-dimensional manifold using deep autoencoders and *post-hoc* generation of treatment/control cohorts based on proximity in the lower-dimensional space results in better matching of confounding variables compared to classical propensity score matching (PSM) in the original high-dimensional space ($P < 0.0001$) and performs similarly to PSM models constructed by experts with prior knowledge of the underlying pathology when evaluated on predicting risk ratios from real-world clinical data. Thus, in cases when the underlying problem is poorly understood and the data is high-dimensional in nature, matching in the autoencoder latent space might be of particular benefit.

Keywords: artificial intelligence; autoencoders; cohort matching; data visualization; deep learning; manifold learning

1. Introduction

Controlled experimentation is one of two fundamental methods of scientific inquiry due to its ability to identify causal mechanisms and to reduce the impact of confounding factors on the final outcome. For these reasons, randomized controlled trials (RCTs) are widely regarded as the gold standard for inquiry in the medical sciences [1]. However, the majority of medical research is observational since this is the most accessible, and often only, means of gathering information. Some studies cannot be randomized for either ethical or pragmatic considerations, which encourages further reliance on observational data for medical decision-making. In other scenarios involving environmental exposures or trauma, the data is necessarily retrospective and often riddled with confounding factors [2].

Some research suggests that observational studies could achieve outcomes nearly equivalent to controlled experimentation if analyzed in a way which accounts for confounders and mitigates their impact (causal inference) [3, 4]. The most common techniques for controlling for confounding factors include multivariable regression analyses and post-hoc matching of treated and control groups, which are often combined in the form of matching propensity scores and performing multivariable regression to estimate average treatment effect on the treated (ATT) [4]. Methodological choices are often made based on the dimensionality of the data and the scientific question. Notably, techniques like stratification, direct matching, and propensity scoring become intractable with increasing dimensionality. Despite this limitation, the ease of implementation has lead to propensity score matching (PSM) becoming one of the most popular methods for analyzing observational data across a range of fields [5, 6].

We hypothesized that embedding high-dimensional input into a lower-dimensional space using deep autoencoders and matching cohorts based on proximity in the lower-dimensional space would result in better matching of confounding variables compared to PSM, which matches cohorts based on propensity scores in the high-dimensional space.

Autoencoders (AEs) are a type of unsupervised machine learning model designed to reconstruct their inputs from a reduced representation and, in doing so, learn a lower dimensional manifold of the data [7]. This technique has become increasingly useful with the advent of deep neural networks and deep AEs that are capable of capturing complex, nonlinear relationships [8]. The basic architecture of an AE consists of: 1) an encoder, which learns relevant features of the input data, 2) a bottleneck, which embeds the features into a compressed ("latent") representation, and 3) a decoder, which reconstructs the original input based on the compressed representation. Key considerations in designing deep AE architectures include setting the capacity of the encoder and decoder, and setting the degree of compression enforced by the bottleneck.

We utilized a deep autoencoder of fixed depth and dimension across all our experiments in order to maintain consistency (see Figure S1). Additionally, to address the issue of model interpretability and to identify whether or not the autoencoder learned semantically meaningful information, we visualized the latent space, using Uniform Manifold Approximation and Projection (UMAP) and t-Distributed Stochastic Neighbor Embedding (t-SNE) to create nonlinear projections onto a 2D manifold [9, 7]. UMAP is an algorithm for dimensionality reduction based on principles in topological data analysis and Riemannian geometry that works by generating a fuzzy topological representation of the data in a high dimensional space and finding a lower dimensional one that matches that as best as possible. t-SNE is a similarly nonlinear dimensionality reduction technique that maintains data similarity in the higher dimensional space when mapping to the lower dimensional manifold by minimizing the

Kullback-Leibler divergence between the higher and lower dimensional distributions.

More broadly, there is increasing interest in developing deep learning techniques for causal inference [10, 11, 12, 13]. In a related work, Johansson et al. introduced a novel balancing neural network, and demonstrated an improved performance on subsequent regression with weighted features [12]. Other deep learning methods generally rely on a combination of representation learning, feature balancing, and in some cases directly optimizing for average treatment effect estimation [10, 12, 13].

1.1. Contribution

We present a novel deep learning approach for cohort matching, latent space cohort matching, which matches patients based on feature vector similarity in the low-dimensional latent space learned by deep autoencoders. In contrast to the current standard of propensity score matching, latent space cohort matching is computationally tractable in the high-dimensional setting and is less easily distracted by uninformative features. Our approach using machine learning to construct the latent space has the advantage over classical dimensionality reduction techniques in being able to capture complex, nonlinear relationships in high-dimensional data. While other deep learning approaches for causal inference yield encouraging results [10, 11, 12, 13], they tend to be complicated and are neither practically implemented nor empirically validated on real world data. In contrast, we present a straightforward autoencoder implementation and empirically validate our latent-space matching algorithm on synthetic and real-world clinical datasets. We thus contribute a practical tool for estimating the causal effect of treatment in real-world high-dimensional observational data. Indeed, our results show that the non-isometric, non-linear embedding generated by the straightforward application of AEs—without the use of novel neural network architectures, loss functions, or regularization strategies—is in itself a good starting point for cohort matching, which we attribute to the fact that AEs learn a representation of the underlying datasets' density, helping to make latent space matching on a lower dimensional manifold a successful technique [14].

2. Materials and methods

2.1. Overview

Algorithm performance was benchmarked on 48 synthetically generated high-dimensional datasets with a known mixture of informative/redundant/random features. Performance was assessed by comparing the number of relevant input features that were matched between the algorithm-derived treatment/control cohorts. While our DeepMatch algorithm (DM) does not require pre-selection of features because it reduces input dimensionality by design, PSM becomes computationally intractable in high-dimensional settings and thus requires pre-selection of features in order to perform well. Consequently, we include in our analysis PSM performance when the dimension of the input data is reduced, either by randomly selecting a limited number of features or by selecting subsets of "expertly-chosen" features (informative only, informative + redundant).

Each experiment was composed of the following elements: cohort generation (with and without intervention), matching with random sampling 500 times as a control, matching with a series of PSM experiments with a variable number of features used for propensity-score generation (repeated 500 times to adequately sample the feature space), and matching based on features in the autoencoder latent space (DeepMatch).

Features were considered to have matched if the differences between treatment/control cohorts were not statistically significant (P >0.05) by either a t-test for ordinal data or chi-square test for categorical data.

Additionally, we leveraged publicly available data in order to extract three suitable observational datasets for assessing algorithm performance on real-world clinical data (Figure 1A,B). Algorithm performance was assessed by comparing the algorithm-derived risk ratios to the ground truth ATT. When available (2/3 datasets), the ground truth estimate of ATT was derived from RCTs conducted on similar populations; in the dataset with no available RCTs, the ground truth estimate of ATT was derived from the odds ratios reported by multiple large registry studies. For each real-world clinical dataset, we compared DM to 1) PSM with pre-selected, expert-driven features and 2) to PSM with a curated set of features including confounder variables that PSM needed to account for when matching treatment/control cohorts (Table S4).

2.2. *Synthetic datasets*

We generated 48 unique synthetic datasets to benchmark DM and PSM performance. Given that features in medical datasets can have variable range in values and be either categorical (e.g., presence or stage of hypertension) or continuous (e.g., blood pressure reading), half of these datasets treated features as inherently categorical and the other half of the datasets used ordinal features. Each of the 24 datasets in each arm was different in the maximum range of values each feature could take, which ranged from 2 (i.e., binary) to 25 (i.e., 25 classes). Varying the range was particularly useful for assessing the impact that the number of classes of a feature can have on matching performance, which can be a limiting factor for PSM.

Each synthetic dataset consisted of 100 k samples. Assignment to the intervention or control group (a binary event) was predicated on a set of 500 features. In order to emulate how an instrumental variable may appear relative to its causal and confounding factors in real-world medical data, we designed three categories of features: informative (50 features), redundant (250 features), and random (200 features). In causal inference terms, informative features serve as instrumental variables or direct causal associations with the outcome of interest, whereas redundant features serve as confounders and random features as noise.

Each intervention is assigned a cluster of points generated by sampling a Gaussian distribution placed on an n-dimensional hypercube, where n is the number of informative features in the dataset. Once sampled, these informative features are then randomly linearly combined with one another in each cluster in order to add intracluster covariance and are placed on the vertices of the hypercube to increase their separation. Redundant features were generated as linear combinations of informative features, and random features were generated by sampling from a uniform distribution. Datasets were generated using the *make_classification* function in *scikit-learn* [15]. Details on the hyperparameters for dataset generation are provided in Table S1.

2.3. *Clinical datasets*

In order to validate the ability of our algorithm to derive risk ratios from real-world observational data, we utilized the patient data in the Nationwide/National Inpatient Sample (NIS) calendar years

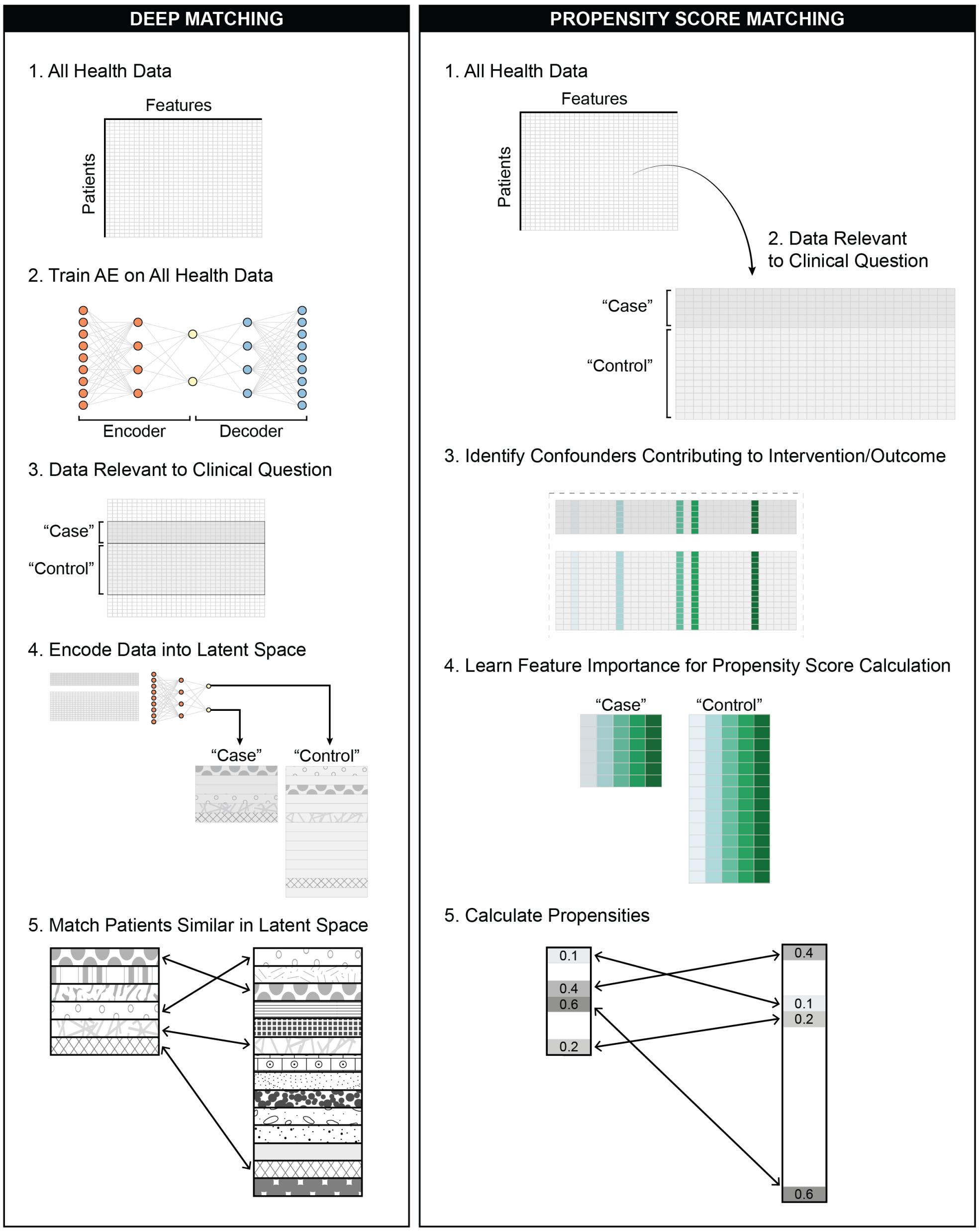
DEEP MATCHING
1. All Health Data
Features
Patients
2. Train AE on All Health Data
Encoder
Decoder
3. Data Relevant to Clinical Question
"Case"
"Control"
4. Encode Data into Latent Space
"Case"
"Control"
5. Match Patients Similar in Latent Space
PROPENSITY SCORE MATCHING
1. All Health Data
Features
Patients
2. Data Relevant to Clinical Question
"Case"
"Control"
3. Identify Confounders Contributing to Intervention/Outcome
4. Learn Feature Importance for Propensity Score Calculation
"Case"
"Control"
5. Calculate Propensities
0.1
0.4
0.6
0.2
0.4
0.1
0.2
0.6

Figure 1a. Summary of the basic concept of "deep matching"—matching patients on the lower dimensional manifold generated by an autoencoder (AE) as compared to propensity score matching (PSM). PSM matches samples with similar probabilities of being assigned to the intervention group, conditioned on the baseline characteristics. It is typically modeled as a linear equation. AE, however, compresses samples into more compact representations, and matches samples based on their similarity in the compact representation.

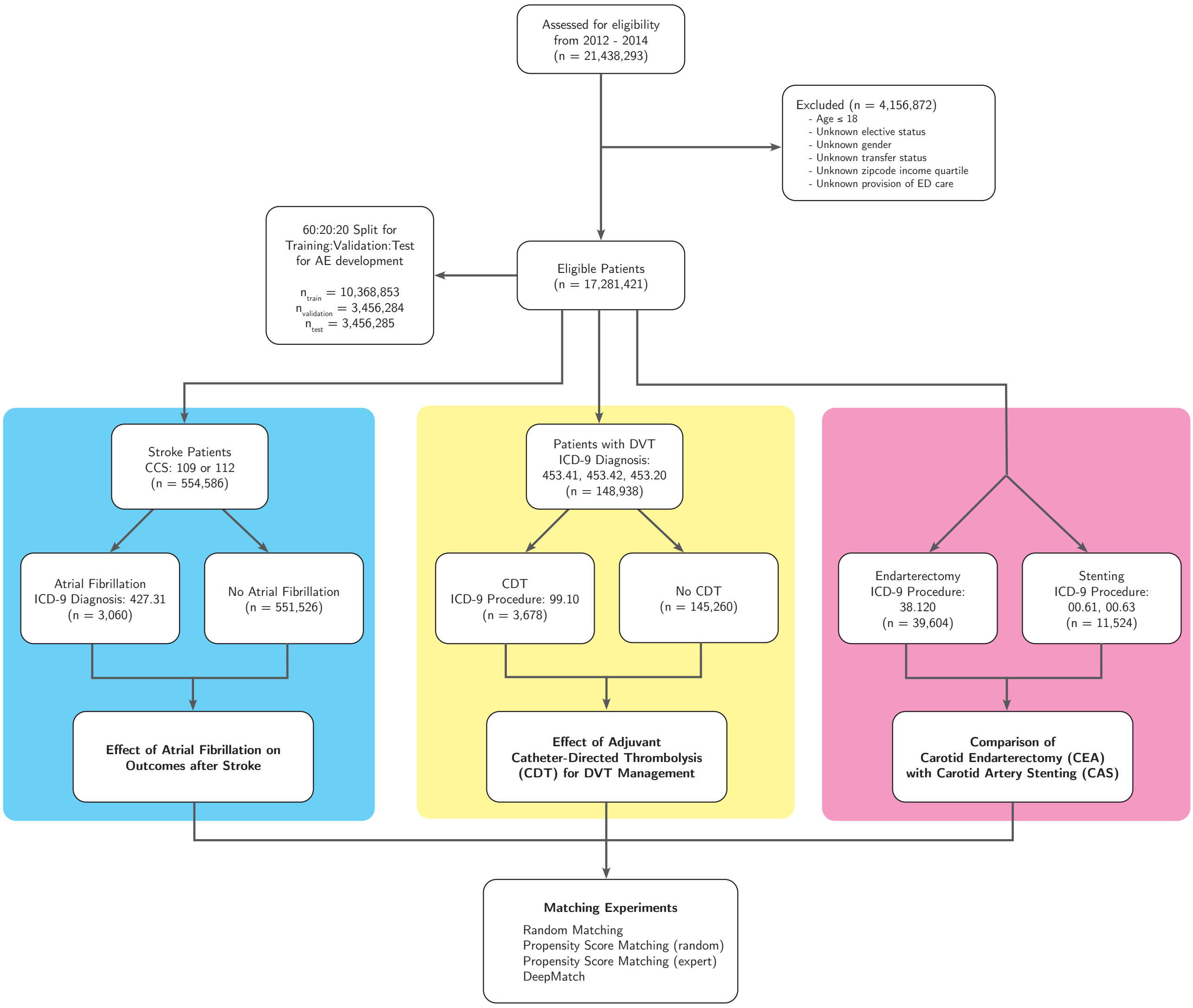

Figure 1b. Summary of our overall approach to selecting medical studies based on existing RCTs and cohorts sampled from the Nationwide Inpatient Sample (NIS) of 2012-2014.

2012–2014. The NIS database contains de-identified information on a sample of all inpatient hospital admissions across the United States from 1988 to 2017; it is maintained by the Agency for Healthcare Research and Quality (AHRQ) under the Healthcare Cost and Utilization Project (HCUP) for the intended use of analyzing hospital utilization, charges, quality, and outcomes [16, 17].

NIS is designed through a systematic sampling procedure to represent all inpatient admissions in the US, and the chosen years included 21,438,293 weighted inpatient admissions representing over 100 million admissions. While data elements contained in the NIS have changed over time, they generally include patient demographics (e.g., age, ethnicity, income quartile by patient zip code), hospital admission (e.g., elective or emergency admission, route of admission, type of hospital being admitted to), pre-existing and in-hospital diagnoses (as defined by the International Classification of Diseases (ICD) system), and outcomes (e.g., patient death during hospital admission, total charge of hospitalization).

We used data from the NIS database from the start of 2012 to the third quarter of 2015 because of the homogeneity in reporting elements during that time period. We excluded data from patients who were pediatric (age < 18) or had missing values for the following variables: gender, zip code, income quartile, transfer status, elective admission, admission month, whether the admission took place on a weekend or not, and urban/rural classification scheme.

From the NIS database, we collected 3 clinical datasets suitable for investigating:

1) the effect of adjuvant catheter-directed thrombolysis (CDT) versus medical management on developing post-thrombotic syndrome for patients with deep vein thrombosis (DVT),

2) the effect of carotid endarterectomy (CEA) versus carotid artery stenting (CAS) on mortality, and

3) the effect of having atrial fibrillation as a co-morbidity on patient outcomes after ischemic stroke.

NIS Dataset #1:

Deep vein thrombosis (DVT) is a common medical condition affecting 1–2 per 1000 Americans per year [18]. This dataset included all patients from NIS with a DVT (ICD-9 diagnosis codes 453.41, 453.42, 453.20; n = 148,938). Those who underwent catheter-directed thrombolysis (treatment group, n = 3678) were seperated from those who did not (control group, n = 145,260) based on the presence of the ICD-9 procedure code, “99.10”. A recent RCT consisting of 700 patients with DVTs demonstrated a relative risk of 0.89 (p = 0.83) associated with the use of CDT as compared to medical management [18], which we take to be our “ground truth” ATT benchmark. The expert-driven PSM model generates propensity scores based on age, sex, history of obesity (‘CM_OBESE’ NIS data element), peripheral vascular disease, smoking, coagulopathy, whether it was a proximal DVT (ICD-9 diagnosis code “453.41”), a patient history of DVT (ICD-9 diagnosis code “V12.51”), and history of anticoagulant use (ICD-9 diagnosis code “V58.61”) [18, 19].

NIS Dataset #2:

This dataset included all patients from NIS with a procedure code for carotid endarterectomy (ICD-9 procedure code “38.120”) and carotid artery stenting (ICD-9 procedure codes “00.61” and “00.63”). We thus obtained two cohorts of size 39,604 (CEA) and size 11,524 (CAS). Treatment results of carotid arterial stenting have been compared to carotid endarterectomy for carotid atherosclerotic disease in multiple studies [20], with no statistically significant difference found. The expert-driven PSM model generates propensity scores based on age, sex, obesity (‘CM_OBESE’ NIS data element), and a history of any of the following: hypertension, dyslipidemia, coronary artery disease, myocardial infarction, coronary artery bypass grafting, congestive heart failure, atrial fibrillation, arrhythmia (“cardiac conduction disorders”), stroke, peripheral vascular disease, diabetes, and smoking [20, 21, 22].

NIS Dataset #3:

This dataset included 554,586 patients admitted with ischemic stroke (CCS code 109 or 112). Those who had atrial fibrillation (n = 3060) were separated from those who did not (n = 551,526) based on the presence of the diagnosis code, “427.31”. While there are no underlying RCTs, several large registry studies have estimated ORs from 1.7–3.3 [20, 23, 24]. The expert-driven PSM model generates propensity scores based on age, sex, history of obesity (‘CM_OBESE’ NIS data element), chronic kidney disease, coagulopathy, and CHADS-VASC score [23, 24, 25].

2.4. *DeepMatch algorithm*

For each dataset, a separate AE was trained using the fixed architecture discussed below and shown in Figure S2. For training, we used a 60: 20: 20 train/valid/test dataset split in order to prevent model overfitting.

To match patients, the newly trained AE model embeds the treatment/control cohorts into the latent space, from which an approximate nearest neighbor tree (Annoy library [26]) on the larger cohort with the Euclidean norm to represent the distance between neighbors is generated. The matching process, which is inspired by the Gale-Shapley algorithm, attempts to iteratively match each point in the smaller cohort by querying the index tree for 64 closest neighbors for each point and matching the smallest pairwise relationships. While this algorithm fails to account for the *net* pairwise distance between matched points, its current structure is more computationally tractable and ensures that the result of this paper’s experiments are a lower bound on the potential benefits of DeepMatch.

2.5. *Autoencoder implementation*

We utilized a deep autoencoder of fixed depth and dimension for all experiments. The encoder block is composed of two dense blocks, each containing a linear, fully-connected layer (256, 64), a leaky ReLU activation layer, and a batch normalization layer. The decoder block is equivalent to the encoder block, but in reverse. Fully-connected layers with a sigmoid activation were placed in parallel after the decoder output to reconstruct the one-hot encoded form of each category of features (e.g., procedure codes, gender). See Figure S1.

We designed our synthetic datasets so that half of them treated features as inherently categorical, while the other half treated features as inherently ordinal. Categorical variables in the AE architecture were represented as one-hot encoded vectors, and the model was trained to minimize the binary cross entropy loss for each feature being reconstructed. Ordinal features do not require one-hot encoding of the features prior to model entry, which enables mean-squared error loss to be used instead.

The clinical datasets posed several challenges for the AE architecture. First, to address the sparsity present in the features (where patients may have a small but variable number of diagnostic codes assigned to them), we took inspiration from collaborative filtering and natural language processing to use embedding layers to reduce the sparse, one-hot encoded form of diagnosis codes, chronic conditions and associated body systems, procedure codes, and external cause of injury codes into a lower-dimensional continuous vector representation that improved training efficiency. Features with a low range of values, such as age binned into deciles, gender, income quartile of patient's zip code, transfer status, and elective admission, were one-hot encoded and directly concatenated with the outputs of these embedded layers. Second, to account for the heterogeneity in data representation, we created a custom loss function formulated as the sum of the cross-entropy loss for each category of features except for age with the sum of the mean-square error loss of the reconstructed age: Loss(original input, reconstructed input) = MSELoss(patient age bucket, reconstructed age bucket) + BCEWithLogitsLoss(patient feature value, reconstructed feature value) for feature ≠ to age bucket. MSE loss was used for age bucket as age decile was treated as an ordinal feature; BCE loss was used for the other features as they were boolean-valued.

The following list of elements was used for training the autoencoder: age ("AGE"), gender ("FEMALE"), evidence of emergency department services on admission ("HCUP_ED"), whether the patient was transferred from another facility ("TRAN_IN"), whether the admission was elective or not ("ELECTIVE"), the income quartile associated with the patient's zip code ("ZIPINC_QRTL"), all ICD-9 diagnosis codes associated with the patient encounter ("DX1", . . . , "DX30"), all ICD-9 external causes of injury codes ("ECODE1", . . . , "ECODE4"), all ICD-9 procedure codes ("PR1", . . . , "PR30"), all ICD-9 codes associated with pre-existing comorbidities ("CHRON1", . . . , "CHRON30"), and the body system associated with the ICD-9 code ("CHRONB1", . . . , "CHRONB30"). All ICD-9 codes (diagnosis, procedure, and external causes of injury) were mapped from their native format (eg "410.3") into an ordinal list of unique elements ranging from 1 to N_{unique}, where N_{unique} is the number of unique codes.

Each autoencoder model was trained using minibatches of 512 patients from 80% of the dataset and allowed to learn for 100 epochs using the Adam optimizer (initial learning rate = 5e-4, ß1 = 0.9, ß2 = 0.999) coupled to a learning rate scheduler that reduced the learning rate by a factor of 10 to no more than 1e-7 when the model loss was within 1e-3 on successive epochs.

3. Results

3.1. Synthetic datasets

We compared the performance of DM vs. PSM on matching cohorts based on 500 features (50 informative, 250 redundant, 200 random). As PSM performance degrades on high-dimensional data, we included in our comparison PSM performance on selected subsets of features (informative only, informative + redundant).

Without prior selection of features, DM outperforms PSM by a statistically significant average improvement of 62 matched features (P <0.0001). When PSM is applied to the subset of informative features only, performance is comparable between the two algorithms: DM results in a statistically insignificant average improvement of 2.4 matched features (P = 0.84, paired sample t-test). These results were similar both when assessing DM performance on categorical and ordinal features (Figure 2A).

In addition, we assessed the impact that the number of classes a feature can have on matching performance. While PSM performance declined as the data range increased, we observed that DM matching on the underlying latent space was effective for both features with smaller numbers of classes

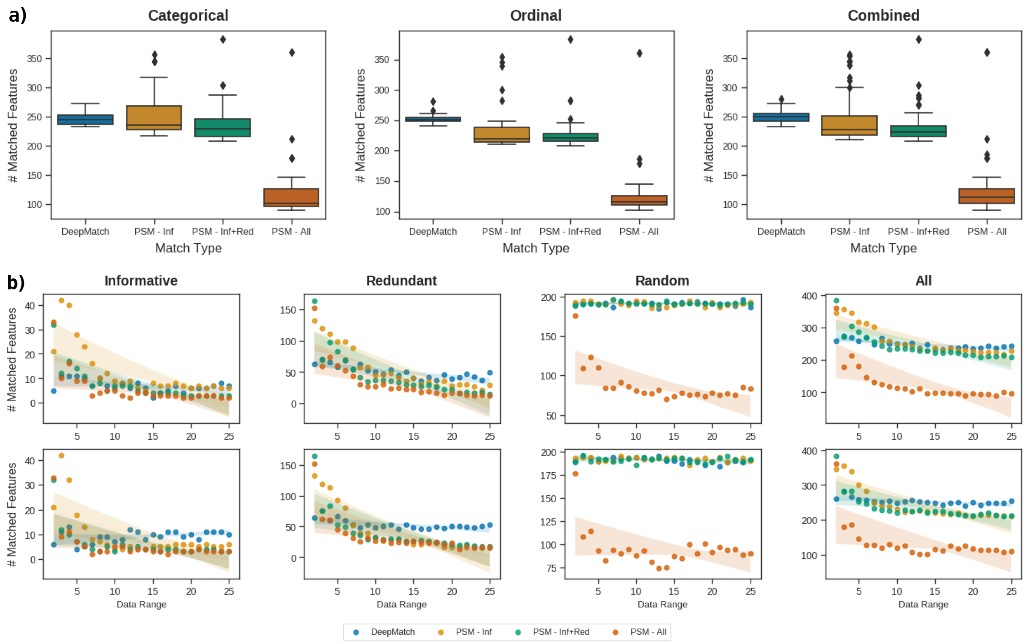

Figure 2. a: Prototyping of latent space matching on a synthetic dataset. A system of linear equations was utilized to generate synthetic data consisting of feature dimensions of varying degrees of informativeness. When we compared different matching approaches without prior selection of features, there was a difference of 62 successfully matched features when compared to propensity score matching (P <0.0001). b: Categorical features (top row), continuous features (bottom row). Matching on the underlying latent space was effective for both smaller numbers of low dimensional features and was increasingly effective as the number of classes for each feature increased.

and was increasingly effective as the range of classes increased for both categorical and continuous features (Figure 2B).

3.2. Real-world clinical datasets

1) Evaluating the DM algorithm on the observational dataset of patient outcomes with and without catheter directed thrombolysis treatment, we found that the cohort matched on the latent space variables had a RR of 0.89, which was similar to PSM using expertly chosen features with a RR of 0.87 (see Appendix for a list of features), despite poor performance on matching presumably clinically relevant features (Figure 3A). Notably, in the absence of an expertly defined feature set, PSM had increasingly worse results as further features were added to the model (Figure 3B). In addition, we assessed the performance at 1% of the original dataset size (30 instead of 3000 admissions, randomly sampled) and noted that the DM algorithm results continued to be robust with regards to the underlying RR estimate (0.67) compared to PSM (1.99), although the underlying matching suffered (Figure 3D).

2) Although multiple studies found no statistically significant difference between carotid artery stenting and carotid endarterectomy for treating carotid atherosclerotic disease, both the expert-driven PSM and the latent-space DM matching on our NIS dataset predict the same relative risk of 3.7. (Figure 3E).

3) Propensity score matching, including the use of expert features, estimated the RR of having atrial fibrillation as a co-morbidity after ischemic stroke as 0.5–0.6, suggesting a protective effect. On the other hand, DM matching in the latent space results in a predicted RR of 1.1, which more closely reflects the results of existing large registry studies with ORs from 1.7–3.3 [20, 23, 24].

3.3. Model interpretability

To address the issue of model interpretability, we used uniform manifold approximation and representation (UMAP) to visualize the underlying latent space (manifold) projected against semantically meaningful features including socioeconomic status, elective medical care, and age. Parameters used for fitting the data for generating the embedding matrix are shown in Tables S2 and S3. We used GPU-accelerated implementations of the [27] t-SNE and UMAP [27, 28] for these visualization studies. Given computational limitations, only 10% of the full dataset (1,736,983 patients) was used for generating the t-SNE mapping and only 20% of the full dataset (3,410,873 patients) was used for generating the UMAP mapping.

On visualization, the distinct clusters show that differences in these features are indeed represented in the latent space (Figure 4). Conceivably, visualizations of this kind could be used not only to check intuition, but also to give rise to new hypotheses about relationships between features.

4. Discussion

We present the first investigation of observational medical data within the lower dimensional space of an autoencoder trained on a population scale dataset. This is one of the few empirical studies demonstrating a novel approach to working with observational data compared to existing statistical techniques

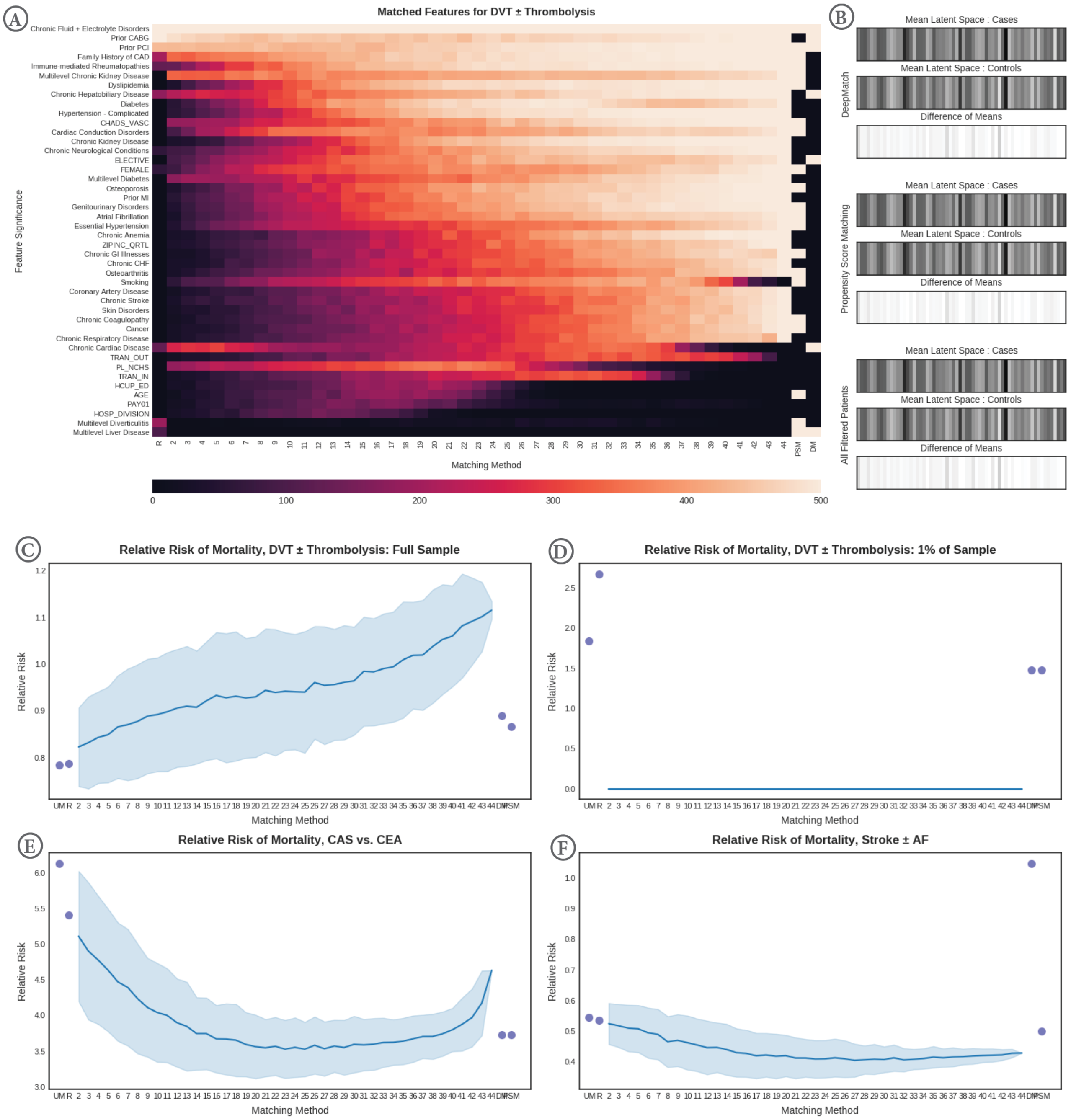

Figure 3. a: Notably, in the absence of an expertly defined feature set, PSM had increasingly worse results as further features were added to the model. b: Using the latent space of the autoencoder we visualized the relative differences between case and control cohorts matched by different algorithms as clinical barcodes. c: Using a cohort of 3678 admissions with DVTs who underwent CDT compared to a control group of 145,260 admissions who underwent best medical management, matching on the latent space (RR = 0.89) was similar to PSM using expertly chosen features (RR = 0.87). d: Using random sampling, we assessed the performance at 1% of the original dataset size (30 instead of 3000 admissions) and noted that the results continued to be robust with regards to the underlying RR estimate (0.67) compared to PSM (1.99), although the underlying matching suffered. We investigated two additional cohorts to further test the robustness of our findings. e: We found that both PSM and latent space matching achieved the same relative risk of 3.7 for carotid stenting (CAS) vs endarterectomy (CEA). f: Lastly, we compared 3060 admissions with ischemic stroke to investigate the risk of inpatient stroke with and without comorbid atrial fibrillation. Several large registry studies have estimated ORs from 1.7–3.3. Propensity scores and matching methods including the use of expert features all estimated ORs from 0.5 to 0.6, suggesting a protective effect, while latent space matching alone estimated an OR of 1.1 which more closely reflects existing observational studies.

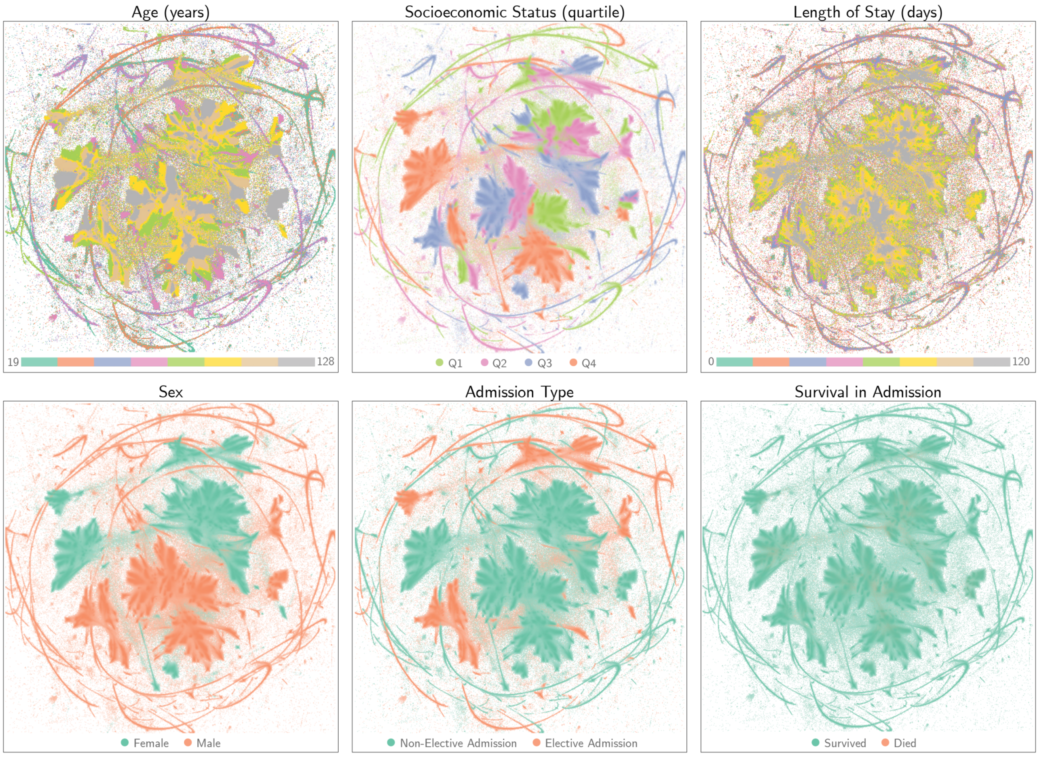

Figure 4. Uniform manifold approximation and representation (UMAP) projection derived from the Nationwide Inpatient Sample (NIS). A randomly selected subset from the 2012–2014 NIS consisting of 20% of the 21,438,293 cases was used to learn the matrix for projection of samples in the latent space into two dimensions using UMAP. All 21,438,293 samples were then projected into a 2D plane and visualized using features of interest, with distinct patterns emerging corresponding to several social determinants of care including elective admission, socioeconomic status, and age.

and the only one using the latest techniques from manifold learning and deep learning. We further show that simple visual inspection of the underlying latent space can be helpful for hypothesis generation and understanding basic relationships in the medical data. A key concern in observational data analysis is controlling for potential confounding variables, as differences in the treatment/control cohorts could bias the estimate of average treatment effect. We lastly show that balance of confounding variables between the treatment/control cohorts is enhanced by latent space cohort matching as compared to propensity score matching. Compared to the conventional technique of PSM, latent space matching is less easily distracted by uninformative features and performs similarly to models constructed by experts with prior knowledge of the underlying pathology. Given the robust average treatment effect (i.e., mortality) relative to our ground truth (i.e., RCTs) despite poor matching on expert clinical features (Figure 3A), our low-dimensional manifold may capture a more global context using nonlinear relationships that may not be apparent when evaluating similarity by feature-by-feature statistical comparisons. In cases when there is little or poor understanding of the underlying problem and the data is high dimensional in nature, matching in the latent space might be of particular benefit to researchers looking to optimize their analysis of observational data without having to resort to controlled experimentation.

This study has several limitations worth mentioning. First, we did not perform an exhaustive search of AE architectures or training methodologies. We opted for a simple AE architecture and simple approach to focus attention on our underlying point—that the lower dimensional embedding of the data contains the information necessary and sufficient for analysis. It is quite possible that a novel AE variant that preserves the underlying dataset's topology or distances may yield superior results. Therefore, our present work with a simple denoising AE may in fact be a lower bound on how well an AE can perform for the purposes of creating a latent space for subsequent matching. We have also only assessed our results on a synthetically generated dataset and three clinical datasets. It is quite possible that edge cases exist where this technique is not of benefit. For example, as with RCTs, the ability of DM to accurately estimate causal effects requires careful consideration of study design: the autoencoder should be trained and the DM matching algorithm performed only after the study population has been well-defined, adequate time for follow-up assessed, and missing data scrutinized. Furthermore, as treatment/control cohort randomization is simulated based on *known features*, there may be unobserved patient characteristics that are not controlled for even with perfect matching of all features recorded in the dataset. However, we are encouraged that in all of our experiments deep matching has performed *no worse* than propensity score matching, and is of particular benefit for high dimensional data.

Second, using deep learning for causal inference raises the issue of model interpretability. While visualization techniques are helpful in addressing this issue, we must remain cognizant that projecting the latent space down to 2D or 3D for visualization introduces warping, so that hypothesis generation based on cluster size or inter-cluster distances are not necessarily accurate. Additionally, while we confirmed that our model learned semantically meaningful features by projecting the latent representation against *a priori* relevant features, such sanity checks may not be possible when there is a poor understanding of the relevant features to begin with.

In conclusion, in a world with increasing quantities of observational data, techniques such as latent space matching may prove useful to scientists seeking to perform causal inference and maximize the use of their data in cases where controlled experimentation is not possible.

Acknowledgments

The authors declare there are no outside funding sources. Data utilized in this study are publicly accessible here: https://www.hcup-us.ahrq.gov/db/nation/nis/nisdbdocumentation.jsp. Source code for this work is publicly available here: https://github.com/nyuolab/deepmatch.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. H. Sacks, T. C. Chalmers, H. S. Jr, Randomized versus historical controls for clinical trials, *Am. J. Med.*, **72** (1982), 233–240. https://doi.org/10.1016/0002-9343(82)90815-4

2. V. Butsic, D. J. Lewis, V. C. Radeloff, M. Baumann, T. Kuemmerle, Quasi-experimental methods enable stronger inferences from observational data in ecology, *Basic Appl. Ecol.*, **19** (2017), 1–10. https://doi.org/10.1016/j.baae.2017.01.005

3. J. Concato, N. Shah, R. I. Horwitz, Randomized, controlled trials, observational studies, and the hierarchy of research designs, *N. Engl. J. Med.*, **342** (2000), 1887–1892. https://doi.org/10.1056/NEJM200006223422507

4. E. A. Stuart, Matching methods for causal inference: A review and a look forward, *Stat. Sci.*, **25** (2010), 1–21. https://doi.org/10.1214/09-STS313

5. J. Pearl, The foundations of causal inference, *Sociol. Methodol.*, **40** (2010), 75–149. https://doi.org/10.1111/j.1467-9531.2010.01228.x

6. A. Abadie, G. W. Imbens, Large sample properties of matching estimators for average treatment effects, *Econometrica*, **74** (2006), 235–267. https://doi.org/10.1111/j.1468-0262.2006.00655.x

7. G. E. Hinton, R. R. Salakhutdinov, Reducing the dimensionality of data with neural networks, *Science*, **313** (2006), 504–507. https://doi.org/10.1126/science.1127647

8. M. Atzmon, A. Gropp, Y. Lipman, Isometric autoencoders, preprint, arXiv:2006.09289.

9. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al., Scikit-learn: Machine learning in python, *J. Mach. Learn. Res.*, **12** (2011), 2825–2830. https://doi.org/10.48550/arXiv.1201.0490

10. N. Kallus, DeepMatch: Balancing deep covariate representations for causal inference using adversarial training, in *Proceedings of the 37th International Conference on Machine Learning*, **119** (2020), 5067–5077.

11. N. Kallus, Optimal a priori balance in the design of controlled experiments, *J. R. Stat. Soc.*, **80** (2018), 85–112. https://doi.org/10.1111/rssb.12240

12. F. D. Johansson, U. Shalit, D. Sontag, Learning representations for counterfactual inference, in *Proceedings of the 33rd International Conference on Machine Learning*, **48** (2016), 3020–3029.

13. A. J. Averitt, N. Vanitchanant, R. Ranganath, A. J. Perotte, The counterfactual χ-GAN: Finding comparable cohorts in observational health data, *J. Biomed. Inform.*, **109** (2020), 103515. https://doi.org/10.1016/j.jbi.2020.103515

14. G. Alain, Y. Bengio, What regularized Auto-Encoders learn from the Data-Generating distribution, *J. Mach. Learn. Res.*, 2012. https://doi.org/10.48550/arXiv.1211.4246

15. Scikit-learn, scikit-learn/scikit-learn, https://github.com/scikit-learn/scikit-learn

16. Hcup, Agency for healthcare research and quality, healthcare cost and utilization project HCUP-US NIS overview, https://www.hcup-us.ahrq.gov/nisoverview.jsp, 2012.

17. Hcup, Agency for healthcare research and quality, healthcare cost and utilization project, NIS database documentation, https://www.hcup-us.ahrq.gov/db/nation/nis/nisdbdocumentation.jsp, 2012.

18. S. Vedantham, S. Z. Goldhaber, J. A. Julian, S. R. Kahn, M. R. Jaff, D. J. Cohen, et al., Pharmacomechanical Catheter-Directed thrombolysis for Deep-Vein thrombosis, *N. Engl. J. Med.*, **377** (2017), 2240–2252. https://doi.org/10.1056/NEJMoa1615066

19. M. Alkhouli, C. J. Zack, H. Zhao, I. Shafi, R. Bashir, Comparative outcomes of catheter-directed thrombolysis plus anticoagulation versus anticoagulation alone in the treatment of inferior vena caval thrombosis, *Circ. Cardiovasc. Interv.*, **8** (2015), e001882. https://doi.org/10.1016/j.jvs.2015.07.046

20. H. S. Gurm, J. S. Yadav, P. Fayad, B. T. Katzen, G. J. Mishkel, T. K. Bajwa, et al., Long-term results of carotid stenting versus endarterectomy in high-risk patients, *N. Engl. J. Med.*, **358** (2008), 1572–1579. https://doi.org/10.1056/NEJMoa0708028

21. L. K. Kim, D. C. Yang, R. V. Swaminathan, R. M. Minutello, P. M. Okin, M. K. Lee, et al., Comparison of trends and outcomes of carotid artery stenting and endarterectomy in the united states, 2001 to 2010, *Circ. Cardiovasc. Interv.*, **7** (2014), 692–700. https://doi.org/10.1161/CIRCINTERVENTIONS.113.001338

22. J. L. Mas, G. Chatellier, B. Beyssen, A. Branchereau, T. Moulin, J. P. Becquemin, et al., Endarterectomy versus stenting in patients with symptomatic severe carotid stenosis, *N. Engl. J. Med.*, **355** (2006), 1660–1671. https://doi.org/10.1056/NEJMoa061752

23. K. Kimura, K. Minematsu, T. Yamaguchi, Japan Multicenter Stroke Investigators' Collaboration (J-MUSIC), Atrial fibrillation as a predictive factor for severe stroke and early death in 15,831 patients with acute ischaemic stroke, *J. Neurol. Neurosurg. Psychiatry*, **76** (2005), 679–683. https://doi.org/10.1136/jnnp.2004.048827

24. K. Keller, L. Hobohm, P. Wenzel, T. Münzel, C. Espinola-Klein, M. A. Ostad, Impact of atrial fibrillation/flutter on the in-hospital mortality of ischemic stroke patients, *Heart Rhythm*, **17** (2020), 383–390. https://doi.org/10.1016/j.hrthm.2019.10.001

25. H. S. Jørgensen, H. Nakayama, J. Reith, H. O. Raaschou, T. S. Olsen, Acute stroke with atrial fibrillation. the copenhagen stroke study, *Stroke*, **27** (1996), 1765–1769. https://doi.org/10.1161/01.STR.27.10.1765

26. Spotify, spotify/annoy, https://github.com/spotify/annoy

27. CannyLab, CannyLab/tsne-cuda, https://github.com/CannyLab/tsne-cuda

28. Rapidsai, rapidsai/cuml, https://github.com/rapidsai/cuml

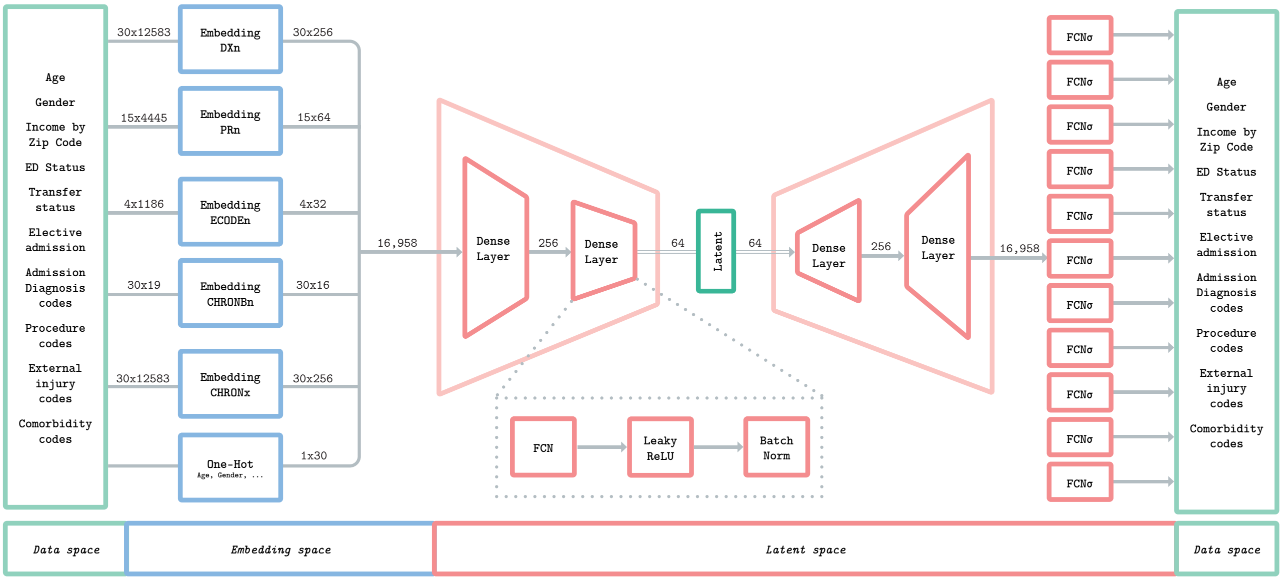

Figure S1. Model architecture used to train the NIS dataset.

Table 1. Synthetic dataset generator parameters.

Parameter	Value
Number of samples	100,000
Informative features	50
Redundant features	250
Random features	200
Repeated features	0
Number of classes	2
Number of clusters per class	10
Class assignment weight	0.95
Flip class factor	0.01
Random state	0

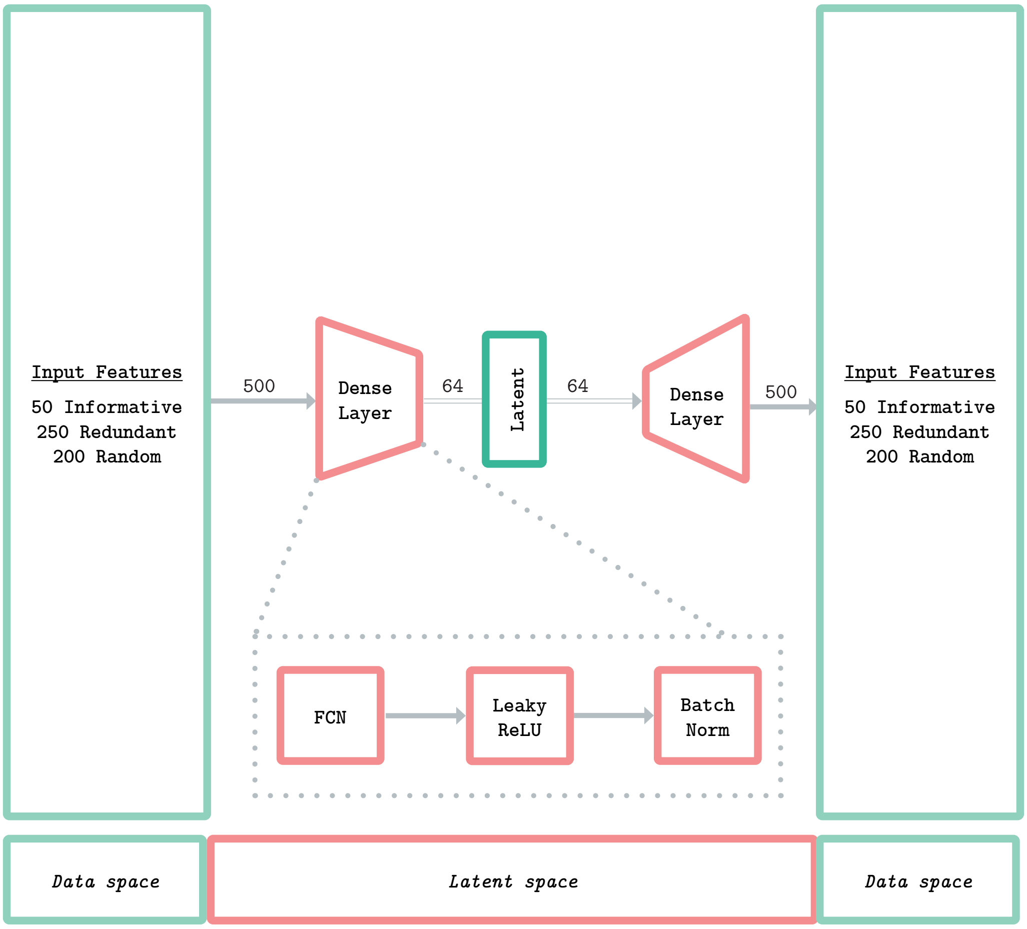

Figure S2. Fixed Autoencoder architecture.

Table 2. t-SNE parameters.

Parameter	**Value**
Perplexity	30
Early Exaggeration	12.0
Number of components	2
Distance metric	Euclidean norm
Initialization	Random

Table 3. UMAP parameters.

Parameter	**Value**
Number of neighbors	15
Number of components	2
Distance metric	Euclidean norm
Minimal distance	0.1
Spread	1.0
Initialization	Spectral embedding of a fuzzy 1D skeleton

Table 4. Curated features for PSM experiments.

Feature	Specification	Comment
AGE	NIS Data Element	Bucketed into deciles.
ELECTIVE	NIS Data Element	Whether the current admission is elective or not.
FEMALE	NIS Data Element	Patient sex.
HCUP_ED	NIS Data Element	Evidence of emergency room services provided on this admission.
HOSP_DIVISION	NIS Data Element	Region current hospital falls into.
PAY01	NIS Data Element	Expected primary payer
PL_NCHS	NIS Data Element	Patient location by local area population.
TRAN_IN	NIS Data Element	Whether current admission was transferred from another facility.
TRAN_OUT	NIS Data Element	Whether current admission was transferred out to another facility.
ZIPINC_QRTL	NIS Data Element	Income quartile associated with that patient's zip code.
Essential Hypertension	CCS: 98	Boolean
Complicated Hypertension	CCS: 99	Boolean
Dyslipidemia	CCS: 53	Boolean
Atrial Fibrillation	ICD-9: 427.31	Boolean
Coronary Artery Disease	CCS: 101	Boolean
Diabetes	CCS: 49, 50	Boolean
Smoking	ICD-9: 305.10, V15.82	Boolean
Prior MI	ICD-9: 412.00	Boolean
Prior PCI	ICD-9: V45.82	Boolean
Prior CABG	ICD-9: V45.81	Boolean
Family history of Coronary Artery Disease	ICD-9: V17.30	Boolean
History of a chronic cardiac illness	CCS: 96-97, 100-101, 103-108	Boolean
Long-standing congestive heart failure	CCS: 108	Boolean
History of stroke	CCS: 109, 112	Boolean
History of arrhythmia	CCS: 105-107	Boolean
History of a chronic respiratory illness	CCS: 127-128, 131-134	Boolean
Any cancer	CCS: 11-47	Boolean
History of fluid or electrolyte disorders	CCS: 55	Boolean
History of anemia	CCS: 59-61	Boolean
History of coagulopathy	CCS: 62	Boolean
History of neurological conditions	CCS: 76-84, 93, 95	Boolean
History of GI Illness	CCS: 135, 138-148, 152-155	Boolean
History of hepatobiliary disease	CCS: 6, 149, 151	Boolean
Chronic kidney disease (any)	CCS: 158	Boolean
Immune-mediated rheumatopathies	CCS: 202, 210-211	Boolean
Osteoporosis	CSC: 206	Boolean
Osteoarthritis	CCS: 203	Boolean
History of dermatological illness	CCS: 197-200	Boolean
History of genitourinary disorders	CCS: 163-172	Boolean
History of peripheral vascular disease	ICD-9: 412.00 CCS: 114-116	Boolean
Liver disease (multi-level)	1: ICD-9 = (70.22, 70.23, 70.32, 70.33, 70.44, 70.54, 70.6, 70.9, 570.00, 571.00, 571.10, 572.20, 573.30, 574.40, 574.41, 574.42, 574.43, 574.44, 573.30, 573.40, 573.80, 573.90) 2: ICD-9 = (456.00, 456.10, 456.20, 456.21, 572.20, 572.30, 572.40, 572.80)	Categorical (0, 1, 2)
Diverticulitis (multi-level)	1: ICD-9 = (562.11, 561.13) 2: ICD-9 = (569.83, 567.22, 569.50, 567.31, 567.38, 614.50, 567.21, 567.29)	Categorical (0, 1, 2)
Chronic kidney disease (multi-level)	1: ICD-9 = 585.10 2: ICD-9 = 585.20 3: ICD-9 = 585.30 4: ICD-9 = 585.40 5: ICD-9 = 585.50 6: ICD-9 = 585.60	Categorical (0, 1, 2, 3, 4, 5, 6)
Complicated vs. Uncomplicated Diabetes (multi-level)	1: CCS = 49 2: CCS = 50	Categorical (0, 1, 2)
CHADS-VASC	Generated as the sum of the following: +2 for AGE ≥ 75 +1 for 75 >AGE ≥ 65 +1 for (FEMALE = 1) +1 for (CM_OBESE = 1) +1 for ('Long-standing congestive heart failure' = 1) +1 for ('History of stroke' = 1) +1 for ('History of peripheral vascular disease' = 1) +1 for ('Diabetes' = 1)	Categorical (0, 1, 2, 3, 4, 5, 6, 7)