npj | Digital Medicine

www.nature.com/npjdigitalmed

ARTICLE OPEN

Check for updates

Deep representation learning of electronic health records to unlock patient stratification at scale

Isotta Landi[1,2], Benjamin S. Glicksberg[3,4,5], Hao-Chih Lee[4,5], Sarah Cherng[4,5], Giulia Landi[6], Matteo Danieletto[3,4,5], Joel T. Dudley[4,5], Cesare Furlanello[1,7,8] and Riccardo Miotto[3,4,5,8]✉

Deriving disease subtypes from electronic health records (EHRs) can guide next-generation personalized medicine. However, challenges in summarizing and representing patient data prevent widespread practice of scalable EHR-based stratification analysis. Here we present an unsupervised framework based on deep learning to process heterogeneous EHRs and derive patient representations that can efficiently and effectively enable patient stratification at scale. We considered EHRs of 1,608,741 patients from a diverse hospital cohort comprising a total of 57,464 clinical concepts. We introduce a representation learning model based on word embeddings, convolutional neural networks, and autoencoders (i.e., ConvAE) to transform patient trajectories into low-dimensional latent vectors. We evaluated these representations as broadly enabling patient stratification by applying hierarchical clustering to different multi-disease and disease-specific patient cohorts. ConvAE significantly outperformed several baselines in a clustering task to identify patients with different complex conditions, with 2.61 entropy and 0.31 purity average scores. When applied to stratify patients within a certain condition, ConvAE led to various clinically relevant subtypes for different disorders, including type 2 diabetes, Parkinson's disease, and Alzheimer's disease, largely related to comorbidities, disease progression, and symptom severity. With these results, we demonstrate that ConvAE can generate patient representations that lead to clinically meaningful insights. This scalable framework can help better understand varying etiologies in heterogeneous sub-populations and unlock patterns for EHR-based research in the realm of personalized medicine.

npj Digital Medicine (2020)3:96 ; https://doi.org/10.1038/s41746-020-0301-z

INTRODUCTION

Electronic health records (EHRs) are collected as part of routine care across the vast majority of healthcare institutions. They consist of heterogeneous structured and unstructured data elements, including demographic information, diagnoses, laboratory results, medication prescriptions, free-text clinical notes, and images. EHRs provide snapshots of a patient's state of health and have created unprecedented opportunities to investigate the properties of clinical events across large populations using data-driven approaches and machine learning. At the individual level, patient trajectories can foster personalized medicine; across a population, EHRs can provide a vital resource to understand population health management and help make better decisions for healthcare operation policies[1].

Personalized medicine focuses on the use of patient-specific data to tailor treatment to an individual's unique health characteristics. However, even seemingly simple diseases can show different degrees of complexity that can create challenges for identification, treatment, and prognosis, despite equivalence at the diagnostic level[2,3]. Heterogeneity among patients is particularly evident for complex disorders, where the etiology is due to an amalgamation of multiple genetic, environmental, and lifestyle factors. Several different conditions have been referred to as complex, such as Parkinson's disease (PD)[4], multiple myeloma (MM)[5], and type 2 diabetes (T2D)[6]. Patients with complex disorders may differ on multiple systemic layers (e.g., different clinical measurements or comorbidity landscape) and in response to treatments, making these conditions difficult to model. Multiple data types in patient longitudinal EHR histories offer a way to examine disease complexity and present an opportunity to refine diseases into subtypes and tailor personalized treatments. This task is usually referred to as "EHR-based patient stratification". This follows a common approach in clinical research, where attempts to identify latent patterns within a cohort of patients can contribute to the development of improved personalized therapies[7].

From a computational perspective, patient stratification is a data-driven, unsupervised learning task that groups patients according to their clinical characteristics[8]. Previous work in this domain aggregates clinical data at a patient level, representing each patient as multi-dimensional vectors, and derives subtypes within a disease-specific population via clustering (e.g., in autism[9]) or topological analysis (e.g., for T2D[10]). Deep learning has been applied to derive more robust patient representations to improve disease subtyping[8,11]. Baytas et al.[8] used time-aware long short-term memory (LSTM) networks to leverage stratification of longitudinal data of PD patients. Similarly, Zhang et al.[11] used LSTM to identify three subgroups of patients with idiopathic PD that differ in disease progression patterns and symptom severity. These studies, however, only focused on curated and small disease-specific cohorts, with ad hoc manually selected features. This approach not only limits scalability and generalizability, but also hinders the possibility to discover unknown patterns that might characterize a condition. Because EHRs tend to be incomplete, using a diverse cohort of patients to derive disease-specific subgroups can adequately capture the features of

[1]Bruno Kessler Institute, Povo, TN, Italy. [2]Department of Psychology and Cognitive Science, University of Trento, Rovereto, TN, Italy. [3]Hasso Plattner Institute for Digital Health at Mount Sinai, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [4]Institute for Next Generation Healthcare, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [5]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [6]Department of Mental Health and Pathological Addiction, Azienda USL Centro "Santi", Parma, Italy. [7]HK3 Lab, Milan, Italy. [8]These authors jointly supervised this work: Cesare Furlanello, Riccardo Miotto. ✉email: riccardo.miotto@mssm.edu

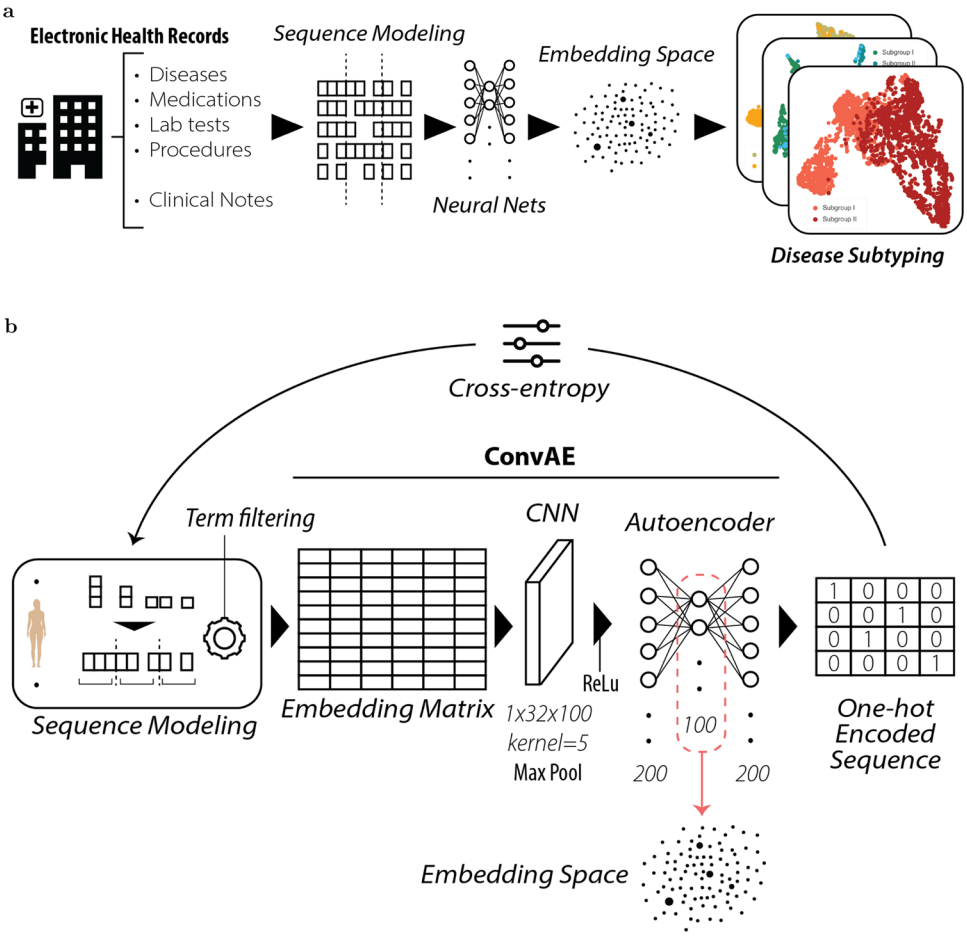

Fig. 1 Patient stratification framework and ConvAE architecture. a Framework enabling patient stratification analysis from deep unsupervised EHR representations; **b** Details of the ConvAE representation learning architecture.

heterogeneity within the disease of interest[12]. However, it is challenging to create large-scale computational models from EHRs because of data quality issues, such as high dimensionality, heterogeneity, sparseness, random errors, and systematic biases. Advances in machine learning, specifically in representation learning[13] and deep learning[14], are introducing different computational models to leverage EHRs for personalized healthcare[15,16]. This work fits into this landscape by presenting an unsupervised patient stratification pipeline that aims to automatically detect clinically meaningful subtypes within any condition by using patient representations learned from a heterogeneous and large cohort of EHRs.

In particular, this paper proposes a general framework for identifying disease subtypes at scale (see Fig. 1a). We first propose an unsupervised deep learning architecture to derive vector-based patient representations from a large and domain-free collection of EHRs. This model (i.e., ConvAE) combines (1) embeddings to contextualize medical concepts, (2) convolutional neural networks (CNNs) to loosely model the temporal aspects of patient data, and (3) autoencoders (AEs) to enable the application of an unsupervised architecture. Second, we show that ConvAE-based representations learned from real-world EHRs of ~1.6M patients from the Mount Sinai Health System in New York improve clustering of patients with different disorders compared to several commonly used baselines. Last, we demonstrate that ConvAE leads to effective patient stratification with minimal effort. To this end, we used the encodings learned from domain-free and heterogeneous EHRs to derive subtypes for different complex disorders and provide a qualitative analysis to determine their clinical relevance.

This architecture enables patient stratification at scale by eliminating the need for manual feature engineering and explicit labeling of events within patient care timelines, and processes the whole EHR sequence regardless of the length of patient history. By generating disease subgroups from large-scale EHR data, this architecture can help disentangle clinical heterogeneity and identify high-impact patterns within complex disorders, whose effect may be masked in case–control studies[17]. The specific properties of the different subgroups can then potentially inform personalized treatments and improve patient care.

RESULTS

We first evaluated the extent to which ConvAE-based patient representations can be used to identify different clinical diagnoses in the EHRs (i.e., disease phenotyping[18]). To this end, we performed clustering analysis using patients with the following eight complex disorders: T2D, MM, PD, Alzheimer's disease (AD), Crohn's disease (CD), breast cancer (BC), prostate cancer (PC), and attention deficit hyperactivity disorder (ADHD). We used SNOMED—CT (Systematized Nomenclature of Medicine—Clinical Terms)[19] to find all patients in the data warehouse diagnosed with these conditions; see Supplementary Table 2 and the "Multi-disease clustering analysis" subsection in "Methods" for more details.

Evaluation was organized as a 2-fold cross-validation experiment to show model generalizability and to assess replication of the stratification results. To this aim, we randomly split the dataset in half, obtaining two independent cohorts of ~800,000 patients that we used to train and test the models (and vice versa). While

Table 1. Multi-disease clustering performances of ConvAE configurations and baselines.

	Entropy[a]	Purity[a]	Disease number[b]
ConvAE 1-layer CNN	2.61 (0.04, [2.58, 2.67])***	0.31 (0.02, [0.31, 0.35])***	6.50 (0.62)***
ConvAE 2-layer CNN	2.75 (0.02, [2.74, 2.78])	0.26 (0.01, [0.26, 0.29])	5.93 (0.50)
ConvAE multikernel CNN	2.66 (0.03, [2.64, 2.70])	0.30 (0.02, [0.29, 0.33])	5.94 (0.47)
RawCount	2.90 (0.02, [2.88, 2.92])	0.18 (0.01, [0.18, 0.20])	4.76 (0.70)
SVD-RawCount	2.90 (0.01, [2.90, 2.92])	0.19 (0.01, [0.18, 0.20])	5.13 (0.79)
SVD-TFIDF	2.85 (0.02, [2.84, 2.87])	0.21 (0.01, [0.21, 0.23])	5.83 (0.76)
Deep Patient	2.81 (0.02, [2.80, 2.84])	0.24 (0.01, [0.23, 0.25])	5.96 (0.74)

The scores reported are averaged over a 2-fold cross-validation experiment. ConvAE 1-layer CNN significantly outperforms all other configurations and baselines on all measures. Multiple pairwise *t* tests with Bonferroni correction are used to compare performances.
CNN convolutional neural network, *SVD* singular value decomposition, *TFIDF* term frequency-inverse document frequency.
***$p < 0.001$.
[a]Mean (s.d., CI).
[b]Mean (standard deviation).

we used all patients in each cohort for training, in the test sets we retained only the patients diagnosed with one of the eight disorders under consideration, obtaining ~94,000 test patients per fold (see the "Dataset" subsection in "Methods" for more details).

Table 1 shows the results using hierarchical clustering for different ConvAE architectures (one, two, and multikernel CNN layers) and baselines in terms of entropy and purity scores averaged over the 2-fold cross-validation experiment. ConvAE performed significantly better than other models largely used in healthcare for representation learning, including Deep Patient[20], for both entropy and purity scores ($p_s < 0.001$, *t* tests comparisons with Bonferroni correction). The configuration with one CNN layer yielded the best overall performance and the learned encodings produced clusters associated with the largest number of distinct diseases (i.e., 6.50, based on purity score analysis). It is worth saying that, without a predictive theory of clustering[21,22], validation metrics frequently fail to correlate with clustering errors[23]. However, such theoretic structure is not applicable in this context because the heterogeneity of the external complex disorder classes do not provide a reliable probabilistic framework. For this reason, we used, rather than estimation error analysis, transparent external metrics, such as entropy and purity scores, which evaluate cluster composition and also account for possible subgroups of complex diseases[24].

Figure 2 visualizes the distribution of the different patient representations along with their disease cohort labels obtained using UMAP (uniform manifold approximation and projection for dimension reduction[25]). ConvAE captures hidden patterns of overlapping phenotypes while still displaying identifiable groups of patients with distinct disorders. Figure 3 shows the same patient distribution highlighting clustering labels and purity percentage scores of each cluster dominating disease. These figures refer to only one of the cross-validation splits; results for the second split are similar and are available in Supplementary Figs. 1 and 2). ConvAE (with one CNN layer) also led to better clustering, visually, than all baselines. Patients with ADHD were the most separated and detected with 80% purity by hierarchical clustering. Visible clusters with >50% purity were also identified for T2D, PC, and PD. Comparing the encoding projections (Fig. 2) to the clustering visualization (Fig. 3), we observe that patients whose disease is not correctly identified by clusters tend to not clearly separate in this low-dimensional space. As an example, AD patients were randomly scattered in the plot and did not lead to distinguishable clusters. This might be due to factors such as sex and age, intrinsic biases, or noise, but it might also reflect a shared phenotypic characterization that drives the learning process into displaying these patient EHR progressions closely together irrespective of disease labels.

We then evaluated the use of ConvAE representations for patient stratification at scale and the identification of clinically relevant disease subtypes. We considered six diseases: T2D, PD, AD, MM, PC, and BC. These are all age-related complex disorders with late onset (i.e., averaged increased prevalence after 60 years of age)[26–31]. We decided to focus on these conditions to avoid, to some extent, the confounding effect of age that could affect learning and the evaluation of different subtypes. Figure 4 shows results running hierarchical clustering on the ConvAE-based patient representations of each different disease cohort. To determine the optimal number of clusters, we empirically selected the smallest number of clusters that minimizes the increase in explained variance (i.e., Elbow method). We were able to identify different subtypes for each disease with no additional feature selection and using representations derived from a domain-free cohort of patients. Supplementary Table 3 reports the number of patients in each cohort and the number of subgroups identified. Similar results were obtained for the second split and are reported in Supplementary Fig. 3.

In the following sections, we present the clinical characterization of T2D, PD, and AD subgroups via enrichment analysis of medical concept occurrences (see Supplementary Material for the characterization of the other conditions). We compare T2D and PD results to related studies based on ad hoc cohorts[10,11]. Conversely, there are no published EHR-based stratification studies for AD, MM, PC, and BC to use for comparison. All subtypes were reviewed by a clinical expert to highlight meaningful descriptors and we used multiple pairwise chi-squared tests to assess group differences. For each disease, we list sex and age statistics of the cohort (between group comparisons are performed via multiple pairwise chi-squared tests and *t* tests), as well as the five most frequent diagnosis, medications, laboratory tests, and procedures, ordered according to in-group and total frequencies, in Supplementary Tables 4–9. The results for the second split are reported in Supplementary Tables 10–15.

Type 2 diabetes

Patients with T2D clustered into three different subgroups that relate to different stages of progression for the disease (see Fig. 4a and Supplementary Table 4 for details).

Subgroup I included 18,325 patients and represents the mild symptom severity cohort, characterized by common T2D symptoms (e.g., metabolic syndrome), which were treated with Metformin, an oral hypoglycemic medication. Moreover, it also included patients exposed to lifestyle risk factors, such as obesity[6].

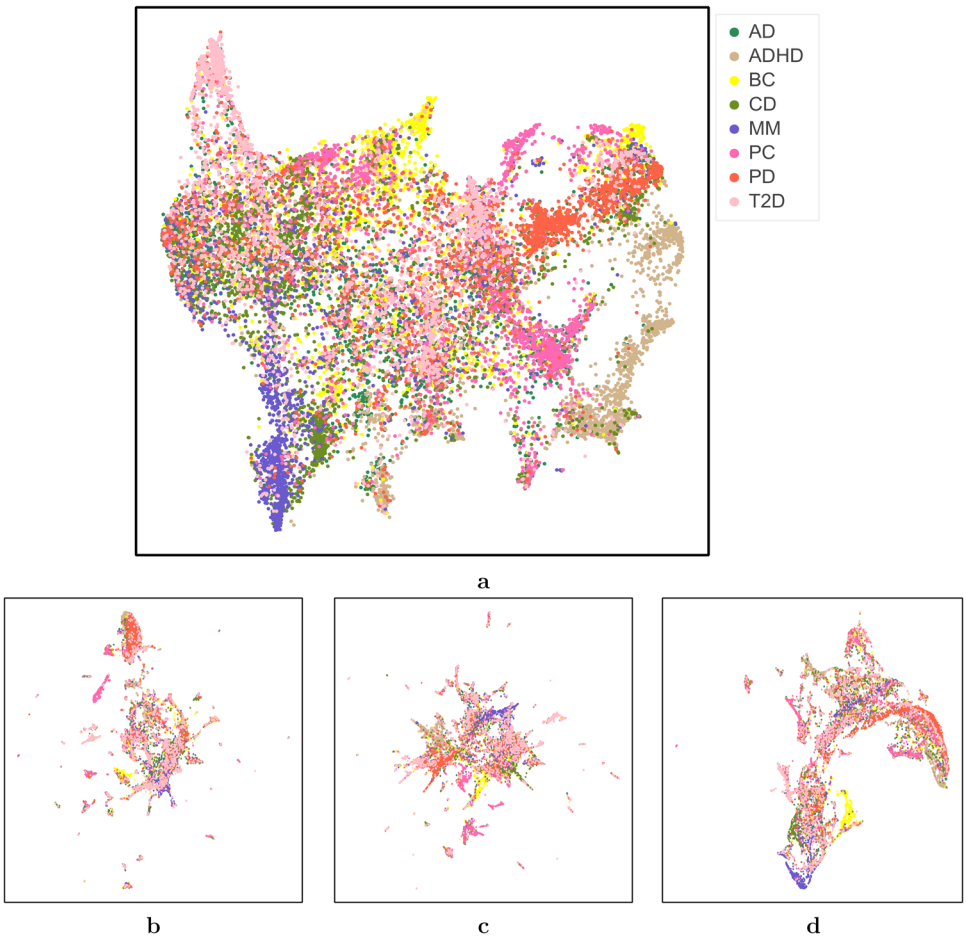

Fig. 2 Uniform manifold approximation and projection (UMAP) encoding visualization. a ConvAE 1-layer CNN; **b** SVD-RawCount; **c** SVD-TFIDF; **d** Deep Patient. AD Alzheimer's disease, ADHD attention deficit hyperactivity disorder, BC breast cancer, CD Crohn's disease, MM multiple myeloma, PC prostate cancer, PD Parkinson's disease, T2D type 2 diabetes.

Subgroups II/III, which were composed by 22,659 and 7704 patients, respectively, showed concomitant conditions associated to T2D progression and worsening symptoms. Specifically, subgroup II clustered patients characterized by microvascular problems, such as diabetic nephropathy, neuropathy, and/or peripheral artery disease. The significant presence of creatinine and urea nitrogen laboratory tests, which estimate renal function, suggests monitoring of kidney diseases, which are often related to T2D[32]. The presence of pain in limb, combined with analgesic drugs (i.e., paracetamol, oxycodone), indicates the presence of vascular lesions at the peripheral level, manifested as ischemic rest pain or ulceration. This was confirmed by peripheral vascular disease diagnoses, which accounts for 50% of terms in the T2D cohort.

Subgroup III showed severe cardiovascular problems, identified by a significant presence of medical concepts related to coronary artery diseases, for example, coronary atherosclerosis, angina pectoris, which are serious risk factors for heart failure. These subjects were often treated with antiplatelet therapy (i.e., acetylsalicylic acid, clopidrogel) to prevent cardiovascular events (e.g., stroke) and were likely to receive invasive procedures to treat severe arteriopathy. For instance, 30% of patients in subgroup III underwent percutaneous transluminal coronary angioplasty, a procedure to open up blocked coronary arteries.

Our results confirm, in part, what was observed by Li et al.[10], which used topology analysis on an ad hoc cohort of T2D patients and identified three distinct subgroups characterized by (1) microvascular diabetic complications (i.e., diabetic nephropathy, diabetic retinopathy); (2) cancer of bronchus and lungs; and (3) cardiovascular diseases and psychiatric disorders. In particular, we detected the same microvascular and cardiovascular disease groups, which are consequences of T2D. In contrast, we were unable to detect a subgroup significantly characterized by cancer, an epiphenomenon that can be caused by secondary immunodeficiency in patients with T2D[33,34]. See Supplementary Material for further description and a clustering comparison via Fowlkes–Mallows index.

Parkinson's disease

Individuals diagnosed with PD divided into two groups (Fig. 4b and Supplementary Table 5): one dominated by motor symptoms (1368 patients) and another (1684 patients) characterized by non-motor/independent features and longer course of disease.

Subgroup I is characterized as a tremor-dominant cohort (i.e., manifested by motor symptoms) because of the significant presence of diagnosis such as essential tremor, anxiety state, and dystonia. It is interesting to note that motor clinical features likely led to a common misdiagnosis of essential tremor, which is an action tremor that typically involves the hands. Parkinsonian tremor, on the contrary, although it can be present during postural maneuvers and action, is much more severe at rest and decreases with purposeful activities. However, when the tremor is severe, it is difficult to distinguish action tremor from resting tremor, leading to the aforementioned misdiagnosis[35]. Moreover, anxiety states, emotional excitement, and stressful situations can exacerbate the tremor, and lead to a delayed PD diagnosis. Brain

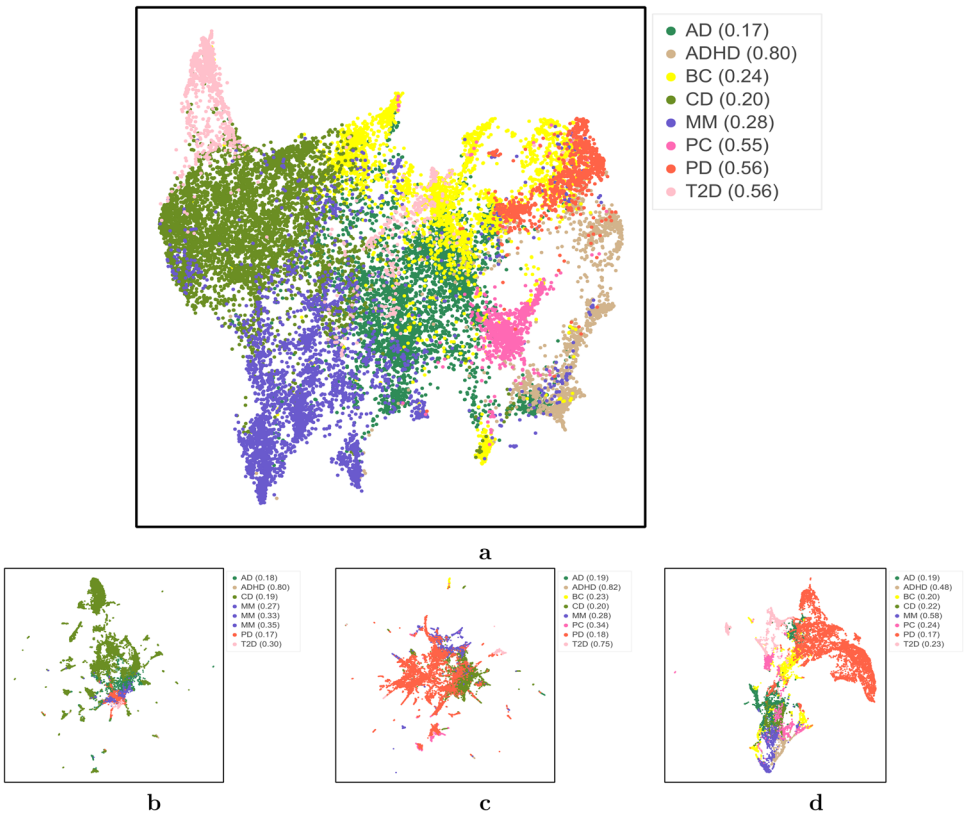

Fig. 3 Uniform manifold approximation and projection (UMAP) clustering visualization. a ConvAE 1-layer CNN; **b** SVD-RawCount; **c** SVD-TFIDF; **d** Deep Patient. AD Alzheimer's disease, ADHD attention deficit hyperactivity disorder, BC breast cancer, CD Crohn's disease, MM multiple myeloma, PC prostate cancer, PD Parkinson's disease, T2D type 2 diabetes.

magnetic resonance imaging (MRI), usually nondiagnostic in PD, was ordered for several patients in this subgroup (13%), suggesting its use for differential diagnosis, for example, to investigate the presence of chronic/vascular encephalopathy.

Subgroup II included non-motor and independent symptoms, such as constipation and fatigue. Patients in subgroup II were significantly diagnosed with coronary artery disease that is prevalent in older patients (>50 years old). Constipation and fatigue are among the most common non-motor problems related to autonomic dysfunction, diminished activity level, and slowed intestinal transit time in PD[36,37].

In their study about PD stratification with PPMI (Parkinson's progression markers initiative) data, Zhang et al.[11] identified three distinct subgroups of patients based on severity of both motor and non-motor symptoms. In particular, one subgroup included patients with moderate functional decay in motor ability and stable cognitive ability; a second subgroup presented with mild functional decay in both motor and non-motor symptoms; and the third subgroup was characterized by rapid progression of both motor and non-motor symptoms. EHRs do not quantitatively capture PD symptom severity; therefore, our analyses cannot replicate these findings. However, unlike Zhang et al.[11], we can discriminate between specific motor and non-motor symptoms and also suggest a longer, but not necessarily more severe, disease course for the non-motor symptom subgroup.

Alzheimer's disease

Patients with AD separated into three subgroups marked by AD onset, disease progression, and severity of cognitive impairment (see Fig. 4c and Supplementary Table 6).

Subgroup I is characterized by 399 patients with early-onset AD, that is, patients whose dementia symptoms have typically developed between the age of 30 and 60 years, and initial neurocognitive disorder. Early-onset AD affects 5% of the individuals with AD in the United States[38], and, because clinicians do not usually look for AD in younger patients, the diagnostic process includes extensive evaluations of patient symptoms. In particular, given that a certain AD diagnosis can only be provided postmortem through brain examination, clinicians first rule out other causes that can lead to early-onset dementia (i.e., differential diagnosis). We find evidence of this practice in this subgroup, which includes postmenopausal women, identifiable by mean age >50, osteoporosis diagnosis with calcium supplement therapy, and menopausal hormone treatment (i.e., estradiol). Patients in this group are also tested for infectious diseases (e.g., HIV, syphilis, hepatitis C, chlamydia/gonorrhea) that are possible causes of early-onset dementia[39], and screened via structural neuroimaging, for example, MRI/positron emission tomography brain. As cognitive dysfunctions that may be mistaken for dementia can also be caused by depression and other psychiatric conditions, the presence of psychiatric service/procedure suggests psychiatric evaluations to exclude depressive pseudodementia. After the differential diagnosis process and the exclusion of other possible causes, eventually these patients received a diagnosis of AD.

Subgroup II includes 1170 patients with late-onset AD, mild neuropsychiatric symptoms, and cerebrovascular disease. Here, the absence of behavioral disturbances in 39% of patients and their high average age ($M = 84.96$, s.d. $= 9.61$) suggest a late AD onset, with a progression characterized by a slower rate of cognitive ability decline[40]. Moreover, the presence of acetylsalicylic acid, an antiplatelet medication, and intracranial hemorrhage diagnosis indicates the co-occurrence of cerebrovascular

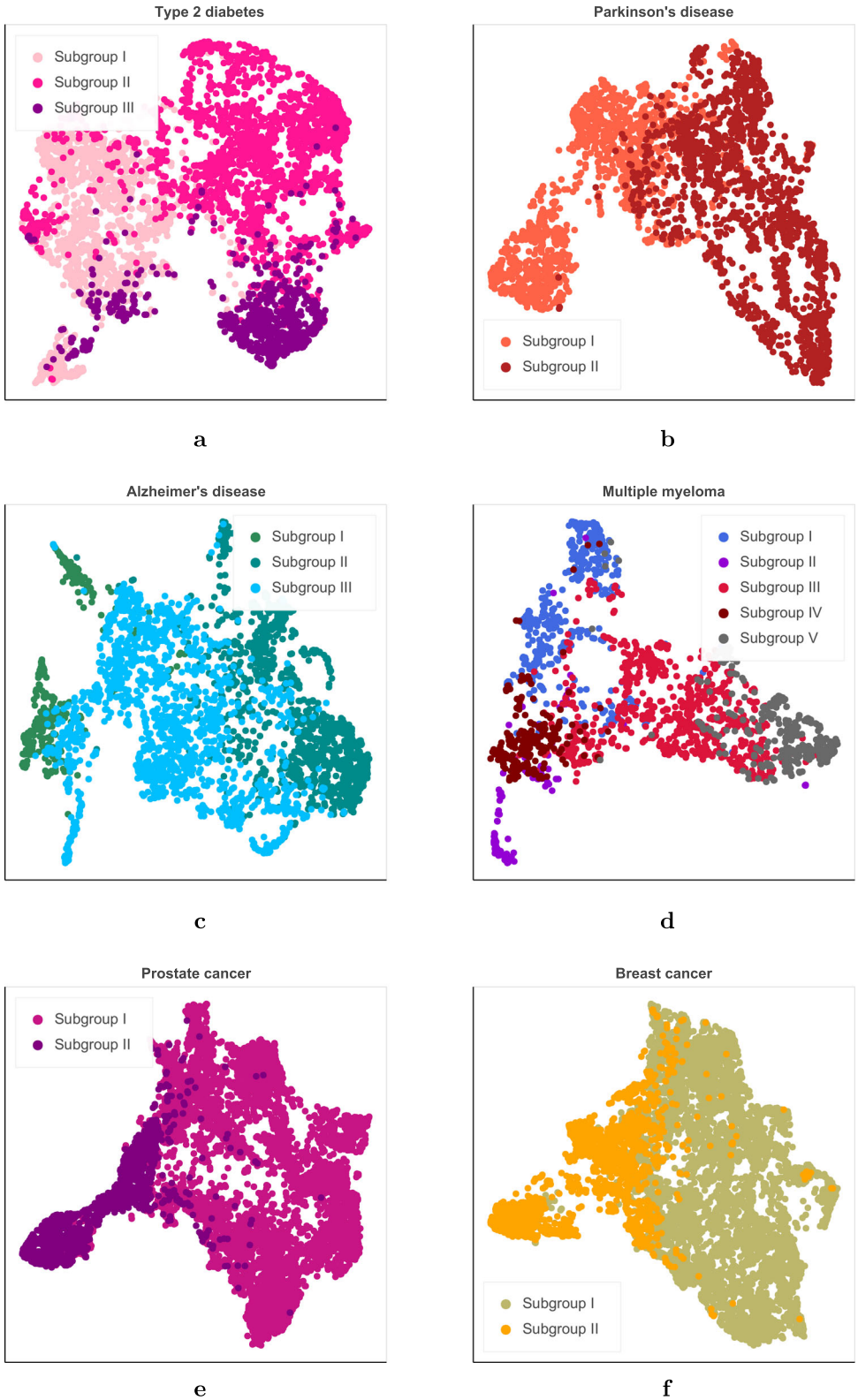

Fig. 4 Complex disorder subgroups. A subsample of 5000 patients with T2D is displayed in **a**. **b**–**f** display patient subtypes for Parkinson's and Alzheimer's disease, multiple myeloma, prostate and breast cancer cohorts, respectively.

disease, which affects blood vessels and blood supply to the brain. Cerebrovascular diseases are common in aging, and can often be associated with AD[41]. In this regard, head computed tomography may have been performed to prevent or identify structural abnormalities related to cerebrovascular disease.

Subgroup III is characterized by 1632 individuals with typical onset and mild-to-moderate dementia symptoms. A cohort of 409 patients was treated with donepezil, a cholinesterase inhibitor, that is a primary treatment for cognitive symptoms and it is usually administered to patients with mild-to-moderate AD,

producing small improvement in cognition, neuropsychiatric symptoms, and activities of daily living[42]. Patients in this subgroup also showed both dementia with and without behavioral disturbances.

DISCUSSION

This study proposes a computational framework to disentangle the heterogeneity of complex disorders in large-scale EHRs through the identification of data-driven clinical patterns with machine learning. Specifically, we developed and validated an unsupervised architecture based on deep learning (i.e., ConvAE) to infer informative vector-based representations of millions of patients from a large and diverse hospital setting, which facilitates the identification of disease subgroups that can be leveraged to personalize medicine. These representations aim to be domain free (i.e., not related to any specific task since learned over a large multi-domain dataset) and enable patient stratification at scale. Results from our experiments show that ConvAE significantly outperformed several baselines on clustering patients with different complex conditions and led to the identification of different clinically meaningfully disease subtypes.

Results identified disease progression, symptom severity, and comorbidities as contributing the most to the EHR-based clinical phenotypic variability of complex disorders. In particular, T2D patients divided into three subgroups according to comorbidities (i.e., cardiovascular and microvascular problems) and symptom severity (i.e., newly diagnosed with milder symptoms). Individuals with PD showed different disease duration and symptoms (i.e., motor, non-motor). AD profiles distinguished early- and late-onset groups and separate patients with mild neuropsychiatric symptoms and cerebrovascular disease from patients with mild-to-moderate dementia. Patients with MM were characterized by different comorbidities (e.g., amyloidosis, pulmonary diseases) that manifest alongside precise typical signs of MM. Patients with PC and BC separated according to disease progression. These findings showed that the features learned by ConvAE describe patients in a way that is general and conducive to identifying meaningful insights into different clinical domains. In particular, this work aims to contribute to the next generation of clinical systems that can (1) scale to include many millions of patient records and (2) use a single, distributed patient representation to effectively support clinicians in their daily activities, rather than multiple systems working with different patient representations derived for different tasks[20].

To this aim, enabling efficient data-driven patient stratification analyses to identify disease subgroups is an important aspect to unlock personalized healthcare. Ideally, when new patients enter the medical system, their health status progression can be tied to a specific subgroup, thereby informing the treating clinician of personalized prognosis and possible effective treatment strategies, or counseling in cases where a certain diagnosis is difficult and a more thorough examination is required (e.g, specific genetic or lab tests). Moreover, the clinical characteristics of the different subtypes can potentially lead to intuitions for novel discoveries, such as comorbidities, side effects, or repositioned drugs, which can be further investigated analyzing the patient clinical trajectories.

Previous studies mostly focused on a specific disease using ad hoc cohorts of patients and features[8–11,43,44]. While these studies obtained relevant clinically meaningful results, the computational framework is hard to replicate for different diseases and it is tied to the specific study and to the specific data. Deep learning has extensively been used to model EHRs for medical analysis[15,16], including clinical prediction, such as disease onset, mortality, and readmission[45–47], and disease phenotyping[20,48]. Because deep learning methods have not yet been leveraged for disease subtyping at scale, ConvAE aims to fill this gap and to provide an architecture that can improve unsupervised EHR preprocessing to favor patient stratification and unveil clinically meaningful and actionable insights. Additionally, unlike previous representation learning methods, which did not consider the temporality of EHRs[20,48], ConvAE uses CNNs in combination with embeddings to specifically capture some of the longitudinal aspects of patient clinical status, leading to more robust representations. CNNs were already used to model EHRs for specific predictive analysis, as part of supervised architectures[49,50]. Differently, we trained CNNs in an unsupervised framework based on AEs to learn general-purpose patient representations. While these representations were used to leverage disease subtype discovery, they can also be fine-tuned and applied to specific supervised tasks, such as disease phenotyping and prediction.

There are several limitations to our study. First, we acknowledge that the lack of any discernible pattern in the multi-disease clustering analysis can also be due to noise and biases in the data, which might affect both learned representations and clustering. In particular, processing EHRs with minimum data engineering, on the one hand, preserves all the available information and, to some extent, prevents systematic biases. On the other hand, it adds hospital-specific biases intrinsic to the EHR structure and noise due to data being redundant and too generic. Improving EHR preprocessing by, for example, better modeling clinical notes and/or improving feature filtering, should help reduce noise and improve performances. Second, we identified patients related to complex disorders using SNOMED—CT codes and this likely led to the inclusion of many false positives that affected the learning algorithms[51]. The use of phenotyping algorithms based on manual rules (e.g., PheKB[52]), or semiautomated approaches (e.g., refs. [53,54]), should help identify better cohorts of patients and, consequently, better disease subtypes. Another limitation comes from the choice, among all possibilities, of the specific complex disorders. This allowed us to test the approach on heterogeneous conditions that affect different biological mechanisms, showing the efficacy of the proposed framework in generalizing to various clinical domains. Nevertheless, the approach should be further evaluated with other typologies of conditions as well, such as multiple sclerosis, autoimmune diseases, and psychiatric disorders. Lastly, we identified relevant concepts in the patient subgroups by simply evaluating their frequency. Adding a semantic modeling component based on, for example, topic modeling[55] or word embeddings[56], might lead to more clinically meaningful patterns.

Future works will attempt to address these limitations and to further improve and replicate the architecture. First, we plan to enable multilevel clustering in order to stratify patients within the subtypes. This should lead to more granular patient stratification, and thus to patterns on a more individual level. Second, we plan to verify ConvAE generalizability by replicating the study on EHRs from different healthcare institutions. Third, we will evaluate the use of disease subtypes as labels for training supervised models that can predict stratified patient risk scores. This, besides further validating the relevance of the results, will also provide an initial and intuitive framework to apply the results of patient stratification to clinical practice. To this aim, we plan to first assess treatment safety and efficacy between subtypes of a specific disease. Finally, to develop more comprehensive disease characterizations, we will include other modalities of data, for example, genetics, into this framework, which will hopefully refine clustering and reveal new etiologies. Multimodal stratified disease cohorts promise to facilitate better predictive capabilities for future outcomes by modeling how molecular mechanisms interact with clinical states.

METHODS

The framework to derive patient representations that enable stratification analysis at scale is based on three steps: (1) data preprocessing; (2)

unsupervised representation learning (i.e., ConvAE); and (3) clustering analysis of disease-specific cohorts (see Fig. 1a). In this section, we report details of this framework as well as the description of the evaluation design.

Dataset

We used de-identified EHRs from the Mount Sinai Health System data warehouse; the study was approved by IRB-19-02369 in accordance with HIPAA guidelines. Mount Sinai Health System is a large and diverse urban hospital located in New York, NY, which generates a high volume of structured, semi-structured, and unstructured data from inpatient, outpatient, and emergency room visits. Patients in the system can have up to 12 years of follow-up data unless they are transferred or move their residence away from the hospital system. We accessed a de-identified dataset containing ~4.2 million patients, spanning the years from 1980 to 2016.

For each patient, we aggregated general demographic details (i.e., age, sex, and race) and clinical descriptors. We included ICD-9 diagnosis codes, medications normalized to RxNorm, CPT-4 procedure codes, vital signs, and lab tests normalized to LOINC. ICD-10 codes were mapped back to the corresponding ICD-9 versions. We preprocessed clinical notes using a tool based on the Open Biomedical Annotator that extracts clinical concepts from the free text[57,58]. The vocabulary V was composed of 57,464 clinical concepts.

We retained all patients with at least two concepts, resulting in a collection of 1,608,741 different patients, with an average of 88.9 records per patient. In particular, the cohort included 900,932 females, 691,321 males, and 16,488 not declared; the mean age of the population as of 2016 was 48.29 years (s.d. = 23.79). Patients were randomly partitioned into half for 2-fold cross-validation to assess model generalizability and replicability of the results. In each train set, we retained 30,000 random patients for tuning the model hyperparameters. Train and test preprocessed sets' details are reported in Supplementary Table 1.

Data preprocessing

Every patient in the dataset is represented as a longitudinal sequence s_p of length M of aggregated temporally ordered medical concepts, that is, $s_p = (w_1, w_2, \ldots, w_M)$, where each w_i is a medical concept from the vocabulary V. Preprocessing includes: (1) filtering the least and most frequent concepts; (2) dropping redundant concepts within fixed time frames; (3) splitting long sequences of records to include the complete patient history while leveraging the CNN framework, which requires fixed-size inputs.

We consider all the EHRs as a document D and each patient sequence s_p as a sentence. For each concept w in V, we first compute the probability of having w in D. We then multiply this by the sum of the probabilities to find w in a sentence s_p for all sentences. In particular, let P be the set of all patients, $\forall\ w \in V$, the filtering score is defined as:

$$\mathbb{P}(w \in D)\sum_{p\in P}\mathbb{P}(w \in s_p) = \frac{\#\{s \in D;\ w \in s\}}{|D|}\sum_{p\in P}\frac{\#\{w_i \in s_p;\ w_i = w\}}{|s_p|}, \tag{1}$$

where $|D|$ is the total number of sentences and $|s_p|$ is the length of a patient sequence. The filtering score combines document frequency, that is, the number of patients with at least one occurrence of w, and term frequency, that is, total number of occurrences of w in a patient sequence. We then drop all concepts with filtering scores outside certain cut-off values to reduce the amount of noise (i.e., not informative concepts that occur multiple times in few patients, or too general concepts that occur in many patients).

A patient may have multiple encounters in their health records that span consecutive days and might include repeated concepts that are often artifacts of the EHR system, rather than new clinical entries. To reduce this bias, we drop all duplicate medical concepts from the patient records within overlapping time intervals of T days. Within the same time window, we also randomly shuffle the medical concepts, given that events within the same encounter are generally randomly recorded[54,59]. Lastly, we eliminate all patients with <3 concepts in their records.

Patient sequences are then chopped into subsequences of fixed length L that are used to train the ConvAE model. Each patient sequence is thus defined as:

$$s_p = [(w_1, \ldots, w_L), (w_{L+1}, \ldots, w_{2L}), \ldots],$$

and subsequences shorter than L are padded with 0 up to length L. For the sake of clarity, in the following section we present the architecture as applied to a general subsequence $s = (w_1, \ldots, w_L)$.

The ConvAE architecture

ConvAE is a representation learning model that transforms patient EHR subsequences into low-dimensional, dense vectors. The architecture consists of three stacked modules (see Fig. 1b). This study proposes to use in combination embedding, CNNs, and AEs to process EHRs and to derive unsupervised vector-based patient representations that can be used for clinical inference and medical analysis.

Given s, the architecture first assigns each medical concept w to an N-dimensional embedding vector v_w to capture the semantic relationships between medical concepts. Specifically, a patient subsequence is represented as an $(L \times N)$ matrix $E = (v_{w_1}, v_{w_2}, \ldots, v_{w_L})^T$, where L is the subsequence length, and N is the embedding dimension. This structure also retains temporal information because the rows of matrix E are temporally ordered according to patient visits.

The architecture is then composed by CNNs, which extract local temporal patterns, and AEs, which learn the embedded representations for each patient subsequence. The CNN applies temporal filters to each embedding matrix. CNN filters applied to EHRs usually perform a one-side convolution operation across time via filter sliding. A filter can be defined as $k \in \mathbb{R}^{h\times N}$, where h is the variable window size and N is the embedding dimension[60,61]. Our approach differs in that it processes embedding matrices as they were RGB images carrying a third "depth" dimension. With this approach, we enable the model filters to learn independent weights for each encoding dimension, thus activating for the most salient features in each dimension of the embedding space. Therefore, we reshape the $(L \times N)$ embedding matrix into $\tilde{E} \in \mathbb{R}^{1\times L\times N}$ and we consider the embedding dimensions as channels. We then apply f filters $\mathbf{k} \in \mathbb{R}^{1\times h\times N}$ to the padded input to keep the same output dimension and learn features that may grasp sequence characteristics. In particular, for each filter j, we obtain:

$$(R)_j = \mathrm{ReLU}(\sum_{i=0}^{N-1}\mathbf{k}_i \star \tilde{\mathbf{e}}_i + \mathbf{b}_j),\ j = 1, \ldots, f, \tag{2}$$

where $R \in \mathbb{R}^{1\times L\times f}$ is the output matrix; $\mathbf{k}_i$ is the h-dimensional weight matrix at depth i; $\tilde{\mathbf{e}}_i \in \mathbb{R}^{1\times L}$ is the ith embedding dimension of the input matrix; $\mathbf{b}$ is the bias vector; and $(\star)$ is the convolution function. We used rectified linear unit (ReLU) as the activation function and max pooling. The output is then reshaped into a concatenated vector of dimension $L \cdot f$. This configuration learns different weights for each embedding dimension to highlight relevant interdependencies of medical concepts, and tune representations of patient histories to identify the most relevant characteristics of their semantic space.

We then use fully dense layers of AEs to derive embedded patient representations that estimate the given input subsequences. Specifically, we extract the hidden representation $\mathbf{y}$, a H-dimensional vector, as the encoded representation of each patient subsequence. Each patient sequence s_p is then transformed into a sequence of encodings s_h that can be post-modeled to obtain a unique vector-based patient representation. Here we simply component-wise average all the subsequence representations.

To train ConvAE, we set up a multi-class classification task that reconstructs each initial input one-hot subsequence of medical terms, from their encoded representations. Given a subsequence of medical concepts s, the ConvAE is trained by minimizing the cross entropy (CE) loss:

$$\mathrm{CE}(\mathrm{Softmax}(O),\ s) = -\frac{1}{L}\sum_{j=1}^{L}\log\left(\mathrm{Softmax}(O^j)_{w_j}\right),$$

where O is the output of ConvAE reshaped into a matrix of dimension $|V| \times L$, w_j is the jth element of sequence s that corresponds to a term indexed in V and

$$\mathrm{Softmax}(O^j)_i = \frac{\exp O_i^j}{\sum_{i=1}^{|V|}\exp O_i^j},\quad i = 1, \ldots, |V|. \tag{3}$$

Since the objective function consists of only self-reconstruction errors, the model can be trained without any supervised training samples.

Clustering analysis for patient stratification

ConvAE-based representations can be used to stratify patients from any preselected cohort without needing additional feature engineering or manual adjustments. To this aim, patients with a specific disease are

selected using, for example, ICD codes, SNOMED—CT diagnosis, or phenotyping algorithms (e.g., refs. [51,53,54]), and clustering is applied to the corresponding representations to identify disease subgroups. Here, specifically, we use SNOMED—CT diagnosis to preselect the disease cohorts and hierarchical clustering with Ward's method and Euclidean distance to derive disease subgroups. We identify the number of subclusters that best disentangles heterogeneity on the disease dataset using the Elbow Method, which empirically selects the smallest number of clusters that minimizes the increase in explained variance.

A systematic analysis of the patients in each subgroup can then automatically identify the medical concepts that significantly and uniquely define each disease subtype. In this work, we rank all the codes by their frequency in the patient sequences. In particular, we compute the percentages of patients whose sequence includes a specific concept both with respect to a subcluster (i.e., in-group frequency) and to the complete disease cohort (i.e., total frequency). Ranking maximizes, first, the in-group percentage, and second, the total percentage. We then analyze the most frequent concepts and we use a pairwise chi-squared test to determine whether the distributions of present/absent concepts with respect to the detected subgroups are significantly different[11].

Implementation details

All model hyperparameters were empirically tuned to minimize the network reconstruction error, while balancing training efficiency and computation time. We tested a large amount of configurations (e.g., time interval T equal to {15, 30}; patient subsequence length L equal to {32, 64}; embedding dimension N spanning {100, 128, 200}). For brevity, we report only the final setting used in the patient stratification experiments. All modules were implemented in Python 3.5.2, using `scikit-learn` and `pytorch` as machine learning libraries[62,63]. Computations were run on a server with an Nvidia Titan V GPU.

We used Eq. (1) to discard terms with a filtering score $<10^{-6}$, that is, document frequency ranging from 1 to 10. Examples of discarded concepts are clotrimazole, an antifungal medication, and torsemide, a medication to reduce extra fluid in the body. We decided to retain all the very frequent concepts as most of them seemed clinically informative (e.g., vital signs). Patients with <3 medical concepts were then discarded. In total, 24,665 medical terms were filtered out, decreasing the vocabulary size to 32,799.

We divided each patient history in consecutive, half-overlapped temporal windows of $T = 15$ days, shuffled unique medical concepts, and dropped redundant terms. Patient sequences were then split in subsequences of length $L = 32$ concepts, obtaining ~3M subsequences of medical concepts for training. This value was chosen to enable efficient training of the autoencoder with GPUs.

We initialized medical concept embeddings using word2vec with the skip-gram model[56]. We considered all the subsequences in the training set as sentences and medical concepts as words[54,59]. We obtained 100-dimensional embeddings for 31,659 medical concepts of the vocabulary. The remaining concepts were initialized randomly; the subsequence padding was initialized as the null vector (i.e., at $\mathbf{0}$). These embedding vectors were then used as input for the ConvAE module and were further refined during the model training.

The CNN module used 50 filters with kernel size = 5 and ReLU activation function. The autoencoder was composed by four hidden layers with 200, 100, 200, and $|V| \times 32$ hidden nodes, respectively, where $|V|$ is the vocabulary size. We used ReLU activation in the first three layers and Softplus activation in the final layer to obtain continuous output. We applied dropout with $p = 0.5$ in the first two layers for regularization. The model was trained using CE loss with the Adam optimizer (learning rate = 10^{-5} and weight decay = 10^{-5})[64] for 5 epochs on all training data and batch size of 128. The size of the patient representations was equal to 100.

We evaluated different CNN configurations composed by 1-layer (i.e., "ConvAE 1-layer CNN"), 2-layers (i.e., "ConvAE 2-layer CNN"), and one multikernel layer (i.e., "ConvAE multikernel CNN"). All hyperparameters were the same, except the number of filters in the second CNN of the 2-layer configuration that was set to 25. Multikernel CNN performs parallel training of distinct CNNs with different kernel sizes, and concatenates the final outputs. We used kernel dimensions equal to 3, 5, and 7.

Baselines

We compared ConvAE with the following representation learning algorithms: "RawCount", "singular value decomposition (SVD)-RawCount", "SVD-TFIDF (term frequency-inverse document frequency)", and "Deep Patient". All baseline derived vector-based patient encodings of size 100.

RawCount is a sparse representation where each patient is encoded into a count vector that has the length of the vocabulary. More specifically, each individual health history s_p is represented as an integer vector $\mathbf{x} \in \mathbb{Z}^{|V|}$, where each element is the frequency of the corresponding clinical concept in the patient longitudinal history, that is, $x_i = \#\{w_i; w_i \in s_p\}$.

SVD-RawCount applies truncated SVD to the RawCount matrix to compute the largest singular values of the raw count encodings, which define the dense, lower-dimensional representations.

SVD-TFIDF transforms the raw count encodings using the TFIDF weighting schema and applies truncated SVD to the resulting matrix. We considered the patient EHR sequences as documents, the entire dataset as corpus and we derived TFIDF scores for all medical concepts. Each patient is then represented as a vector of length $|V|$, with the corresponding TFIDF weight for each concept, and the matrix obtained is reduced via truncated SVD.

Deep Patient transforms the raw count matrix using a stack of denoizing AEs as proposed by Miotto et al.[20]. We used the implementation details presented in the paper, with batch size = 32, corruption noise = 5%, and 5 training epochs.

Multi-disease clustering analysis

We evaluated all the representation learning approaches in a clustering task to determine how they were able to disentangle patients with different conditions. We chose eight complex disorders: T2D, MM, PD, AD, CD, PC, BC, and ADHD. We retrieved all the corresponding patients in the test sets using SNOMED—CT codes after verifying that at least one correspondent ICD-9 code was present in a patient EHRs. In particular, we looked for: "type 2 diabetes mellitus" (250.00) for T2D; "multiple myeloma without mention of having achieved remission" (203.00) for MM; "paralysis agitans" (332.0) for PD; "Alzheimer's disease" (331.0) for AD; "regional enteritis of unspecified site" (555.9) for CD; "malignant neoplasm of prostate" (185) for PC; "malignant neoplasm of female breast" (174.9) for BC; and "attention deficit disorder with hyperactivity" (314.01) for ADHD. We discarded all patients with comorbidities within the selected diseases to facilitate the clustering interpretation. We then performed hierarchical clustering with $k = 8$ clusters (i.e., same as the different diseases) for all the representations to evaluate if patients with the same condition were grouping together. The final test sets were composed by ~94,000 patients per fold but were unbalanced, with disease cohorts ranging from ~1900 to 50,000 patients (see Supplementary Table 2). To use balanced datasets and improve the efficacy of the experiment, we sub-sampled 5000 random patients for the highly populated diseases, and we iterated this subsampling process 100 times, obtaining 100 different clustering per test set.

We used entropy and purity scores averaged across the 100 experiments of each fold to measure to what extent the clusters matched the different diseases. In particular, for each cluster j, we define the probability that a patient in j has disease i as:

$$p_{ij} = \frac{m_{ij}}{m_j}, \tag{4}$$

where m_j is the number of patients in cluster j and m_{ij} is the number of patients in cluster j with a diagnosis of disease i. Entropy for each cluster is defined as:

$$E_j = -\sum_i p_{ij} \log_2 p_{ij}, \tag{5}$$

and conditional entropy H(disease|cluster) is then computed as:

$$H(\text{disease}|\text{cluster}) = \sum_j \frac{m_j}{m} E_j,$$

where m is the total number of elements in the complex disease dataset.

Purity identifies the most represented disease in each cluster. For a cluster j, purity P_j is defined as $P_j = \max_i p_{ij}$, where p_{ij} is computed as before. The overall purity score is then the weighted average of P_j for each cluster j. The perfect clustering obtains averaged entropy and purity scores = 0 and 1, respectively.

Disease subtyping analysis

We evaluated the usability of ConvAE representations to discover disease subtypes for different and diverse conditions (i.e., patient stratification at

scale). In particular, we selected a cohort of patients with T2D, PD, AD, MM, PC, and BC and ran hierarchical clustering on the ConvAE-based patient representations. These are all age-related complex disorders with late onset (i.e., increased prevalence after 60 years of age[26–31]). We focused only on these conditions to attempt reducing confounding age effects that could affect the analysis of the subtypes (as it could happen on CD and ADHD cohorts, where a common onset age is less defined). To reduce noise in the sequence encodings, we averaged all patient subsequence representations fr"om the first diagnosis forward, and we dropped sequences shorter than three concepts. We ranged the number of clusters from 2 to 15 and we used the Elbow Method to empirically select the smallest number of clusters that minimize the increase in explained variance. We then performed a qualitative analysis of each subtype, similarly to Zhang et al.[11], to identify which medical concepts characterized the specific group of patients. We further verified the various subgroups in the medical literature and with the support of a practicing clinician.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

DATA AVAILABILITY

The data used for this study are available from the Mount Sinai Health System (NYC), but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are, however, available from the authors upon reasonable request and with the permission of Mount Sinai Health System.

CODE AVAILABILITY

Code is available at: https://github.com/landiisotta/convae_architecture.

Received: 9 April 2020; Accepted: 17 June 2020;
Published online: 17 July 2020

REFERENCES

1. Jensen, P. B., Jensen, L. J. & Brunak, S. Mining electronic health records: towards better research applications and clinical care. *Nat. Rev. Genet.* **13**, 395 (2012).
2. Cutting, G. R. Cystic fibrosis genetics: from molecular understanding to clinical application. *Nat. Rev. Genet.* **16**, 45–56 (2014).
3. Alexandrov, V. et al. Large-scale phenome analysis defines a behavioral signature for Huntington's disease genotype in mice. *Nat. Biotechnol.* **34**, 838–44 (2016).
4. Langston, J. W. The Parkinson's complex: Parkinsonism is just the tip of the iceberg. *Ann. Neurol.* **59**, 591–596 (2006).
5. de Mel, S., Lim, S. H., Tung, M. L. & Chng, W. J. Implications of heterogeneity in multiple myeloma. *BioMed Res. Int.* 1–12, https://doi.org/10.1155/2014/232546 (2014).
6. Pearson, E. R. Type 2 diabetes: a multifaceted disease. *Diabetologia* **62**, 1107–1112 (2019).
7. Dugger, S. A., Platt, A. & Goldstein, D. B. Drug development in the era of precision medicine. *Nat. Rev. Drug Discov.* **17**, 183–196 (2017).
8. Baytas, I. M. et al. Patient subtyping via time-aware LSTM Networks. In *Proc. 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (eds Matwin S, S., Yu, S. & Farooq, F.) 65–74 (ACM, New York, 2017).
9. Doshi-Velez, F., Ge, Y. & Kohane, I. Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis. *Pediatrics* **133**, e54–e63 (2013).
10. Li, L. et al. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Sci. Transl. Med.* **7**, 311ra174 (2015).
11. Zhang, X. et al. Data-driven subtyping of Parkinson's disease using longitudinal clinical records: a cohort study. *Scientific Rep.* **9**, 797 (2019).
12. Chen, D. et al. Deep learning and alternative learning strategies for retrospective real-world clinical data. *npj Dig. Med.* **2**, 1–5 (2019).
13. Bengio, Y., Courville, A. & Vincent, P. Representation learning: a review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 1798–1828 (2013).
14. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
15. Miotto, R., Wang, F., Wang, S., Jiang, X. & Dudley, J. T. Deep learning for healthcare: review, opportunities and challenges. *Brief Bioinform.* **19**, 1236–1246 (2017).
16. Xiao, C., Choi, E. & Sun, J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *J. Am. Med. Inform. Assoc.* **25**, 1419–1428 (2018).
17. Manchia, M. et al. The impact of phenotypic and genetic heterogeneity on results of genome wide association studies of complex diseases. *PLoS ONE* **8**, e76295 (2013).
18. Banda, J. M., Seneviratne, M., Hernandez-Boussard, T. & Shah, N. H. Advances in electronic phenotyping: from rule-based definitions to machine learning models. *Annu. Rev. Biomed. Data Sci.* **1**, 53–68 (2018).
19. Cote, R. A. & Robboy, S. Progress in medical information management: systematized nomenclature of medicine (snomed). *JAMA* **243**, 756–762 (1980).
20. Miotto, R., Li, L., Kidd, B. A. & Dudley, J. T. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific Rep.* **6**, 26094 (2016).
21. Dougherty, E. R. & Brun, M. A probabilistic theory of clustering. *Pattern Recogn.* **37**, 917–925 (2004).
22. Dalton, L. A., Benalcázar, M. E. & Dougherty, E. R. Optimal clustering under uncertainty. *PLoS ONE* **13**, https://doi.org/10.1371/journal.pone.0204627 (2018).
23. Brun, M. et al. Model-based evaluation of clustering validation measures. *Pattern Recogn.* **40**, 807–824 (2007).
24. Amigó, E., Gonzalo, J., Artiles, J. & Verdejo, F. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Inform. Retrieval* **12**, 461–486 (2009).
25. McInnes, L., Healy, J., Saul N., & Grossberger, L. UMAP: Uniform Manifold Approximation and Projection for dimension reduction. *J Open Source Softw* **3**, 861, https://doi.org/10.21105/joss.00861 (The Open Journal, 2018).
26. Cowie, C. C., Casagrande, S. S. & Geiss, L. S. Prevalence and incidence of type 2 diabetes and prediabetes. In *Diabetes in America* 3rd edn (eds Barrett-Connor, E. et al.) 3–1 (National Institutes of Health, Bethesda, 2018).
27. de Lau, L. M. L. & Breteler, M. M. B. Epidemiology of Parkinson's disease. *Lancet Neurol.* **5**, 525–535 (2006).
28. Qiu, C., Kivipelto, M. & von Strauss, E. Epidemiology of alzheimeras disease: occurrence, determinants, and strategies toward intervention. *Dialog. Clin. Neurosci.* **11**, 111 (2009).
29. Kazandjian, D. Multiple myeloma epidemiology and survival: a unique malignancy. In *Seminars in Oncology*, Vol. **43** (eds Ahn I. E. & Mailankody, S.) 676–681 (Elsevier, 2016).
30. Cancer Stat Facts: Prostate Cancer. https://seer.cancer.gov/statfacts/html/prost.html (2019).
31. Cancer Stat Facts: Female Breast Cancer. https://seer.cancer.gov/statfacts/html/breast.html (2019).
32. Vallon, V. & Komers, R. Pathophysiology of the diabetic kidney. *Compr. Physiol.* **1**, 1175–1232 (2011).
33. Malaguarnera, L., Cristaldi, E. & Malaguarnera, M. The role of immunity in elderly cancer. *Crit. Rev. Oncol. Hematol.* **74**, 40–60 (2010).
34. Delamaire, M. et al. Impaired leucocyte functions in diabetic patients. *Diabetic Med.* **14**, 29–34 (1997).
35. Jain, S., Lo, S. E. & Louis, E. D. Common misdiagnosis of a common neurological disorder. *Arch. Neurol.* **63**, 1100–1104 (2006).
36. Alves, G., Wentzel-Larsen, T. & Larsen, J. P. Is fatigue an independent and persistent symptom in patients with Parkinson disease? *Neurology* **63**, 1908–1911 (2004).
37. Siciliano, M. et al. Fatigue in Parkinson's disease: a systematic review and meta-analysis. *Mov. Disord.* **33**, 1712–1723 (2018).
38. Alzheimer's association. Younger/Early-Onset Alzheimer's. https://www.alz.org/alzheimers-dementia/what-is-alzheimers/younger-early-onset (2019).
39. Manji, H., Jäger, H. R. & Winston, A. HIV, dementia and antiretroviral drugs: 30 years of an epidemic. *J. Neurol. Neurosurg. Psychiatry* **84**, 1126–1137 (2013).
40. Lyketsos, C. G. et al. Prevalence of neuropsychiatric symptoms in dementia and mild cognitive impairment. *JAMA* **288**, 1475–1483 (2002).
41. Snyder, H. M. et al. Vascular contributions to cognitive impairment and dementia including Alzheimer's disease. *Alzheimers Dement.* **11**, 710–717 (2015).
42. Birks, J. S. & Harvey, R. J. Donepezil for dementia due to Alzheimer's disease. *Cochrane Database Syst. Rev.* **6**, CD001190 (2018).
43. Lombardo, M. V. et al. Unsupervised data-driven stratification of mentalizing heterogeneity in autism. *Scientific Rep.* **6**, 35333 (2016).
44. Stevens, E. et al. Identification and analysis of behavioral phenotypes in autism spectrum disorder via unsupervised machine learning. *Int. J. Med. Inform.* **129**, 29–36 (2019).
45. Choi, E., Bahadori, M. & Sun, J. Doctor AI: predicting clinical events via recurrent neural networks. In *Proc. Machine Learning for Healthcare*, Vol. 56 (eds Doshi-Velez, F. et al.) (PMLR, 2016).
46. Pham, T., Tran, T., Phung, D. & Venkatesh, S. DeepCare: A deep dynamic memory model for predictive medicine. In *Advances in Knowledge Discovery and Data Mining* (eds Bailey, J. et al.) 30–41 (Springer International Publishing, 2016).

47. Rajkomar, A. et al. Scalable and accurate deep learning with electronic health records. *npj Dig. Med.* **1**, 18 (2018).
48. Beaulieu-Jones, B. K. et al. Semi-supervised learning of the electronic health record for phenotype stratification. *J. Biomed. Inform.* **64**, 168–178 (2016).
49. Nguyen, P., Tran, T., Wickramasinghe, N. & Venkatesh, S. Deepr: a convolutional net for medical records. *IEEE J. Biomed. Health Inform.* **21**, 22–30 (2017).
50. Suo, Q. et al. Deep patient similarity learning for personalized healthcare. *IEEE Trans. NanoBiosci.* **17**, 219–227 (2018).
51. Wei, W. et al. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. *J. Am. Med. Inform. Assoc.* **23**, e20–e27 (2015).
52. Kirby, J. C. et al. Phekb: a catalog and workflow for creating electronic phenotype algorithms for transportability. *J. Am. Med. Inform. Assoc.* **23**, 1046–1052 (2016).
53. Halpern, Y., Horng, S., Choi, Y. & Sontag, D. Electronic medical record phenotyping using the anchor and learn framework. *J. Am. Med. Inform. Assoc.* **23**, 731–740 (2016).
54. Glicksberg, B. S. et al. Automated disease cohort selection using word embeddings from Electronic Health Records. In *Biocomputing 2018* (eds Altman, R. B. et al.) 145–156, https://doi.org/10.1142/9789813235533_0014 (World Scientific, 2017).
55. Blei, D., Ng, A. & Jordan, M. Latent dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–1022 (2003).
56. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient estimation of word representations in vector space. Preprint at https://arxiv.org/abs/1301.3781 (2013).
57. Jonquet, C., Shah, N. H. & Musen, M. A. The open biomedical annotator. In *AMIA Summits on Translational Science Proceedings* (ed American Medical Informatics Association) 56–60 (American Medical Informatics Association, Bethesda, MD, 2009).
58. Lependu, P., Iyer, S. V., Fairon, C. & Shah, N. H. Annotation analysis for testing drug safety signals using unstructured clinical notes. *J. Biomed. Seman.* **17**, s5 (2012).
59. Choi, Y., Chiu, C. Y. I. & Sontag, D. Learning low-dimensional representations of medical concepts. In *AMIA Summits on Translational Science Proceedings* (ed American Medical Informatics Association) 41–50 (American Medical Informatics Association, Bethesda, MD, 2016).
60. Zhu, Z. et al. Measuring patient similarities via a deep architecture with medical concept embedding. In *2016 IEEE 16th International Conference on Data Mining* (eds Bonchi, E. et al.) 749–758 (IEEE, 2016).
61. Suo, Q. et al. Personalized disease prediction using a CNN-based similarity learning method. In *2017 IEEE International Conference on Bioinformatics and Biomedicine* (eds Hu, X. et al.) 811–816 (IEEE, 2017).
62. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
63. Paszke, A. et al. Automatic differentiation in pytorch. In (eds Wiltschko, A., van Merriënboer, B. & Lamblin, P.) *NeurIPS Autodiff Workshop*, https://autodiff-workshop.github.io/ (2017).
64. Kingma, D. & Adam, J. B. Adam: a method for stochastic optimization. In *Proc. 3rd International Conference on Learning Representations* (eds Bengio, Y. & LeCun, Y.) 1–15, https://dblp.org/db/conf/iclr/iclr2015 (2015).

ACKNOWLEDGEMENTS

R.M. would like to thank the support from the Hasso Plattner Foundation, the Alzheimer's Drug Discovery Foundation and a courtesy GPU donation from Nvidia. I.L. acknowledges the support from the Bruno Kessler Institute.

AUTHOR CONTRIBUTIONS

I.L. and R.M. conceived and designed the work. I.L. conducted the research and the experimental evaluation, and drafted the manuscript. R.M. created the dataset, supervised and supported the research, and substantially edited the manuscript. B.S.G. substantially edited the manuscript and created the architecture figures. H.-C.L. and S.C. advised on methodological choices and critically revised the manuscript. G.L. provided clinical validation of the results and critically revised the manuscript. M.D. revised the manuscript and contributed to the interpretation of the data. J.T.D. and C.F. supported the research and revised the manuscript. All the authors gave final approval of the completed manuscript version and are accountable for all aspects of the work.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information is available for this paper at https://doi.org/10.1038/s41746-020-0301-z.

Correspondence and requests for materials should be addressed to R.M.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.