Inferring multimodal latent topics from electronic health records

Yue Li1✉, Pratheeksha Nair1, Xing Han Lu1, Zhi Wen1, Yuening Wang1, Amir Ardalan Kalantari Dehaghi1, Yan Miao1, Weiqi Liu1, Tamas Ordog2, Joanna M. Biernacka3,4, Euijung Ryu3, Janet E. Olson3, Mark A. Frye4, Aihua Liu5, Liming Guo5, Ariane Marelli5, Yuri Ahuja6, Jose Davila-Velderrain6 & Manolis Kellis6,7✉

Electronic health records (EHR) are rich heterogeneous collections of patient health information, whose broad adoption provides clinicians and researchers unprecedented opportunities for health informatics, disease-risk prediction, actionable clinical recommendations, and precision medicine. However, EHRs present several modeling challenges, including highly sparse data matrices, noisy irregular clinical notes, arbitrary biases in billing code assignment, diagnosis-driven lab tests, and heterogeneous data types. To address these challenges, we present MixEHR, a multi-view Bayesian topic model. We demonstrate MixEHR on MIMIC-III, Mayo Clinic Bipolar Disorder, and Quebec Congenital Heart Disease EHR datasets. Qualitatively, MixEHR disease topics reveal meaningful combinations of clinical features across heterogeneous data types. Quantitatively, we observe superior prediction accuracy of diagnostic codes and lab test imputations compared to the state-of-art methods. We leverage the inferred patient topic mixtures to classify target diseases and predict mortality of patients in critical conditions. In all comparison, MixEHR confers competitive performance and reveals meaningful disease-related topics.

1 School of Computer Science and McGill Centre for Bioinformatics, McGill University, Montreal, Quebec H3A0E9, Canada. 2 Department of Physiology and Biomedical Engineering and Division of Gastroenterology and Hepatology, Department of Medicine, and Center for Individualized Medicine, Mayo Clinic, Rochester, MN, USA. 3 Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA. 4 Department of Psychiatry and Psychology, Mayo Clinic, Rochester, MN, USA. 5 McGill Adult Unit for Congenital Heart Disease Excellence (MAUDE Unit), Montreal QC H4A 3J1 Quebec, Canada. 6 Computer Science and Artificial Intelligence Lab, Massachusetts Institute of Technology, 32 Vassar St, Cambridge, MA 02139, USA. 7 The Broad Institute of Harvard and MIT, 415 Main Street, Cambridge, MA 02142, USA. ✉email: yueli@cs.mcgill.ca; manoli@mit.edu

https://doi.org/10.1038/s41467-020-16378-3

© Springer Nature Limited. All rights reserved.
The broad adoption of electronic health record (EHR) systems has created unprecedented resources and opportunities for conducting health informatics research. Hospitals routinely generate EHR data for millions of patients, which are increasingly being standardized by using systematic codes, such as Logical Observation Identifiers Names and Codes (LOINC) for lab tests, RxNorm for prescriptions; and Systematized Nomenclature of Medicine (SNOMED), Diagnosis-Related Groups (DRG), and International Classification of Diseases (ICD-9) for diagnoses (albeit for billing purposes). In the USA, for example, the number of nonfederal acute care hospitals with basic digital systems increased from 9.4% to 96% over the 7-year period between 2008 and 2015. Furthermore, the amount of comprehensive EHR data recording multiple data types, including clinical notes, increased from only 1.6% in 2008 to 40% in 2015. With the aid of effective computational methods, these EHR data promise to define an encyclopedia of diseases, disorders, injuries, and other related health conditions, uncovering a modular phenotypic network.

Distilling meaningful concepts from the raw EHR data presents several challenges, and it is often unfeasible to directly model the joint distribution over the entire EHR feature space. On the other hand, it is possible to formulate a latent topic model over discrete data, in analogy to automatic text categorization, considering each patient as a document and each disease meta-phenotype as a topic (grouping recurrent combinations of individual phenotypes). Our task then is to learn a set of meaningful disease topics, and the probabilistic mixture memberships of patients for each disease topic, representing the combination of meta-phenotypes inferred for each patient.

In this paper, we formalize this analogy by introducing a latent topic model, MixEHR, designed to meet the challenges intrinsic to heterogeneous EHR data. The main objective of our approach is twofold: (1) distill meaningful disease topics from otherwise highly sparse, biased, and heterogeneous EHR data; and (2) provide clinical recommendations by predicting undiagnosed patient phenotypes based on their disease mixture membership. Importantly, we aim for our model to be interpretable in that it makes not only accurate predictions but also intuitive biological sense. MixEHR can simultaneously model an arbitrary number of EHR categories with separate discrete distributions. For efficient Bayesian learning, we developed a variational inference algorithm that scales to large-scale EHR data.

MixEHR builds on the concepts of collaborative filtering and latent topic modeling. In particular, our method is related to the widely popular text-mining method Latent Dirichlet Allocation (LDA). However, LDA does not account for missing data and NMAR mechanism. Our method is also related to several EHR phenotyping studies focusing on matrix factorization and model interpretability. Recently developed deep-learning methods primarily focus on prediction performance of target clinical outcomes. Apart from these methods, Graph-based Attention Model (GRAM) uses the existing knowledge graph to embed ICD code and make prediction of diagnostic codes at the next admission using the graph embedding. Deep patient uses a stacked stochastic denoising autoencoder to model the EHR code by its latent embedding, which are used as features in a classifier for predicting a target disease. More detailed review are described in Supplementary Discussion.

We apply MixEHR to three real-world EHR datasets: (1) Medical Information Mart for Intensive Care (MIMIC)-III; a public dataset to date containing over 50,000 ICU admissions; (2) Mayo Clinic EHR dataset containing 187 patients, including with 93 bipolar disorder and 94 controls; (3) the Québec Congenital Heart Disease Dataset (Quebec CHD Database) on over 80,000 patients being followed for congenital heart disease (CHD) for over 28 years (1983–2010) at the McGill University Health Centre (MUHC) in Montreal. In these applications, we find that the clinical features emerging across EHR categories under common disease topics are biologically meaningful, revealing insights into disease co-morbidities. The inferred topics reveal sub-categories of bipolar disorder, despite the small sample size. The inferred patient topic mixtures can be used to effectively predict diagnostic code in patient’s future hospital visits from outpatient data, impute missing lab results, and predict future next-admission patient mortality risk.

**Results**

**MixEHR probabilistic model.** We take a probabilistic joint matrix factorization approach, by projecting each patient’s high-dimensional and heterogeneous clinical record onto a low-dimension probabilistic meta-phenotype signature, which reflects the patient’s mixed memberships across diverse latent disease topics. The key innovation that enables us to handle highly heterogeneous data types is that we carry out this factorization at two levels. At the lower level, we use data-type-specific topic models, learning a set of basis matrices for each data type. Using the MIMIC-III data as an example, we learned seven basis matrices corresponding to clinical notes, ICD-9 billing codes, prescriptions, DRG billing codes, CPT procedural codes, lab tests, and lab results from the MIMIC-III dataset. We link these seven basis matrices at the higher level using a common loading matrix that connects the multiple data types for each patient (Fig. 1a).

To model the lab data, we assign common latent topic to both the lab test and lab test result such that they are conditionally independent given their topic assignments (Fig. 1b). We then updates the lab topic distributions using both the observed lab test results and the missing/imputed lab test results weighted by their topic assignment probabilities for each patient. This differs from other latent topic models, which assume that data are either completely observed or missing at random. To learn the model, we use a variational Bayes to update each model parameters in a coordinate ascent (Fig. 1c). The inference algorithms are detailed in “Methods” and “Supplementary Methods”. After learning the model, we can obtain three pieces of clinically relevant information: (1) to impute missing EHR code for each patient, we take average over the disease topics; (2) to infer patients’ meta-phenotypes as their mixed disease memberships, we average over the phenotype dimension; and (3) to infer latent disease topic distribution (i.e., distributions over all clinical variables), we average over the patient dimension.

**Multimodal disease topics from MIMIC-III.** We applied MixEHR to MIMIC-III data, which contain ~39,000 patients each with a single admission and ~7500 patients each with multiple admissions (Supplementary Table 1). We used MixEHR to jointly model six data categories, including unstructured text in clinical notes, ICD-9, DRG, current procedural terminology (CPT), prescriptions, and lab tests, together comprising of ~53,000 clinical features and ~16 million total clinical observations (“Methods”). The frequency of observed EHR code over all admissions is low (Supplementary Fig. 27). Majority of the variables including the lab tests were observed in less than 1% of the 58,903 admissions (i.e., greater than 99% missing rates for most variables). Therefore, the data are extremely sparse underscoring the importance of integrating multimodal data information to aid the imputation.

To choose the optimal number of topics (K), we performed fivefold cross-validation to evaluate models with different K by the averaged predictive likelihood on the held-out patients (Supplementary Fig. 1a). We set K = 75 for subsequent analyses as it gave the highest predictive likelihood. We then examined the
clinical relevance of our learned disease topics by qualitatively assessing the coherence of the 5 most probable EHR codes for each topic. Taking a column with common index $k$ from the topic distribution matrices for each of the six EHR data types (i.e., notes, ICD-9, CPT, DRG, lab tests, and prescription) gives us a single disease topic distribution over clinical terms across those data types. We annotated all of the 75 topics based on the top EHR codes and found that majority of the inferred topics are specific to distinct diseases (Supplementary Table 2).

For the purpose of demonstration, we paid special attention to the disease topics exhibiting the highest likelihoods over the broad set of ICD-9 codes with prefixes {250, 296, 331, 042, 205, 415, 571}, representing diabetes, psychoses, neurodegeneration, HIV, myeloid leukemia, acute pulmonary heart disease, and chronic liver disease and cirrhosis, respectively. The top terms are displayed in alternative word cloud representations in Supplementary Fig. 3. We observed salient intra-topic consistency and inter-topic contrast with respect to each topic’s highest-scoring EHR features across diverse EHR data categories (colored bars on the right; Fig. 2a). Notably, if we were to use LDA and thereby model observations across all data categories as draws from the same multinomial distribution, the learned disease topics would have been dominated by clinical notes, given that these contain far more features (i.e., words) than the other categories. We
### Part A: Top Clinical Features from Select Disease Topics

| Disease Topics | EHR Codes |
|----------------|-----------|
| Cirrhosis      | Liver failure (5712), Alcohol use (2530) |
| PulmEm         | Pulmonary embolism (4984), Infiltrative effects (4992) |
| HIV            | Human immunodeficiency virus (042), HIV-related conditions (4223) |
| Diabetes       | Insulin resistance (01740), Diabetes mellitus (3072) |
| Leukemia       | Acute leukemia (APR,8904), Neutrophil count (9625) |
| Dementia       | Dementia (APR,3310), Mental disorder (9625) |
| Psycho         | Poisoning (APR,8124), Schizophrenia (APR,7502) |

### Part B: Topics Correlated with Age

| Disease Topics | EHR Codes |
|----------------|-----------|
| Heart failure  | Coronary artery (4704), Heart failure (APR,1954) |
| Cardiovascular | Circulatory disorders (APR,8996), Myocardial infarction (APR,9031) |
| Dementia       | Mental disorder (APR,1574), Dementia (APR,3310) |
| Neonate        | Neonatal death (APR,5012), Complications of pregnancy (APR,5237) |
| Preterm birth  | Neonatal death (APR,5012), Preterm birth (APR,5237) |

### Part C: Topics Correlated with Psychiatric Traits

#### Schizophrenia-associated codes

- Delirium: Schizophrenia nos.-unspec (29590)
- Dementia: Schizophrenia (29590)
- Poisoning: Alcohol use (2530)

#### PTSD-associated codes

- PTSD: Ptsd-disorder (29280)
- Anxiety: Panic disorder (30800)
- Depression: Major depression (30800)
- Alcohol use: Alcohol use (2530)

### Figure 2: Disease Topics from MIMIC-III Dataset

- **a**: Top five EHR codes for select disease topics from the 75-topic MixEHR model. The rows are the top features, and columns are the topics. The color intensities indicate the inferred probabilities of these features belonging to each topic.
- **b**: Age-correlated disease topics. Top three EHR codes were displayed for the top three most and least age-correlated topics.
- **c**: Top correlated EHR codes with schizophrenia and PTSD.
illustrate this aspect in the application of mortality prediction below.

Moreover, we observed interesting connections between abnormal lab results and specific diseases (Supplementary Fig. 4). For instance, topic M31 is enriched for alcoholic cirrhosis billing codes and ascites-based lab tests typical for cirrhosis patients; in M42, the code for HIV infection is grouped with abnormal counts of immune cells, including CD4 lymphocytes; diabetes codes tend to rank highly alongside high blood acetone in M48, and topic M62 is enriched for opioid abuse along with valproic acid and lithium prescriptions—common treatments for bipolar disorder. Interestingly, topic M59—associated with neurodegenerative diseases such as Alzheimer’s disease (AD)—is strongly associated with vitamin B12, which was recently shown to exhibit differential pattern in AD patients. We also correlated the 75-top distributions with patients’ age. The results clearly show that the top positively age-correlated topics are related to heart failure (M20), cardiovascular (M44), and dementia (M59). In contrast, the top three negatively age-correlated topics are all related with newborn (Fig. 2b; Supplementary Fig. 5).

**Disease comorbidity network in MIMIC-III data.** We reasoned that the similarity of the annotated disease (ICD-9 codes) distributions over topics might help uncover nontrivial associations between diseases. In order to infer a disease comorbidity network, we used the 75 learned disease topics and calculated the pairwise correlations between ICD-9 diagnostic codes. We included 424 ICD-9 codes observed in at least 100 patients, and not in the external injuries (ICD-9 code starting with V) or supplementary classification (ICD-9 code starting with E) sections. For comparison, we calculated the correlations between the same 424 ICD-9 codes using their sparse distributions over patients in the raw data. Despite focusing on the ICD-9 codes with at least 100 patients, on average each code is observed in only 0.8% of patients (min: 0.3% and max 2%). Consequently, due to the high sparsity of diseases among patients, we observed very weak correlations based on the observed data ranging between $-0.02$ and 0.06 (Supplementary Fig. 6a). However, when we calculated the ICD-9 codes using our inferred topic embeddings, we observed much stronger correlations. More importantly, diseases of similar physiology form modules consistent with clinical intuition (Supplementary Fig. 6b). Thus, the projection of multimodal data to the inferred topic embeddings space enables the discovery of strong disease association, not directly measurable from the raw EHR data.

Using these topic embeddings, we can also obtain a comorbidity network centering on a specific disease of interest by correlating the disease’s diagnostic code with all other diagnostic codes. In particular, we obtained such a comorbidity network for schizophrenia, post traumatic stress disorder (PTSD), Alzheimer’s disease, and bipolar disorder (Fig. 2c; Supplementary Figs. 7–14). The size of the terms are proportional to their Pearson correlation with the disease of interest. To control the false discovery rate, we randomly shuffled the topic probabilities for the target phenotype to calculate the background correlation. Only phenotypes with permutation $p$-value $<0.01$ are displayed. We observed meaningful related phenotypes in all of the four examples.

**Patient risk prioritization using MIMIC-III data.** Besides gaining insights from the disease topics (i.e., basis matrices), we can also exploit the disease topic mixture memberships along the patients dimension. As an example, we took the 50 patients with the highest proportions for topics M31, M35, and M50, which as we noted above are associated with alcoholic cirrhosis, pulmonary embolism, and leukemia, respectively (Fig. 3a). We observed a clear enrichment for the expected ICD-9 codes—571, 415, and 205, corresponding to the diagnoses of chronic liver disease and cirrhosis, acute pulmonary heart disease, and myeloid leukemia, respectively—among these patients (Fig. 3b). However, not all of the patients were classified by their expected diagnostic codes.

We then hypothesized that these unclassified patients perhaps exhibit other markers of the relevant diseases. To further investigate these high-risk patients, we examined the 10 highest-scoring EHR codes under these topics (Fig. 3c). We highlighted some evidence as why the patients with the absence of ICD-9 code were prioritized as high risk by our model. Most patients under the leukemia topic M31, including those missing ICD-9 code 205, received bone marrow biopsies (CPT code 4131) and infusion of a cancer chemotherapeutic substance (CPT code 9925)—clear indicators of leukemia diagnosis and treatment. Likewise, several patients under the pulmonary embolism topic (M35) that are missing ICD-9 code 415 nevertheless have the DRG billing code for pulmonary embolism, a prescription for enoxaparin sodium (which is used to treat pulmonary embolism), or the related ICD-9 code 453 for other venous embolism and thrombosis. Moreover, all patients under the cirrhosis topic M50 have the key word cirrhosis mentioned in their clinical notes. Notably, although the procedural code for esophageal variceal banding is not particularly prevalent, MixEHR predicts reasonably high probability of undergoing this procedure among cirrhosis-topic patients—a clinically reasonable prediction. This suggests that our model is not only intuitively interpretable but also potentially useful for suggesting patient’s future medical interventions.

**Classification of Mayo Clinic bipolar disorder patients.** To further demonstrate the utility of our approach in discovering meaningful multimodal topics, we applied MixEHR to a separate dataset containing 187 patients, including 93 bipolar disorder cases and 94 age- and sex-matched controls from Mayo Clinic. Despite the small sample size, the patients were deeply phenotyped: there are in total 7731 heterogeneous EHR features across five different data categories, including ICD-9 codes, procedure codes, patient provided information (PPI), lab tests, and prescription codes. In total, there are 108,390 observations.

We evaluated our model by fivefold cross-validation. Here, we trained MixEHR or the unsupervised baseline models and a logistic regression (LR) classifier that uses the patient topic mixture derived from MixEHR or embeddings derived from other methods to predict the bipolar disorder label (Fig. 4a; “Methods”). We observed superior performance of MixEHR+LR compared with LDA+LR and RBM+LR in both the area under the ROC and the area under the precision-recall curves (Fig. 4b). We quantified the predictive information of the 20 topics based on the linear coefficients of the LR classifier. We observed that topics M19 and M7 have the highest positive coefficients for bipolar disorder label (Fig. 4c). To confirm the finding, we visualized the 187 patient topic mixtures side-by-side with their BD diagnosis ICD-9 296 code (Fig. 4d). Indeed, we find that M19 and M7 are highly correlated with the BD diagnosis label, and M19 and M7 may represent two distinct subgroups of BD patients. To check whether these two topics were driven by demographic information, we compared the topic mixture membership with sex and age for each patient. We did not observe significant difference between sexes for M19 mixture probabilities or M7 mixture probabilities (Supplementary Fig. 28). There is also no significant correlation between age and M19 or M7 topic mixture (Supplementary Fig. 31).
The results as an exploratory purpose only. Results were not corrected for multiple testing, and thus we used note that these phenome-wide association studies (PheWAS) for M19 or M7, to assess positive enrichment for the topic. We associated with both the M7 and M19 topics, we also psychotic behavior (296.53) are significant for bipolar type I disorder with the most recent severe depressed episode without history of psychiatric condition (V17.0), and bipolar type I (or current) depressed, moderate) codes for BD (Fig. 5). Certain lab tests (e.g., SLC6A4, HTR2A, and cytochrome P450 enzymes) are also associated with M19, but not M7. Lithium lab test is also associated with high M19 mixture probabilities, but not M7. Thus it is possible that M19 patients may have had more severe symptoms and needed an increased use of pharmacogenically guided treatment. PheWAS for the PPI questions are displayed in Supplementary Fig. 32. Interpretation on the PPI PheWAS must be taken with caution (“Methods”). With these data, we are able to demonstrate a potential utility of our MixEHR approach to classify BD patients into potentially clinically meaningful categories that require further investigation in larger dataset.

**EHR code prediction in MIMIC-III data.** Retrospective EHR code prediction has its value in diagnosing the code entered in the existing EHR patient records and making suggestions about the potential missing code and incorrectly entered code based on the expectation of those EHR codes. To predict EHR codes, we formulated a k-nearest neighbor approach (Supplementary Fig. 16a; “Methods”). We evaluated the prediction accuracy based on fivefold cross-validation. The predicted EHR codes match consistently with the observed frequency of the diagnostic codes (Supplementary Fig. 15). Overall, the median prediction accuracy of MixEHR (multimodal) is 88.9%, AUROC is 85%, and AUPRC is 9.6%. In contrast, the baseline model that ignores distinct data types (attenuated; i.e., LDA) obtained only 83.3% accuracy, 77.1% AUROC, and 6.7% AUPRC, respectively (Supplementary Fig. 16b). The prediction performances vary among distinct data types (Supplementary Fig. 17) and also among different ICD-9 disease groups (Supplementary Fig. 18), which are attributable to the rareriness of the diseases, complexity of the disease codes, and the strength and limitations of our model-based assumption. One caveat in this experiment is that we are potentially using highly related code to predict the current target code. Therefore, we should distinguish this experiment from predicting diagnostic code in the future admissions of inpatient data or future visits in the outpatient data.

**Fig. 3 Prioritize patients by disease mixture from MIMIC-III dataset.** a Word clouds of the three topics with colors indicating different data types and font size proportional to the probabilities of the EHR codes under that topic. b We selected top 50 patients under three of the learned topics for illustration. These topics are displayed as word clouds where the size of the word is proportional to their probabilities under the topic. The heatmap indicates patient mixture memberships for the topics. The color bar on the right of the heatmap indicates the actual diagnoses of the diseases related to each of the topics, namely leukemia, pulmonary embolism, and cirrhosis. c Top EHR codes of the high-risk patients. The top ten EHR codes (rows) from each topic were displayed for the top 50 patients (columns) under each topic. We highlighted some evidence (asterisks) as why some of the patients in absence of ICD-9 codes were prioritized as high risk by our method based on their relevance to the diseases of interests (i.e., leukemia, pulmonary embolism, and cirrhosis).
Fig. 4 Classification of bipolar disorder patients using the Mayo Clinic dataset. a Workflow to evaluate the classification accuracy of bipolar disorder in a fivefold cross-validation (CV). b Classification accuracy of fivefold CV in ROC and precision-recall curves. c Linear coefficients of the 20 topics ranked by decreasing order. d Inferred 20-topic mixture of 187 patients. The patients with and without bipolar disorder diagnosed were colored in red and green, respectively, on the left side of the dendrogram.
In the MIMIC-III data, there are 7541 patients who were admitted to the hospital more than once. This allows us to evaluate how well we can predict the diagnostic codes in the next admission based on the information from the previous admission(s) for the same patients. To this end, we developed a pipeline that combines MixEHR topics with recurrent neural network (RNN) with Gated Recurrent Unit (GRU) (Fig. 6a; “Methods”). By leveraging the correlation structure information between rarely observed codes and commonly observed codes via the latent topics, we hypothesize that our approach can achieve comparable prediction accuracy as GRAM, which relies on the accuracy of the existing ICD-9 knowledge graph or taxonomy.

We observe that MixEHR-RNN confers the highest overall accuracy among the three methods, although the difference between MixEHR+RNN and GRAM is small (Fig. 6b). We also generated the prediction accuracy on the codes binned by their frequencies as before (Supplementary Fig. 35). MixEHR+RNN performs the best in four out of the five ranges and falls short by only 2% in the last range (80–100) compared with GRAM. Both MixEHR+RNN and GRAM outperform Doctor AI by a large margin. Doctor AI did not do well in this application because of the small sample size and short medical history of the MIMIC-III data (11 years) (Fig. 6b).

**Longitudinal EHR code prediction using Quebec CHD data.**

We evaluated the code prediction accuracy on the 28-year longitudinal outpatient data from the CHD database using a MixEHR+RNN architecture (Fig. 6c; “Methods”). Compared to our model with the baseline RNN, we observed a significant improvement in terms of both AUPRC (Wilcoxon signed-rank tests p-value < 0.0408) and AUROC (Wilcoxon signed-rank tests p-value < 5.35e-55) (Fig. 6d). Therefore, adding MixEHR significantly improves the EHR code prediction over the baseline RNN model. We also observed that the learned weights connected to the 50-topic mixture exhibit higher magnitude than the learned weights connected to the concatenated dense layer embedding (Supplementary Fig. 19). This means that the network relies heavily on the topic mixture to make accurate predictions.

**Lab results imputation using MIMIC-III data.** We then evaluated the imputation accuracy of missing lab results. We first performed an extensive simulation by subsampling from MIMIC-III data to compare lab imputation accuracy between MixEHR and MICE (Supplementary Note). Our approach demonstrated superior performance compared with MICE and the model that assumes missing-at-random (MAR) lab results (Supplementary Fig. 25). We then demonstrated our approach on imputing the real MIMIC-III lab results. Here, we took a similar approach as the k-nearest neighbor approach in the EHR code prediction (Fig. 7a; “Methods”). We observed that MixEHR achieves

---

**Fig. 5 Phenome-wide association studies (PheWAS) of the 6776 EHR codes on the two bipolar disorder mixture topics (M7 and M19) based on the Mayo Clinic dataset.** For each EHR code, we tested the difference between patient groups with and without that code in terms of M19 and M7 topic membership probabilities. The labeled EHR codes are the ones with Wilcoxon signed-rank one-sided test p-values < 0.001. The red, blue, and black color indicate the codes that are significant in only M19, in only M7, and in both M19 and M7, respectively. PheWAS for the PPI questions are displayed in Supplementary Fig. 32.

**Fig. 6a** illustrates the architecture of MixEHR + GRU. **Fig. 6b** shows the prediction accuracy on the codes binned by their frequencies as before (Supplementary Fig. 35). MixEHR+RNN performs the best in four out of the five ranges and falls short by only 2% in the last range (80–100) compared with GRAM. Both MixEHR+RNN and GRAM outperform Doctor AI by a large margin. Doctor AI did not do well in this application because of the small sample size and short medical history of the MIMIC-III data (11 years) (Fig. 6b).
significantly higher accuracy compared to CF-RBM (Fig. 7b, c; Wilcoxon test $p$-value $< 0.00013$, Kolmogorov–Smirnov (KS) test $p$-value $< 1.15 \times 10^{-5}$). This is attributable to two facts: (1) MixEHR accounts for NMAR by jointly modeling the distribution both the lab missing indicators and lab results; (2) MixEHR is able to leverage more information than CF-RBM: it models not only the lab data but also other administrative data and clinical notes. We also observed significantly higher accuracy for our MixEHR imputation strategy compared to a simple averaging approach (Supplementary Fig. 34; “Methods”).

**Mortality-risk prediction using the MIMIC-III data.** Predicting patient’s mortality risk in the Intensive Care Unit (ICU) is vital to assessing the severity of illness or trauma and facilitating subsequent urgent care. Because we are able to distill meaningful disease topics from heterogeneous EHR data, we sought to assess the predictive capacity of these disease topics for mortality predictions in future admissions. Here, we predicted the mortality outcome in the last admission based on the second-last admission of the patient within 6 months. We obtained the highest 31.62% AUPRC from MixEHR (50 topics) among all methods (Fig. 8a).
We also experimented with two different classifiers namely logistic regression with L2-norm (Supplementary Fig. 29) and random forest (Supplementary Fig. 30). We observed overall best results with logistic regression followed by second highest AUROC of 65.69% (Supplementary Fig. 22).

Furthermore, we designed another experiment that used the earliest/second admission to predict the mortality in the last admission. The value of this application is to identify patients who are discharged too early and therefore to provide a measure on whether the patients should be discharged from the ICU based on their current condition. Once again, MixEHR outperformed other baseline methods (Supplementary Fig. 33). Here, we focused our analysis on the mortality in the last admission based on the information in the second-last admission.

Notably, LDA obtained similar performance to MixEHR, whereas PCA, SiCNMF, and EN performed a lot worse on this task. This suggests that the topic models (i.e., LDA and MixEHR) are generally more suitable to modeling the discrete count data in the EHR. Based on the elastic net linear coefficients, the three most positively predictive mortality topics are normal newborn (topic M52 with the most positive coefficient), aneurysm (M8), and drug poisoning (M73), respectively (Fig. 8c). Interestingly, the three most negatively predictive mortality topics are normal newborn (topic M52 with the most negative coefficient), aneurysm (M8), and drug poisoning (M73), respectively. We note that each MixEHR topic is represented by the top EHR features from diverse data types. In contrast, all of the LDA topics are overwhelmed by the category that contains a lot more features than the other categories. Qualitatively, this greatly reduces the interpretability of the topics compared to our MixEHR.
We also ran DeepPatient$^{29}$ on the MIMIC-III data and did not obtain good performance (“Methods”; Supplementary Fig. 23). This is perhaps due to the small sample size and the lack of carefully designed preprocessing pipeline$^{29}$ as the authors demonstrated on their own EHR data. In addition, previous work has shown the benefits of using demographic information in representing EHR codes and patient visits$^{35}$. Our preliminary results show little benefit of adding demographic information including sex, age, religion, ethnicity, and marital status in the mortality prediction (Supplementary Fig. 24). However, we acknowledge its importance and will continue exploring this in our future work.

**Discussion**

We present MixEHR as an unsupervised Bayesian framework to efficiently model the distribution of heterogeneous high-dimensional EHR data. We envision MixEHR to be a step toward several promising translational venues, such as (1) facilitating clinicians’ decision-making, (2) directing further/deeper phenotyping, (3) enabling personalized medical recommendations, and (4) aiding biomarker discovery. Specifically, MixEHR can infer the expected phenotypes of a patient conditioned only on a subset of clinical variables that are perhaps easier and cheaper to measure. Because new data are constantly entering the...
EHR systems, our model needs to adapt to incoming data without refitting to the entire dataset—in other words, to go online—a topic for future development.

In the current framework, MixEHR assumes that the input data are a set of two-dimensional matrices of patients by measurements (lab tests, diagnoses, treatments, etc.), as delineated in the structure of our datasets. However, one could envision EHR data being represented as higher-dimensional data objects—for instance, as a three-dimensional object of patients by lab tests by diagnoses, patients by diagnoses by treatments, or patients by diagnoses by time. To model such high-dimensional EHR data (as it becomes available), we can extend MixEHR to a probabilistic tensor-decomposition (TD) framework. Recently developed non-Bayesian models such as Marble36, Rubik37, and non-negative TD21 are TD-based methods that show great promise. MixEHR is an ideal modeling core to serve as the basis for such upcoming modeling high-dimensional challenges.

It is challenging to interpret associations with patient provided information (PPI), compared with the other data types we used here. The presence of PPI data is confounded by factors such as the point of care (e.g., primary care versus psychiatric department), gender, demographics, and other social determinants of health. We will revisit this challenge in our future work.

Our basic MixEHR framework is cross-sectional. Nonetheless, longitudinal EHR learning is an important research area as it promises to forecast patients’ disease outcomes based on their entire medical history records while taking into account the temporal trajectory of patient states27,38,39. To this end, we present a pipeline that provides the MixEHR-inferred topic mixture at each admission as an input to a recurrent neural network in order to model the longitudinal aspect of the EHR data. It is more challenging to model irregularly sampled time points as commonly observed in the outpatient visits while modeling high-dimensional variables. In addition, in our lab imputation analysis, we did not model the sequential event of lab test results within the same admission, which requires a more sophisticated model that takes into account NMAR and the irregularly measured lab tests during the patient’s stay. We leave these as future extensions to MixEHR.

In the mortality prediction, we used the information at the early admission to predict mortality in the last admission. Meanwhile, we understand the value of using the first 24 or 48 h data to predict in-hospital mortality within the same admission (e.g. ref. 17). This is different from our application. We did not carry out this experiment because not all of the EHR data in MIMIC-III have a time stamp (i.e., CHARTTIME) within the same admission. In particular, although clinical notes and lab tests have chart time that records the time they were taken during the in-hospital admission, all of the rest of the data including ICD-9 diagnostic code, prescription, CPT, DRG code do not have a chart time and are usually entered upon patient’s discharge. This makes it difficult to design such an experiment that uses all of the information. We acknowledge that one caveat in our analysis is that some patients who are terminally ill (e.g., at the metastatic cancer stage) may choose to die at home and therefore our prediction may reflect not only the patients’ health states but also their mental and social states.

In summary, our data-driven approach, using a large number of heterogeneous topics over long observation periods, has the potential of leading to clinically meaningful algorithms that will help clinicians anticipate high disease-burden events in clinical practice. Our findings can inform the growing emphasis on proactive rather than reactive health management as we move into an era of increasing precision medicine.

Methods

MixEHR topic model (MixEHR). We model EHR data using a generative latent topic model (Fig. 1). Notations are summarized in Supplementary Table 3. Suppose there are K latent disease topics, each topic \( k \in \{1, \ldots, K\} \) under data type \( t \in \{1, \ldots, T\} \) prescribes a vector of unknown weights \( \phi_k^{(t)} = (\phi_{k,1}^{(t)}, \ldots, \phi_{k,V}^{(t)}) \) for \( V \) EHR features, which follows a Dirichlet distribution with unknown hyperparameter \( \beta_{k,\cdot} \).

In addition, each topic is also characterized by a set of unknown Dirichlet-distributed weights \( \eta_k = \{\eta_{k,v}\}_{v=1}^V \) with \( V \) distinct values (e.g., a lab test \( l \) with \( V_l \) distinct lab result values). For each patient \( j \in \{1, \ldots, D\} \), the disease mixture membership \( \theta_j \) is generated from the K-dimensional Dirichlet distribution \( \text{Dir}(\cdot) \) with unknown asymmetric hyperparameters \( \alpha_k \). To generate EHR observation \( i \) for patient \( j \), a latent topic \( z_{ij}^{(t)} \) under data type \( t \) is first drawn from multinomial distribution with rate \( \theta_j \). Given the latent topic \( z_{ij}^{(t)} \), a clinical feature \( x_{ij}^{(t)} \) is then drawn from multinomial distribution with rate equal to \( \phi_{z_{ij}^{(t)}}^{(t)} \).

For lab data, we use a generative process where each patient has a variable for the result for every lab test regardless whether the results is observed. One important assumption we make here is that the lab results \( y_{ij}^{(t)} \) and lab observation \( r_{ij} \) for patient \( j \) are conditionally independent given the latent topic \( h_{ij} \). Namely that the probability of taking the test and the value of that test both depend on the latent disease topic, but not on each other. In terms of the generative process, we first sample the latent variable \( y_{ij} \) from the multinomial distribution with rate \( h_{ij} \). Conditioned on the latent variable, we then concurrently sample (1) the lab result \( y_{ij}^{(t)} \) from \( V_l \)-dimensional Dirichlet distribution \( \eta_{ij}^{(t)} \) with hyperparameters \( \zeta_l \) over \( V_l \) values and (2) the lab observation indicator \( r_{ij} \) from Binomial distribution \( \text{Bin}(\cdot) \) with rate \( y_{ij}^{(t)} \), which follows a Beta distribution with unknown shape and scale hyperparameters \( a_l \) and \( b_l \). The hyperparameters \( a_l, b_l, \alpha_k, \zeta_l, a_l, b_l \) follow unknown Gamma distributions.

Formally, we first generate global variables as the parameters for the disease topic distribution:

\[ \phi_k^{(t)} \sim \text{Dir}(\beta_k) \]
\[ \eta_k \sim \text{Dir}(\zeta) \]
\[ y_{ij}^{(t)} \sim \text{Bin}(\eta_{ij}^{(t)}) \]
\[ x_{ij}^{(t)} \sim \text{Mul}(\phi_k^{(t)}) \]

and the local variables for the lab data, including latent topic, lab result, and observation indicator:

\[ h_{ij} \sim \text{Mul}(\theta_j) \]
\[ y_{ij} \sim \text{Mul}(\eta_{ij}) \]
\[ r_{ij} \sim \text{Bin}(y_{ij}) \]

where \( \text{Gam}(\cdot, \cdot), \text{Dir}(\cdot), \text{Bern}(\cdot), \text{Mul}(\cdot), \) and \( \text{Bin}(\cdot) \) denote gamma, Dirichlet, beta, multinomial, and Binomial distributions, respectively.

Marginalized likelihood. Treating the latent variables as missing data, we can express the complete joint likelihood based on our model as follows:

\[ p(y, r, x, z, h, \theta, \phi, \psi, \eta, \gamma) = p(\theta_{\text{heatmap}}|z) p(h|\theta)p(x|\theta)p(z|\phi)p(r|\psi)p(y|\psi)p(\eta|\gamma)p(\gamma|x, y, z, r) \]

At this point, we could formulate a variational inference algorithm with mean-field factorization by optimizing an evidence lower bound of the marginal likelihood with respect to the model parameters40. However, that would be inaccurate and computationally expensive for the following reasons. The mean-field variational distribution assumes that all of the latent variables are independent. On the other hand, we know from our model that these independence assumptions are unrealistic. In particular, the latent topic assignments depend on the patient mixtures, and the disease latent topic distribution and topic assignment are dependent given the observed lab tests or diagnostic code (i.e., the \( V \) shape structure in the probabilistic graphical model). In addition, the variational updates require extensive usages of exponential function and digamma function for every latent topic assignment of every patient, which are computationally expensive to compute for larger number of patients.

We can achieve more accurate and efficient inference by first analytically integrating out the patient mixtures and latent topic distributions\(^{33,34}\). To do so, we exploit the respective conjugacy of Dirichlet variables \( \phi, \theta \) to the multinomial likelihood variables \( x, y, z, h \), and the conjugacy of Beta variable \( \psi \) to the binomial lab observation indicator variable \( r \). This way we can simplify the complete joint
likelihood distribution of the data and the latent variables $\{z, h\}$ as follows:

$$p(x, y, t, z, h|a, \beta, \xi, a, b) = \int p(x, y, t|\theta)p(\theta|a)dt \int p(x|z, \phi)p(\phi|b)d\phi$$

$$= \prod_l \Gamma(\sum_{n_l} \alpha_l) \prod_l \Gamma\left(\alpha_l + n_{l,k}^0 + m_{l,k}\right) \prod_l \Gamma\left(\sum_{n_l} \beta_l \right) \prod_l \Gamma\left(\beta_l + m_{l,k}^0\right) \prod_l \Gamma\left(\sum_{n_l} \gamma_l \right) \prod_l \Gamma\left(\gamma_l + m_{l,k}\right)$$

$$\quad \Pi_l \Gamma\left(\sum_{n_l} \gamma_l \right) \Gamma\left(n_{l,k}\right) \Pi_l \Gamma\left(\gamma_l + n_{l,k}\right) \Pi_l \Gamma\left(\sum_{n_l} \gamma_l \right) \Gamma\left(n_{l,k}\right)$$

where the sufficient statistics are:

$$n_{l,k} = \sum_{i=1}^{M_l} z_{l,i}^{(k)} = k, \quad n_{l,k}^0 = \sum_{i=1}^{M_l} z_{l,i} = w, \quad z_{l,i}^{(k)} = k$$

$$m_{l,k} = \sum_{i=1}^{V} \sum_{v=1}^{V} y_{l,i} | h_{l,k} = k, \quad m_{l,k}^0 = \sum_{i=1}^{V} y_{l,i} | h_{l,k} = 0$$

$$= \sum_{i=1}^{V} \sum_{v=1}^{V} y_{l,i} | h_{l,k} = k, \quad \eta_{l,i,v} = \sum_{i=1}^{V} \sum_{v=1}^{V} y_{l,i} | h_{l,k} = 0$$

Note that we use $y_{l,i}$ to denote the frequency of observing the lab test $l$ of outcome $v$ for patient $j$ and $\eta_{l,i,v} = v$ as binary indicator of a single test. Detailed derivations are in Supplementary Methods.

**Joint collapsed variational Bayesian inference.** The marginal likelihood can be approximated by evidence lower bound (ELBO):

$$\log p(x, y, t|a, \beta, \xi, a, b) = \log \sum_{x, h} p(x, y, t, z, h|a, \beta, \xi, a, b) q(z, h)$$

$$= \sum_{x, h} q(z, h) \log p(x, y, t, z, h|a, \beta, \xi, a, b) - \sum_{x, h} q(z, h) \log q(z, h)$$

$$= \mathbb{E}_{Q(z, h)} \log p(x, y, t, z, h|a, \beta, \xi, a, b) - \mathbb{E}_{Q(z, h)} \log q(z, h)$$

Maximizing (6) with respect to the variational parameter $y_{l,i,v}^{(k)}$ and $h_{l,k}$ is equivalent to calculating the expectation of $z_{l,i}^{(k)} = k$ and $h_{l,k} = k$ with respect to all of the other latent variables:

$$\log q(z_{l,i}^{(k)}) = \mathbb{E}_{Q(z_{l,i})} \log p(x, z_{l,i})$$

Maximizing (7) with respect to the variational parameter $y_{l,i,v}^{(k)}$ and $h_{l,k}$ is equivalent to exponentiating and normalizing the distribution of $z_{l,i}^{(k)} = k$ and $h_{l,k} = k$ gives

$$y_{l,i,v}^{(k)} = \frac{\exp(\mathbb{E}_{Q(z_{l,i})} \log p(x, z_{l,i} = k))}{\sum_{k=1}^{K} \exp(\mathbb{E}_{Q(z_{l,i})} \log p(x, z_{l,i} = k))}$$

$$h_{l,k} = \frac{\exp(\mathbb{E}_{Q(h_{l,k})} \log p(y, h_{l,k} = k))}{\sum_{k=1}^{K} \exp(\mathbb{E}_{Q(h_{l,k})} \log p(y, h_{l,k} = k))}$$

We can approximate these expectations by first deriving the conditional distribution for $p(z_{l,i}^{(k)} = k|x, y, t)$ and $p(h_{l,k} = k|x, y, t)$ (Supplementary Methods) and then using Gaussian distributions to approximate the sufficient statistics by the summation of the variational parameters in zero-order Taylor expansion:

$$y_{l,i,v}^{(k)} \approx \left( a_{l,i} + \tilde{n}_{l,v}^{(k)} + \tilde{m}_{l,k}^{(k)} \right) \frac{\tilde{n}_{l,k}^{(k)} + \tilde{m}_{l,k}^{(k)}}{\sum_{k=1}^{K} \tilde{n}_{l,k}^{(k)} + \tilde{m}_{l,k}^{(k)}}$$

To infer $p(h_{l,k} = k|x, y, t)$, we will need to separately consider whether the lab test $l$ is observed or missing. In particular, we can infer the topic distribution $\lambda_{l,k} \propto \mathbb{E}_Q \log p(h_{l,k} = k|x, y, t)$ of an observed lab test $y_l$ at value $v$ as:

$$\lambda_{l,k} \propto \left( a_{l,k} + \tilde{n}_{l,k}^{(k)} + \tilde{m}_{l,k}^{(k)} \right) \frac{\tilde{n}_{l,k}^{(k)} + \tilde{m}_{l,k}^{(k)}}{\sum_{k=1}^{K} \tilde{n}_{l,k}^{(k)} + \tilde{m}_{l,k}^{(k)}}$$

For unobserved lab tests, we infer the joint distribution of latent topic and lab result $\pi_{l,k} \propto \mathbb{E}_Q \log p(h_{l,k} = k, y_l = v|x, y, t)$:

$$\pi_{l,k} \propto \left( a_{l,k} + \tilde{n}_{l,k}^{(k)} + \tilde{m}_{l,k}^{(k)} \right) \frac{\tilde{n}_{l,k}^{(k)} + \tilde{m}_{l,k}^{(k)}}{\sum_{k=1}^{K} \tilde{n}_{l,k}^{(k)} + \tilde{m}_{l,k}^{(k)}}$$

where the notation $\tilde{n}_{l,k}^{(k)}$ indicates the exclusion of variable index $l, f$. The sufficient statistics in the above inference have closed-form expression conditioned on the other latent variables:

$$\tilde{n}_{l,k}^{(k)} = \sum_{i=1}^{V} \sum_{v=1}^{V} \tilde{y}_{l,i,v} | h_{l,k} = k, \quad \tilde{n}_{l,k}^{(k)} = \sum_{i=1}^{V} \sum_{v=1}^{V} \tilde{y}_{l,i,v} | h_{l,k} = 0$$

Furthermore, we update the hyperparameters by maximizing the marginal likelihood under the variational expectations via empirical Bayes fixed-point updates:

$$\beta_{l,k} = \frac{a_{l,k} + \tilde{n}_{l,k}^{(k)} + \tilde{m}_{l,k}^{(k)}}{\sum_{k=1}^{K} \tilde{n}_{l,k}^{(k)} + \tilde{m}_{l,k}^{(k)}}$$

Other hyperparameters updates are similar. The learning algorithm therefore follows expectation-maximization: E-step infers $y_{l,i,v}^{(k)}, \lambda_{l,k}, \eta_{l,i,v}$; M-step updates the above sufficient statistics of model parameters and hyperparameters. Details and time complexity analysis are described in Supplementary Methods.

**Stochastic joint collapsed variational Bayesian inference.** To learn MixEHR from massive-scale EHR data, we propose a stochastic collapsed variational Bayesian (SCVB) algorithm. The main objective of our SCVB algorithm is to avoid keeping track of all of the latent topic assignments $y_{l,i,v}^{(k)}$ and $h_{l,k}$'s while maintaining accurate model inference. Specifically, we first calculate the intermediate variational updates from randomly sampled mini-batches of $D'$ patients:

$$\tilde{n}_{l,k}^{(k)} = \frac{\sum_{i=1}^{D'} \sum_{v=1}^{V} \tilde{y}_{l,i,v} | h_{l,k} = k, \tilde{n}_{l,k}^{(k)} = \sum_{i=1}^{D'} \sum_{v=1}^{V} \tilde{y}_{l,i,v} | h_{l,k} = 0$$

We then update the latent disease topics using natural gradients, which are the intermediate updates weighted by a fixed learning rate:

$$n_{l,k}^{(k)} = \left( 1 - \rho \right) n_{l,k}^{(k)} + \rho \tilde{n}_{l,k}^{(k)}$$

Details are described in Supplementary Methods. We observed that the SCVB algorithm achieves comparable performance as the full-batch learning with only a fraction of time and constant memory (Supplementary Fig. 2).

**Predictive likelihood.** To evaluate model learning and monitor empirical convergence, we performed fivefold cross-validation. For each patient in the validation fold, we randomly selected 50% of their EHR features to infer their disease mixtures and then used the other 50% of the features to evaluate the log predictive likelihood—a common metric to evaluate topic models:

$$\frac{1}{D} \sum_{l=1}^{D} \log \sum_{k=1}^{K} \tilde{y}_{l,i,v} | h_{l,k} = k, \tilde{n}_{l,k}^{(k)} = \sum_{l=1}^{L} \log \sum_{k=1}^{K} \tilde{y}_{l,i,v} | h_{l,k} = 0$$

where we inferred the variational expectations of the disease mixture and disease
topics as:
\[
\delta_{jk} = \frac{a_{jk} + \frac{1}{c}}{\sum_{k'} a_{jk} + \frac{1}{c}} \phi_{tjk}^{(i)},
\]
\[
\psi_{jk} = \frac{a_{jk} + \frac{1}{c}}{a_{jk} + \frac{1}{c} + \beta_{jk} + \frac{1}{c}} - \psi_{jk}\sum_{k'} a_{jk} + \frac{1}{c} \frac{\delta_{jk}}{\psi_{jk}}.
\]
Sensible models should demonstrate improved predictive likelihood on held-out patients. We evaluated the predictive log likelihood of models with \( K \leq 100 \), and ultimately set \( K = 75 \) as it gave the highest predictive likelihood (Supplementary Fig. 1). Nonetheless, models with \( K \) within the range between 50 and 100 have similar performance.

### Retrospectively predict EHR codes by a k-nearest neighbor approach

To impute missing data in an individual-specific way, we here describe a k-nearest neighbor approach. As illustrated in Supplementary Fig. 16a, the prediction can be divided into three steps:

- **Train MixEHR on training set to learn the EHR code by disease topic matrices \( \Psi \) across data types and infer the disease topic mixtures \( \theta_{ij}^{\text{mix}} \) for each training patient data point;**

- **To infer the probability of an unknown EHR code \( t \) for a test patient \( f \), use MixEHR and the learnt disease topic matrices \( \Psi \) to infer the disease topic mixture \( \theta_{ij}^{\text{mix}} \) for the \( k \) most similar training patients \( S_f \). Here, the patient–patient similarity matrix is calculated based on the Euclidean distance between their disease topic mixtures:**

\[
d_{ij} = \sqrt{\sum_{c}(\theta_{ij} - \theta_{jk})^2}
\]

(24)

where \( \theta_{ij} \) and \( \theta_{jk} \) are the mixture memberships of the test patient \( f \) and training patient \( j \), respectively. Finally, we take the average of the EHR code \( t \) over these \( k \)-nearest neighbor patients as the prediction for the target code \( t \) for test patient \( f \):

\[
x_{tf} = \frac{1}{k} \sum_{j} x_{tjk}\]

(25)

We empirically determined the number of nearest neighbors \( k \) to be 100.

### Evaluating retrospective EHR code prediction

To evaluate the prediction performance, we chose a set of EHR codes from each data category based on the following criteria. For clinical notes, we chose the key words that are recorded in more than 10 patients but fewer than 500 patients to focus on disease-specific words with sufficient number of patients. For other data categories, we chose target codes that are observed in test patient data type for at least 100 patients but no more than 50% of the training cohort \( (\sim 19,000 \text{ patients}) \). We then selected the top 100 codes per data type in terms of the number of patients except for the ICD-9 diagnostic code, for which we evaluated prediction on each of the 499 codes in order to assess the performance differences by ICD-9 disease groups (Supplementary Fig. 1B). In total, we obtained 976 target EHR codes (Supplementary Table 4).

For unbiased evaluation, we performed a fivefold cross-validation. For training fold \( n \), we trained MixEHR to learn the disease topic matrices \( \Psi \) from training fold \( n \) and patient disease topic mixtures \( \theta_{ij}^{\text{mix}} \) in training fold \( n \). For the corresponding validation fold, we evaluated the prediction of each target code \( t \) for each test patient \( f \) as follows. We first removed the target code \( t \) from the test patient \( f \) if it is observed for the test patient \( f \). We then inferred the mixture code for the test patient \( f \) and inferred the probability of target code \( t \) for test patient \( f \) using the \( k \)-nearest neighbor approach. We repeated this procedure for the five validation folds to obtain the predictions of each target EHR code.

For comparison, we evaluated the performance of the baseline topic model which ignores distinct data types in the EHR data. Notably, such model is essentially equivalent to LDA. To this end, we modified the training and validation datasets to have the same data type for all of the 53432 EHR codes and gave each code a unique ID \( (\text{i.e.,} 1 \text{ to } 53,432) \). We then extracted the 976 target codes for fivefold cross-validation predictions. We ran MixEHR on such data-type-flattened data. We evaluated the predictions based on area under the ROC curve (AUROC) and precision–recall curve (AUPRC) as well as the overall accuracy by thresholding the prediction probability by 1/\( k \), where \( k = 100 \) is the empirically determined number of nearest neighbors.

### Implementations of the existing methods

For experiments comparing our method with GRAM\(^{22} \) and Doctor AI\(^{25} \), we used the open source implementation made available by the authors of the papers found at https://github.com/mp2893/gram and https://github.com/mp2893/dotoral, respectively. GRAM and Doctor AI were both set to predict the medical codes (as defined in their papers) in the next admission for the MIMIC-III. This task is also known as sequential diagnosis prediction. The same Classical Classification Software (CCS) medical codes used in the original implementation of both methods were used in the comparisons as well. Note that Doctor AI can also be used to predict the time duration of next visit, but we did not compare our method with it in this aspect. For both Doctor AI and GRAM, we used the default recommended settings provided in their code bases. Both our models were implemented in Theano. To compare both our models with Doctor AI and GRAM, we used Accuracy@20 as described in the GRAM paper for different percentile of frequencies of the medical codes.

In our comparison with Deep Patient, since there was no published code available, we implemented the most recent version of the code and evaluated its performance according to the user guide of the code. We closely followed the steps mentioned in the paper by the authors. We did not follow the elaborate data preprocessing pipeline used in the paper as our dataset (MIMIC-III) is different from the one used by\(^{29} \). Our implementation had three layers of autoencoders with 500 hidden units (as in the original paper) and was coded using Keras library. We used the Deep Patient representation as input to a classifier \( \text{(i.e., logistic regression)} \) to predict mortality and compared the accuracy with MixEHR in terms of the area under the precision–recall (AUPRC) and area under the ROC (AUROC) (Supplementary Fig. 23).

Conditioned Factored Restricted Boltzmann Machine (CF-RBM) was implemented by adapting the code from https://github.com/elip2/cfrbm. CF-RBM was originally designed for Netflix movie recommendation. Each movie has a five-star rating. We modified the code to make it work with two-state lab results \( \text{(i.e., normal and abnormal)} \). We used 100 hidden units of CF-RBM to train on the randomly sampled 80% of the admissions and tested the CF-RBM on the remaining 20% of admissions. We ensured that the training and test sets were the same for both CF-RBM and MixEHR. We evaluated the trained CF-RBM as follows. We iterated over the admissions in the test set, and for every lab test in that admission, we masked its test result and made a prediction based on the remaining lab tests of the same admission. This procedure was repeated for every lab test in every admission. We then recorded the predicted lab results and the true lab results for evaluation purpose. The same evaluation procedure was also used for MixEHR.

SiCNMF was implemented using the Python code from author’s Github repository https://github.com/sugnasekar/SiCNMF\(^{40} \). The model was trained on the MIMIC-III dataset with six data types as the six collective matrices until convergence.

Unless mentioned otherwise, all of the other unsupervised learning and classification methods such as PCA, random forest, elastic net, logistic regression, and LDA were implemented using the Scikit-Learn Python library.

### MIMIC-III data description

The methods were performed in accordance with the PhysioNet data user agreement. All of the MIMIC-III data received simple and minimal preprocessing of the raw comma separated value (CSV) files provided from MIMIC-III critical care database (mimic.physionet.org). For all data types, we removed nonspecific EHR code that were commonly observed among admissions based on their inflection points (Supplementary Fig. 26). The inflection point was defined as the percentile of frequencies of the medical codes.

For unbiased evaluation, we performed a fivefold cross-validation. For training fold \( n \), we trained MixEHR to learn the disease topic matrices \( \Psi \) from training fold \( n \) and patient disease topic mixtures \( \theta_{ij}^{\text{mix}} \) in training fold \( n \). For the corresponding validation fold, we evaluated the prediction of each target code \( t \) for each test patient \( f \) as follows. We first removed the target code \( t \) from the test patient \( f \) if it is observed for the test patient \( f \). We then inferred the mixture code for the test patient \( f \) and inferred the probability of target code \( t \) for test patient \( f \) using the \( k \)-nearest neighbor approach. We repeated this procedure for the five validation folds to obtain the predictions of each target EHR code.

For comparison, we evaluated the performance of the baseline topic model which ignores distinct data types in the EHR data. Notably, such model is essentially equivalent to LDA. To this end, we modified the training and validation datasets to have the same data type for all of the 53432 EHR codes and gave each code a unique ID \( (\text{i.e.,} 1 \text{ to } 53,432) \). We then extracted the 976 target codes for fivefold cross-validation predictions. We ran MixEHR on such data-type-flattened data. We evaluated the predictions based on area under the ROC curve (AUROC) and precision–recall curve (AUPRC) as well as the overall accuracy by thresholding the prediction probability by 1/\( k \), where \( k = 100 \) is the empirically determined number of nearest neighbors.

### Mayo Clinic data description

The 187 patients \( (94 \text{ bipolar disorder cases and } 93 \text{ controls}) \) were selected from the Mayo Clinic Bipolar Disorder Biobank and the Mayo Clinic Biobank, respectively. They had consented for research, including research using their EHR data\(^{27,48} \). This study was reviewed and approved by Mayo Clinic Institutional Review Board and by the access committees from both biobanks. The EHR information for each of the patients was extracted from both databases including ICD-9 codes (2224 ICD-9-CM codes), procedure codes (2200 CPT codes), patient provided information (955 text questions), lab tests (1742 lab test codes), medication prescriptions (610 prescription order codes). In total, there are 7731 codes and 108,390 observations (Supplementary Table 5). For each case, one subject from the Mayo Clinic Biobank was matched on age and sex, after excluding subjects with psychiatric conditions in self-reports and/or EHR. Supplementary Table 6 summarizes age-sex distribution by cases and controls.
Quebec CHD data description. The dataset was derived from the EHR documented Quebec Congenital Heart Disease (CHD) database with 84,498 patients and 288,101 EHRs from 1983 to 2010. We use the study universe databases as data sources. No participant is enrolled. The study was approved by Research Ethics Board of McGill University Health Centre. For each patient, the patient history can be traced back to the later of birth or 1983, and followed up to the end of the study or death of the patient, whichever came earlier. The data set includes two data types: ICD-9 or ICD-10 billing code and intervention code. The raw dataset consists of individual records of each patient’s medical visit. Each visit lasted less than a day. Thus, they were not considered as admissions, and are categorized as outpatient data. The informed consent was obtained from all participants.

By using the time stamp of those visits, we aggregated the diagnoses and intervention codes by the months into visit blocks. Thus, each visit block represents the record of a patient in that time interval as a set of diagnoses and a set of interventions. Since the patients’ visit history was recorded as sequential events over the 28 years, we represented each patient’s complete medical history as a time series of diagnoses and intervention sets. The length of the sequence depends on the total number of months. We only kept the months in which a patient has visited a medical institution at least once. The diagnoses given were represented as a four- or five-digit ICD-9 or ICD-10 code. All of the codes were converted to ICD-9 using a board-approved conversion table. In order to generate the labels, we took the first three digits of each code. The processed data are summarized in Supplementary Table 3.

Bipolar disorder Mayo Clinic patients classification details. Among the five data categories, lab tests and PPI have two separate data features unlike the other three binary data categories. Specifically, they consist of (1) presence vs. absence of the data in the EHR (e.g., whether a lab test was done or whether a patient answered a particular PPI question) and (2) the actual value for the variable, i.e., test results for a lab test, or patient’s answer for a PPI question. Given the small sample size, we chose to model the first data feature (whether the data exists in the EHR) to mimic the data structure used in the rest of the data categories, recognizing that we may miss important information by not analyzing the actual lab test results or the answers for particular PPI questions. We speculate that the missing pattern may be related to the disease classification and disease subgroups, especially for the lab codes. However, we note that the missing pattern for the PPI data might be strongly confounded by other nonclinical characteristics such as age, sex, and social determinants of health.

We randomly split the 80,000 patients into 80% for training and 20% for the patients of testing. For MixEHR+RNN, we first trained a baseline RNN model, we used a 20-topic MixEHR to model the distinct distribution of these six data types. Our preliminary analysis showed that more than 20 topics produced degenerate models. First, we sought to quantitatively examine whether the patient topic mixtures provide useful information for accurate classification of bipolar disorder. To this end, we divided the 187 patients into five folds (Fig. 4a). We trained MixEHR on the four folds and then a logistic regression (LR) classifier that uses the patient topic mixture in the four folds to predict the bipolar disorder label. Given the small sample size, we chose to model the first data feature (whether the data exists in the EHR) to mimic the data structure used in the rest of the data categories, recognizing that we may miss important information by not analyzing the actual lab test results or the answers for particular PPI questions. We speculate that the missing pattern may be related to the disease classification and disease subgroups, especially for the lab codes. However, we note that the missing pattern for the PPI data might be strongly confounded by other nonclinical characteristics such as age, sex, and social determinants of health.

We evaluated the performance in terms of the code prediction improvement. For the ease of reference, we call this approach as MixEHR+RNN to distinguish it from the baseline RNN that takes only the MixEHR as input (i.e., Doctor AI). For the baseline, we used 100-dimensional latent embedding (i.e., fully connected dense layer) to encode the observed EHR code in the input. For the MixEHR+RNN model, we used 50-dimensional latent embedding to encode the observed EHR code augmented with 50-topic MixEHR for the topic mixture embedding (Supplementary Fig. 21). The results are presented in Table 2, Supplementary Table 2, and Figure 7a.

We then compared our model to the clinical gold standard, i.e., the true diagnostic code. For evaluation, we randomly split the 80,000 patients into 80% of the patients for training and 20% of the patients for testing. For MixEHR+RNN, we first trained a baseline RNN model, we used a 20-topic MixEHR to model the distinct distribution of these six data types. Our preliminary analysis showed that more than 20 topics produced degenerate models. First, we sought to quantitatively examine whether the patient topic mixtures provide useful information for accurate classification of bipolar disorder. To this end, we divided the 187 patients into five folds (Fig. 4a). We trained MixEHR on the four folds and then a logistic regression (LR) classifier that uses the patient topic mixture in the four folds to predict the bipolar disorder label. Given the small sample size, we chose to model the first data feature (whether the data exists in the EHR) to mimic the data structure used in the rest of the data categories, recognizing that we may miss important information by not analyzing the actual lab test results or the answers for particular PPI questions. We speculate that the missing pattern may be related to the disease classification and disease subgroups, especially for the lab codes. However, we note that the missing pattern for the PPI data might be strongly confounded by other nonclinical characteristics such as age, sex, and social determinants of health.

We evaluated the performance in terms of the code prediction improvement. For the ease of reference, we call this approach as MixEHR+RNN to distinguish it from the baseline RNN that takes only the MixEHR as input (i.e., Doctor AI). For the baseline, we used 100-dimensional latent embedding (i.e., fully connected dense layer) to encode the observed EHR code in the input. For the MixEHR+RNN model, we used 50-dimensional latent embedding to encode the observed EHR code augmented with 50-topic MixEHR for the topic mixture embedding (Supplementary Fig. 21). The results are presented in Table 2, Supplementary Table 2, and Figure 7a.

We then compared our model to the clinical gold standard, i.e., the true diagnostic code. For evaluation, we randomly split the 80,000 patients into 80% of the patients for training and 20% of the patients for testing. For MixEHR+RNN, we first trained a baseline RNN model, we used a 20-topic MixEHR to model the distinct distribution of these six data types. Our preliminary analysis showed that more than 20 topics produced degenerate models. First, we sought to quantitatively examine whether the patient topic mixtures provide useful information for accurate classification of bipolar disorder. To this end, we divided the 187 patients into five folds (Fig. 4a). We trained MixEHR on the four folds and then a logistic regression (LR) classifier that uses the patient topic mixture in the four folds to predict the bipolar disorder label. Given the small sample size, we chose to model the first data feature (whether the data exists in the EHR) to mimic the data structure used in the rest of the data categories, recognizing that we may miss important information by not analyzing the actual lab test results or the answers for particular PPI questions. We speculate that the missing pattern may be related to the disease classification and disease subgroups, especially for the lab codes. However, we note that the missing pattern for the PPI data might be strongly confounded by other nonclinical characteristics such as age, sex, and social determinants of health.

We evaluated the performance in terms of the code prediction improvement. For the ease of reference, we call this approach as MixEHR+RNN to distinguish it from the baseline RNN that takes only the MixEHR as input (i.e., Doctor AI). For the baseline, we used 100-dimensional latent embedding (i.e., fully connected dense layer) to encode the observed EHR code in the input. For the MixEHR+RNN model, we used 50-dimensional latent embedding to encode the observed EHR code augmented with 50-topic MixEHR for the topic mixture embedding (Supplementary Fig. 21). The results are presented in Table 2, Supplementary Table 2, and Figure 7a.

We then compared our model to the clinical gold standard, i.e., the true diagnostic code. For evaluation, we randomly split the 80,000 patients into 80% of the patients for training and 20% of the patients for testing. For MixEHR+RNN, we first trained a baseline RNN model, we used a 20-topic MixEHR to model the distinct distribution of these six data types. Our preliminary analysis showed that more than 20 topics produced degenerate models. First, we sought to quantitatively examine whether the patient topic mixtures provide useful information for accurate classification of bipolar disorder. To this end, we divided the 187 patients into five folds (Fig. 4a). We trained MixEHR on the four folds and then a logistic regression (LR) classifier that uses the patient topic mixture in the four folds to predict the bipolar disorder label. Given the small sample size, we chose to model the first data feature (whether the data exists in the EHR) to mimic the data structure used in the rest of the data categories, recognizing that we may miss important information by not analyzing the actual lab test results or the answers for particular PPI questions. We speculate that the missing pattern may be related to the disease classification and disease subgroups, especially for the lab codes. However, we note that the missing pattern for the PPI data might be strongly confounded by other nonclinical characteristics such as age, sex, and social determinants of health.

We evaluated the performance in terms of the code prediction improvement. For the ease of reference, we call this approach as MixEHR+RNN to distinguish it from the baseline RNN that takes only the MixEHR as input (i.e., Doctor AI). For the baseline, we used 100-dimensional latent embedding (i.e., fully connected dense layer) to encode the observed EHR code in the input. For the MixEHR+RNN model, we used 50-dimensional latent embedding to encode the observed EHR code augmented with 50-topic MixEHR for the topic mixture embedding (Supplementary Fig. 21). The results are presented in Table 2, Supplementary Table 2, and Figure 7a.

We then compared our model to the clinical gold standard, i.e., the true diagnostic code. For evaluation, we randomly split the 80,000 patients into 80% of the patients for training and 20% of the patients for testing. For MixEHR+RNN, we first trained a baseline RNN model, we used a 20-topic MixEHR to model the distinct distribution of these six data types. Our preliminary analysis showed that more than 20 topics produced degenerate models. First, we sought to quantitatively examine whether the patient topic mixtures provide useful information for accurate classification of bipolar disorder. To this end, we divided the 187 patients into five folds (Fig. 4a). We trained MixEHR on the four folds and then a logistic regression (LR) classifier that uses the patient topic mixture in the four folds to predict the bipolar disorder label. Given the small sample size, we chose to model the first data feature (whether the data exists in the EHR) to mimic the data structure used in the rest of the data categories, recognizing that we may miss important information by not analyzing the actual lab test results or the answers for particular PPI questions. We speculate that the missing pattern may be related to the disease classification and disease subgroups, especially for the lab codes. However, we note that the missing pattern for the PPI data might be strongly confounded by other nonclinical characteristics such as age, sex, and social determinants of health.

We evaluated the performance in terms of the code prediction improvement. For the ease of reference, we call this approach as MixEHR+RNN to distinguish it from the baseline RNN that takes only the MixEHR as input (i.e., Doctor AI). For the baseline, we used 100-dimensional latent embedding (i.e., fully connected dense layer) to encode the observed EHR code in the input. For the MixEHR+RNN model, we used 50-dimensional latent embedding to encode the observed EHR code augmented with 50-topic MixEHR for the topic mixture embedding (Supplementary Fig. 21). The results are presented in Table 2, Supplementary Table 2, and Figure 7a.
To further compare our proposed approach, we took a simpler imputation strategy by averaging. For each lab test, we took the average of the observed lab results over the training admissions. We then used the average frequency as the predictions for the lab results on the 20% testing admissions.

MIMIC-III mortality prediction details. We first trained MixEHR on the ~39K patients with single-admissions to learn the disease topics. We used the trained MixEHR to infer topic mixture of each admission for the patients who had more than one admission. We took the last two admissions that are within 6 months apart, which gave us around 4000 multi-admission patients. We used the topic mixture inferred for the second-last admission as an input to a classifier to predict the mortality outcome at the discharge of the last admission (i.e., as indicated by the EXPIRED flag in the Discharge location of the last admission). To realistically evaluate our model, we filtered out Discharge Summary from the clinical notes in all of the data (under CATEGORY of the notes) because they usually contain conclusion about patients’ critical conditions. We then performed a 5-fold cross-validation to evaluate the prediction accuracy. We experimented with K = 50 and 75 topics. As baseline models, we tested LDA on a data types, Principal Component Analysis (PCA) with 50 Principal Components (PCs), Sparsity-inducing Collected Non-negative Matrix Factorization (SiCNMF) on the same input data16. We ran SiCNMF with the default settings with lower rank set to 20. Lastly, we trained elastic net directly on the raw input data from the ~4000 patients’ second-last admission to predict the mortality in their last admissions upon discharge.

Moreover, we designed another experiment that used the earliest/first admission to predict the mortality in the last admission. We removed patients if their earliest admission lasted longer than 48 h based on the difference between the admission time and discharge time within the same admission. This also gave us 4040 patients whose first in-hospital stay was shorter than 2 days. Same as the above, we performed a fivefold CV on these patients as above using the pre-trained unsupervised models on the single-admission patients’ data to generate topic mixtures on the first admissions as input to a supervised classifier (i.e., elastic net), which then predicts the mortality in the last admission upon discharge.

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

Data availability

The MIMIC-III data analyzed in this paper are publicly available through PhysioNet (http://mimic.physionet.org). Mayo Clinic and Quebec CHD data are not publicly accessible due to restricted user agreement.

Code availability

We implemented MixEHR in C++ as a standalone Unix command-line software using OpenMP for multi-threaded inference. It allows an arbitrary number of discrete data types and discrete states per EHR feature. The software is available as Supplementary Software 1 or at https://github.com/llab-mcgill/mixehr.

Received: 23 July 2019; Accepted: 23 April 2020;
Published online: 21 May 2020

References

1. Johnson, A. E. W. et al. MIMIC-III, a freely accessible critical care database. Sci. Data 3, 160035–1600359 (2016).
2. Charles, D., Gabriel, M. & Furukawa, M. F. Adoption of electronic health record systems among US non-federal acute care hospitals: 2008-2012. ONC Data Brief, 9, 1–9 (2013).
3. Henry, J., Pylypchuk, Y., Searcy, T. & Patel, V. Adoption of electronic health record systems among US non-federal acute care hospitals: 2008-2015. ONC data brief, 35, 1–9 (2016).
4. Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022 (2003).
5. Salakhutdinov, R., Mnih, A. & Hinton, G. Restricted Boltzmann machines for collaborative filtering. Proceedings of the 24th International Conference on Machine Learning. In ACM Press (eds. Ghahramani, Z.) 791–798 (New York, 2007).
6. Mnih, A. & Salakhutdinov, R. R. Probabilistic matrix factorization. Advances in Neural Information Processing Systems, In MIT Press (eds Platt, J. C., Koller, D., Singer, Y. & Roweis, S. T.) 1257–1264 (2008).
7. Hernandez-Iobar, J. M., Houbly, N. & Ghahramani, Z. Probabilistic matrix factorization with non-random missing data. Proceedings of the 31th International Conference on Machine Learning. In JMLR (eds Xing, E. P. & Jебraz, T. S.) 32: 1512–1520 (2014).
8. Marin, B. M. & Zemel, R. S. Collaborative prediction and ranking with non-random missing data. The Proceedings of the third ACM conference on Recommender systems. In ACM Press (eds Burke, R., Felfernig, A. & Schmidt-Thieme. L.) 5–12 (New York, 2009).
9. Fraser, G. & Yan, R. Collaborative filtering and the missing at random assumption. Empir. Softw. Eng. 19, 1–9 (2014).
10. Mcaliffe, J. D. & Blei, D. M. Supervised topic models. In Advances in Neural Information Processing Systems (eds Platt, J. C., Koller, D., Singer, Y. & Roweis, S. T.) Vol. 20, 121–128 (Curran Associates., Inc., 2008).
11. Blei, D. M. Probabilistic topic models. Commun. ACM 55, 77–84 (2012).
12. Griffiths, T. L. & Steyvers, M. Finding scientific topics. Proc. Natul. Acad. Sci. USA 103, 15026–15030 (2006).
13. Teh, Y. W., Newman, D. & Welling, M. A collapsed variational bayesian inference algorithm for latent dirichlet allocation. In Advances in Neural Information Processing Systems (eds Schöklop, B., Platt, J. C. & Hoffman, T.) Vol. 19, 1533–1560 (MIT Press, 2007).
14. Asuncion, A., Welling, M., Smyth, P. & Teh, Y. W. On smoothing and inference for topic models. Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence UAI’09, In AUAI Press (eds Bühles, J. & Ng. A.) 27–34 (Arlington, VA, 2009).
15. Little, R. J. A. & Rubin, D. B. Statistical Analysis with Missing Data (John Wiley & Sons, 2014).
16. Halpern, Y., Horng, S. & Sontag, D. (eds Doshi-Velez, F., Fackler, J., Kale, D., Wallace, R. & Wiens, J.) in Proceedings of the 1st Machine Learning for Healthcare Conference. PMLR 56, 209–225 (2016).
17. Joshi, S., Gunasekar, S., Sontag, D. & Ghosh, J. Identifiable phenotyping using constrained non-negative matrix factorization. Proceedings of the 1st Machine Learning for Healthcare Conference. In PMLR (eds Doshi-Velez, F., Fackler, J., Kale, D., Wallace, R. & Wiens, J.) in Proceedings of the 1st Machine Learning for Healthcare Conference. PMLR 56, 209–225 (2016).
18. Pivovarov, R. et al. Learning probabilistic phenotypes from heterogeneous EHR data. J. Biomed. Inform. 58, 156–165 (2015).
19. Gunasekar, S. et al. Phenotyping using structured collective matrix factorization of multi-source EHR data. Preprint at https://arxiv.org/abs/1609.04466 (2016).
20. Flaherty, M., Giaever, G., Kumm, J., Jordan, M. I. & Arkin, A. P. Latent variable model for chemogenomic profiling. Bioinformatics 21, 3268–3269 (2005).
21. Zhao, J. et al. Detecting time-evolving phenotypic topics via tensor factorization on electronic health records cardiovascular disease case study. J. Biomed. Inform. 98, 103270 (2019).
22. Wang, Y. et al. Unsupervised machine learning for the discovery of latent disease clusters and patient subgroups using electronic health records. J. Biomed. Inform. 102, 103364 (2020).
23. Wang, L., Tong, L., Davis, D., Arnold, T. & Esposito, T. The application of unsupervised deep learning in predictive models using electronic health records. BMJ Med. Res. 20, 1–9 (2020).
24. Razavian, N. & Sontag, D. Temporal convolutional neural networks for diagnosis from lab tests. Preprint at https://arxiv.org/abs/1511.07938 (2015).
25. Cheng, Y., Wang, F., Zhang, P. & Hu, J. Risk prediction with electronic health records: a deep learning approach. In 2016 SIAM International Conference. (eds Venkatasubramanian, S. & Meira, W.) 432–440 (SIAM, 2016).
26. Lipton, Z. C., Kale, D. C., Elkan, C. & Weitk, R. R. Learning to diagnose with LSTM recurrent neural networks. Preprint at https://arxiv.org/abs/1511.03677 (2015).
27. Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F. & Sun, J. Doctor AI: predicting clinical events via recurrent neural networks. JMLR Workshop Conf. Proc. 56, 301–318 (2016).
28. Nguyen, P., Tran, T., Wickramasinghe, N. & Venkatesh, S. Deepr: A Convolutional Net for Medical Records. In IEEE Journal of Biomedical and Health Informatics. 21, 22–30 (2017).
29. 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. Sci. Rep. 6, 1–10 (2016).
30. Suresh, H., Szolovits, P. & Ghassami, M. The use of autoencoders for discovering patient phenotypes. Preprint at https://arxiv.org/abs/1703.07004 (2017).
31. Rajkomar, A. et al. Scalable and accurate deep learning with electronic health records. npj Digital Med. 1, 18 (2018).
32. Choi, E., Bahadori, M. T., Song, L., Stewart, W. F. & Sun, J. GRAM: graph-based attention model for healthcare representation learning. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. In ACM Press (eds Matwin, S., Yu, S. & Farooq, F.) 787–795 (Sutter Health, Sacramento, New York, NY, 2017).
33. Kuchibhotla, K., Bergeron, A., Wedman, J., Poritz-Katz, B. S. & Abarbanel, J. M. Neuropsychology of vitamin B12 deficiency in elderly dementia patients and control subjects. J. Geriatr. Psychiatry Neurol. 18, 33–38 (2005).
34. van Buuren, S. & Groothuis-Oudshoorn, K. Mice: multivariate imputation by chained equations in R. J. Stat. Softw. 45, 1–67 (2011).
35. Choi, E. et al. Multi-layer representation learning for medical concepts. In ACM SIGKDD International Conference. In ACM Press (eds Aggarwal, C. & Smola, A.) 1495–1504 (New York, NY, 2016).
36. Ho, J. C., Ghosh, J. & Sun, J. Marble: high-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD ’14. In ACM Press (eds MacKassy, S. & Perlich, C.) 115–124 (New York, NY, 2014).
37. Wang, Y., Chen, R. Ghosh, J. & Kho, A. Rubik: knowledge guided tensor factorization and completion for health data analytics. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. In ACM Press (eds Cao, L. & Zhang, C.) 265–1274 (New York, NY, 2015).
38. Schulam, P. & Saria, S. A framework for individualizing predictions of disease trajectories by exploiting multi-resolution structure. In Advances in Neural Information Processing Systems (eds Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M. & Garnett, R.) Vol. 28, 748–756 (Curran Associates, Inc., 2015).
39. Schulam, P. & Saria, S. Integrative analysis using coupled latent variable models for individualizing prognoses. J. Mach. Learn. Res. 17, 1–35 (2016).
40. Jordan, M. I., Ghahramani, Z., Jaakkola, T. S. & Saul, L. K. An introduction to variational methods for graphical models. Learning in Graphical Models. In MIT Press (eds Jordan, M., Kleinberg & J., Scholkopf, B.) 461–474 (2006).
41. Griffiths, T. L. & Steyvers, M. Finding scientific topics. Proc. Natl Acad. Sci. USA 101 (Suppl 1), 5228–5235 (2004).
42. Minka, T. Estimating a Dirichlet distribution. Technical Report (MIT, 2000).
43. Hoffman, M. D., Blei, D. M., Wang, C. & Paisley, J. W. Stochastic variational inference. J. Mach. Learn. Res. 14, 1303–1347 (2013).
44. Foulds, J., Boyles, L., Dubois, C., Smyth, P. & Welling, M. Stochastic collapsed variational bayesian inference for latent dirichlet allocation. Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. In ACM Press (eds Grossman, R. & Uthursamy, R.) 446–454 (New York, NY, 2013).
45. Feinerer, I., Hornik, K. & Meyer, D. Text mining infrastructure in R. J. Stat. Softw. 25, 1–54 (2008).
46. Frye, M. A. et al. Development of a bipolar disorder biobank: differential phenotyping for subsequent biomarker analyses. Int. J. Bipolar Disord. 3, 30 (2015).
47. Olson, J. E. et al. The Mayo Clinic Biobank: a building block for individualized medicine. Mayo Clin. Proc. 88, 952–962 (2013).
48. Marelli, A. J., Mackie, A. S., Ionescu-Ittu, R., Rahme, E. & Pilote, L. Congenital heart disease in the general population: changing prevalence and age distribution. Circulation 115, 163–172 (2007).
49. Marelli, A. J. et al. Lifetime prevalence of congenital heart disease in the general population from 2000 to 2010. Circulation 130, 749–756 (2014).

Acknowledgements

Y.L. is supported by Natural Sciences and Engineering Research Council (NSERC) Discovery Grant (RGPIN-2019-0621), Fonds de recherche Nature et technologies (FRQNT) New Career (NC-268592), and Canada First Research Excellence Fund Healthy Brains for Healthy Life (HBHL) initiative New Investigator award (G249591). The funding for the Mayo Clinic bipolar disorder study is provided by Marriott Foundation. Mayo Clinic Center is funded by Individualized Medicine. The Quebec CHD Database is funded by Dr. Marelli’s Canadian Institute of Health Research Foundation Grant Award #35223. We thank Mathieu Blanchette for the helpful comments on the paper.

Author contributions

Y.L. and M.K. conceived the initial idea. Y.L. developed and implemented the method, processed and analyzed the data, and wrote the initial draft of the paper. P.N. ran the experiments on the EHR code prediction and mortality prediction using MIMIC-III data. Y.W. and A.A.K.D. helped running part of the lab imputation experiments. T.O., J.M.B., J.E.O., and M.A.F. provided the Mayo Clinic dataset. E.R. performed the data extraction and de-identification of the EHR data for both the bipolar and the control cohorts. W.L., Y.M., and Z.W. helped running the Mayo Clinic part of the experiment. T.O., J.M.B., and E.R. helped analyze the results. A.M. provided the data from Quebec CHD Database. A.L. and L.G. processed the data. X.H.L. implemented the MixEHR+RNN model and ran the experiments on the Quebec CHD Database with the help from Z.W. A.L. and A.M. helped interpreting the results from the Quebec CHD Database. Y.A. and J.D. performed the initial data processing and helped interpreted the initial results. M.K. edited the initial draft of the paper. Y.L. and M.K. supervised the project. All of the authors contributed to the final writing of the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41467-020-16378-3.

Correspondence and requests for materials should be addressed to Y.L. or M.K.

Peer review information Nature Communications thanks Yanshan Wang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

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.

© The Author(s) 2020.