Journal of Biomedical Informatics 102 (2020) 103364

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Unsupervised machine learning for the discovery of latent disease clusters and patient subgroups using electronic health records

Yanshan Wang[a,*], Yiqing Zhao[a], Terry M. Therneau[b], Elizabeth J. Atkinson[b], Ahmad P. Tafti[a], Nan Zhang[a], Shreyasee Amin[c], Andrew H. Limper[d], Sundeep Khosla[e], Hongfang Liu[a]

[a] Division of Digital Health Sciences, Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA
[b] Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA
[c] Division of Rheumatology, Department of Medicine, Mayo Clinic, Rochester, MN, USA
[d] Division of Pulmonary and Critical Care Medicine, Department of Internal Medicine, Mayo Clinic, Rochester, MN, USA
[e] Division of Endocrinology and Kogod Center on Aging, Department of Internal Medicine, Mayo Clinic, Rochester, MN, USA

ARTICLE INFO

Keywords:
Unsupervised Machine learning
Artificial intelligence
Electronic health records
Epidemiology
Aging

ABSTRACT

Machine learning has become ubiquitous and a key technology on mining electronic health records (EHRs) for facilitating clinical research and practice. Unsupervised machine learning, as opposed to supervised learning, has shown promise in identifying novel patterns and relations from EHRs without using human created labels. In this paper, we investigate the application of unsupervised machine learning models in discovering latent disease clusters and patient subgroups based on EHRs. We utilized Latent Dirichlet Allocation (LDA), a generative probabilistic model, and proposed a novel model named Poisson Dirichlet Model (PDM), which extends the LDA approach using a Poisson distribution to model patients' disease diagnoses and to alleviate age and sex factors by considering both observed and expected observations. In the empirical experiments, we evaluated LDA and PDM on three patient cohorts, namely Osteoporosis, Delirium/Dementia, and Chronic Obstructive Pulmonary Disease (COPD)/Bronchiectasis Cohorts, with their EHR data retrieved from the Rochester Epidemiology Project (REP) medical records linkage system, for the discovery of latent disease clusters and patient subgroups. We compared the effectiveness of LDA and PDM in identifying disease clusters through the visualization of disease representations. We tested the performance of LDA and PDM in differentiating patient subgroups through survival analysis, as well as statistical analysis of demographics and Elixhauser Comorbidity Index (ECI) scores in those subgroups. The experimental results show that the proposed PDM could effectively identify distinguished disease clusters based on the latent patterns hidden in the EHR data by alleviating the impact of age and sex, and that LDA could stratify patients into differentiable subgroups with larger p-values than PDM. However, those subgroups identified by LDA are highly associated with patients' age and sex. The subgroups discovered by PDM might imply the underlying patterns of diseases of greater interest in epidemiology research due to the alleviation of age and sex. Both unsupervised machine learning approaches could be leveraged to discover patient subgroups using EHRs but with different foci.

1. Introduction

The rapid adoption of electronic health records (EHRs) has enabled the use of the EHR data for primary and secondary purposes, such as clinical process optimization, clinical decision support, treatment outcome improvement, clinical research, and epidemiological monitoring of the nation's health [1]. An emerging use on the EHR data is to develop advanced machine learning models, primarily supervised learning, to discover new interconnections between diseases, facilitate precise prediction of health status, and help effectively prevent diseases or disabilities [2–5]. As opposed to supervised learning, unsupervised machine learning has been introduced to identify new patterns and relations in the irregularly-sampled data without using human created labels [6,7], mostly for predictive modeling, such as the prediction of patient health status [3], disease progression trajectory prediction [8], or phenotypes prediction [9,10]. In this paper, we investigate the use of unsupervised machine learning in the discovery of latent disease clusters and patient subgroups using EHRs.

The problem of discovering potential disease clusters and patient subgroups is extremely important for the study of aging. According to

* Corresponding author.
E-mail addresses: Wang.Yanshan@mayo.edu (Y. Wang), Liu.Hongfang@mayo.edu (H. Liu).

https://doi.org/10.1016/j.jbi.2019.103364
Received 14 May 2019; Received in revised form 16 December 2019; Accepted 23 December 2019
Available online 28 December 2019

the United Nations Population Division, the global share of older people (⩾60-year-old) increased from 8% in 1950 and 9% in 1990 to 12 % in 2013, and will continue to grow to an estimated 21% in 2050 [11]. The development of chronic illness plays an important role in this demographic shift. The older people who survive with chronic illnesses are more likely to develop additional chronic illnesses [12]. Traditionally, most comorbidities have been studied separately [13], however, it is common for most older people to have two or more chronic morbidities. Therefore, discovering disease clusters will help in the systematic examination of all comorbidities for associations with a specific condition and improve risk assessment and future prediction. Moreover, identifying new disease clusters that may reflect underlying mechanisms ("latent traits") would help define new domains of risk in a population. Discovering patient subgroups with the similar underlying disease patterns could facilitate diagnosis and treatment decision making, and epidemiological analysis and research [14].

In this study, we utilized Latent Dirichlet Allocation (LDA), an unsupervised generative probabilistic model, and proposed a novel model named Poisson Dirichlet Model (PDM), which extends the LDA approach for the EHR data. The proposed PDM uses a Poisson distribution to model patients' disease diagnoses by considering both observed and expected observations that could alleviate the impact of age and sex for identifying differentiated patient subgroups. In the experiments, we evaluated LDA and PDM on the EHR data of three patient cohorts, namely Osteoporosis, Delirium/Dementia, and Chronic Obstructive Pulmonary Disease (COPD)/Bronchiectasis Cohorts, retrieved from the Rochester Epidemiology Project (REP) medical records linkage system, for the discovery of latent disease clusters and patient subgroups. We compared the effectiveness of LDA and PDM in identifying latent disease clusters through the visualization of disease representations learned by two approaches in a two dimensional scattered plot. We tested the performance of LDA and PDM in differentiating patient subgroups through survival analysis, as well as statistical analysis of demographics and Elixhauser Comorbidity Index (ECI) scores in those subgroups.

2. Background

The creation of medical record linkage systems that connect EHR data from multiple institutions could capture the entire health care experience of a geographically defined population. The REP is a pioneer linkage system developed through a collaboration between health care providers in southeastern Minnesota, and involves Olmsted Medical Center, Mayo Clinic, Rochester Family Medicine Clinic and other medical care providers in southeastern Minnesota [15–17]. It is a unique infrastructure for epidemiology and outcomes research that links the medical records of local health care providers to community residents. Enabled by the REP, previous studies evaluated morbidity occurrences one diagnosis at a time [18] and used traditional analytic techniques (e.g., tree models/recursive partitioning) to define disease clusters [19]. However, these approaches do not adequately address the co-occurrence of multiple disease states within an individual.

A set of methods relevant to studying multimorbidities has arisen in the field of document processing under the rubric of "topic models". Latent Dirichlet Allocation (LDA) is a widely used topic modeling method proposed by Blei et al. [20]. LDA categorizes all words in a collection of textual documents into a set of distinct "topics", while simultaneously classifying each document by the topics it contains [21]. A given word may be associated with multiple topics, and multiple topics may appear in a given document. Since a patient could also be represented by a set of diagnosed diseases that share similar undiscovered interrelations, we simply used the analogy "words" = "diseases", "documents" = "patients", "topics" = "disease topics", and applied LDA to discover disease clusters in our previous work [22]. Although our previous study showed the potential of LDA in leveraging hidden pattern information from EHR data, we failed to observe dichotomized disease clusters from the results of LDA. Moreover, another shortcoming of LDA, especially in a cohort that spans a large age range, is that it will identify clusters that are due primarily to age and/or sex, disease associations that are already known and thus not very interesting. Examples would be athletic injuries or vaccinations in the young and joint ailments in the old. LDA uses a Multinomial distribution to simulate the generative process of each single disease, which leads LDA to focus on the proportions of various diseases in the cohort. Of greater interest in epidemiology research is the prediction or clustering of *excess* risk, event rates that are over and above what would be expected for a given age and sex. Furthermore, we didn't investigate the potential of topic models in discovering patient subgroups for a defined cohort. Stratifying patients into subgroups with similar characteristics and risks will not only facilitate epidemiological analysis and research, but enable personalized care that will improve the efficiency and effectiveness of disease prevention, diagnosis, and treatment.

3. Methods

3.1. Mathematical modeling

Suppose D denotes a disease diagnosis code set $D = \{d_1, d_2, ..., d_V\}$ with size V, a cohort C is represented as a group of M patients $C = \{\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_M\}$, and a patient $\mathbf{w}_m$ is represented as a sequence of N disease diagnoses $\mathbf{w}_m = (w_{m,1}, w_{m,2}, ..., w_{m,N})$ where $w_{m,n} \in D$, $n = 1, 2, ..., N$ is a disease diagnosis code from D for patient $\mathbf{w}_m$. Given these notations, we describe LDA and the proposed PDM in this subsection.

Let z denote the disease topics, which is akin to topics in LDA. Suppose K denotes the dimensionality of z, θ a K-dimensional Dirichlet random variable, α a K-dimensional parameter with $\alpha_i > 0$, and β a $K \times V$ matrix parameter, we can define LDA as the following generative process for each patient $\mathbf{w}$ in a cohort C:

1. For each of K disease topics:
   (a) Choose $\phi_k \sim$ Dirichlet(β)
2. For each of M patients in C:
   (a) Choose $\theta_m \sim$ Dirichlet(α).
   (b) For each of N diseases that patient $\mathbf{w}_m$ has been diagnosed:
      i. Choose a latent disease cluster $z_{m,n} \sim$ Multinomial(θ_m).
      ii. Choose a disease $w_{m,n} \sim$ Multinomial($\phi_{z_{m,n}}$).

LDA can be represented as a graphical model at the upper-left of Fig. 1. By applying Gibbs sampling, we can learn parameters α and β, and hyper-parameters θ and ϕ in LDA. We could leverage ϕ, the probability of a disease in a latent topic, to discover latent comorbidities that appear in the same disease cluster.

In this study, we also propose a novel unsupervised machine learning model, Poisson Dirichlet Model (PDM), which extends LDA for discovering disease clusters of excess risk. Based on a patient's age, sex, and length of follow-ups we can compute an expected number of diagnoses $e_{m,n}$ for subject m and diagnosis n, and compare the observed count $y_{m,n}$ to this expectation. We hypothesize that the PDM could be sensitive to patterns of excess disease risk, which will be different than overall risk, and that these patterns could identify more distinguished disease clusters than LDA.

With the same notations used for LDA, PDM assumes the following generative process for each patient in the cohort:

1. For each of K disease topics:
   (a) Choose $\phi \sim$ Dirichlet(β)
2. For each of M patients in C:
   (a) Choose $\gamma_m \sim$ Gamma(ξ, δ), where $E(\gamma) = \xi \cdot \delta = 1$.
   (b) Choose $\theta_m \sim$ Dirichlet(α).
   (c) For each of N diseases that patient $\mathbf{w}_m$ has been diagnosed:
      i. Choose a latent disease cluster $z_{m,n} \sim$ Multinomial(θ_m).

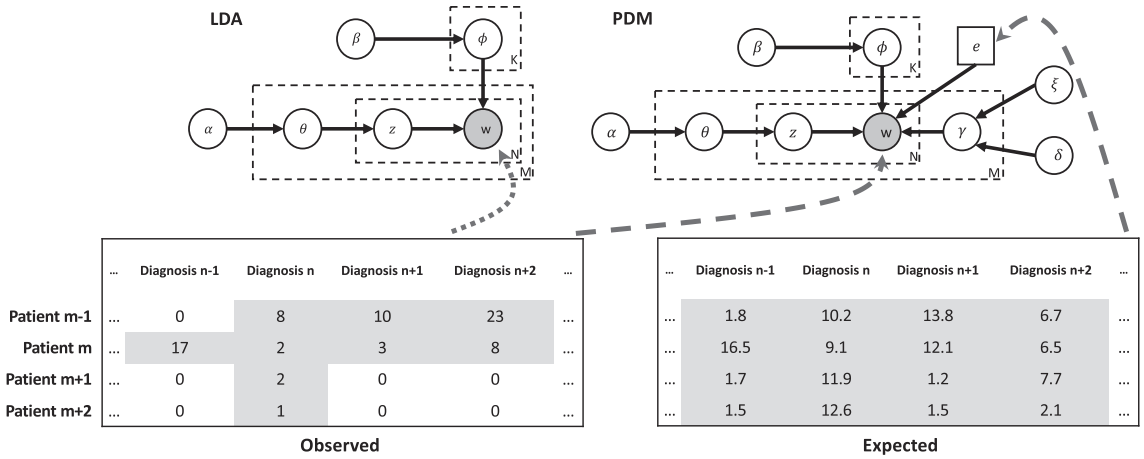

Fig. 1. The graphical model representation of LDA (upper-left) and PDM (upper-right). Circles represents random variables, gray-shaded circles represents observed states, the dashed plates represents replication over a set of variables, and the solid squares represent constants, which specify fixed-valued variables. *M* is the number of patients in a cohort, *N* is number of diagnosis for a patient, and *K* is the number of disease groups. The proposed PDM leverages both observed and expected number of diagnosis for each patient in the population, which alleviates the difficulty of LDA in dealing with missing data that were not collected during residents' absence in the healthcare system, particularly in the medical EHRs, as well as incorporating the epidemiological characteristics of the population.

ii. Choose a disease count $y_{m,n} \sim \text{Poisson}(\phi_{z_{m,n}} \cdot e_{m,n} \cdot \gamma_m)$.

The proposed PDM model is represented as a probabilistic graphical model at the upper-right of Fig. 1. In this generative process, $e_{m,n}$ denotes the expected number of diagnoses for disease $w_{m,n}$ that is computed by a simple rate model fit to the overall data. Specifically, the follow-ups for each subject in the cohort were divided into bins based on single years of age and sex. For each disease diagnosis, the counts (number of occurrences) were modeled using a generalized additive model separately for males and females, assuming a Poisson error structure. Age was fit using a smoothing spline with 4 degrees of freedom and the log of person-years in each bin were treated as an offset. The expected event rates for each person were estimated using predictions from these models based on the age, sex, and follow-up of each person. After dividing each patient's follow-up by age and sex, $e_{m,n}$ can be derived mathematically by:

$$e_{m,n} = \mathbb{E}(Y|X_{w_{m,n}}) = \beta + log(X_{1,w_{m,n}}) + X_{2,w_{m,n}}$$

where β is a coefficient, $log(X_{1,w_{m,n}})$ is an offset of the person-years, $X_{2,w_{m,n}}$ is age, and Y follows a Poisson distribution.

Poisson distribution is utilized to generate the total number of each disease that has been diagnosed for a patient. γ_m is a positive multiplier for patient m generated from a Gamma distribution with mean equal to 1. Since some patients may have a significantly larger number of diagnoses (e.g., sicker or better insurance) than others, γ_m is utilized as a normalizer on the number of patients' diagnoses so that the parameters of Poisson distribution are not learned towards the extreme cases.

In plain language, we use PDM to generate diseases using two steps for each patient in a cohort: (1) each patient has different portions of the disease topics with different proportion; (2) each disease a patient has been diagnosed with is drawn from one of the disease topics, where the selected disease topic is chosen from the per-document distribution over disease topics. PDM defines a disease topic to be a distribution over a fixed number of diagnosis codes representing diseases. For example, the rheumatology topic has the disease osteoarthritis (e.g., CCS 203: Osteoarthritis) with high probability while the psychiatry disorder topic has the bipolar disorder with high probability (e.g., CCS 657.1: Mood disorders: Bipolar disorders). We assume that the disease topics are specified and hidden from the data. PDM is a statistical model using the intuition that each patient with multi-morbidities exhibits multiple disease topics. Since PDM is a data-driven approach without using prior medical knowledge, the disease topics indicate the inner-connections of diseases hidden in the data that could be potentially used to uncover novel disease relations.

Fig. 1 also illustrates that LDA models the observed number of disease diagnoses while PDM takes advantage of the combination of observed and expected number of diagnoses. The use of expected data for each patient in PDM alleviates the impact of age and sex on finding different patient subgroups, and lessens the difficulty of LDA in handling missing data that were not collected during residents' absence in the healthcare system, particularly in medical EHRs, while simultaneously incorporates the epidemiological characteristics of the population. Since the experimental data are from Olmsted County, we used the expected risk table of the Minnesota population. Since $e_{m,n}$ is precalculated before training PDM, it is treated as a fixed-valued constant in the parameter learning process.

3.2. Parameter estimation

Markov Chain Monte Carlo (MCMC) methods are usually used to generate random samples that can be used in estimating the parameters of posterior distributions in a probabilistic machine learning model. MCMC constructs a Markov chain to converge to the target distribution, and generates samples from that Markov chain. Since the Dirichlet priors are conjugate to the Multinomial distributions, Gibbs sampling, a widely adopted MCMC algorithm, is utilized for inference of the LDA model [23]. However, it cannot be applied to PDM since the Poisson distributions are not conjugate to the Dirichlet priors. Therefore, a more general MCMC sampling method, Metropolis-Hastings (MH) algorithm [24], was applied to approximate the distributions and learn the parameters of PDM. The MH sampling algorithm creates a Markov chain based on a proposal distribution and corrects the wrong density through an acceptance-rejection step, in comparison with the Gibbs sampling algorithm that always accepts the proposal distribution. We implemented the parameter estimation algorithm using JAGS[1] and rJAGS[2], which automatically choose the proposal distribution for the sampling process. We utilized two sampling chains with a burn-in of 500 iterations followed by 1000 iterations for inference. The models

[1] http://mcmc-jags.sourceforge.net/
[2] https://cran.r-project.org/package=rjags.

were ran on a computing server with 80 Intel Xeon(R) E5-4650 2.4 GHz CPUs and a total of 794 Gb memory. The implementation codes for PDM are publicly available[3].

3.3. Applications on EHRs

3.3.1. Discovering latent disease clusters

Given a cohort of patients diagnosed with a certain disease, unsupervised machine learning approaches allow us to discover latent comorbidity clusters for that disease, which could help define new domains of risk. LDA and the proposed PDM model represent diseases in a latent topic space. The estimated parameter β in LDA or PDM is a disease-topic matrix indicating the probability of a disease occurring in a latent topic. We hypothesize that the diseases with similar characteristics would be automatically clustered in the latent topic space, which is called a latent disease cluster. In order to verify the effectiveness of LDA and PDM for identifying latent disease clusters, we qualitatively visualized the disease representation in the latent topic space using the disease-topic matrix by applying a machine learning visualization method, t-SNE [25], which mapped the disease representation in the high-dimensional topic space into a two-dimensional space that enables visualization. By doing so, we could qualitatively evaluate the potential of unsupervised machine learning in discovering disease clusters.

We adopted the Hopkins statistic [26] on the disease representations learned by LDA and PDM to quantitatively assess clustering results. The Hopkins statistic is a sampling test that measures cluster tendency by comparing the distances between disease data points and their nearest neighbors to the distances from a sample of pseudo data points. The value of Hopkins statistic ranges from 0 to 1 and a greater value indicates a better clustering method.

3.3.2. Discovering patient subgroups

An interesting question in the epidemiology study is whether we can stratify patients into subgroups so that the patients in the same subgroup have similar health characteristics and risks. Population-based evidence has been shown to be a major source of support for medical decision making for an individual [27]. Using the estimated parameters of LDA or PDM, we can calculate the posterior probability of a latent disease cluster for a given patient, i.e., $p(z|\mathbf{w}_m, \alpha, \beta) = \sum_i^N p(z|w_{m,i}, \alpha, \beta)$, which is a patient-topic probability matrix. Each row of this matrix performs like a feature vector for each patient projected by LDA and PDM that could facilitate effective patient subgroup clustering. In our experiments, we tested three clustering algorithms using these feature vectors, namely hierarchical clustering [28], K-means clustering [29], and Birch clustering algorithms [30], with five different numbers of subgroups, ranging from 2 to 6 subgroups.

To evaluate these subgroups, we carried out survival analysis on each patient subgroup. We used the Log-rank statistical test to compare the difference between the survival curves. In addition to the survival analysis, we conducted statistical analysis on the demographics and number of diagnoses for the subgroups. We also used the widely adopted comorbidity measure, Elixhauser Comorbidity Index (ECI) [31], to compare patient subgroups in each cohort. We compared ECI scores between the subgroups, and 29 ECI categories, including congestive heart failure, cardiac arrhythmias, valvular disease, pulmonary circulation disorders, peripheral vascular disease, hypertension, paralysis, other neurological disorders, chronic pulmonary disease, diabetes (uncomplicated), diabetes (complicated), hypothyroidism, renal failure, liver disease, peptic ulcer disease (excluding bleeding), lymphoma, metastatic cancer, solid tumour (without metastasis), rheumatoid arthritis/collaged vascular disease, coagulopathy, obesity, weight loss, fluid and electrolyte disorders, blood loss anaemia, deficiency anaemia, alcohol abuse, drug abuse, psychoses, and depression. Our goal is to evaluate whether the patient subgroups discovered by the proposed PDM model could differentiate patients in a defined cohort.

4. Datasets

In this section, we describe the datasets that have been used in the empirical experiments of evaluating LDA and the proposed PDM model. Three test cohorts extracted from the REP, namely the Osteoporosis Cohort, the Delirium/Dementia Cohort, and the Chronic Obstructive Pulmonary Disease (COPD)/Bronchiectasis Cohort, were utilized. These cohorts were retrieved from a REP aging cohort that consisted of a total of 72,000 patients who were 50 years of age and older during the interval January 1, 1995 through December 31, 2011 from Olmsted County. The number of diagnoses of the REP cohort ranges from 0 to 9198 (mean = 286, std = 416, see supplementary material for disease frequency distribution of the population). We utilized the patients from the REP cohort whose total number of disease diagnoses were between three hundred and five hundred and who had at least thirty diagnoses of osteoporosis (CCS code: 206), delirium/dementia (CCS code: 653), and COPD/bronchiectasis (CCS code: 127) for the three cohorts. Since this is a proof-of-concept study and we are not aiming to make clinical conclusions or statements, we arbitrarily chose these numbers to ensure that the cohorts had an appropriate number of patients and diagnoses so that the computation was feasible to verify the machine learning approaches. These diseases have been shown to be related to aging with complicated comorbidities. We used Clinical Classifications Software (CCS)[4] as the diagnosis taxonomy for each patient since the granularity of the International Classification of Diseases (ICD) classification is too detailed for many clinical practice cases [22]. Since the REP had created a longitudinal record spanning each subject's entire period of residency in the community, we retrieved all their ICD-9 diagnostic codes from the REP linkage system. We collapsed the ICD-9 codes into 285 CCS categories. We followed the intuition of stopword removal for text preprocessing, and removed the rare CCS codes if the total number of patients with the codes was less than or equal to two in each cohort. Table 1 lists the basic demographics of the three cohorts, including the number of patients (male and female), median age (male and female), and median observed number of diagnoses for the three cohorts. LDA and PDM were trained on the data of these cohorts, with numbers of disease topics ranging from 10 to 30 with step size of 5 (i.e., $K = 10, 15, 20, 25, 30$). We empirically chose $K = 20$ for both approaches based on the superior Hopkins statistics (see Table S1 in the supplementary materials for the different values of Hopkins statistics with different numbers of disease topics).

5. Results

5.1. Visualization of latent disease clusters

We first compared the visualization of disease representations in the latent topic space learned by LDA and the proposed PDM model for three cohorts, as depicted in Fig. 2. We tested different perplexity settings in t-SNE, namely 5, 10, 15, 20, for both LDA and PDM, and chose 10 for LDA and 20 for PDM due to the better visualization with data points spreading out (see supplemental materials for the different perplexity settings in the Osteoporosis Cohort). We chose 5000 as the number of iterations for both approaches since that is generally enough for convergence [25]. We can see that the disease clusters are explicitly dichotomized by PDM while almost no disease clusters could be observed by LDA. The Hopkins statistics of disease representations by PDM are 0.923, 0.958, 0.936 for osteoporosis, delirium/dementia, and

[3] https://github.com/yanshanwang/poissondirichlet.

[4] https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp.

Table 1
Demographics and basic statistics of the study cohorts.

Cohort	Osteoporosis Cohort	Delirium/Dementia Cohort	COPD/Bronchiectasis Cohort
# Patients	388	304	685
# Male (%)	21 (5.4%)	95 (31.2%)	337 (49.2%)
# Female (%)	367 (94.6%)	209 (68.8%)	348 (50.8%)
Median Age	74.4	83.6	73.2
Median Age (Male)	74.7	85.0	75.1
Median Age (Female)	68.8	81.6	71.1
Median Pearson-Years	18	14.4	15.1
Median Observed # Diagnosis	406	387.5	402

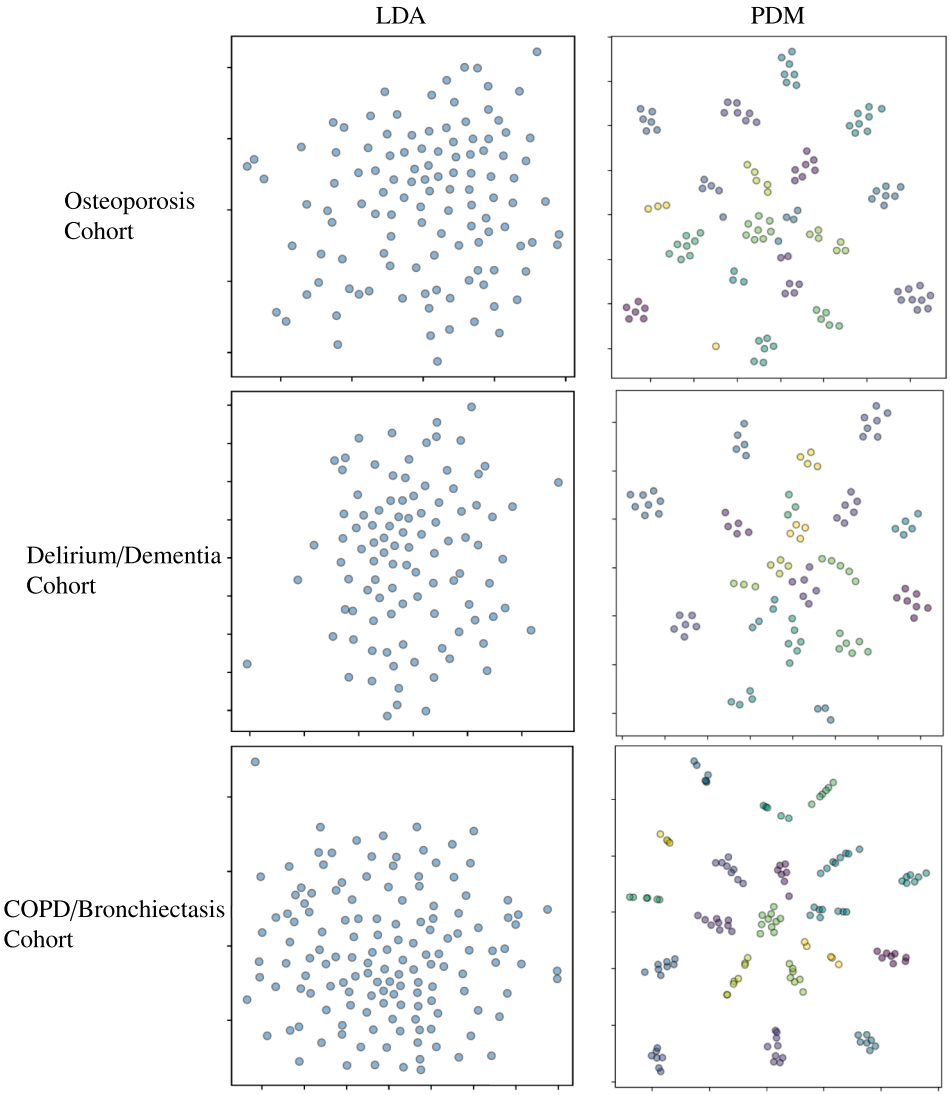

Fig. 2. Comparison of the visualization of diseases (represented by CCS) in the latent topic space learned by LDA (left column) and PDM (right column) for three cohorts.

COPD, respectively, while the Hopkins statistics of LDA are 0.876, 0.890, 0.828. The Hopkins statistic results quantitatively verified that the disease representations learned by PDM have better cluster tendency than those by LDA.

Since LDA failed to identify latent disease clusters, we only list the latent comorbidities identified by PDM that appeared in the clusters of osteoporosis, delirium/dementia, and COPD in Table 2. In order to verify the results, we tried to find evidence by searching PubMed[5] articles' titles or abstracts using keywords from the target disease and the latent comorbidities, in combination with the term "comorbidity" or "comorbidities". For example, we found 31 PubMed research articles for *osteoporosis* and *implant or graft*, 52 articles for *dementia* and *osteoporosis*, and 30 articles for *COPD* and *cerebrovascular disease*. The latent comorbidities are not highly related to age and sex and thus the prediction or clustering of *excess* risk would be of more interest for epidemiological analysis. This result shows that the proposed PDM model is able to learn the latent patterns hidden in the EHR data that differentiate disease clusters by alleviating the impact of age and sex on the

[5] https://www.ncbi.nlm.nih.gov/pubmed/.

Table 2
Comorbidities appeared in the clusters of osteoporosis, delirium/dementia, and COPD/bronchiectasis by the proposed PDM approach.

Disease	Latent Comorbidities in the Same Cluster
Osteoporosis (CCS: 206)	Headache; including migraine (CCS: 84) Nonspecific chest pain (CCS: 102) Diverticulosis and diverticulitis (CCS: 146) Complication of device; implant or graft (CCS: 237) Abdominal pain (CCS: 251)
Delirium/dementia (CCS: 653)	Immunizations and screening for infectious disease (CCS: 10) Conduction disorders (CCS: 105) Osteoporosis (CCS: 206) Fracture of lower limb (CCS: 230)
COPD/bronchiectasis (CCS: 127)	Peri-; endo-; and myocarditis; cardiomyopathy (except that caused by tuberculosis or sexually transmitted disease) (CCS:97) Aortic; peripheral; and visceral artery aneurysms (CCS: 115) Fracture of lower limb (CCS: 230) Late effects of cerebrovascular disease (CCS: 113) Complications of surgical procedures or medical care (CCS: 238) Biliary tract disease (CCS: 149) Other female genital disorders (CCS: 175)

diseases.

We provided the top ten diagnoses with the greatest probability in each topic, the corresponding probabilities, and overall proportion of each topic for each method and each cohort in the supplementary material. We have some interesting findings according to the top diagnoses in each topic. For example, we found that CCS206: osteoporosis had the greatest probability in Topic 0 by PDM and has the second greatest probability in Topic 13 by LDA for the Osteoporosis Cohort. However, those topics represent different disease clusters. For example, Topic 0 by PDM contains osteoporosis, COPD, peripheral and visceral artery aneurysms as top diseases while Topic 13 by LDA contains disorders of teeth and jaw and osteoporosis as top diseases. Explaining the clinical meaning of topics and disease clusters and their use cases in healthcare is subject to the future work.

5.2. Validation of patient subgroups

In this section, we demonstrate the experimental results of patient subgroups discovered by LDA and the proposed PDM model for the three cohorts. As aforementioned, we tested three clustering algorithms (i.e., hierarchical clustering, K-means clustering, and Birch clustering algorithms) and five different numbers of patient subgroups (i.e., 2, 3, 4, 5, 6) on the patient-topic matrix computed by LDA and PDM. We carried out survival analysis on these subgroups for each cohort based on patients' death information. The p-values of the Log-rank test on the survival curves are listed in Table 3. From the table, we observe that LDA generally produces patient subgroups with smaller p-values than PDM. When LDA was utilized, the K-means clustering algorithm generated differentiated patient subgroups with statistical significance for both Osteoporosis and COPD/Bronchiectasis Cohorts when the number of patient subgroups was 2, and achieved the best p-values ($p = 0.0051$) for the Delirium/Dementia Cohort when the number of patient subgroups was 3. The p-values of patient subgroups generated by PDM were much higher than those by LDA: K-means clustering algorithm achieved p-values of 0.071 and 0.00023 with the number of patient subgroups equal to 6 and 2 for the Delirium/Dementia and COPD Cohorts, respectively; and the Birch clustering algorithm produced a p-value of 0.0085 for the Osteoporosis Cohort when the number of patient subgroups was 2. Overall, the K-means clustering algorithm outperformed other clustering algorithms in identifying patient subgroups. The patient subgroup results with the best p-values from LDA and PDM with the smallest number of patient subgroups are chosen for further survival and comorbidity analysis. We also note that using p-value in this study is not optimal as many p-values indicate statistical significance and a smaller p-value doesn't necessarily mean the subgroups are more significantly different. Having said that, using the p-value provides a means to select a relatively better number of subgroups and clustering algorithms.

5.2.1. Survival curves

Kaplan-Meier survival curves of the selected patient subgroups discovered by LDA and PDM are depicted in Fig. 3 for three cohorts.

Survival analysis of patient subgroups discovered by LDA for the Osteoporosis Cohort depicted in Fig. 3 LDA-(a) showed significant difference between the survival curves of patient subgroups with $p<0.0001$. The patients in Subgroup 1 had a distinguished worse survival rate than those in Subgroup 2. Thus, at the point of clinical care, the patients in Subgroup 1 should receive more attention than those in Subgroup2. Fig. 3 LDA-(b) shows significant difference at level $p < 0.01$ between the survival curves of patient subgroups for the Delirium/Dementia Cohort. Three identified patient subgroups are distinguishable with disparate survival curves. The patients in Subgroup 2 have a better survival rate than those in Subgroups 1 and 3. The patients in Subgroup 1 have the lowest survival rate across the survival time distribution. Fig. 3 LDA-(c) shows a significant difference between the survival curves of patient subgroups with $p<0.0001$ for the COPD/Bronchiectasis Cohort. The patients in Subgroup 2 have a prominent lower risk and better survival than those in Subgroup 1.

Figs. 3 PDM-(a), PDM-(b), and PDM-(c) show the survival analysis on the patient subgroups discovered by PDM for three cohorts. Fig. 3 PDM-(a) showed difference between the survival curves of patient subgroups at the level of 0.01 for the Osteoporosis Cohort. The survival curves are similar to those by LDA. The patients in Subgroup 2 have a better survival rate than those in Subgroup 1 when survival time is approximately < 6 years. The survival probability of both subgroups drops dramatically when survival time is between 6 years and 8 years. The survival probability of patients in Subgroup 2 decreases more rapid than those in Subgroup 1 after survival time > 11 years. Though Fig. 3 PDM-(b) does not show significant difference between the survival curves of patient subgroups for the Delirium/Dementia Cohort, patient subgroups are distinguishable with disparate survival curves. For example, the patients in Subgroups 1, 3, and 6 have similar survival curves when survival time is approximately < 5 years. However, the patients in Subgroup 6 have a longer survival time than those in Subgroups 1 and 3. The patients in Subgroups 2 and 5 have better survival than other subgroups across the survival time distribution, and those in Subgroup 2 have better survival than those in Subgroup 5 when survival time is > 7 years. We provided pairwise p-values between six subgroups in the supplemental materials. Fig. 3 PDM-(c) shows that the survival curves of patient subgroups are different at the level of < 0.001 for the COPD Cohort. The patients in Subgroup 1 have a noticeably better survival rate than those in Subgroup 2. These results explicitly show the potential of applying unsupervised machine learning models to stratify patients into groups with different risks.

5.2.2. Statistical analysis

Tables 4–6 list the statistics of demographics, number of diagnoses, and ECI scores for the patient subgroups identified by LDA and PDM for the Osteoporosis, Delirium/Dementia, and COPD/Bronchiectasis Cohorts, respectively. Statistical significance is based on the Kruskal-Wallis Test ($p < 0.001$). Also reported are the ECI categories that are different among patient subgroups with statistical significance (p-value<0.001) in each cohort. The complete analysis for each ECI category can be found in the supplemental material.

We first analyze the patient subgroups identified by LDA. As shown in Table 4 for the Osteoporosis Cohort, the patients in Subgroup 1 are

Table 3
P-values of the Log-rank test on survival curves of patient subgroups using unsupervised machine learning models and three clustering algorithms with different number of subgroups for the three cohorts. The subgroups with bolded p-values by PDM and LDA are chosen for further survival and comorbidity analysis.

Number of subgroups	2		3		4		5		6	
Unsupervised Methods	PDM	LDA	PDM	LDA	PDM	LDA	PDM	LDA	PDM	LDA
Osteoporosis Cohort										
Hierarchical clustering	0.16	0.00017	0.28	0.00081	0.44	0.00075	0.41	<0.00073	0.41	<0.0001
K-means clustering	0.39	**<0.0001**	0.58	<0.0001	0.38	<0.0001	0.38	<0.0001	0.6	0.00032
Birch clustering	**0.0085**	0.002	0.015	0.0038	0.0098	0.00078	0.011	0.0021	0.018	0.0035
Delirium/Dementia Cohort										
Hierarchical clustering	0.10	0.41	0.083	0.23	0.16	0.40	0.26	0.54	0.28	0.61
K-means clustering	0.34	0.012	0.49	**0.0051**	0.50	0.011	0.42	0.029	**0.071**	0.0057
Birch clustering	0.14	0.071	0.34	0.19	0.32	0.34	0.46	0.50	0.12	0.033
COPD/Bronchiectasis Cohort										
Hierarchical clustering	0.017	0.0045	0.014	<0.0001	0.035	<0.0001	0.027	<0.0001	0.021	<0.0001
K-means clustering	**0.00023**	**<0.0001**	0.00032	<0.0001	0.0017	<0.0001	0.086	<0.0001	0.0026	<0.0001
Birch clustering	0.15	0.00045	0.10	<0.0001	0.15	<0.0001	0.12	<0.0001	0.2	<0.0001

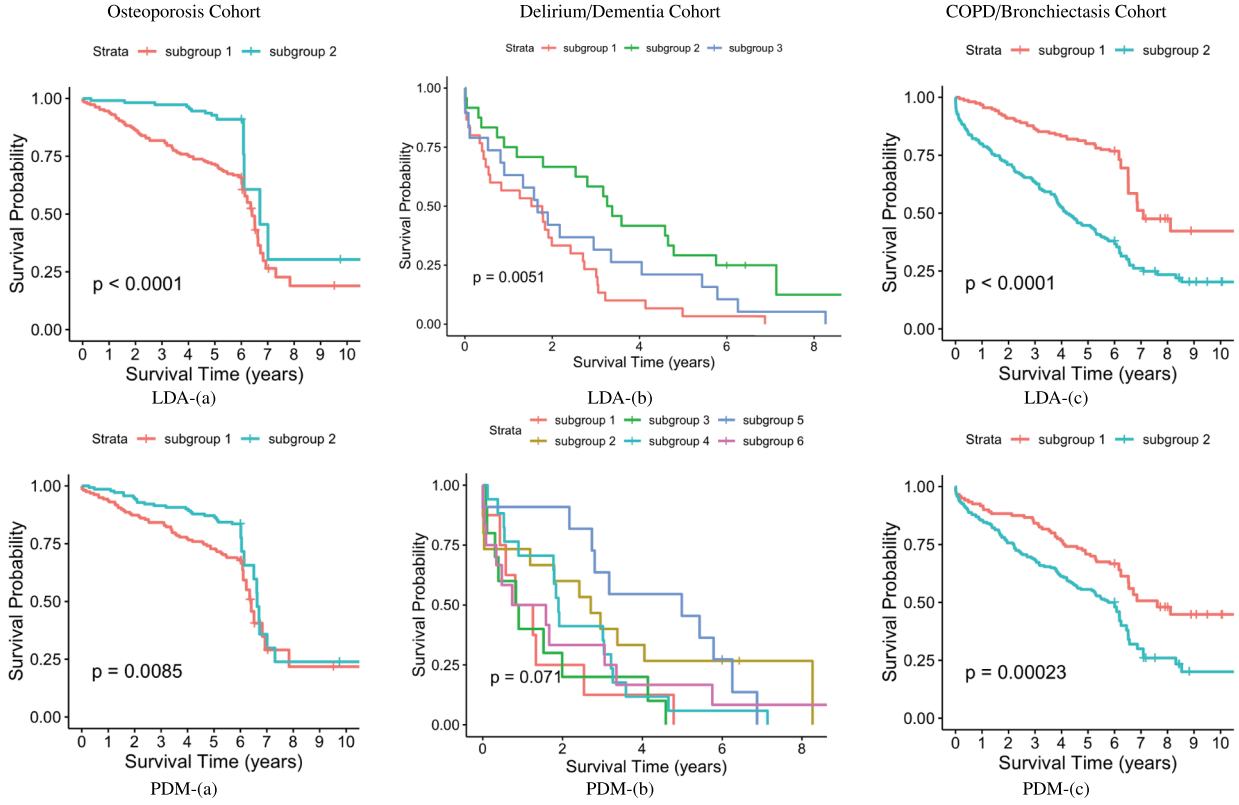

Fig. 3. Survival analysis of the patient subgroups discovered by LDA and PDM from the Osteoporosis, Delirium/Dementia, and COPD/Bronchiectasis Cohorts.

older than those in Subgroup 2 with statistical significance at the level of 0.001, which might be the reason that the patients of Subgroup 1 have a worse survival rate in Fig. 3 (a). Median ECI scores between the two subgroups are statistically different at the level of 0.001. The patients in Subgroup 1 have a higher median ECI score, which means they have more comorbidities. 73.4% of patients in Subgroup 1 also have a larger number of comorbidites in terms of ECI (5+). The result is consistent with their survival analysis in Fig. 3 (a). Seven ECI categories, including congestive heart failure, pulmonary circulation disorders, hypertension, other neurological disorders, weight loss, fluid and electrolyte disorders, and psychoses, are statistically different between two patient subgroups at the level of 0.001. The diseases in these ECI categories potentially contributed to differentiate patients into subgroups for the Osteoporosis Cohort.

For the patient subgroups identified by LDA from the Delirium/Dementia Cohort, there is statistically significant difference in age but not in sex among the three subgroups. The fact that Subgroup 2 has better survival than Subgroup 1 is mainly due to the younger age of Subgroup 2. This result is consistent with previous findings that age is a strong risk factor for dementia [32]. No statistical significance is found in the number of diagnoses at the level of 0.001. The patients in Subgroup 1 and those in Subgroup 2 have a similar number of comorbidities in terms of median number of ECI and ECI (5+). Hypertension is the ECI category that has a statistically significant difference between the three subgroups, which is also consistent with the outcome of large observational studies that that hypertension plays a role in dementia and Alzheimer's disease [33].

There is also statistically significant difference in age but not in sex for the patient subgroups identified by LDA from the COPD/Bronchiectasis Cohort. The patients in Subgroup 1 have a higher

Table 4
Demographics and statistics of patient subgroups identified by LDA and PDM from the Osteoporosis Cohort. Statistically significance is based on the Kruskal-Wallis Test ($p < 0.001$). The complete analysis for all ECI categories is provided in the supplemental material.

Methods	LDA				PDM		
Subgroups	Subgroup 1	Subgroup 2	p-value		Subgroup 1	Subgroup 2	p-value
# Patients (%)	271 (69.8%)	117 (30.2%)			216 (55.7%)	172 (44.3%)	
Sex			0.034				0.096
# Male (%)	19 (7.0%)	2 (1.7%)			8 (3.7%)	13 (7.6%)	
# Female (%)	252 (93.0%)	115 (98.3%)			208 (96.3%)	159 (92.4%)	
Median Age	78.3	66.3	**<0.001**		71.1	74.6	0.002
Median # Diagnosis	414.0	390.0	0.007		405	408	0.815
Median ECI	6.0	4.0	**<0.001**		6.0	5.0	0.013
ECI Groups			**<0.001**				0.331
# Patients in ECI (0–1)	7 (2.6%)	9 (7.7%)			9 (4.2%)	7 (4.1%)	
# Patients in ECI (2–4)	65 (24.0%)	64 (54.7%)			65 (30.1%)	64 (37.2%)	
# Patients in ECI (5+)	199 (73.4%)	44 (37.6%)			142 (65.7%)	101 (58.7%)	
ECI Categories with p<0.001							
Congestive heart failure	48 (17.7%)	5 (4.3%)	**<0.001**	Psychoses	65 (30.1%)	23 (13.4%)	**<0.001**
Pulmonary circulation disorders	28 (10.3%)	1 (0.9%)	**<0.001**				
Hypertension	209 (77.1%)	64 (54.7%)	**<0.001**				
Other neurological disorders	85 (31.4%)	16 (13.7%)	**<0.001**				
Weight loss	97 (35.8%)	16 (13.7%)	**<0.001**				
Fluid and electrolyte disorders	139 (51.3%)	30 (25.6%)	**<0.001**				
Psychoses	85 (31.4%)	3 (2.6%)	**<0.001**				

median ECI score and a much larger number of comorbidities in terms of ECI (5+) than those in Subgroup 2 with a statistically significant difference at p<0.001. This result is consistent with the survival analysis in Fig. 3 (c). Nine ECI categories are statistically different between two patient subgroups at the level of 0.001. These categories include congestive heart failure, cardiac arrhythmias, pulmonary circulation disorders, renal failure, obesity, weight loss, fluid and electrolyte disorders, deficiency anaemia, and psychoses. Interestingly, obesity appears to be associated with better outcome (15.2% in Subgroup 1 and 26.3% in Subgroup 2) and weight loss with worse outcome (38.6% in Subgroup 1 and 19.5% in Subgroup 2). That is likely related to the fact that being underweight in COPD/bronchiectasis might be a bad prognostic factor since the respiratory muscles lose strength during severe weight loss, which leads to respiratory failure. Being extremely obese might also worsen COPD/bronchiectasis, but is not addressed in our analysis.

The first observation from the results of patient subgroups identified by PDM is that, unlike LDA, PDM does not differentiate patients into subgroups based on age and sex. No statistical significance is found in age and sex for the patient subgroups of the three cohorts. This characteristic of PDM is very important for epidemiological studies. For example, a 68-year-old man and a 50-year-old woman with heart failure and stroke may have a similar risk of fracture when all two comorbidities are considered together. We expect that an appreciation of these different clusters of comorbid conditions will potentially enhance our understanding of the latent traits underlying disease risk and thus provide more insights in developing new treatments. This result validates the ability of PDM to remove the factors of age and sex for discovering patient subgroups, which enables analysis of hidden patterns of diseases that are of greater interest in epidemiology research. For the Osteoporosis Cohort, no statistical significance is found in the number of diagnoses or the number of comorbidities. Only one ECI category, psychoses, is statistically different between two patient subgroups at the level of 0.001. For the Delirium/Dementia Cohort, no ECI category was found different with statistical significance among patient subgroups. The superior survival curves of Subgroup 5 to those of other subgroups, as shown in Fig. 3 PDM-(b), might be related to the positive effect of the treatment for multi-morbidities on the survival of this subgroup. This result may support inferences of cause and effect or indicate potential associations between diseases to improve patient care by conducting further cross-sectional studies or case-control studies. We will investigate how to make meaningful clinical statements using the results in our future work. The conventional comorbidity taxonomy could not identify significant factors that cause differential patient subgroups discovered by PDM. In other words, PDM might leverage other undiscovered hidden disease patterns to discover these patient

Table 5
Demographics and statistics of patient subgroups identified by LDA and PDM from the Delirium/Dementia Cohort. Statistically significance is based on the Kruskal-Wallis Test ($p < 0.001$). The complete analysis for all ECI categories is provided in the supplemental file.

Methods	LDA				PDM						
Subgroups	Subgroup 1	Subgroup 2	Subgroup 3	p-value	Subgroup 1	Subgroup 2	Subgroup 3	Subgroup 4	Subgroup 5	Subgroup 6	p-value
# Patients (%)	167 (55.0%)	72 (23.7%)	65 (21.3%)		47 (15.5%)	41 (13.5%)	34 (11.2%)	62 (20.4%)	56 (18.4%)	64 (21.1%)	
Sex				0.114						0.445	
# Male (%)	44 (26.3%)	28 (38.9%)	23 (35.4%)		16 (34.0%)	15 (36.6%)	9 (26.5%)	19 (30.6%)	12 (21.4%)	24 (37.5%)	
# Female (%)	123 (73.7%)	44 (61.1%)	42 (64.6%)		31 (66.0%)	26 (63.4%)	25 (73.5%)	43 (69.4%)	44 (78.6%)	40 (62.5%)	
Median Age	86.4	79.7	82.9	**<0.001**	82.4	84.0	84.9	84.3	85.7	82.3	0.623
Median # Diagnosis	385.0	414.0	376.0	0.009	383.0	399.0	385.5	409.0	411.5	368.5	0.007
Median ECI	8.0	8.0	7.0	**<0.001**	8.0	8.0	7.0	8.0	8.0	8.0	0.130
ECI Groups				**<0.001**						0.874	
# Patients in ECI (0–1)	0 (0.0%)	0 (0.0%)	0 (0.0%)		0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	
# Patients in ECI (2–4)	5 (3.0%)	2 (2.8%)	12 (18.5%)		2 (4.3%)	2 (4.9%)	2 (5.9%)	3 (4.8%)	4 (7.1%)	6 (9.4%)	
# Patients in ECI (5+)	162 (97.0%)	70 (97.2%)	53 (81.5%)		45 (95.7%)	39 (95.1%)	32 (94.1%)	59 (95.2%)	52 (92.9%)	58 (90.6%)	
ECI Categories with p<0.001											
Hypertension	140 (83.8%)	58 (80.6%)	39 (60.0%)	**<0.001**							

Table 6
Demographics and statistics of patient subgroups identified by LDA and PDM from the COPD/Bronchiectasis Cohort. Statistically significance is based on the Kruskal-Wallis Test ($p < 0.001$). The complete analysis for all ECI categories is provided in the supplemental file.

Methods	LDA				PDM		
	Subgroup 1	Subgroup 2	p-value		Subgroup 1	Subgroup 2	p-value
# Patients (%)	495 (72.3%)	190 (27.7%)			207 (30.2%)	478 (69.8%)	
Sex			0.005				0.489
# Male (%)	260 (52.5%)	77 (40.5%)			106 (51.2%)	231 (48.3%)	
# Female (%)	235 (47.5%)	113 (59.5%)			101 (48.8%)	247 (51.7%)	
Median Age	76.0	66.8	**<0.001**		71.1	74.6	0.002
Median # Diagnosis	403.0	402.0	0.791		409.0	399.5	0.526
Median ECI	9.0	6.0	**<0.001**		7.0	8.0	**<0.001**
ECI Groups			**<0.001**				0.016
# Patients in ECI (0–1)	0 (0.0%)	2 (1.1%)			2 (1.0%)	0 (0.0%)	
# Patients in ECI (2–4)	15 (3.0%)	35 (18.4%)			21 (10.1%)	29 (6.1%)	
# Patients in ECI (5+)	480 (97.0%)	153 (80.5%)			184 (88.9%)	449 (93.9%)	
ECI Categories with p<0.001							
Congestive heart failure	264 (53.3%)	27 (14.2%)	**<0.001**	Fluid and electrolyte disorders	111 (53.6%)	337 (70.5%)	**<0.001**
Cardiac arrhythmias	350 (70.7%)	102 (53.7%)	**<0.001**				
Pulmonary circulation disorders	154 (31.1%)	23 (12.1%)	**<0.001**				
Renal failure	123 (24.8%)	19 (10.0%)	**<0.001**				
Obesity	75 (15.2%)	50 (26.3%)	**<0.001**				
Weight loss	191 (38.6%)	37 (19.5%)	**<0.001**				
Fluid and electrolyte disorders	380 (76.8%)	68 (35.8%)	**<0.001**				
Deficiency anaemia	110 (22.2%)	21 (11.1%)	**<0.001**				
Psychoses	140 (28.3%)	19 (10.0%)	**<0.001**				

subgroups that could be used to discover novel disease associations and treatments. For the COPD/Bronchiectasis Cohort, the patients in Subgroup 2 are older than those in Subgroup 1 but without statistical significance at the level of 0.001. ECI scores between the two subgroups are statistically different with $p<0.001$. This result, similar to that of LDA, indicates that age portends worse prognosis as does greater comorbidities (ECI). Only one ECI category, i.e., fluid and electrolyte disorders, is statistically different between two patient subgroups at the level of 0.001.

6. Discussion and conclusion

In this study, we investigate the applications of unsupervised machine learning approaches in discovering latent disease clusters and patient subgroups using the EHR data. We utilized LDA, an unsupervised probabilistic generative model in the rubric of topic models, and proposed a novel unsupervised machine learning approach, named PDM. PDM extends the conventional LDA and uses a Poisson distribution to model patients' disease diagnoses and to alleviate age and sex factors by considering both observed and expected observations. We applied LDA and PDM in two clinical use cases: discovery of latent disease clusters and patient subgroups using EHRs.

Both approaches were evaluated on the diagnostic EHR data of three cohorts, namely the Osteoporosis Cohort, the Delirium/Dementia Cohort, and the COPD/Bronchiectasis Cohort, retrieved from the REP medical linkage system. We verified the effectiveness of discovering latent disease clusters through the visualization of disease representation in the latent topic space. The 2-D scattered plot shows that PDM discovered explicitly dichotomized disease clusters based on the latent patterns hidden in the EHR data than LDA. This result implies that we could utilize these disease clusters to identify multiple latent comorbities, which could be used to calculate excess risk above what would be expected for a given age and sex.

Furthermore, we applied LDA and PDM to discover patient subgroups, and carried out survival analysis on these subgroups. The experimental results show that LDA could stratify patients into more differentiable subgroups than PDM in terms of p-values. However, those subgroups identified by LDA are highly associated with patients' age and sex. Though the difference between the subgroups discovered by PDM has worse p-values, these patient subgroups might imply the underlying patterns of diseases of greater interest in epidemiology research by alleviating the impact of age and sex. For example, in the study for the Osteoporosis Cohort, the ECI categories in LDA are mostly correlated with age, such as hypertension and heart failure, while the ECI category of psychoses identified by PDM might imply the substantial difference in the survival curves. Evidence has been found in a research study [34] that risedronate, which is a medication commonly used for the treatment of osteoporosis, could trigger psychiatric side effects and result in termination of treatment for osteoporosis, which have the potential to reduce survivals. Relationships between the study disease and the ECI categories identified by PDM might of interest to researchers to discover hidden links between comorbidities caused by disease mechanism, treatment, procedure, etc. that are not correlated with age and sex. Therefore, the proposed PDM might be a better option than LDA for studying latent disease patterns in aging cohorts for which we would like to alleviate the impact of age and sex since they are major drivers of aging-associated diseases.

We also provided the dominant topics in each subgroup in Table S2 in the supplementary materials. The dominant topics make sense as ways to differentiate patients. For example, for the Osteoporosis Cohort, Topic 10 is an obvious cancer disease topic containing several cancer diseases and is the dominant topic for Subgroup 1 that might lead to the worse survival curve. In our future work, we will study whether these disease topics are clinically meaningful and how to utilize them in healthcare applications.

Due to the similarity of unsupervised machine learning and human learning, unsupervised learning is more closely aligned with artificial intelligence (AI), where a computer is expected to learn to identify complex processes and patterns without a human's guidance. Compared to the supervised machine learning models that lack generalizability and suffer from infeasibility of discovering novel patterns from EHRs [3], the unsupervised machine learning techniques utilized to discover particular comorbidity clusters from EHRs that may reflect underlying mechanisms ("latent traits") would help define new domains of risk. The disease clusters discovered by PDM might contain new potential risks for diseases. Moreover, unsupervised machine learning approaches could be used to stratify patients into subgroups with similar characteristics and risks. Discovering differentiated patient subgroups will not only facilitate epidemiological analysis and research, but enable personalized care that will improve the efficiency and effectiveness of disease prevention, diagnosis, and treatment. We believe that both approaches could be leveraged to create an AI platform for exploiting the rich data resources of the REP, and likewise serve as a model for use by others with EHRs.

This study has a few limitations. First, since this is a proof-of-concept study, we are not attempting to make any clinical statements and conclusions. In future work, we will apply the methods on carefully selected case and control cohorts to seek potential interesting clinical results. Second, LDA and the proposed PDM model were tested on cohorts with a relatively small number of patients and diagnoses due to the computational cost of the MH algorithm. The average computation time for LDA and PDM is 4.1 h and 2.6 h, respectively, and it increases exponentially when the numbers of patients and diagnoses increase. The detailed computation time of LDA and PDM for each cohort is summarized in Table S3 in the supplementary materials. Faster MCMC methods should be considered in future work so that the proposed model could be scaled up to larger cohorts. Third, the time stamp of diagnosis, which is an important feature for epidemiology, was not considered in LDA and PDM models. For example, a 60-year-old man with osteoporosis and heart disease will lead to different results if those diagnoses are from 15 years ago versus from the last 3 years. Fourth, PDM was only evaluated on the EHR data from the REP. In the future, we will evaluate the generalizability of the proposed model on the EHR data from other institutions. Finally, unsupervised machine learning approaches other than probabilistic generative models, such as deep neural network based models [3,35,36], and supervised models, such as LASSO-based models [37], were not compared in this study. In the future, we will compare the proposed approaches with the deep neural network based models as well as supervised approaches in discovering latent disease clusters and patient subgroups using EHRs.

CRediT authorship contribution statement

Yanshan Wang: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing - original draft, Writing - review & editing. **Yiqing Zhao:** Formal analysis. **Terry M. Therneau:** Conceptualization, Methodology. **Elizabeth J. Atkinson:** Data curation, Formal analysis. **Ahmad P. Tafti:** Formal analysis. **Nan Zhang:** Formal analysis. **Shreyasee Amin:** Writing - review & editing. **Andrew H. Limper:** Writing - review & editing. **Sundeep Khosla:** Funding acquisition. **Hongfang Liu:** Writing - review & editing, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by NIH grants P01AG004875, R01GM102282, UL1TR002377, U01TR002062, and R01LM011934, Mayo Clinic internal grants, and made possible by the Rochester Epidemiology Project (R01AG034676) and the U.S. Public Health Service.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.jbi.2019.103364.

References

[1] W.R. Hersh, Adding value to the electronic health record through secondary use of data for quality assurance, research, and surveillance, Am. J. Managed Care 13 (6) (2007) 277–279.
[2] Z. Obermeyer, E.J. Emanuel, Predicting the future?big data, machine learning, and clinical medicine, New Engl. J. Med. 375 (13) (2016) 1216.
[3] R. Miotto, L. Li, B.A. Kidd, J.T. Dudley, Deep patient: an unsupervised representation to predict the future of patients from the electronic health records, Sci. Rep. 6 (2016) 26094.
[4] C. Xiao, E. Choi, J. Sun, Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review, J. Am. Med. Inform. Assoc. 25 (10) (2018) 1419–1428.
[5] Y. Wang, L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, S. Liu, Y. Zeng, S. Mehrabi, S. Sohn, et al., Clinical information extraction applications: a literature review, J. Biomedical Informatics 77 (2018) 34–49.
[6] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436.
[7] T.Q. Chen, Y. Rubanova, J. Bettencourt, D.K. Duvenaud, "Neural ordinary differential equations," in Advances in Neural, Inf. Process. Syst. (2018) 6572–6583.
[8] X. Wang, D. Sontag, F. Wang, "Unsupervised learning of disease progression models," in, Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2014, pp. 85–94.
[9] R. Pivovarov, A.J. Perotte, E. Grave, J. Angiolillo, C.H. Wiggins, N. Elhadad, Learning probabilistic phenotypes from heterogeneous ehr data, J. Biomedical Informatics 58 (2015) 156–165.
[10] J.H. Son, G. Xie, C. Yuan, L. Ena, Z. Li, A. Goldstein, L. Huang, L. Wang, F. Shen, H. Liu, et al., Deep phenotyping on electronic health records facilitates genetic diagnosis by clinical exomes, Am. J. Human Genetics 103 (1) (2018) 58–73.
[11] D. o. E. United Nations and P.D. Social Affairs, World population ageing 2013, United Nations, New York, 2013.
[12] M.J. Divo, C.H. Martinez, D.M. Mannino, Ageing and the epidemiology of multimorbidity, 2014.
[13] L.E. Vanfleteren, M.A. Spruit, M. Groenen, S. Gaffron, V.P. van Empel, P.L. Bruijnzeel, E.P. Rutten, J. Opt Roodt, E.F. Wouters, F.M. Franssen, Clusters of comorbidities based on validated objective measurements and systemic inflammation in patients with chronic obstructive pulmonary disease, Am. J. Respiratory Critical Care Med. 187 (7) (2013) 728–735.
[14] P.M. Schnell, Q. Tang, W.W. Offen, B.P. Carlin, A bayesian credible subgroups approach to identifying patient subgroups with positive treatment effects, Biometrics 72 (4) (2016) 1026–1036.
[15] L.J. Melton III, History of the rochester epidemiology project, Mayo Clin. Proc. 71 (3) (1996) 266–274.
[16] W.A. Rocca, B.P. Yawn, J.L.S. Sauver, B.R. Grossardt, L.J. Melton, History of the rochester epidemiology project: half a century of medical records linkage in a us population, Mayo Clinic Proc. 87 (12) (2012) 1202–1213.
[17] J.L. St Sauver, B.R. Grossardt, B.P. Yawn, L.J. Melton III, J.J. Pankratz, S.M. Brue, W.A. Rocca, Data resource profile: the rochester epidemiology project (rep) medical records-linkage system, Int. J. Epidemiol. 41 (6) (2012) 1614–1624.
[18] R. Melton, L.J.S. Achenbach, E. Atkinson, T.M. Therneau, S. Amin, Long-term mortality following fractures at different skeletal sites: a population-based cohort study, Osteoporos. Int. 24 (5) (2013) 1689–1696.
[19] R. Savica, B.R. Grossardt, J.H. Bower, J.E. Ahlskog, W.A. Rocca, Risk factors for parkinson's disease may differ in men and women: an exploratory study, Hormones Behav. 63 (2) (2013) 308–314.
[20] D.M. Blei, Probabilistic topic models, Commun. ACM 55 (4) (2012) 77–84.
[21] W. Zhao, W. Zou, J.J. Chen, Topic modeling for cluster analysis of large biological and medical datasets, in: BMC Bioinformatics, vol. 15, no. 11. BioMed Central, 2014, p. S11.
[22] D.C. Li, T. Therneau, C. Chute, H. Liu, Discovering associations among diagnosis groups using topic modeling, AMIA Summits Transl. Sci. Proc. 2014 (2014) 43.
[23] T.L. Griffiths, M. Steyvers, Proc. Natl. Acad. Sci. 2004, pp. 5228–5235.
[24] W.K. Hastings, Monte carlo sampling methods using markov chains and their applications, 1970.
[25] L.V.D. Maaten, G. Hinton, Visualizing data using t-sne, J. Mach. Learn. Res. 9 (Nov) (2008) 2579–2605.
[26] B. Hopkins, J.G. Skellam, A new method for determining the type of distribution of plant individuals, Ann. Bot. 18 (2) (1954) 213–227.
[27] C.S. Ledbetter, M.W. Morgan, Toward best practice: leveraging the electronic patient record as a clinical data warehouse, J. Healthcare Inf. Manage. 15 (2) (2001) 119–132.
[28] J.H. Ward Jr, Hierarchical grouping to optimize an objective function, J. Am. Stat. Assoc. 58 (301) (1963) 236–244.
[29] J.A. Hartigan, M.A. Wong, Algorithm as 136: A k-means clustering algorithm, J. Roy. Stat. Soc. Ser. C (Appl. Stat.) 28 (1) (1979) 100–108.
[30] T. Zhang, R. Ramakrishnan, M. Livny, Birch: an efficient data clustering method for very large databases, in: ACM Sigmod Record, vol. 25, no. 2. ACM, 1996, pp. 103–114.
[31] A. Elixhauser, C. Steiner, D.R. Harris, R.M. Coffey, Comorbidity measures for use with administrative data, Medical care (1998) 8–27.
[32] S. Gao, H.C. Hendrie, K.S. Hall, S. Hui, The relationships between age, sex, and the incidence of dementia and alzheimer disease: a meta-analysis, Arch. Gen. Psychiatry 55 (9) (1998) 809–815.
[33] C. Tzourio, Hypertension, cognitive decline, and dementia: an epidemiological perspective, Dialogues Clin. Neuroscience 9 (1) (2007) 61.
[34] S. Hirschmann, A. Gibel, I. Tsvelikhovsky, A. Lisker, Late-onset psychosis and risedronate treatment for osteoporosis: a case report, Clin. Schizophrenia Related Psychoses 9 (1) (2013) 36–39.
[35] E. Choi, M.T. Bahadori, A. Schuetz, W.F. Stewart, J. Sun, Doctor ai: Predicting clinical events via recurrent neural networks, in: Machine Learning for Healthcare Conference, 2016, pp. 301–318.
[36] E. Choi, M.T. Bahadori, L. Song, W.F. Stewart, J. Sun, Gram: graph-based attention model for healthcare representation learning, Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2017, pp. 787–795.
[37] N.M. Ballarini, G.K. Rosenkranz, T. Jaki, F. König, M. Posch, Subgroup identification in clinical trials via the predicted individual treatment effect, PloS One 13 (10) (2018) e0205971.