Journal of Biomedical Informatics 60 (2016) 210–223

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Modeling healthcare data using multiple-channel latent Dirichlet allocation

Hsin-Min Lu [a,*], Chih-Ping Wei [a], Fei-Yuan Hsiao [b,c,d]

[a] Department of Information Management, College of Management, National Taiwan University, Taipei 106, Taiwan
[b] Graduate Institute of Clinical Pharmacy, College of Medicine, National Taiwan University, Taipei 100, Taiwan
[c] School of Pharmacy, College of Medicine, National Taiwan University, Taipei 100, Taiwan
[d] Department of Pharmacy, National Taiwan University Hospital, Taipei 100, Taiwan

ARTICLE INFO

Article history:
Received 11 September 2015
Revised 29 January 2016
Accepted 3 February 2016
Available online 16 February 2016

Keywords:
Healthcare data mining
Health informatics
Multiple-channel latent Dirichlet allocation
Diagnosis–medication associations
Medication prediction
Diagnosis prediction

ABSTRACT

Information and communications technologies have enabled healthcare institutions to accumulate large amounts of healthcare data that include diagnoses, medications, and additional contextual information such as patient demographics. To gain a better understanding of big healthcare data and to develop better data-driven clinical decision support systems, we propose a novel multiple-channel latent Dirichlet allocation (MCLDA) approach for modeling diagnoses, medications, and contextual information in healthcare data. The proposed MCLDA model assumes that a latent health status group structure is responsible for the observed co-occurrences among diagnoses, medications, and contextual information. Using a real-world research testbed that includes one million healthcare insurance claim records, we investigate the utility of MCLDA. Our empirical evaluation results suggest that MCLDA is capable of capturing the comorbidity structures and linking them with the distribution of medications. Moreover, MCLDA is able to identify the pairing between diagnoses and medications in a record based on the assigned latent groups. MCLDA can also be employed to predict missing medications or diagnoses given partial records. Our evaluation results also show that, in most cases, MCLDA outperforms alternative methods such as logistic regressions and the *k*-nearest-neighbor (KNN) model for two prediction tasks, i.e., medication and diagnosis prediction. Thus, MCLDA represents a promising approach to modeling healthcare data for clinical decision support.

1. Introduction

Healthcare institutions routinely collect healthcare records that include diagnoses, medications, patient demographics, and other related information. Large amounts of healthcare records provide researchers and practitioners with a promising opportunity to develop novel analytical methods for improving healthcare quality, safety, and efficiency. Among various potential research directions, this study aims to develop a probabilistic latent variable model that captures the relations (or associations) among important variables in healthcare data, including diagnoses, medications, and contextual information (e.g., patient demographics). A learned latent variable model can provide the foundation for several valuable applications. First, the model can evaluate the likelihood of a set of diagnoses, medications, and contextual information in a record, thereby helping to identify outliers and improve data quality [1]. Second, the model can be used to identify disease groups (i.e., comorbidity) [2,3] that are valuable for clinical and academic research. Third, the model is able to predict missing diagnoses and medications using incomplete data. Finally, the model can be applied to augment patient problem lists given the prescribed medications, improving the completeness and accuracy of problem lists in medical histories [4].

Several previous studies have investigated the relations between diagnoses and medications in healthcare data. Data mining techniques, such as association rule mining (ARM) [4] and supervised learning [5], have been adopted to extract important associations, which are then used to predict missing medications. Existing approaches, however, have several limitations. First, ARM discovers associations based on item-level co-occurrences and does not provide abstractions at higher levels, thereby limiting its usefulness for data exploration and decision support. Second, direct applications of supervised learning techniques tackle the data modeling problem from a prediction perspective. Although a

* Corresponding author.
E-mail addresses: luim@ntu.edu.tw (H.-M. Lu), cpwei@ntu.edu.tw (C.-P. Wei), fyshsiao@ntu.edu.tw (F.-Y. Hsiao).

http://dx.doi.org/10.1016/j.jbi.2016.02.003
1532-0464/

previous study has reported promising outcomes using small datasets with fewer than 200 observations [5], supervised learning might provide limited value in non-predictive applications such as disease group identification and patient stratification.

To address these limitations, we propose the modeling of diagnoses, medications, and contextual information in healthcare data using a multiple-channel latent Dirichlet allocation (MCLDA) model. Our proposed MCLDA model assumes that the observed diagnoses, medications, and contextual information are driven by latent health status groups. By estimating the latent health status groups, MCLDA is able to construct the latent relations among diagnoses, medications, and contextual variables in different health status groups, revealing the potential latent confounding that influences the observed co-occurrences. Moreover, MCLDA can be adopted as a dimensional reduction method that summarizes each record using a probability vector over the latent health status group. Because the number of latent health status groups is much smaller than the number of unique diagnoses, medications, and contextual information in typical applications, this probability vector is less susceptible to the data sparsity issue and can be adopted by other data-mining methods as an effective representation of healthcare records.

MCLDA belongs to the family of topic models, a text mining approach that assumes that observed word co-occurrences are governed by latent variables. The well-known latent Dirichlet allocation (LDA) [6,7] and its variants were originally developed to identify latent topics from a set of documents, analyze long-term topical trends, and jointly model words and references in a document [8]. Recently, LDA has been applied to healthcare data to identify phenotypic topics based on ICD-9 codes [9], and to infer patterns of clinical pathways from electronic health records [10]. Our MCLDA model contributes to the literature by applying topic models to healthcare data and developing a novel triple-channel model structure to jointly accommodate three important types of data (i.e., diagnosis, medication, and contextual information) for healthcare data mining.

To evaluate MCLDA, we construct a research testbed that includes one million outpatient visits sampled from the National Health Insurance Research Dataset (NHIRD) in Taiwan. NHIRD covers the insurance claims of a random sample of one million patients from among the entire population of Taiwan since 1998. Using a ten-fold cross-validation method, our experiments show that MCLDA outperforms other baseline models, including LDA, the mixed membership model (MMM), the *k*-nearest-neighbor algorithm (kNN), logistic regression (LR), the token co-occurrence-based method (CoOccur), and prediction based on unconditional probability (Popular). This study contributes to the healthcare data mining literature by (i) successfully applying and extending the topic modeling approach to healthcare data and (ii) providing empirical evidence that MCLDA is able to discover important latent structures that can be applied to prediction tasks for clinical decision support.

The remainder of this paper is organized as follows. Section 2 reviews the related literature. Section 3 describes the proposed MCLDA model. Section 4 presents our experimental results. Finally, Section 5 summarizes our findings and concludes the paper with a brief discussion on future research directions.

2. Literature review

Diagnoses and medications are two key variables in healthcare data. Clinicians record diagnoses that characterize a patient's main problems and prescribe medications to treat the patient. These two variables, together with contextual information, such as patient demographics, summarize the main theme of a patient's visit. The richness and importance of healthcare data have inspired researchers to conduct data mining studies. In the following subsections, we first review previous studies that have conducted data mining on diagnoses and medications using ARM. We then discuss the prediction tasks involving diagnoses and medications. Further, we summarize previous topic modeling studies that use methods closely related to MCLDA. Finally, we conclude this section with a discussion on research gaps.

2.1. Mining diagnosis–medication and diagnosis–diagnosis associations

Several previous studies have focused on the research problem of identifying important associations in healthcare data, including diagnosis–medication associations and diagnosis–diagnosis associations [4,11]. The discovered diagnosis–medication associations are valuable for tasks such as identifying the clinical use of medications [12]. In contrast, the discovered diagnosis–diagnosis associations could represent the network structure among diseases. Roque et al. [13] constructed a disease network using medical records to reveal the comorbidity structures of patients in a healthcare institution.

Previous studies mainly adopted electronic medical records to construct their research testbeds. Diagnoses and medications can typically be extracted from coded structured fields [4]. Although variables constructed from structured fields are usually sufficient for data mining purposes, a previous study has reported that the free text notes in medical records can provide additional useful diagnoses [13]. This idea can be further extended by combining diagnosis–medication pairs in biomedical literature with those identified from medical records [12].

Previous studies generally adopted ARM, a collection of techniques based on co-occurring pairs, to extract association rules involving diagnoses and medications. In the context of healthcare data, a "transaction" is a list of diagnoses and medications recorded by clinicians during a patient's visit. Despite variations in research topics, previous studies often configured ARM to extract special types of association rules, such as Diagnosis X → Diagnosis Y [13], Medication X → Diagnosis Y [4], and Diagnosis Y → Medication X [12]. A previous study has reported that Chi-squared tests perform better than other metrics, such as support, confidence, interest, and conviction [4], in terms of extracting diagnosis–medication associations.

The use of ARM for extracting diagnosis–medication or diagnosis–diagnosis associations from the free text of biomedical literature is possible as long as relevant variables are extractable from the text through suitable natural language processing (NLP) tools [12]. Existing NLP tools, such as MedLEE [14,15] and BioMedLEE [16], are convenient choices for researchers. We refer readers to [17] for a more comprehensive review of NLP issues in healthcare data.

An interesting variation in this line of research is to adopt crowdsourcing to extract useful diagnosis–medication pairs from a large diagnosis–medication link dataset manually created by clinicians. Because the manually created links might be incomplete and error-prone, additional crowdsourcing-based methods can filter pairs based on the proportion of patients receiving a particular drug and with a particular diagnosis for which a link between the drug and the diagnosis has been manually asserted [18].

A common approach to evaluating the identified diagnosis–medication associations is to rank identified pairs by selected measures (e.g., Chi-squared statistics) and then manually inspect the top-ranked pairs. The diagnosis–medication pairs identified by ARM-based approaches are able to achieve a precision (in the top 500) of 89.2% [4]. Crowdsourcing was reported to have a precision of 66% for 11,029 diagnosis–medication pairs from 100 randomly

selected patients [18]. Because of the high cost to evaluate large number of possible diagnosis–medication pairs, an alternative approach is to visualize mining results that support the validity of the study [13].

2.2. Predicting diagnosis and medications

In addition to identifying associations in healthcare data, previous studies have focused on prediction tasks, including predicting diagnoses given medications [19] and predicting missing medications given diagnoses [5]. These prediction tasks have direct applications in clinical decision support systems and might lead to higher quality of healthcare.

Several approaches have been developed for these prediction tasks. For example, collaborative filtering approaches [20], including *k*-nearest-neighbor (kNN), logistic regression, and co-occurrence, have been adopted to identify missing medications [5]. The basic idea is that a missing medication is similar to an item that a consumer might purchase given historical records. The kNN approach identifies missing medications by aggregating the medications of similar records based on cosine-similarity metrics. Logistic regression identifies missing medications by estimating a model with binary outcomes for each medication. Diagnoses and other medications in the same record serve as independent variables in the regression model. The collection of regression models for all medications can then jointly rank medications that might be missing given observed diagnoses and medications. The co-occurrence approach assigns a score to each medication-medication or diagnosis–medication pair according to the number of times the pair co-occurs. Given a medical record to be tested, the score for each potentially missing medication is computed by aggregating individual scores from the observed diagnoses and medications.

Another approach is to develop decision rules based on manually coded rules. Wright et al. [11,19] developed 17 rules involving diagnosis–medication associations, medications, laboratory results, billing codes, and vital signs to infer patients' medical problems in 17 target categories.

The overall evaluations of prediction tasks yield positive results. Manually created decision rules have a sensitivity of 93.9% and a positive predictive value (PPV) of 91.7% based on 17 selected diagnosis groups [11]. A clinical trial also reported that 41.1% of identified patient problems based on ARM and other structured fields were accepted by medical practitioners [19]. The hitrate@10 (i.e., the percentage of testing instances that contain at least one correct answer among the top 10 items generated by an algorithm) for missing medication prediction given other observed diagnoses and medications is between 42% and 65%, depending on the data source [5].

2.3. Topic models

A topic model [6,7] is an algorithm that aims to discover latent structures in large document collections. The proposed MCLDA model is closely related to topic models, such as LDA and mixed membership models (MMM) [8], in several aspects. We briefly summarize these two categories of topic models below. We refer readers to [21] for additional discussions regarding the recent development of topic models.

A topic model defines how words in a document are generated through the control of latent topics. In the context of healthcare data mining, words correspond to tokens that represent diagnoses or medications prescribed during a patient's visit. A document corresponds to a healthcare record that contains these tokens. The latent topic can be interpreted as the latent health status group of a patient. In our subsequent discussion, we employ the terms "record," "token," and "latent health status group" instead of the terms "document," "word," and "latent topic" to better fit the current application domain.

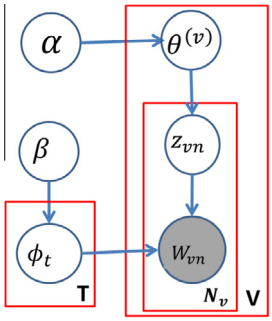


Fig. 1. Latent Dirichlet allocation (LDA).

The LDA model (see Fig. 1) assumes that when a patient visits a healthcare institution, his/her mixture of latent health status groups determines the diagnoses and medications he/she will receive during that visit. If T represents the number of health status groups, his/her health status mixture for the visit v, $\theta^{(v)}$, is drawn from a Dirichlet prior with hyper-parameter α. The variable $\theta^{(v)}$ is a vector of T elements. The element at position t represents the probability that the patient's current health status belongs to group t. All the elements in $\theta^{(v)}$ sum up to 1.

Given $\theta^{(v)}$, we are ready to generate N_v tokens in this record. For each token at position n, LDA first draws $z_{vn} \sim Multinomial(\theta^{(v)})$, and then draws $w_{vn} \sim Multinomail(\phi_{z_{vn}})$. Here, $\phi_{z_{vn}}$ is a probability vector that determines the distribution of diagnoses and medications conditioned on health status group z_{vn}. Note that the length of $\phi_{z_{vn}}$ is the total number of unique diagnoses and medications.

A potential drawback of adopting LDA for healthcare data is that the diagnoses and medications are drawn from a combined distribution $\phi_{z_{vn}}$. As a result, a record v might contain only medications but no diagnosis or only diagnoses but no medication. MMM [8] addresses this issue by creating separate "channels" for diagnoses and medications, both of which are conditioned on the same θ_v. MMM and similar models, such as Link-LDA [22], were originally developed to model words and references in a document. In our context, the "word-reference" co-occurrence becomes a "diagnosis–medication" co-occurrence. MMM generates medications and diagnoses conditioned on $\theta^{(v)}$ separately for a record v. Specifically, for the n-th medication of record v, a latent health status group $z_{vn}^{(m)} \sim Multinomial(\theta^{(v)})$ is first generated, followed by a sampling of the medication $w_{vn}^{(m)} \sim Multinomail(\phi_{m,z_{vn}})$. The vector $\phi_{m,z_{vn}}$ includes the probabilities of every medication in group z_{vn}. A similar procedure is performed to generate $w_{vn}^{(i)}$, the observed diagnoses for this record.

The data generation process of LDA and MMM defines the joint posterior of parameters and latent variables that need to be estimated. Consider, for example, the joint posterior of LDA:

$$p(\theta, \phi, z|w, \alpha, \beta) = p(\theta, \phi, z, w|\alpha, \beta)/p(w|\alpha, \beta)$$
$$\propto p(w|z, \phi)p(z|\theta)p(\theta|\alpha)p(\phi|\beta),$$

where θ, ϕ, z, and w are the collections of all $\theta^{(v)}$, ϕ_t, z_{vn}, and W_{vn}, respectively. The last three terms follow from the random variable dependency structure in LDA (see Fig. 1) and the fact that $p(w|\alpha, \beta)$ is a constant. Finding the posterior mode is a challenging task because each word W_{vn} is associated with a latent variable z_{vn}. Let V, Q_m, and Q_i denote the total number of records, number of unique medications, and number of unique diagnoses, respectively. Then, there are $T \times V$ elements in θ and $T \times (Q_m + Q_i)$ elements in ϕ. The

total number of unknowns would be too large for a moderately sized dataset, with thousands of unique diagnoses and medications.

One popular approach adopted in previous studies to address this challenge is the collapsed Gibbs sampling (CGS) method. CGS is a sampling-based algorithm that constructs a Markov chain whose limiting distribution is the joint posterior of latent variables and parameters [23,24]. CGS integrates out $\theta^{(v)}$ to improve the estimation efficiency. This type of algorithm can approximate the joint posterior to an arbitrary precision, given unlimited computing resources. In practice, a fixed number of iterations will be run to collect sufficient samples for subsequent inference tasks.

2.4. Research gaps

This literature review includes several remarkable observations. First, few previous studies have used healthcare insurance claims in this line of research. Although an insurance claim record contains only a subset of information from a medical record, it is usually collected from many healthcare institutions, implying that a healthcare insurance claim dataset can be a suitable testbed for investigating general healthcare data modeling problems that reflect variations across healthcare institutions.

Second, ARM and supervised learning methods such as kNN are the most popular technical approaches in this line of research. Although some degree of success has been reported in previous studies, existing approaches have several limitations. ARM discovers item-level associations but does not consider other factors such as patient demographics, a fact that may influence the strength of associations. In addition, ARM does not provide higher-level abstractions that are useful for data exploration and decision support. Moreover, previous studies have reported their empirical results based on small samples with fewer than 200 unique medications and therapeutic classes [5]. Because a typical hospital usually hands out thousands of unique diagnoses and drugs (see, e.g., [18]), larger testbeds should provide additional insights into the modeling of healthcare data.

Finally, topic models such as LDA and MMM may provide viable alternatives for healthcare data modeling. However, thus far, LDA has mostly been applied to textual data for latent topic discovery. MMM was developed to handle two types of data—text and references—for similar applications. Extending topic models to the context of healthcare data modeling and jointly incorporating three types of healthcare data (diagnoses, medications, and contextual information) require additional efforts toward model development in order to leverage the full potential of topic models in discovering the latent structures in healthcare data.

3. Multiple channel latent Dirichlet allocation (MCLDA)

We propose an MCLDA model to better represent the co-occurrences among medications, diagnoses, and contextual information in healthcare data. Based on the existing topic models (i.e., LDA and MMM), we have expanded the model structure such that three types of observed variables (medications, diagnoses, and contextual information) can be accommodated jointly. In addition, we have developed a CGS-based model inference method to discover the latent structures and to perform prediction tasks based on the inferred models. First, we summarize the proposed MCLDA model. Then, we present the CGS-based model inference method.

3.1. The model

Fig. 2 shows the model structure of the proposed MCLDA. MCLDA assumes that an observed datapoint in a record v includes three types of data: medications ($W_{vn}^{(m)}$), diagnoses ($W_{vn}^{(i)}$), and contextual information ($W_{vn}^{(h)}$). The subscript n denotes the position of a token in the record. A record v has N_{vm} medications, N_{vi} diagnoses, and N_{vh} contextual information variables. We assume that each data type has at least one token (i.e., $N_{vm} > 0$, $N_{vi} > 0$, and $N_{vh} > 0$). Our model can be generally applied to the degenerate cases when one or more types of variables are missing.

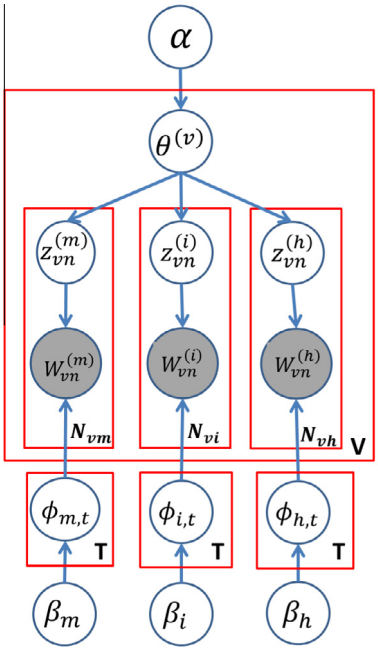


Fig. 2. Multiple-channel latent Dirichlet allocation (MCLDA).

MCLDA defines a data generation process that involves several sets of variables. The first set of variables includes a mixture of latent health status groups $\theta^{(v)}$ of record v. The variable $\theta^{(v)}$ characterizes the record by specifying the probability that each token is associated with a latent health status group. Each health status group is a set of patient visits with similar health conditions. Following previous topic modeling studies, we assume that $\theta^{(v)}$ follows a Dirichlet distribution to simplify subsequent computation [6].

The second set of variables represents the probability of tokens given a health status group. The vector $\phi_{m,t}$ includes the probability of each medication in health status group t. Similar definitions apply to $\phi_{i,t}$ and $\phi_{h,t}$, which are the probabilities of diagnoses and contextual information, respectively, in health status group t. The variables $\phi_{m,t}$, $\phi_{i,t}$, and $\phi_{h,t}$ are assumed to be generated from Dirichlet distributions with parameters β_m, β_i, and β_h, respectively.

To generate an observed medication $W_{vn}^{(m)}$ in record v, we first generate the latent health status group for this token, $z_{vn}^{(m)}$, by drawing $z_{vn}^{(m)} \sim multinomial(\theta^{(v)})$. The corresponding token distribution, $\phi_{m,z_{vn}^{(m)}}$, is then selected to generate the observed token by drawing $W_{vn}^{(m)} \sim multinomial(\phi_{m,z_{vn}^{(m)}})$. The diagnoses and contextual information variables are generated similarly.

We refer to the substructure in MCLDA that generates one type of observed data as a "channel." The proposed model has three channels that generate medications ($W_{vn}^{(m)}$), diagnoses ($W_{vn}^{(t)}$), and contextual information ($W_{vn}^{(h)}$), respectively. Evidently, LDA has only one channel, which becomes a constraint when modeling healthcare data. MMM has two channels and can thus alleviate the issue of cramming different types of data in one channel. MCLDA includes three channels, one for each data type. This design allows MCLDA to (i) better organize the observed data and their

latent structures through the common underlying latent health status groups and (ii) separate vocabulary variables $\phi_{m,t}$, $\phi_{i,t}$, and $\phi_{h,t}$.

3.2. Model inference

We develop a CGS-based inference method for MCLDA. We integrate out $\theta^{(v)}$ and update $z_{vn}^{(m)}$, $z_{vn}^{(i)}$, and $z_{vn}^{(h)}$ individually during model inference. To achieve this objective, we need to compute the posterior of the latent health status group for each token:

$$p\left(z_{vn}^{(m)}=j\middle|z_{\cdot\cdot}^{(m)}_{-vn},z_{\cdot\cdot}^{(i)},z_{\cdot\cdot}^{(h)},w_{\cdot\cdot}^{(m)},w_{\cdot\cdot}^{(i)},w_{\cdot\cdot}^{(h)}\right),$$

$$p\left(z_{vn}^{(i)}=j\middle|z_{\cdot\cdot}^{(m)},z_{\cdot\cdot}^{(i)}_{-vn},z_{\cdot\cdot}^{(h)},w_{\cdot\cdot}^{(m)},w_{\cdot\cdot}^{(i)},w_{\cdot\cdot}^{(h)}\right),\ \text{and}$$

$$p\left(z_{vn}^{(h)}=j\middle|z_{\cdot\cdot}^{(m)},z_{\cdot\cdot}^{(i)},z_{\cdot\cdot}^{(h)}_{-vn},w_{\cdot\cdot}^{(m)},w_{\cdot\cdot}^{(i)},w_{\cdot\cdot}^{(h)}\right).$$

Note that $z_{\cdot\cdot}^{(m)}_{-vn}$ denotes the latent health status groups of all medications except the one for the n-th medication in record v. The variables $w_{\cdot\cdot}^{(m)}$, $w_{\cdot\cdot}^{(i)}$, and $w_{\cdot\cdot}^{(h)}$ denote the collection of all medications, diagnoses, and contextual information, respectively. Our subsequent discussion focuses on $p\left(z_{vn}^{(m)}=j\middle|\cdot\right)$. The other two posteriors can be derived in a similar manner.

The posterior of the latent health group for the n-th medication in record v is

$$\begin{aligned}
&p\left(z_{vn}^{(m)}=j\middle|z_{\cdot\cdot}^{(m)}_{-vn},z_{\cdot\cdot}^{(i)},z_{\cdot\cdot}^{(h)},w_{\cdot\cdot}^{(m)},w_{\cdot\cdot}^{(i)},w_{\cdot\cdot}^{(h)}\right)\\
&\quad\propto p\left(w_{vn}^{(m)}\middle|z_{vn}^{(m)}=j,z_{\cdot\cdot}^{(m)}_{-vn},w_{\cdot\cdot}^{(m)}_{-vn},w_{\cdot\cdot}^{(i)},w_{\cdot\cdot}^{(h)}\right)\\
&\qquad\times p\left(z_{vn}^{(m)}=j\middle|z_{\cdot\cdot}^{(m)}_{-vn},z_{\cdot\cdot}^{(i)},z_{\cdot\cdot}^{(h)},w_{\cdot\cdot}^{(m)}_{-vn},w_{\cdot\cdot}^{(i)},w_{\cdot\cdot}^{(h)}\right)\\
&\quad\propto\frac{\beta_m+c_{-vn,j}^{(w_{vn}^{(m)})}}{Q_m\beta_m+c_{-vn,j}^{(M)}}\frac{\alpha+c_{-vn,j}^{(v_m)}+c_{\cdot,j}^{(v_i)}+c_{\cdot,j}^{(v_h)}}{T\alpha+N_{vm}+N_{vi}+N_{vh}-1}.
\end{aligned}\tag{1}$$

The second and third lines in Eq. (1) follow from the Bayesian theorem; the last line can be achieved by integrating out $\theta^{(v)}$. Note that $c_{-vn,j}^{(w_{vn}^{(m)})}$ denotes the number of medications $w_{vn}^{(m)}$ assigned to health status group j, excluding the assignment at position vn. The variable $c_{-vn,j}^{(M)}$ denotes the number of medications M assigned to health status group j, excluding the assignment at position vn. The first term in the last line of Eq. (1) is the fraction of medications in health status group j that have the value $w_{vn}^{(m)}$, excluding the medication at position vn. The Dirichlet prior smoothens the probability estimation by adding β_m and $Q_m\beta_m$ to the numerator and denominator, respectively.

The second term in the last line of Eq. (1) captures the tendency that the record v belongs to group j. The count $c_{-vn,j}^{(v_m)}$ denotes the number of medications in record v that have been assigned to group j, excluding the medication at position vn. Similarly, $c_{\cdot,j}^{(v_i)}$ ($c_{\cdot,h}^{(v_h)}$) denotes the number of diagnoses (contextual information tokens) in record v that have been assigned to group j. The Dirichlet prior contributes to the smoothing variables α and $T\alpha$. In contrast, the first term in the last line of Eq. (1) captures the tendency that the observed medication $w_{vn}^{(m)}$ belongs to group j.

Following a similar procedure, we derive the posteriors of the latent health status groups for diagnoses and medications:

$$\begin{aligned}
&p\left(z_{vn}^{(i)}=j\middle|z_{\cdot\cdot}^{(i)}_{-vn},z_{\cdot\cdot}^{(m)},z_{\cdot\cdot}^{(h)},w_{\cdot\cdot}^{(m)},w_{\cdot\cdot}^{(i)},w_{\cdot\cdot}^{(h)}\right)\\
&\quad\propto\frac{\beta_i+c_{-vn,j}^{(w_{vn}^{(i)})}}{Q_i\beta_i+c_{-vn,j}^{(I)}}\frac{\alpha+c_{-vn,j}^{(v_i)}+c_{\cdot,j}^{(v_m)}+c_{\cdot,j}^{(v_h)}}{T\alpha+N_{vm}+N_{vi}+N_{vh}-1},
\end{aligned}\tag{2}$$

and

$$\begin{aligned}
&p\left(z_{vn}^{(h)}=j\middle|z_{\cdot\cdot}^{(h)}_{-vn},z_{\cdot\cdot}^{(m)},z_{\cdot\cdot}^{(i)},w_{\cdot\cdot}^{(m)},w_{\cdot\cdot}^{(i)},w_{\cdot\cdot}^{(h)}\right)\\
&\quad\propto\frac{\beta_h+c_{-vn,j}^{(w_{vn}^{(h)})}}{Q_h\beta_h+c_{-vn,j}^{(H)}}\frac{\alpha+c_{-vn,j}^{(v_h)}+c_{\cdot,j}^{(v_m)}+c_{\cdot,j}^{(v_i)}}{T\alpha+N_{vm}+N_{vi}+N_{vh}-1}.
\end{aligned}\tag{3}$$

Eqs. (1)–(3) lay the foundation for the model estimation algorithm. Algorithm 1 in Appendix A summarizes the model inference procedure based on these equations. The initial groups in $z_{vn}^{(m)}$, $z_{vn}^{(i)}$, and $z_{vn}^{(h)}$ are randomly assigned. The latent groups are then sequentially updated using Eqs. (1)–(3). To reduce the impact of randomly assigned initial values, we discard the first B sweeps. The algorithm will record $z_{vn}^{(m)}$, $z_{vn}^{(i)}$, and $z_{vn}^{(h)}$ for every g sweeps to reduce the autocorrelation among estimated latent groups.

The output of Algorithm 1 is used to estimate the parameters in MCLDA. The token probability of each data type conditional on a latent health status group summarizes the latent structures of healthcare data. These parameters, including $\phi_{m,t}$, $\phi_{i,t}$, and $\phi_{h,t}$, can be estimated from the recorded $z_{\cdot\cdot}^{(m)}$, $z_{\cdot\cdot}^{(i)}$ and $z_{\cdot\cdot}^{(h)}$ in Algorithm 1. Let $\hat{\phi}_{m,t}=(\hat{\phi}_{m,t,1},\hat{\phi}_{m,t,2},\ldots,\hat{\phi}_{m,t,Q_m})$ be the medication distribution conditioned on health status group t. Then,

$$\hat{\phi}_{m,t,u}=\frac{\beta_m+c_{\cdot,j}^{(u)}}{Q_m\beta_m+c_{\cdot,j}^{(M)}},\tag{4}$$

where u is an index for medications and $c_{\cdot,j}^{(u)}$ is the number of medications u assigned to group t in a recorded sweep. Similar methods can be used to compute $\hat{\phi}_{i,t,u}$ and $\hat{\phi}_{h,t,u}$.

Another important parameter is $\hat{\theta}^{(v)}$, the mixture of latent health status groups for a record v. Let $\hat{\theta}^{(v)}=\left(\hat{\theta}_1^{(v)},\theta_2^{(v)},\ldots,\hat{\theta}_T^{(v)}\right)$. Then, the posterior probability that a record belongs to group t is $\hat{\theta}_t^{(v)}=\frac{\alpha+c_{\cdot,t}^{(v_m)}+c_{\cdot,t}^{(v_i)}+c_{\cdot,t}^{(v_h)}}{T\alpha+N_{vm}+N_{vi}+N_{vh}}$. Averaging the results from multiple recorded sweeps typically can improve the precision of parameter estimation.

3.3. Prediction using MCLDA

The estimated model can be employed to predict missing diagnoses or medications for a patient visit. To simplify our discussion, we focus on predicting diagnoses given medications and contextual information. This task can arise in a clinical setting when a clinician has entered medications and contextual information but has not yet entered diagnoses into the system. Another similar task, i.e., predicting medications based on observed diagnoses and contextual information, can arise if diagnoses were entered first.

Let $w_{v\cdot}^{(m)}$, $w_{v\cdot}^{(i)}$, and $w_{v\cdot}^{(h)}$ denote the medications, diagnoses, and contextual information in record v. We want to predict $w_{v\cdot}^{(i)}$ based on other observed tokens. To achieve this using a learned MCLDA model, we first estimate $\theta^{(v)}$, the mixture of health status groups, conditioned on other observed tokens. Specifically, the posterior probability that an observed medication belongs to group j is

$$\begin{aligned}
&p\left(z_{vn}^{(m)}=j\middle|\hat{\phi}_m,\hat{\phi}_i,\hat{\phi}_h,z_{\cdot\cdot}^{(m)}_{-vn},z_{v\cdot}^{(h)},w_{v\cdot}^{(m)},w_{v\cdot}^{(h)}\right)\\
&\quad\propto p\left(w_{vn}^{(m)}\middle|z_{vn}^{(m)}=j,\hat{\phi}_m\right)p\left(z_{vn}^{(m)}=j\middle|z_{\cdot\cdot}^{(m)}_{-vn},z_{v\cdot}^{(h)}\right)\\
&\quad=\hat{\phi}_{m,j,w_{vn}^{(m)}}\frac{\alpha+c_{-vn,j}^{(v_m)}+c_{\cdot,j}^{(v_h)}}{T\alpha+N_{vm}+N_{vh}-1}.
\end{aligned}$$

Note that this equation is similar to but not the same as (2) because $\hat{\phi}_m$ is estimated from the training data using (4). A similar equation can compute the posterior of the contextual information variables. Starting with a random assignment, each latent group for medications and contextual information can be updated

sequentially for a predefined number of sweeps. We can then compute the mixture of latent health status groups by counting the number of tokens assigned to a health status group:

$$\hat{\theta}_t^{(v)} = \frac{\alpha + c_{\cdot,t}^{(v_m)} + c_{\cdot,t}^{(v_h)}}{T\alpha + N_{vm} + N_{vh}}.$$

The probability of observing a diagnosis conditioned on the observed medications and contextual information is

$$p\left(w_{vn}^{(i)} | w_{v\cdot}^{(m)}, w_{v\cdot}^{(h)}\right) = \int p\left(w_{vn}^{(i)} | \theta^{(v)}\right) p\left(\theta^{(v)} | w_{v\cdot}^{(m)}, w_{v\cdot}^{(h)}\right) d\theta^{(v)} \approx \frac{1}{S'} \sum_{s=1}^{S'} \sum_{t=1}^{T} \phi_{i,t,w_{vn}^{(i)}}^{(s)} \theta_t^{(v)(s)}, \quad (5)$$

where S' is the total number of recorded sweeps during model training, and the superscript s marks the parameters computed according to a specific recorded sweep. The inner summation computes the weighted average of a diagnosis through an estimated mixture of health status groups, based on the s-th recorded sweep from Algorithm 1. The outer summation averages the result over all recorded sweeps. To provide the top n predictions, we rank $p(w_{vn}^{(i)}|\cdot)$ in descending order and output the first n diagnoses.

4. Experimental results

In this section, we first introduce our research testbed, followed by a detailed discussion regarding prediction tasks, performance evaluation measures, baseline models, and hyper-parameter tuning. We then present the estimation results of MCLDA and compare MCLDA with the baseline models. We computed all prediction performance measures using ten-fold cross-validation.

4.1. Research testbed

We employed the National Health Insurance Research Dataset (NHIRD) of Taiwan for our experiments. NHIRD includes the health insurance claim data of one million randomly sampled Taiwanese citizens since 1998. We dropped records before 2002 because some health institutions did not adopt International Classification of Diseases, Ninth Revision (ICD-9) codes until 2002. Our research testbed consists of one million records of outpatient visits randomly sampled from NHIRD. We employed NHIRD to construct our research testbed because, compared to clinical data that are often institution-specific, NHIRD covers a large population and offers a broad collection of data from different healthcare institutions. Table 1 summarizes the statistics of our research testbed.

We did not consider a diagnosis or medication if it appeared fewer than ten times in our research testbed. Our testbed included 3830 unique diagnoses, 1416 unique medications, and 60 unique contextual information tokens. Each record included at least one diagnosis and at least one medication. In this study, we considered three types of contextual information: patient sex, patient age, and medical specialty. We grouped patient age into intervals of five years. For example, the age group 5 includes patients between the ages of 25 and 30.[1] The medical specialty of an outpatient visit is the specialty of the clinician attending this outpatient visit. Our testbed includes 45 specialties. Examples include general medicine, family medicine, and gastroenterology.

Most records in our research testbed were from different patients. Table 2 summarizes the number of records by patient. Our dataset involves 494,673 patients. Among them, 50.20% of patients contributed only one record per patient, 25.00% of patients contributed two records per patient, and 21.31% of patients contributed three to five records. The remaining 3.508% of patients contributed more than six records per patient. We noted that patients who contributed more records often had a higher median age. To reflect real world scenario and usage, our research testbed allowed multiple records for the same patients. As a result, when conducting a ten-fold cross-validation, records from the same patient may appear in both training and testing folds, which may bias our reported performance because records from the same patient may be similar.

To provide concrete ideas regarding our research testbed, we present two anonymized records in Table 3. Each record in our research testbed consists of diagnoses, medications, and contextual information. A diagnosis is coded with ICD-9, and a medication is

[1] Age group index starts from 0.

Table 1
Summary statistics of our research testbed.

Item	Value
# of records	1,000,000
# of diagnosis tokens	1,680,256
Avg. length of diagnosis tokens	1.68
# of unique diagnoses	3830
# of medication tokens	3,386,315
Avg. length of medication tokens	3.39
# of unique medications	1416
# of contextual information tokens	2,999,994
Avg. length of contextual information tokens	3.00
# of unique contextual information tokens	60

Table 2
Number of records by patient.

# of Records Per Patient	# of Patients	# of Records	Median Age of Patients
1	248,296 (50.20%)	248,296	35
2	123,626 (25.00%)	247,252	39
3–5	105,397 (21.31%)	376,545	48
6–10	16,146 (3.264%)	112,302	61
11–15	1062 (0.215%)	12,946	66
16–20	123 (0.025%)	2123	65
21–25	19 (0.004%)	426	71
26–30	4 (0.001%)	110	46
Total	494,673 (100%)	1,000,000	–

Table 3
Sample records of our research testbed.

Record	Tokens
Sample record 1	
Diagnosis	4781 Other diseases of nasal cavity and sinuses
	4650 Acute laryngopharyngitis
	46410 Acute tracheitis without mention of obstruction
Medication	NOSCAPINE
	DEXCHLORPHENIRAMINE MALEATE
	IBUPROFEN
Contextual information	Age_Group_09
	Sex_F
	Div_Otolaryngology
Sample record 2	
Diagnosis	5715 Cirrhosis of liver without mention of alcohol
	1550 Malignant neoplasm of liver, primary
	07030 Viral hepatitis B without mention of hepatic coma, acute or unspecified, without mention of hepatitis delta
Medication	URSODEOXYCHOLIC ACID
	BIOFLAVINOIDS (=CITRUS FLAVONOID COMPOUNDS)
	ACETAMINOPHEN (=PARACETAMOL)
	PANTOPRAZOLE SODIUM SESQUIHYDRATE
Contextual information	Age_Group_11
	Sex_M
	Div_Gastroenterology

coded with an internal ID that is omitted here. The contextual information includes three tokens that specify the age group, sex, and medical specialty.

4.2. Prediction tasks

We considered two prediction tasks in our experiments: medication prediction and diagnosis prediction. Medication prediction aims to forecast the medications that should appear together with the diagnoses and contextual information in a record. Similarly, diagnosis prediction aims to deduce the assigned diagnosis codes given medications and contextual information. We constructed a testing record for medication (diagnosis) prediction by hiding the medications (diagnoses) in the record. For each run of the cross-validation, we trained a model using 900,000 records and tested it on the remaining 100,000 records.

We selected these two tasks mainly because of their practical significance. Patient problem lists are often inaccurate, incomplete, and out of date [19,25]. The diagnosis prediction task can be applied to the patient's medical history in order to improve the accuracy and completeness of the problem list based on the medications. Moreover, clinicians need to enter diagnoses and medications during a patient's visit. These two tasks reflect the potential recommendations that a clinical decision support system can provide when the data-entry process has been half completed. Because clinicians often order more than one medication or assign more than one diagnosis, we evaluate the prediction performance based on the complete list of diagnoses and medications instead of focusing on diagnosis–medication pairs that may not fully reflect the complexity faced by a clinician when prescribing medicines. Finally, because the ground truth is readily available in the research testbed, we did not need to engage domain experts to create gold standards for our experiments.

4.3. Performance evaluation measures

We considered several performance measures, including normalized discounted cumulative gain (NDCG), mean average precision (MAP), hitrate@n, and perplexity. We selected the first three measures because, when conducting diagnosis or medication prediction, the model outputs an ordered list of predictions that need to be evaluated based on the rank order of the correct answer. The last evaluation measure (perplexity) is a function of hold-out log-likelihood and is often used in latent topic models studies. We briefly discuss these measures below.

NDCG measures the ranking quality of an ordered list of predictions. A list with correct answers appearing earlier in the list is preferred. Following the literature [26], we computed NDCG of an ordered list of predictions by first computing the discounted cumulative gain (DCG):

$$\mathrm{DCG} = \mathrm{rel}_1 + \sum_{i=2}^{n} \frac{rel_i}{\log_2(i)},$$

where $\mathrm{rel}_i \in \{0, 1\}$, $i = 1, 2, \ldots, n$, indicates whether the prediction at position i is correct ($i = 1$) or not ($i = 0$), and n is the total number of possible answers. A correct prediction at position i is discounted with $\log_2(i)$. The best possible DCG, denoted as ideal DCG (IDCG), is an ordered list of predictions that have all correct answers up front. Then, NDCG is the DCG normalized by IDCG:

$$\mathrm{NDCG} = \frac{\mathrm{DCG}}{\mathrm{IDCG}}.$$

NDCG can be interpreted as the normalized prediction performance level of a prediction task. For example, an NDCG of 45% means that the current DCG is 45% of the IDCG. We report the averaged NDCG of the testing records.

MAP is another popular measure for an ordered list of predictions. The basic idea is to compute the averaged precision based on the positions of correct answers for each testing record and then average over all the testing records [27]. For a testing record a, the average precision is $AP(a) = \left(\frac{1}{m_a}\right)\sum_{k=1}^{m_a} precision(rank_k)$, where m_a is the number of correct answers in this testing record, $rank_k$, $k = 1, 2, \ldots, m_a$ are the positions of the correct answers in the ordered list of predictions, starting from the head, and $precision(rank_k)$ is the precision computed according to the top $rank_k$ predictions.

For example, if there are three correct answers in testing record a, and these answers appear in positions 2, 5, and 8 of an ordered list of predictions, then the precision at position 2, $precision(2)$, is $1/2$ because at position 2, there are two predictions and one of them is correct. Similarly, $precision(5) = 2/5$ and $precision(8) = 3/8$. The average precision of this testing record is $\frac{1}{3}\left(\frac{1}{2} + \frac{2}{5} + \frac{3}{8}\right) \approx 0.425$.

MAP is simply the average of all testing records:

$$\mathrm{MAP} = \frac{1}{|T|}\sum_{a \in T} AP(a),$$

where T is the set of testing records.

Hitrate@n is defined as the percentage of records that include at least one correct prediction [28]. Specifically, if there are N testing records, and among these records, the top n predictions are able to correctly guess at least one token in N_C records, then $hitrate@n = N_c/N$. Hitrate@n has been applied to recommender system evaluation [29] and missing medication prediction [5]. We consider $n = 1$, 3, and 5 in our experiments because the top predictions are more relevant in our setting.

Perplexity is an evaluation measure commonly used in topic modeling [7,30]. Given a set of testing tokens W_{test}, the perplexity of W_{test} is $perplexity(W_{test}) = \exp\{-L(W_{test})/C_{W_{test}}\}$, where $C_{W_{test}}$ is the count of testing tokens and $L(W_{test})$ is the log-likelihood of the testing tokens computed according to an estimated model. For MCLDA, the predictive probability of a testing token can be computed using Eq. (5) and $L(W_{test}) = \sum_{i=1}^{C_{W_{test}}} \log \hat{w}_i$, where $\hat{w}_i$ is the estimated probability of a testing token.

Perplexity is the exponential of negative average log-likelihood. A large perplexity, which equates to a low average log-likelihood, suggests poor performance. If a model always assigns a probability of 1.0 to the correct answer, then the resultant perplexity is 0. The perplexity of guessing that a ball is placed in one of G boxes with equal probability is G; this example intuitively shows that the perplexity of a model can be interpreted as the average "difficulty" of guessing a testing token.

4.4. Baseline models

In this study, we considered six baseline models for our prediction tasks. The first two baseline models, MMM and LDA, are topic models that have been used in text mining studies. We applied these two models to the context of healthcare data mining as a way of understanding the performance of existing text mining approaches. MMM takes observed medications and contextual information in one channel, and diagnoses in another channel when conducting diagnosis prediction. Contextual information is merged with diagnosis when conducting medication prediction. LDA is trained on observed medications, diagnoses, and contextual information combined.

In addition to topic models, we considered several prediction methods, including CoOccur, kNN, logistic regression (LR), and

Popular. These methods have been applied in previous studies. We briefly introduce these methods below.

CoOccur [5] ranks candidate tokens based on co-occurrence frequency. For example, in the case of predicting diagnoses given medications $\hat{w}^{(m)} = \left\{w_{x,1}^{(m)}, w_{x,2}^{(m)}, \ldots w_{x,n_x}^{(m)}\right\}$, a diagnosis i is assigned a score as follows:

$$score_{co}(i) = \sum_{g \in \hat{w}^{(m)}} \sum_{r \in D_{train}} I(g \in r \text{ and } i \in r),$$

where D_{train} is the training data, r is a record, and $I(\cdot)$ is an indicator function. The diagnoses with greater scores are ranked higher for prediction.

The kNN [31] method performs prediction based on the k examples in the training dataset that bear the closest similarity to the testing record. Following a previous study [5], we adopted the Ochiai similarity (equivalent to the cosine similarity for records with binary features only) to compute the distance between two records r_i and r_j:

$$dist(r_i, r_j) = \sqrt{\frac{a}{|r_i|}\frac{a}{|r_j|}},$$

where a is the number of tokens co-present in r_i and r_j, and $|r_i|$ and $|r_j|$ are the token lengths of these two records. Consider the case of predicting diagnoses given medications. Using the Ochiai similarity, we can identify k records from the training data that are closest to the testing record and then compute v_i^K, the frequency with which diagnosis i appears in this subset. The prediction is then made based on v_i^K. The parameter k is selected via a parameter-tuning procedure using the training data.

LR is a statistical model that estimates the conditional probability of a token. Consider again the case of predicting diagnoses given observed medications. Then, LR models the diagnosis prediction problem as

$$prob(y_{ji} = 1|x_j) = \frac{1}{1 + \exp(x_j^T b_i)},$$

where x_j is the vector of dummy variables representing the medications and contextual information in record j, and b_i is a parameter vector to be estimated. Since there may be more than one diagnosis in a record, this is a multi-label classification problem [32]. We adopted the binary relevance method and constructed a multi-label classifier using a collection of 3830 binary logistic regression models. In this formulation, each diagnosis i has its own model and can be estimated separately using the training data. For a testing record r_a, we ranked potential diagnoses by $\hat{p}_i = 1/(1 + \exp(x_a^T \hat{b}_i))$.

We considered several variations of logistic regression, including logistic regression with L2 regularization, L1 regularization, and the elastic-net models [33]. Specifically, for diagnosis i, we optimized the penalized log-likelihood $\max_{b_i}\left\{\frac{1}{N}\sum_{a=1}^{N}[I(y_{ai} = 1) \log prob(y_{ai}|x_a) + I(y_{ai} = 0)\log(1 - prob(y_{ai}|x_a))] - \lambda\sum_{g=1}^{fm}\left[\frac{1}{2}(1-\alpha)b_{ig}^2 + \alpha|b_{ig}|\right]\right\}$, where N is the size of the training data, y_{ai} indicates whether diagnosis i is present in record a, and fm is the number of features for this prediction task. We denoted the model with $\alpha = 0$ (L2 regularization) as LR_L2, $\alpha = 0.25$ as LR_ENET25, $\alpha = 0.5$ as LR_ENET50, $\alpha = 0.75$ as LR_ENET75, and $\alpha = 1$ (L1 regularization) as LR_L1. The parameter λ is tuned on the basis of the training data. We employed the glmnet package (https://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html) in R for the LR family model inference.

Our last baseline model, Popular, conducts prediction based on unconditional probabilities in the training data [5,34]. For the case of predicting diagnoses given medications, we estimate the probabilities in the training data and then rank all diagnoses based on this score. This method generates the same prediction for all testing records because it does not consider other information that might exist in a testing record.

4.5. Hyper-parameter tuning

We need to determine several hyper-parameters in MCLDA. The most important hyper-parameter is the number of latent health status groups (T). Following previous studies [6], we conducted a grid search using training data. Fig. 3 plots the perplexity of medication and diagnosis predictions using the training data of the first run of ten-fold cross validation. We found several interesting patterns. First, the perplexity of the diagnosis prediction is lower than that of the medication prediction. This result suggests that, in general, predicting diagnoses given medications is easier than predicting medications given diagnoses. Second, the perplexity of diagnosis prediction reaches the lowest value when the number of groups is 70; it remains low for 70–100 groups and begins to increase when the number of groups increases beyond 100. A similar pattern appears for the perplexity of medication prediction.

Based on the result, we set the number of latent groups to 70. MMM and LDA showed similar patterns, so we decided to set the

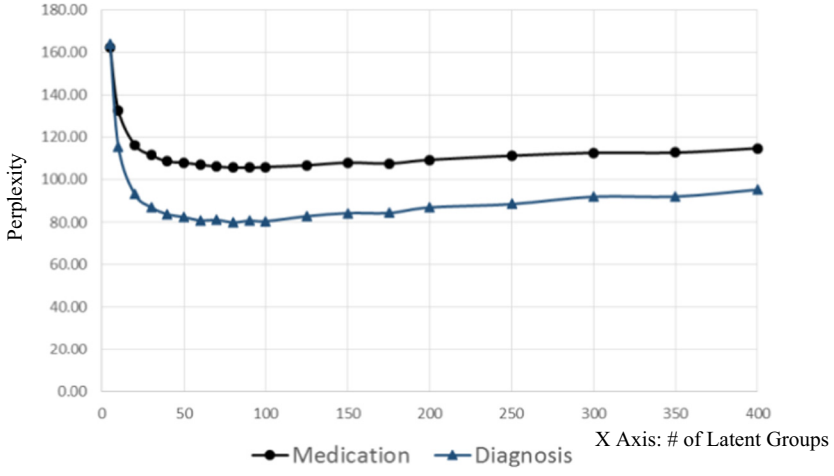


Fig. 3. Perplexity attained by MCLAD for medication and diagnosis prediction with different numbers of latent groups.

Table 4
List of hyper-parameters for the logistic regression family.

Model	α	λ for medication prediction	λ for diagnosis prediction
LR_L2	0	0.0067	0.0016
LR_ENET25	0.25	0.0016	0.0004
LR_ENET50	0.50	0.0008	0.0004
LR_ENET75	0.75	0.0033	0.0033
LR_L1	1.00	0.0004	0.0002

number of latent health status groups to 70 for these two models as well. We also applied Minka's fixed-point approach (see Section 3 of [35]) to update α, β_m, and β_i in MMM and LDA.

We set the hyper-parameter k in the kNN model to 50, based on a grid search. The hyper-parameters for the logistic regression family are listed in Table 4. To save time, we tuned the hyper-parameters using the training data of the first fold during a ten-fold cross-validation. The other nine folds used the same hyper-parameters selected in the first run of the ten-fold cross-validation.

4.6. MCLDA inference results

We set the total number of sweeps to 1500 and discarded the first 1000 sweeps. The thinning coefficient was 20 (i.e., record sampling results for every 20 sweeps). Our preliminary experiments suggest that the inference algorithm converges fairly quickly and that 1500 sweeps are more than sufficient to give a reasonable estimation result. Most of the key observations remain the same if the total number of sweeps increases from 1500 to 15,000.

We employed Eq. (4) to compute the estimated latent health status groups. Table 5 lists the top five diagnoses, medications, and contextual information in a selected health status group. The percentage in front of each diagnosis is the conditional probability of observing this diagnosis in this latent health status group. The conditional probabilities of all diagnoses sum to one. Similar interpretations apply to the medication and contextual information. In this group, there is a 50% probability that the diagnosis is an acute upper respiratory infection (ICD-9 = 4659). Other main diagnoses, such as influenza (ICD-9 = 4871) and acute nasopharyngitis (ICD-9 = 460), are also related to health issues connected to the flu or common cold. The top medications in this group are the common medicines for colds, such as acetaminophen for pain relief and dextromethorphan HBR to suppress coughing. Note that CIMETIDINE H2, a gastroprotective agent, appeared in the top five list mainly due to the common clinical practice in Taiwan of prescribing gastroprotective agents for patients (particularly elderly patients) [36]. The contextual information showed that most patients in this group visit clinicians in the general medicine and family medicine divisions. Moreover, female patients and older patients (age group 14: ages 70–75) are more common in this group. Since the contextual information contains different types of information such as age, sex, and medical specialty, the listed probability should be normalized within each type for correct interpretation. For example, the probability of observing a female patient within this latent health status group is 56% (=$18\% \div (18\% + 14\%)$). The normalization process does not change the relative magnitude of probability within each information type. As a result, comparing the listed probability of contextual information within the same type can provide intuition regarding the patient background in this group. A complete list of latent health status groups can be found in the Appendix B.

Table 5
Selected latent health status group (Group 28).

Diagnosis	Medication	Contextual information
[50%] 4659 Acute upper respiratory infections of unspecified site	[12%] ACETAMINOPHEN (=PARACETAMOL)	[21%] Div. General
[12%] 4871 Influenza with other respiratory manifestations	[4%] DEXTROMETHORPHAN HBR	[18%] Sex_F
[8%] 460 Acute nasopharyngitis [common cold]	[3%] CHLORPHENIRAMINE MALEATE	[14%] Sex_M
[5%] 7840 Headache	[2%] CIMETIDINE H2	[7%] Div. Family
[4%] 4658 Acute upper respiratory infections of other multiple sites	[2%] MEFENAMIC ACID	[5%] Age Group 14

In addition to the parameters for health status groups, MCLDA assigns each diagnosis, medication, and contextual information token a latent health status group during model inference. The assigned latent group can identify the pairing relationships between diagnoses and medications. We have omitted contextual information in the subsequent discussion for brevity.

Consider Record A1 listed in Table 6. Two diagnoses, acute upper respiratory infection of unspecified site (ICD-9 = 4659) and type II diabetes mellitus (ICD-9 = 25000) were assigned to Group 12 and Group 41, respectively. Among the five medications prescribed, glyburide and metformin HCl are antidiabetic drugs and were assigned to group 41, the same group as that of type II diabetes mellitus. The remaining medications, acetaminophen and terfenadine, were assigned to group 12, the same group as that of acute upper respiratory infection. Such group assignments within a record suggest that glyburide is used to treat type II diabetes mellitus while the remaining medicines are related to acute upper respiratory infection. We confirm this speculation by checking the medication properties using Micromedex. The group assignments for individual tokens and the estimated parameters of the entire testbed suggest that MCLDA is able to identify latent grouping structures in healthcare data and can provide useful information for clinical decision support.

The estimated latent health status groups can also provide a bird's-eye view of the training dataset. For this purpose, we visualize the similarity of each pair of groups based on their diagnosis and medication distributions. We adopted the symmetric Kullback–Leibler (KL) divergence to measure distribution similarity [27]. Smaller symmetric KL divergence indicates more similar distributions between the two groups. Specifically, given two diagnosis distributions Q_i and Q_j of latent health status groups i and j, the KL divergence of Q_j from Q_i is $D_{KL}(Q_i \| Q_j) = \sum_a Q_i(a) \ln Q_i(a)/Q_j(a)$. The symmetric KL divergence is defined as $S_{DL}(Q_i, Qj) = D_{KL}(Q_i \| Q_j) + D_{KL}(Q_j \| Q_i)$.

Fig. 4 shows the similarity matrix for the 70 estimated latent health status groups. The upper left region shows pairwise diagnosis similarities and the lower right region shows pairwise medication similarities. We re-ordered the groups such that groups with similar diagnosis and medication distributions were located together. The red grids indicate small symmetric KL divergence

Table 6
Selected record with estimated latent health status groups.

Record A1	
4659 Acute upper respiratory infections of unspecified site; Group = 12	ACETAMINOPHEN; Group = 12
	TERFENADINE; Group = 12
25000 Diabetes mellitus without mention of complication, Type II; Group = 41	GLYBURIDE; Group = 41
	METFORMIN HCl; Group = 41

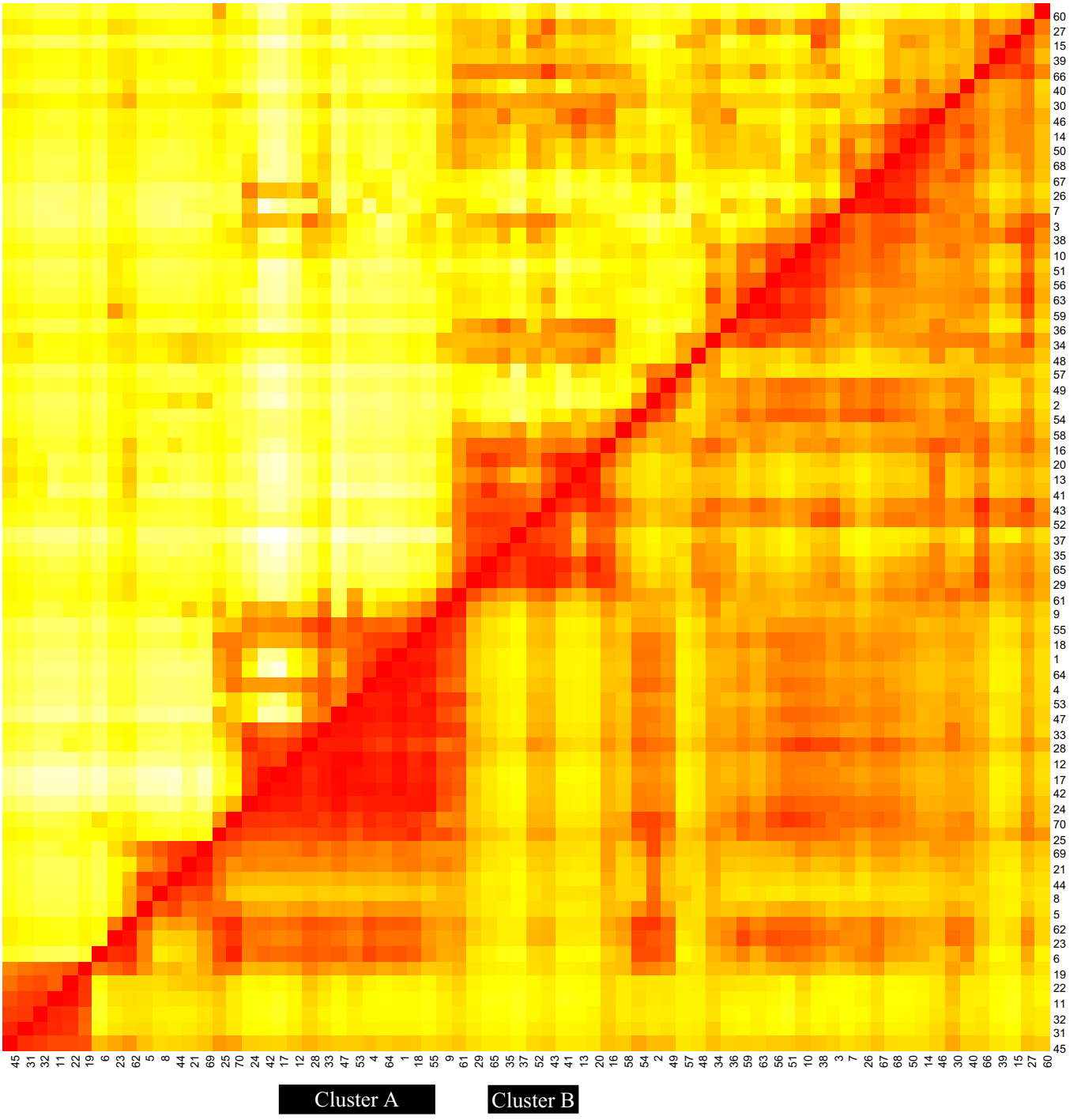


Fig. 4. Visualization of similarity between latent health status groups.

and thus high similarity between groups. The white grids indicate large symmetric KL divergence and thus low similarity between groups. The colors in between indicate different degrees of similarity.

One clear pattern in Fig. 4 is the red line along the anti-diagonal grids running from the lower left region to the upper right region, indicating large self-similarity. Moreover, the lower-right region is "warmer" than the upper-left region, suggesting that medication distributions are often more similar than diagnosis distributions for a pair of groups. Finally, there are several square blocks of "clusters" along the anti-diagonal line. These are clusters of latent health status groups that share similar diagnoses and medications.

For example, cluster "A" (marked on the *x*-axis in Fig. 2) consists of groups 70, 24, 42, 17, 12, 28, 33, 47, 53, 4, 61, 1, 18, and 55. Their diagnoses including acute upper respiratory infections of unspecified site (ICD-9 = 4659), acute tonsillitis (ICD-9 = 463), acute nasopharyngitis (ICD-9 = 460), acute bronchitis (ICD-9 = 4660), and acute sinusitis NOS (ICD-9 = 4619), all of which pertain to acute upper respiratory diseases (or so-called "common cold") [37]. Treatments for these diagnoses are very similar and mostly provide relief of symptoms in clinical settings. Commonly used medications include analgesics such as acetaminophen and non-steroidal anti-inflammatory drugs (NSAIDs) such as mefenamic acid, diclofenac sodium, or ibuprofen to help reduce fever, aches,

and pains; nasal decongestants such as dextromethorphan HBR or pseudoephedrine HCl to facilitate breathing; or mucolytic agents such as bromhexine HCl or ambroxol hydrochloride to dissolve thick mucus and to help relieve respiratory difficulties.

In addition to cluster "A," cluster "B" (marked on the *x*-axis in Fig. 2) identified by MCLDA is related to "chronic diseases." Cluster "B" consists of groups 65, 35, 37, 52, 43, 41, 13, and 20. This cluster describes patients with hypertension (ICD-9 = 4019) with or without concomitant diabetes mellitus (ICD-9 = 25000). This is concordant with the findings of our previous epidemiological study, which suggest that approximately 20% of hypertensive patients have concomitant diabetes mellitus [38]. Treatments for this cluster of patients thus include antihypertensive drugs such as amlodipine (besylate), felodipine, enalapril maleate, atenolol, or valsartan, and antidiabetic drugs such as metformin HCl, gliclazide, glyburide, or glimepiride. Antiplatelets such as aspirin or dipyridamole, which help to prevent future heart attacks or strokes, are also essential to this cluster [39].

4.7. Prediction performance

We evaluated the prediction performance of MCLDA based on NDCG, MAP, hitrate@1, hitrate@3, hitrate@5, and perplexity. The performance was computed using a ten-fold cross-validation. The training and testing divisions for each fold were synced across different approaches, which allowed us to compare performances via pairwise *t*-tests. We reported perplexity in the second half of this section for selected models only, because the computation of perplexity requires a model to assign a probability to each prediction, and some of the baseline models, such as CoOccur, are unable to meet this requirement.

Table 7 summarizes the performance of predicting medications given diagnoses using MCLDA and the ten baseline models. The results clearly show that MCLDA achieved the best performance for medication prediction. MCLDA reached an NDCG of 50.1% and a MAP of 30.6%, both of which were significantly higher than those of the baseline models at a 99% confidence level. CoOccur ranked second with an NDCG of 49.9% and an MAP of 30.5%. The Popular model, which considers only the unconditional medication probability, had an NDCG of only 31.6% and a MAP of 13.7%. The logistic regression family performed slightly worse than kNN. LR_ENET25 delivered a better performance with a NDCG of 48.1% and a MAP of 28.8%. MCLDA also achieved the highest hitrate@1, hitrate@3, and hitrate@5. The hitrate@5 of MCLDA was 70.0%, suggesting that 70.0% of the top 5 medication predictions include at least one correct answer. Although the performance gaps between MCLDA and the other baseline models (i.e., MMM, LDA, CoOccur, kNN, etc.) were not substantial, such differences were statistically significant based on pairwise *t*-tests.

The diagnosis prediction performance shows a similar pattern (see Table 8). MCLDA was the best performer based on NDCG (54.7%), hitrate@1 (32.5%), hitrate@3 (53.3%), and hitrate@5 (62.9%). Most differences were statistically significant at the 99% confidence level. The only exception was that the hitrate@1 difference between MCLDA and MMM was significant at a 90% confidence level. Remarkably, MMM achieved the highest MAP (37.8%), while the MAP of MCLDA was slightly lower (37.5%). MMM and LDA had a similar level of performance compared with MCLDA. Popular was still the worst performer and the logistic regression family was better than CoOccur. We note that CoOccur ranked ninth for diagnosis prediction based on NDCG, whereas CoOccur's NDCG ranked second for medication prediction. One possible reason is that patients with different diagnoses may be treated with similar medication. As a result, it is not easy for simple co-occurrences to provide good diagnosis prediction based on medications.

Overall, MCLDA is clearly the best performer for medication prediction. As for diagnosis prediction, MCLDA performed the best based on most of the performance measures. In addition to MCLDA, MMM, LDA, and kNN showed good performance in both prediction

Table 7
Predicting medications given diagnoses.

Method	NDCG	MAP	Hitrate@1	Hitrate@3	Hitrate@5
MCLDA	**50.1** (0.02)	**30.6** (0.03)	**35.4** (0.06)	**59.5** (0.06)	**70.0** (0.06)
MMM	49.7*** (0.03)	30.1*** (0.03)	34.5*** (0.06)	58.3*** (0.07)	69.1*** (0.08)
LDA	49.3*** (0.02)	29.7*** (0.03)	34.5*** (0.06)	58.2***(0.06)	68.9*** (0.06)
CoOccur	49.9** (0.02)	30.5* (0.02)	35.3* (0.03)	58.7*** (0.05)	69.3*** (0.06)
kNN	49.0*** (0.03)	29.6*** (0.03)	34.1*** (0.04)	58.1*** (0.06)	69.4*** (0.07)
LR_L2	47.8*** (0.02)	28.5*** (0.02)	33.5*** (0.03)	56.5*** (0.06)	66.4*** (0.05)
LR_ENET25	48.1*** (0.01)	28.8*** (0.01)	33.7*** (0.03)	56.9*** (0.04)	67.2*** (0.04)
LR_ENET50	47.4*** (0.01)	28.1*** (0.01)	33.6*** (0.03)	56.7*** (0.04)	67.0*** (0.05)
LR_ENET70	46.0*** (0.02)	26.9*** (0.02)	33.7*** (0.04)	56.5*** (0.05)	65.7*** (0.06)
LR_L1	45.1*** (0.03)	26.0*** (0.03)	32.2*** (0.04)	54.5*** (0.03)	64.4*** (0.06)
Popular	31.6*** (0.02)	13.7*** (0.02)	21.3*** (0.05)	30.4*** (0.06)	37.7*** (0.06)

Performance computed via a ten-fold cross-validation. Standard errors are inside the parentheses. Bold values indicate the best performer in a column.
* Indicate the difference between a method and MCLDA is statistically significant at the 90% confidence level based on pairwise *t*-tests.
** Indicate the difference between a method and MCLDA is statistically significant at the 95% confidence level based on pairwise *t*-tests.
*** Indicate the difference between a method and MCLDA is statistically significant at the 99% confidence level based on pairwise *t*-tests.

Table 8
Predicting diagnoses given medications.

Method	NDCG	MAP	Hitrate@1	Hitrate@3	Hitrate@5
MCLDA	**54.7** (0.03)	37.5 (0.03)	**32.5** (0.06)	**53.3** (0.07)	**62.9** (0.07)
MMM	54.3*** (0.04)	**37.8*** (0.04)	32.3* (0.05)	52.9*** (0.06)	62.2*** (0.06)
LDA	52.4*** (0.03)	36.1*** (0.03)	31.0*** (0.06)	50.5*** (0.07)	59.5*** (0.07)
CoOccur	47.7*** (0.04)	30.8*** (0.04)	25.1*** (0.05)	42.6*** (0.08)	51.5*** (0.08)
kNN	54.1*** (0.03)	37.8*** (0.04)	31.9*** (0.07)	53.1*** (0.06)	62.7*** (0.07)
LR_L2	50.8*** (0.03)	34.9*** (0.04)	30.5*** (0.05)	49.2*** (0.05)	57.6*** (0.04)
LR_ENET25	50.6*** (0.03)	34.7*** (0.04)	30.3*** (0.06)	48.9*** (0.05)	57.2*** (0.05)
LR_ENET50	50.2*** (0.03)	34.4*** (0.04)	30.3*** (0.06)	48.7*** (0.05)	57.0*** (0.05)
LR_ENET70	47.9*** (0.04)	32.8*** (0.05)	30.0*** (0.06)	47.1*** (0.07)	54.3*** (0.06)
LR_L1	46.5*** (0.04)	30.9*** (0.04)	27.3*** (0.05)	44.6*** (0.05)	52.2*** (0.06)
Popular	30.8*** (0.03)	15.4*** (0.03)	10.5*** (0.04)	20.6*** (0.05)	25.8*** (0.06)

Performance computed via a ten-fold cross-validation. Standard errors are inside the parentheses. Bold values indicate the best performer in a column.
* Indicate the difference between a method and MCLDA is statistically significant at the 90% confidence level based on pairwise *t*-tests.
** Indicate the difference between a method and MCLDA is statistically significant at the 95% confidence level based on pairwise *t*-tests.
*** Indicate the difference between a method and MCLDA is statistically significant at the 99% confidence level based on pairwise *t*-tests.

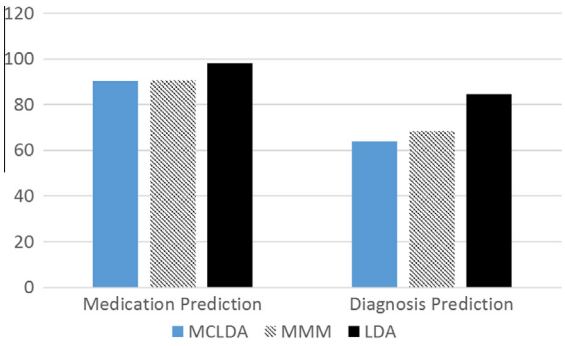


Fig. 5. Perplexity comparison of MCLDA, MMM, and LDA.

tasks. CoOccur achieved satisfactory good performance only for medication prediction.

We evaluated the perplexity for MCLDA, MMM, and LDA. Fig. 5, which shows the perplexity of these models for the two prediction tasks, provides several remarkable insights. First, for the task of predicting diagnoses (the three bars on the right), MCLDA had the lowest perplexity, whereas LDA had the highest perplexity. The perplexity differences were statistically significant at a 99% confidence level. Second, for the task of predicting medications (the three bars on the left), MCLDA also had the lowest perplexity, followed by MMM and LDA. However, the perplexity difference between MCLDA and MMM was not statistically significant while the difference between MMM and LDA was statistically significant. Since all three models adopted the same information set, the insignificant gap between MCLDA and MMM implies that diagnosis depends upon contextual information such as sex, age, and medical specialty while the prescriptions given for patients' problems are less influenced by contextual information. These results suggest that, consistent with the performance measured by NDCG, MCLDA outperformed other topic models evaluated in this study for diagnosis prediction, and was better than LDA for medication prediction.

4.8. Prediction performance on subsamples

Although MCLDA led the other approaches when they were trained on large datasets, we were also interested in knowing how the performance may change when the models are trained on smaller datasets and for more challenging prediction tasks. One reason to consider the performance of a small training dataset is that applications at a single institution may greatly limit the sample size. Moreover, medication and diagnosis predictions are most valuable when the underlying records involve several diagnoses (and thus are more challenging).

For each fold of the ten-fold cross-validation, we constructed a small training dataset by randomly sampling 1000 records from the original training dataset and discarding the remaining 899,000 records. The original testing dataset was filtered, and only records with three or more diagnoses were kept. We considered a training dataset of 1000 records small because there were 3830 unique diagnoses and 1416 unique medications.

Table 9 summarizes the medication prediction performance. MCLDA was still the best performer while the NDCG dropped about 10 percentage points to 40.0%. The Hitrate@5 was 54.0% while the full sample was 70.0%. The NDCG gap between MCLDA and kNN was 7.0 percentage points while the gap was only 1.1 percentage points in the full sample. The NDCG of CoOccur was 2.2 percentage points lower than that of MCLDA, while the gap was only 0.2 percentage points in the full sample. The performance differences between MCLDA and the baseline models, including LDA, CoOccur, kNN, the logistic regression family, and popular, were statistically significant based on pairwise *t*-tests. While MCLDA has significantly higher NDCG and MAP compared to those of MMM, the performance differences measured by Hitrate@1, Hitrate@3, and Hitrate@5 were not statistically significant. The logistic regression family performed the worst among the baseline models except for Popular. One possible reason is the sparsity problem becomes more severe in a small training dataset.

Table 9
Sub-sample performance: Predicting medication given diagnosis using 1000 training records and testing records with three or more diagnoses.

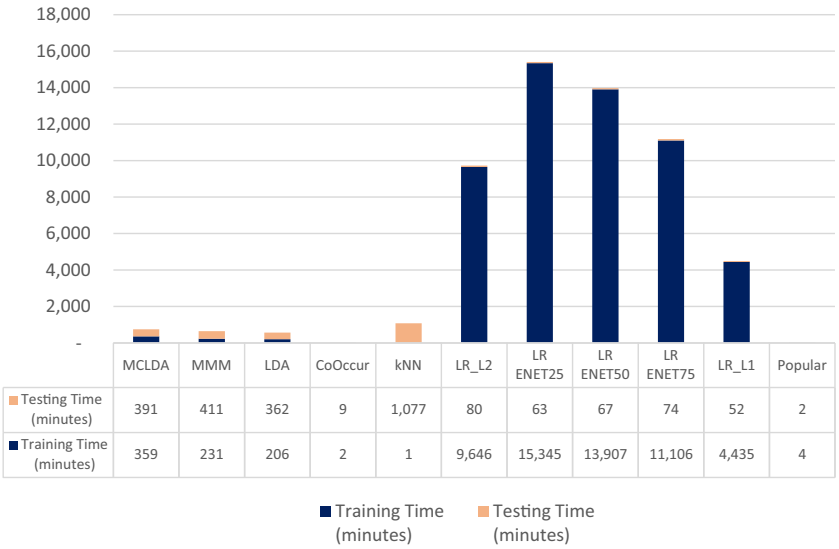
Method	NDCG	MAP	Hitrate@1	Hitrate@3	Hitrate@5
MCLDA	**40.0**	**20.7**	**25.7**	44.6	**54.0**
	(0.14)	(0.13)	(0.27)	(0.37)	(0.24)
MMM	36.4***	19.5***	23.8	**44.9**	53.6
	(0.10)	(0.07)	(0.22)	(0.26)	(0.26)
LDA	37.6***	17.1***	23.7***	42.1***	52.2***
	(0.21)	(0.2)	(0.46)	(0.47)	(0.45)
CoOccur	37.8***	17.4***	24.9***	43.0***	52.6***
	(0.09)	(0.07)	(0.11)	(0.18)	(0.2)
kNN	33.0***	13.1***	19.2***	32.9***	41.6***
	(0.09)	(0.06)	(0.12)	(0.38)	(0.3)
LR_L2	30.7***	10.8***	13.1***	27.3***	37.6***
	(0.17)	(0.12)	(0.11)	(0.62)	(0.57)
LR_ENET25	32.6***	12.7***	17.2***	34.0***	44.0***
	(0.03)	(0.03)	(0.16)	(0.1)	(0.1)
LR_ENET50	32.4***	12.5***	17.0***	33.6***	43.6***
	(0.03)	(0.03)	(0.17)	(0.09)	(0.12)
LR_ENET70	33.4***	13.5***	19.1***	36.0***	45.7***
	(0.07)	(0.04)	(0.11)	(0.18)	(0.18)
LR_L1	32.2***	12.3***	16.9***	33.3***	43.0***
	(0.03)	(0.05)	(0.26)	(0.09)	(0.12)
Popular	27.2***	8.4***	12.8***	19.0***	24.6***
	(0.1)	(0.07)	(0.1)	(0.23)	(0.14)

Bold values indicate the best performer in a column.
*** Statistically significant (at the 99% confidence level) difference from a pairwise *t*-test comparison between the method and MCLDA. Performance computed via a ten-fold cross-validation. Standard errors are inside the parentheses.

Table 10
Sub-sample performance: Predicting diagnoses given medications using 1000 training records and testing records with three or more diagnoses.

Method	NDCG	MAP	Hitrate@1	Hitrate@3	Hitrate@5
MCLDA	36.0	19.1	**24.6**	43.3	51.9
	(0.22)	(0.19)	(0.34)	(0.48)	(0.56)
MMM	38.1***	17.7***	24.6	**43.3**	**53.2***
	(0.14)	(0.13)	(0.56)	(0.26)	(0.32)
LDA	**41.9*****	**25.6*****	20.5***	36.1***	45.9***
	(0.11)	(0.14)	(0.25)	(0.3)	(0.16)
CoOccur	33.4***	16.4***	18.7***	34.5***	43.1***
	(0.09)	(0.08)	(0.12)	(0.22)	(0.13)
kNN	32.1***	15.2***	17.6***	32.1***	41.5***
	(0.09)	(0.03)	(0.31)	(0.35)	(0.34)
LR_L2	30.7***	13.8***	16.0***	32.9***	40.7***
	(0.09)	(0.06)	(0.13)	(0.14)	(0.21)
LR_ENET25	31.1***	14.2***	18.7***	34.3***	41.1***
	(0.14)	(0.13)	(0.12)	(0.3)	(0.37)
LR_ENET50	30.8***	14.0***	18.5***	33.9***	40.4***
	(0.13)	(0.12)	(0.12)	(0.27)	(0.36)
LR_ENET70	31.5***	14.8***	20.2***	35.3***	42.0***
	(0.15)	(0.14)	(0.12)	(0.33)	(0.4)
LR_L1	30.6***	13.7***	18.2***	33.3***	39.7***
	(0.13)	(0.11)	(0.08)	(0.28)	(0.33)
Popular	27.3***	9.3***	4.6***	24.5***	31.8***
	(0.1)	(0.08)	(0.07)	(0.34)	(0.16)

Performance computed via a ten-fold cross-validation. Standard errors are inside the parentheses. Bold values indicate the best performer in a column.
* Indicate the difference between a method and MCLDA is statistically significant at the 90% confidence level based on pairwise *t*-tests.
** Indicate the difference between a method and MCLDA is statistically significant at the 95% confidence level based on pairwise *t*-tests.
*** Indicate the difference between a method and MCLDA is statistically significant at the 99% confidence level based on pairwise *t*-tests.



	MCLDA	MMM	LDA	CoOccur	kNN	LR_L2	LR ENET25	LR ENET50	LR ENET75	LR_L1	Popular
Testing Time (minutes)	391	411	362	9	1,077	80	63	67	74	52	2
Training Time (minutes)	359	231	206	2	1	9,646	15,345	13,907	11,106	4,435	4

Fig. 6. Comparison of training and testing time to complete medication prediction and diagnosis prediction using 900,000 training records and 100,000 testing records. Experiments conducted on a machine with an Intel Core i5-4590 CPU and 32 GB of RAM.

For the diagnosis prediction task (Table 10), MCLDA had the best hitrate@1 (24.6%) while LDA had the best NDCG (41.9%) and MAP (25.6%). MMM had the best hitrate@3 (43.3%) and hitrate@5 (53.2%). This pattern suggests that, compared to LDA, the first few diagnosis predictions made by MCLDA and MMM were more accurate, but the predictions ranked lower were less accurate. The performance gap between MCLDA and kNN was 3.9 percentage points for NDCG and 17.6 percentage points for hitrate@5; both were much wider than the case of the full sample. The NDCG of CoOccur was 2.6 percentage points lower than that of MCLDA, a performance gap that became narrower compared to the results for the full sample. The narrower gap suggests that CoOccur was less influenced by the size of the training dataset. However, the full sample results also suggest that CoOccur is not an attractive choice if the training dataset is large.

Overall, the evaluation results on subsamples suggest that prediction performance of all approaches decreased. At the same time, the performance differences between different approaches became larger. Two potential reasons may have contributed to the wider performance gaps. First, smaller training datasets exhibited more severe data sparsity problem. Second, we evaluated performance on more challenging test sets (i.e., records with three or more diagnoses). Methods such as MCLDA and LDA seemed to handle these two problems better and had better performances compared to other approaches.

4.9. Comparison of computational costs

Fig. 6 summarizes the computational cost as measured by training and testing time. We report the training and testing time required to finish medication prediction and diagnosis prediction using 900,000 training records and 100,000 testing records running a commercial PC with an Intel Core i5-4590 CPU and 32 GB of RAM. MCLDA took 359 min to train and 391 min to conduct the two prediction tasks. The average time needed to make a prediction for a testing record was 0.117 ($391 \div 100000 \div 2 \times 60$) seconds, which is fast enough to be deployed for real-time prediction tasks in a clinical setting. MMM and LDA had a similar level of computational cost as that of MCLDA.

Popular and CoOccur finished training and testing within 11 min. kNN took 1 min to prepare the data (i.e., training) but required 1077 min to finish the prediction tasks. The result is not surprising because kNN needs to scan the whole training dataset in order to locate the k nearest neighbors for each prediction. The logistic regression families (LR_L2, LR_ENET25, LR_ENET50, LR_ENET75, and LR_L1) took much longer to train compared to the other approaches. The main reason is that medication prediction and diagnosis prediction are multiple-label prediction problems [32]. Using the binary relevance method based on logistic regression required training 3830 binary logistic regressions for diagnosis prediction and 1416 binary logistic regressions for medication prediction. The testing time, however, was short among all approaches considered.

5. Concluding remarks

In this study, we developed and evaluated an MCLDA model that is able to identify the latent health status groups in healthcare data. MCLDA is able to include diagnoses, medications, and contextual information as separate channels that allow better model learning and prediction. We developed a Gibbs sampling-based algorithm for efficient model inference. Using one million records from a healthcare insurance claim dataset, we evaluated MCLDA and compared the proposed model with other methods, including MMM, LDA, CoOccur, kNN, the logistic regression family, and Popular. The experimental results showed that, in most cases, MCLDA is the best performer for both diagnosis and medication prediction tasks. Moreover, MCLDA is able to identify the pairing relations among diagnoses and medications in a record based on the latent group assignments.

Our study contributes to the healthcare data mining literature by applying and extending existing topic models to healthcare data. Our experiments suggest that MCLDA is a strong candidate for researchers and practitioners who need to identify hidden structures in healthcare data and conduct predictions involving diagnoses and medications. Our experiments also suggest that the performance of existing methods vary according to the prediction tasks. For example, CoOccur is a good choice for medication prediction, while kNN is able to deliver good performance for diagnosis prediction.

We plan to pursue several research directions in the future. First, we plan to evaluate the drug indication relations predicted by MCLDA against an existing knowledge base such as Lexi-Comp.

Second, we plan to study the utility of the identified co-morbidities by incorporating them into well-known problems such as patient risk stratification. Third, we are interested in extending the proposed MCLDA model such that patients' preexisting medical conditions can be considered by the model. Last, we plan to investigate the potential of modeling patient health status progression based on the latent health status groups identified by MCLDA. Existing prognostic models often aggregate diagnoses and medications through manually defined groups. MCLDA might be able to provide a better approach for handling the data sparsity issues in healthcare data.

Conflict of interest

The authors declare that there are no conflicts of interest associated in this article.

Acknowledgments

This work is supported in part by the National Science Council of Taiwan, under Grant NSC101-3114-Y-002-003, and by the Ministry of Science and Technology, Taiwan, under Grants MOST 103-2410-H-002-108-MY3, MOST 103-2410-H-002-110-MY3, and MOST 104-2410-H-002-225-MY3.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jbi.2016.02.003.

References

[1] K. Chen, H. Chen, N. Conway, J.M. Hellerstein, T.S. Parikh, Usher: improving data quality with dynamic forms, IEEE Trans. Knowl. Data Eng. 23 (2011) 1138–1153.
[2] H. Salmasian, D.E. Freedberg, C. Friedman, Deriving comorbidities from medical records using natural language processing, J. Am. Med. Inform. Assoc. (2013).
[3] D. Nikovski, Constructing Bayesian networks for medical diagnosis from incomplete and partially correct statistics, IEEE Trans. Knowl. Data Eng. 12 (2000) 509–516.
[4] A. Wright, E.S. Chen, F.L. Maloney, An automated technique for identifying associations between medications, laboratory results and problems, J. Biomed. Inform. 43 (2010) 891–901.
[5] S. Hasan, G.T. Duncan, D.B. Neill, R. Padman, Automatic detection of omissions in medication lists, J. Am. Med. Inform. Assoc. 18 (2011) 449–458.
[6] T.L. Griffiths, M. Steyvers, Finding scientific topics, Proc. Natl. Acad. Sci. USA 101 (2004) 5228–5235.
[7] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[8] E. Erosheva, S. Fienberg, J. Lafferty, Mixed-membership models of scientific publications, Proc. Natl. Acad. Sci. 101 (2004) 5220–5227.
[9] Y. Chen, J. Ghosh, C.A. Bejan, C.A. Gunter, S. Gupta, A. Kho, et al., Building bridges across electronic health record systems through inferred phenotypic topics, J. Biomed. Inform. 55 (2015) 82–93.
[10] Z. Huang, W. Dong, L. Ji, C. Gan, X. Lu, H. Duan, Discovery of clinical pathway patterns from event logs using probabilistic topic models, J. Biomed. Inform. 47 (2014) 39–57.
[11] A. Wright, J. Pang, J.C. Feblowitz, F.L. Maloney, A.R. Wilcox, H.Z. Ramelson, et al., A method and knowledge base for automated inference of patient problems from structured data in an electronic medical record, J. Am. Med. Inform. Assoc. 18 (2011) 859–867.
[12] E.S. Chen, G. Hripcsak, H. Xu, M. Markatou, C. Friedman, Automated acquisition of disease-drug knowledge from biomedical and clinical documents: an initial study, J. Am. Med. Inform. Assoc. 15 (2008) 87–98.
[13] F.S. Roque, P.B. Jensen, H. Schmock, M. Dalgaard, M. Andreatta, T. Hansen, et al., Using electronic patient records to discover disease correlations and stratify patient cohorts, PLoS Comput. Biol. 7 (2011) e1002141.
[14] C. Friedman, G. Hripcsak, L. Shagina, H. Liu, Representing information in patient reports using natural language processing and the Extensible Markup Language, J. Am. Med. Inform. Assoc. 6 (1999) 76–87.
[15] C. Friedman, L. Shagina, Y. Lussier, G. Hripcsak, Automated encoding of clinical documents based on natural language processing, J. Am. Med. Inform. Assoc. 11 (2004) 392–402.
[16] L. Chen, C. Friedman, Extracting phenotypic information from the literature via natural language processing, Medinfo 11 (2004) 758–762.
[17] S.M. Meystre, G.K. Savova, K.C. Kipper-Schuler, J.F. Hurdle, Extracting information from textual documents in the electronic health record: a review of recent research, IMIA Yearbook 2008: Access Health Inform. 3 (2008) 128–144.
[18] A.B. McCoy, A. Wright, A. Laxmisan, M.J. Ottosen, J.A. McCoy, D. Butten, et al., Development and evaluation of a crowdsourcing methodology for knowledge base construction: identifying relationships between clinical problems and medications, J. Am. Med. Inform. Assoc. 19 (2012) 713–718.
[19] A. Wright, J. Pang, J.C. Feblowitz, F.L. Maloney, A.R. Wilcox, K.S. McLoughlin, et al., Improving completeness of electronic problem lists through clinical decision support: a randomized, controlled trial, J. Am. Med. Inform. Assoc. 19 (2012) 555–561.
[20] B. Sarwar, G. Karypis, J. Konstan, J. Riedl, Item-based collaborative filtering recommendation algorithms, in: Proceedings of the 10th International Conference on World Wide Web. ACM, Hong Kong, Hong Kong, 2001, pp. 285–295.
[21] D.M. Blei, Probabilistic topic models, Commun. ACM 55 (2012) 77–84.
[22] R.M. Nallapati, A. Ahmed, E.P. Xing, W.W. Cohen, Joint latent topic models for text and citations, in: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, Las Vegas, Nevada, USA, 2008, pp. 542–550.
[23] J. Besag, Spatial interaction and the statistical analysis of lattice systems, J. Roy. Stat. Soc.: Ser. B (Methodol.) 36 (1974) 192–236.
[24] S. Geman, D. Geman, Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, IEEE Trans. Pattern Anal. Mach. Intell. 6 (1984) 721–741.
[25] D.M. Kaplan, Clear writing, clear thinking and the disappearing art of the problem list, J. Hosp. Med. 2 (2007) 199–202.
[26] B. Croft, D. Metzler, T. Strohman, Search Engines: Information Retrieval in Practice, Addison-Wesley Publishing Company, 2009.
[27] C.D. Manning, P. Raghavan, H. Schutze, An Introduction to Information Retrieval, Cambridge University Press, Cambridge, England, 2009.
[28] M. Deshpande, G. Karypis, Item-based top-N recommendation algorithms, ACM Transact. Inform. Syst. 22 (2004) 143–177.
[29] S. Kabbur, X. Ning, G. Karypis, FISM: factored item similarity models for top-N recommender systems, in: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, Chicago, Illinois, USA, 2013, pp. 659–667.
[30] M. Rosen-Zvi, C. Chemudugunta, T. Griffiths, P. Smyth, M. Steyvers, Learning author-topic models from text corpora, ACM Transact. Inform. Syst. 28 (2010) 1–38.
[31] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.
[32] G. Tsoumakas, I. Katakis, I. Vlahavas, Mining Multi-label Data, in: O. Maimon, L. Rokach (Eds.), Data Mining and Knowledge Discovery Handbook, Springer, US, 2010, pp. 667–685.
[33] H. Zou, T. Hastie, Regularization and variable selection via the Elastic Net, J. Roy. Stat. Soc. B 67 (2005) 301–320.
[34] F. Ricci, L. Rokach, B. Shapira, P.B. Kantor, Recommender Systems Handbook, Springer-Verlag, Inc., New York, 2010.
[35] T.P. Minka, Estimating a Dirichlet distribution. Technical Report 2000; <http://research.microsoft.com/en-us/um/people/minka/papers/dirichlet/minka-dirichlet.pdf>.
[36] T.J. Chen, L.F. Chou, S.J. Hwang, Prevalence of anti-ulcer drug use in a Chinese cohort, World J. Gastroentero. 9 (2003) 1365–1369.
[37] K.H. Huang, Y.C. Hsieh, C.T. Hung, F.Y. Hsiao, Off-label antibiotic use in the pediatric population: a population-based study in Taiwan, J. Food Drug. Anal. 20 (2012) 597–602.
[38] C.I. Hsu, F.Y. Hsiao, F.L.L. Wu, L.J. Shen, Adherence and medication utilisation patterns of fixed-dose and free combination of angiotensin receptor blocker/thiazide diuretics among newly diagnosed hypertensive patients: a population-based cohort study, Int. J. Clin. Pract. 69 (2015) 729–737.
[39] Y.W. Tsai, Y.W. Wen, W.F. Huang, P.F. Chen, K.N. Kuo, F.Y. Hsiao, Cardiovascular and gastrointestinal events of three antiplatelet therapies: clopidogrel, clopidogrel plus proton-pump inhibitors, and aspirin plus proton-pump inhibitors in patients with previous gastrointestinal bleeding, J. Gastroenterol. 46 (2011) 39–45.