Wang et al. BMC Medical Informatics and Decision Making 2019, 19(Suppl 3):69
https://doi.org/10.1186/s12911-019-0779-y

BMC Medical Informatics and Decision Making

RESEARCH Open Access

Discovering associations between problem list and practice setting

Liwei Wang, Yanshan Wang, Feichen Shen, Majid Rastegar-Mojarad and Hongfang Liu*

From The Sixth IEEE International Conference on Healthcare Informatics (ICHI 2018)
New York, NY, USA. 4-7 June 2018

Abstract

Background: The Health Information Technology for Economic and Clinical Health Act (HITECH) has greatly accelerated the adoption of electronic health records (EHRs) with the promise of better clinical decisions and patients' outcomes. One of the core criteria for "Meaningful Use" of EHRs is to have a problem list that shows the most important health problems faced by a patient. The implementation of problem lists in EHRs has a potential to help practitioners to provide customized care to patients. However, it remains an open question on how to leverage problem lists in different practice settings to provide tailored care, of which the bottleneck lies in the associations between problem list and practice setting.

Methods: In this study, using sampled clinical documents associated with a cohort of patients who received their primary care at Mayo Clinic, we investigated the associations between problem list and practice setting through natural language processing (NLP) and topic modeling techniques. Specifically, after practice settings and problem lists were normalized, statistical χ^2 test, term frequency-inverse document frequency (TF-IDF) and enrichment analysis were used to choose representative concepts for each setting. Then Latent Dirichlet Allocations (LDA) were used to train topic models and predict potential practice settings using similarity metrics based on the problem concepts representative of practice settings. Evaluation was conducted through 5-fold cross validation and Recall@k, Precision@k and F1@k were calculated.

Results: Our method can generate prioritized and meaningful problem lists corresponding to specific practice settings. For practice setting prediction, recall increases from 0.719 (k = 2) to 0.931 (k = 10), precision increases from 0.882 (k = 2) to 0.931 (k = 10) and F1 increases from 0.790 (k = 2) to 0.931 (k = 10).

Conclusion: To our best knowledge, our study is the first attempting to discover the association between the problem lists and hospital practice settings. In the future, we plan to investigate how to provide more tailored care by utilizing the association between problem list and practice setting revealed in this study.

Keywords: Problem list, Practice setting, Topic modeling, Statistical χ^2 test, TF-IDF and enrichment analysis

Background

Since its enactment in 2009, the Health Information Technology for Economic and Clinical Health Act (HITECH) has greatly accelerated the adoption of electronic health records (EHRs) with the promise of better clinical decisions and patients' outcomes. According to the Centers for Medicare & Medicaid Services (CMS), "meaningful use" of EHRs refers to the use of EHRs to achieve significant improvements in care. One of the core criteria for "Meaningful Use" of EHRs is to have a codified up to date problem list that lists the most important health problems faced by a patient [1–4]. The problem list was first introduced by Weed in 1968 in his promotion for a Problem-Oriented Medical Record (POMR) [5]. Since then it has been widely used and become a key component in patient records. In the Health Level Seven International's Electronic Health Record System Functional Model (EHR-S FM), a problem list "may include, but is not limited to chronic conditions, diagnoses, or symptoms, functional limitations, visit or stay-specific conditions, diagnoses, or symptoms" [6].

* Correspondence: liu.hongfang@mayo.edu
Division of Digital Health Sciences, Department of Health Sciences Research, Mayo Clinic, Rochester, MN 55905, USA

Ideally, physicians could benefit from an accurate problem list to track a patient's status and progress, to maintain continuity of patient care and to organize clinical reasoning and documentation [7]. Accurate problem lists could also be used for the improvement of the quality of care, the realization of clinical decision support, and the facilitation of research and quality measurement [8]. The problem list can serve a variety of uses in diverse healthcare settings by providing a succinct view of a patient's health status and therefore should be used and maintained to meet different needs. For example, a primary care physician concerns chronic and acute conditions while a specialty provider may focus only on a subset of problems relevant to that area of medicine. An emergency provider may address only the critical acute presenting problems. Other clinicians may use the problem list for tracking conditions that should be addressed for specific care delivery goals. Extensive studies have been conducted to assess the usefulness of problem lists, for example, through the exploration of the use pattern of problem lists [9], the detection of problem list gaps in recording patients' problems [10, 11], the creation and maintenance of a problem list using natural language processing [12–14], and the use of problem list for decision making support [15]. However, due to the inconsistent use across providers as well as the lack of the consensus of what should be documented in the problem lists [16], problem lists are frequently inaccurate and out-of-date [15]. It remains an open question how to leverage the problem list to provide tailored care at different practice settings (e.g., primary care, cardiology, or emergency) and for different care providers (e.g., clinicians, nurses, or social workers), of which the bottleneck lies in the associations between problem list and practice setting.

In this study, we aim to investigate the associations between the problem list and practice settings using the longitudinal EHR data from Mayo Clinic by mapping problems and practice settings to standard representations and assessing the associations between them using topic modeling [17] and clustered imaging map (CIM) [18].

Methods

Figure 1 illustrates the overall workflow in this study. Our method used natural language processing (NLP) to normalize problem list and manually aggregated practice settings (step 1–3), where "Cardiovascular" is the

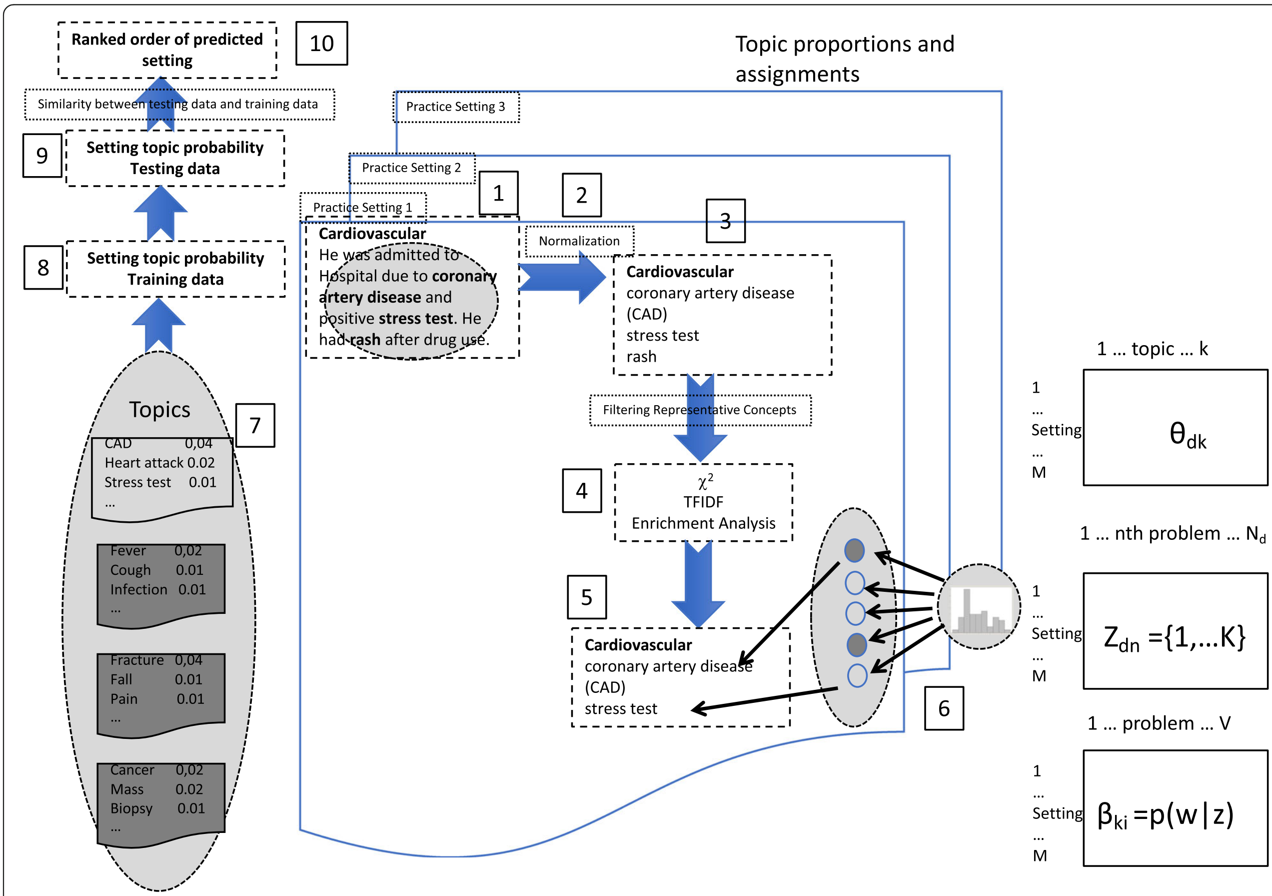

Fig. 1 The overall workflow of the proposed method

practice setting of the problems like "coronary artery disease". Representative concepts were then filtered using χ2, term frequency-inverse document frequency (TF-IDF) and enrichment analysis based on the Semantic Medline (step 4–5). Subsequently Latent Dirichlet Allocations (LDA) [19] were used to train topic models and predict potential practice settings using similarity metrics based on the problem list (step 6–10). Finally 5-fold cross validation was utilized for evaluation, while cluster image map [20] revealing setting similarity from all randomly chosen data was used for visualization.

Data sources

The collection of clinical documents used in our analysis consists of clinical notes for a cohort of patients receiving their primary care at Mayo Clinic, spanning a period of 15 years (1998–2013), and covering both inpatient and outpatient settings. Problems in those documents are generally itemized entries as either phrases (e.g., "*Allergic rhinitis/vasomotor rhinitis*") or short sentences (e.g, "*Her asthma appeared to be very mild*"). After normalization of settings and problem list, we randomly selected 1000 notes (documents) for each of 64 settings as the input for filtering, in total 64,250 notes was used as input for the step 4 to choose representative concepts. Then 60,345 notes were kept for training topic model in step 6. We then randomly selected 200 notes from each setting as testing data, in total 13,498 notes was used as input for step 9 to test the predicted settings.

The latest version of Semantic Medline Database (SemMedDB) has more than 84.6 million semantic associations from 25,582,462 Medline citations up to Dec 312,015 from 1865, based on the natural language processing tool SemRep and Unified Medical Language System (UMLS) [21]. Among eight tables, the most comprehensive PREDICATION_AGGREGATE (PA) table contains all available information from the SemMedDB, including subject concepts, object concepts, sentence ID, PubMed IDs (PMIDs), and so on. Article level co-occurrences among subject-object concepts, i.e., 1,164,352 total co-occurrences of concepts from all practice settings were used in enrichment analysis for statistically significant concepts associated with each setting extracted from clinical notes in the SemMedDB.

Normalization of settings

As a large volume of clinical documents has been generated in the context of EHRs, the HL7/LOINC Document Ontology (DO) was developed to support a range of use cases (e.g., retrieval, organization, display, and exchange) [22]. It contains a hierarchical structure comprising five axes: Kind of Document (KOD), Type of Service (TOS), Setting, Subject Matter Domain (SMD) and Role. Each axis contains a set of values. Some studies explored the applicability of DO in document representation and mapping [23, 24], and use of LOINC codes for document exchange in the clinical scenario [25, 26]. Other studies have focused on the improvement of axes of SMD [27], TOS [28], and Setting [29], mainly through increasing the coverage of each axis to make it more comprehensively representative. For example, Rajamani et al. proposed extended values for Settings of Care from 20 to 274, that fall into 14 main classes, such as Inpatient, Outpatient, Public Health, Community, and Mobile [29]. Currently the settings in Mayo clinic notes are relatively refined. First, locations are usually used for differentiating settings of the same practice (e.g., Family Medicine BA, Family Medicine KA, where BA and KA indicated locations). Second, more detailed classifications have been generated under specific specialties (e.g., "Ped Neonatology-I" and "Psych Ped SMH", (SMH is a location of Mayo Clinic)). In this way, names of settings could provide plentiful information on subjects, specialties and locations. Such refinement could facilitate targeted treatment. However, it results in a large number of settings, e.g., during the study period, there are more than 1000 settings in clinical notes. This brings hurdles for the meaningful use of problem lists in different settings.

In this paper, we studied the settings associated with more than 4500 clinical notes based on proposed extended values for Settings of Care [29] for setting aggregation. Two steps were taken to aggregate various settings into more general ones. First, for practice settings with the same practice and various locations, we kept the subject and removed locations. For example, "Family Medicine BA" and "Family Medicine KA" were merged into "Family Medicine". Second, for those settings with similar specialties, we aggregated them into the general settings. For example, "Ped Neonatology-I" and "Psych Ped SMH" were aggregated into "Pediatrics". In total, 64 settings were aggregated corresponding to 266 practice settings.

Normalization of problem list

With a good coverage of frequently used terms in problem lists [30], the CORE Problem List Subset has been created to align with the meaningful use requirement and better implement Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) in electronic health records (EHR) [30]. In a previous study [31], we assessed the coverage of SNOMED CT for codifying problem lists in narrative format by extracting itemized entries from clinical notes and normalize them to the Unified Medical Language System (UMLS) [32] concepts. In this study, we applied the same methodology but kept UMLS concepts that can be mapped to the CORE Problem List Subset codes (the August 2015 version of The CORE Problem List Subset of SNOMED CT was used). Only diagnosis related sections were kept for further study, e.g., "History of Present Illness" and "Diagnosis".

Filtering representative concepts for each setting

In order to choose representative concepts among randomly selected notes for each setting, first statistical χ^2 test was conducted, then TF-IDF and enrichment analysis for co-occuring concepts in each setting performed based on Semantic Medline. The purpose of χ^2 test is to find concepts having significant association with practice settings. TF-IDF helps to remove concepts that appear in most practice settings and can't demonstrate their unique value for specific practice setting. In enrichment analysis, we used an external data source, Semantic Medline to verify if the concepts in each setting after χ^2 and TF-IDF filtering were overrepresented in the large-scale Semantic Medline. More details will be discussed in the following paragraphs.

After NLP and setting aggregation, each document had a corresponding setting and contains a list of normalized Concept Unique Identifiers (CUIs) for problems. In our pilot experiment, we randomly selected 1000 notes (documents) for each of 64 settings for 5 times. We found that out of total 4573 normalized problems, around 3630 are covered by randomly selected notes, account for 79.4%. We can infer from these results that 1000 notes (documents) could represent the corresponding practice setting.

Therefore randomly selected 64,250 training documents were used as input for calculation of χ^2, deriving χ^2 values for 240,110 concept and practice setting pairs. After choosing those cocept and pracice setting pairs with $\chi^2 > 6.64$ ($P < 0.01$), 17,180 were kept. TF-IDF for these pairs was calculated using the Eq. 1.

$$\mathrm{TFIDF} = \frac{\mathrm{Fc, s}}{\log(1 + \mathrm{N/nc})} \tag{1}$$

where *Fc,s* denotes the frequency of the concept c in the setting s, *N* the total number of settings, and *nc* the number of settings containing the concept. Fourteen thousand, one hundred and sixty concept and practice setting pairs with TF-IDF greater than 1 were kept for further enrichment analysis.

Enrichment analysis, primarily based on Gene Ontology, has been used for summarizing and profiling a gene set [33]. Recently, a few studies explored different sources, i.e., the Medical Subject Headings for enrichment analysis [34, 35]. As one of repositories for semantic predications processed from the Medline, Semantic Medline has been employed for the discovery of relationships among biological entities [36]. In this study, we proposed to leverage the abundant entities and semantic associations in the Semantic Medline for concept co-occurring enrichment analysis to verify if the concepts in each setting after χ^2 and TF-IDF filtering were overrepresented in the large-scale Semantic Medline.

Specifically we calculated the enrichment fold of concept-setting pairs. Enrichment fold means to what extent is the rate that co-occurring concepts from each setting actually appear in the Semantic Medline more than the average rate of all possible concept pairs in the Semantic Medline. Higher enrichment fold indicates higher possibility that the co-occurring concepts from each setting occur in the Semantic Medline more frequently than the average co-occurring rate in the Semantic Medline. Enrichment analysis for co-occurring concepts was performed in the SemMedDB using the Eqs. 2 and 3 to ultimately obtain the Enrichment Fold (Eq. 4).

$$\mathrm{ProbExpSet} = \frac{\mathrm{TotalSemOcc}}{\mathrm{TotalPairNum}} \tag{2}$$

where TotalPairNum refers to the total number of possible concept pairs (e.g., (C0011849, C0015967)), among the concept collection from all randomly selected notes, TotalSemOcc refers to the total co-occurrence in the SemMedDM of all concept pairs from the concept collection. ProbExpSet calculates the average co-occurrence in the SemMedDB of all concept pairs from the concept collection, i.e., the expected probability for co-occurrence of concepts from each setting.

$$\mathrm{ProbObsSet} = \frac{\mathrm{TotalSemOccSet}}{\mathrm{TotalPairNumSet}} \tag{3}$$

where TotalPairNumSet refers to the total number of possible concept pairs in each setting, TotalSemOccSet refers to the total co-occurrence in the SemMedDB of all concept pairs from this setting. ProbObsSet calculates the average co-occurrence in the SemMedDB of all concept pairs from each setting, i.e., the observed probability for co-occurrence of concepts from each setting.

$$EnrichFold = \frac{\mathrm{ProbObsSet}}{\mathrm{ProbExpSet}} \tag{4}$$

The representative concepts for each setting was filtered with a threshold of enrichment fold over one.

Topic modeling

In order to investigate the associations between problem list and practice setting, probabilistic topic modeling could serve as an effective method. Topic modeling has been useful to discover high-level knowledge and a broad range of themes from large collections of text documents. In biomedical domain, it has been applied in various aspects, such as discovering relevant clinical concepts and relations between patients [37], mining treatment patterns in Traditional Chinese Medicine (TCM) clinical cases [38], revealing clinical risk stratification from a large volume of electronic health records [39], clustering long-term biomedical time series such as electrocardiography (ECG) and electroencephalography (EEG) signals [40]. As a type of topic modeling, Latent Dirichlet Allocations (LDA) [19] has gained popularity in diverse fields since it holds great promise as a means of gleaning actionable insight from the text or image datasets. Howes et al. applied unsupervised LDA to analyze clinical dialogues as a higher-level measure of content [41]. Wang et al. developed BioLDA for the application in complex biological relationships in recent PubMed articles [42]. Flaherty et al. rank gene-drug relationships in biomedical literatures based on the LDA [43]. Chen et al. extended LDA by including background distribution to study microbial samples [44]. All these studies amplified the usability of topic modeling and LDA in biomedical field.

In this study, R package "topicmodels" [45] was used to build topic models for both setting similarity calculation and prediction purposes. Instead of using existing evaluation metrics [46–49], we chose the optimal number of topics in our data using log likelihood [50–52]. We calculated the log likelihood values with the number of topics varied from 5 to 150 by 5, and then investigated the performance by comparing the log likelihood value, of which the highest indicates the optimal number of topics. Additional file 1 shows the result of log likelihood method for choosing the optimal number of topics.

Then we fit an LDA model with the optimal number of topics using Gibbs sampling with a burn-in of 1000 iterations. To obtain the posteriors in the LDA analysis, we used collapsed Gibbs sampling because of relatively large number of topics in our study [53]. After we obtained the posteriors, we calculated the log-likelihood of the whole collection of problem settings by integrating all the latent variables.

To obtain setting similarity, the topic modeling was built first using all randomly sampled data, i.e., 1000 notes with chosen representative concepts per each setting, then setting topic probability of training sets was calculated based on the term topic probability derived from the topic models, specifically term topic probability associated with specific setting identified through representative concepts (terms) was extracted to calculate the average topic probability related to each setting. Pearson correlation coefficients among settings were calculated based on topic probabilities in settings using R3.2.1. Clustered Image Maps was then generated for visualization. Clustered Image Maps (i.e., heat maps) represent "high-dimensional" data sets by clustering of the axes to bring similar things together to create patterns of color [18]. To assess relationships between settings and problems, we generated clustered image maps [20] by: i) forming a matrix of the Pearson correlation coefficient among settings from all randomly sampled data, ii) clustering rows and columns of the resulting matrix, and iii) quantile-color coding of the resulting matrix.

To predict settings, all randomly sampled data was divided into 5 parts. Each part in turn was used to evaluate the settings derived by analysis of the other four parts, in the usual n-fold cross-validation manner. Setting prediction was conducted as follows:

- Setting topic probability of training sets was calculated based on the term topic probability for each setting derived from the topic models.
- Test data were predicted using the posterior function of the topic model derived from corresponding training data to obtain the setting topic probability using predicted term topic probability.
- Based on the setting topic probability, similarity was calculated among settings from training data and every one setting from testing data iteratively, so as to get the ranking order of the predicted settings based on Pearson correlation coefficient.

Evaluation
Predicted settings for each tested setting were ranked according to their similarity. In order to evaluate the predicted performance, precision@k and recall@k (k = 2, 4, 6, 8, 10) were used for evaluation (Eqs. 5 and 6). For example, TP@k was calculated as the number of correctly predicted settings from top 1 to top k. FP@k was the unique number of correctly predicted settings from top 1 and top k reducing TP@k, FN@k was the total gold standard setting number (64) reducing TP@k. Based on the precision@k and recall@k, F1@k has also been derived (Eq. 7). We conducted a 5-fold cross validation, mean values were taken as the final evaluation results.

$$\text{Precision@k} = \frac{\text{TP@k}}{\text{TP@k} + \text{FP@k}} \tag{5}$$

$$\text{Recall@k} = \frac{\text{TP@k}}{\text{TP@k} + \text{FN@k}} \tag{6}$$

$$\text{F1@k} = \frac{2^{*}\text{TP@k}}{2^{*}\text{TP@k} + \text{FP@k} + \text{FN@k}} \tag{7}$$

Results
There were 3.3 million notes containing problems in an itemized format with a total of 18.9 million phrases or short sentences that are mapped to 4701 unique problem concepts. There were a total of 1265 settings out of which 266 were aggregated into 64 settings, consisting of 2.4 million notes (73% of normalized notes), and 113 thousand patients with 4573 normalized problems.

Results showed that enrichment folds are between 2.1 and 19.2 after TF-IDF and χ^2 screening. As mentioned before, the threshold of enrichment fold more than 1 was used to filter representative concepts for each setting. These results indicated all concept pairs in each setting from randomly selected notes are significantly co-occurring in the Semantic Medline. We then used these concepts as the representative concepts for each practice setting.

Figure 2 is a word cloud figure developed using the open source software Kumo [54]. It showed the representative concepts from randomly selected four settings where larger font size means higher frequency. These concepts revealed the major themes of corresponding settings. For example, Addiction setting is featured by e.g., alcohol and nicotine, Cardiovascular setting by e.g., coronary and artery, Dermatology setting by e.g., skin and rash, while Urology setting by, e.g., urinary and bladder. Additional file 2 showed the frequency of top 10 concepts for each setting.

Figure 3 showed the clustered image map [18] where a positive correlation (red color) indicates that problems in one setting or group of settings are similar to those in another setting or another group and a negative correlation (blue color) indicates that problems in one setting or group are different from those in another setting or group. From Fig. 3, we can see that some settings are highly

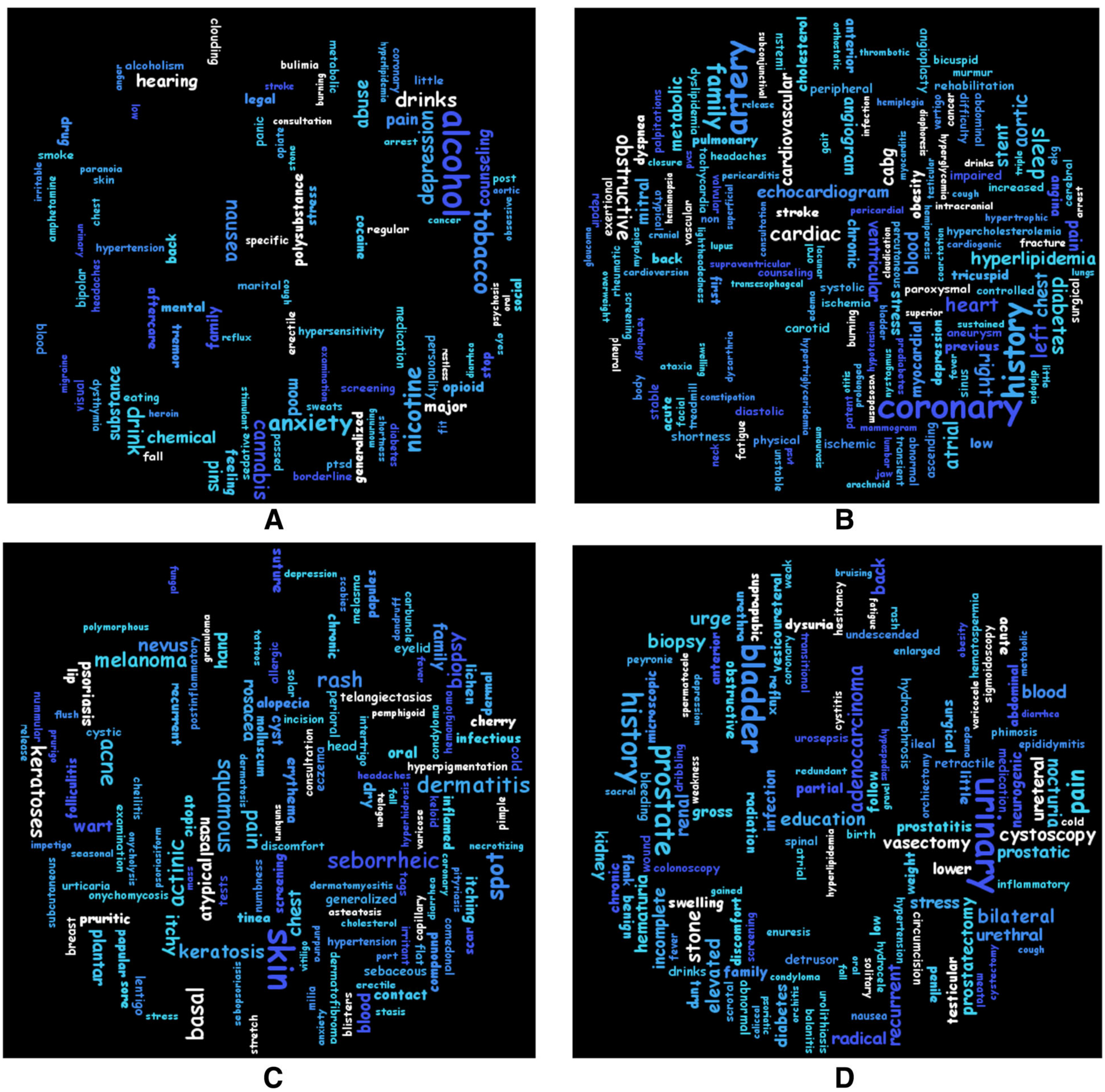

Fig. 2 Concepts in various practice settings. **a**. Addiction setting. **b**. Cardiovascular setting. **c**. Dermatology setting. **d**. Urology setting. Font and color are assigned randomly, and font size is proportional to their significance, i.e., larger font means higher frequency of the term

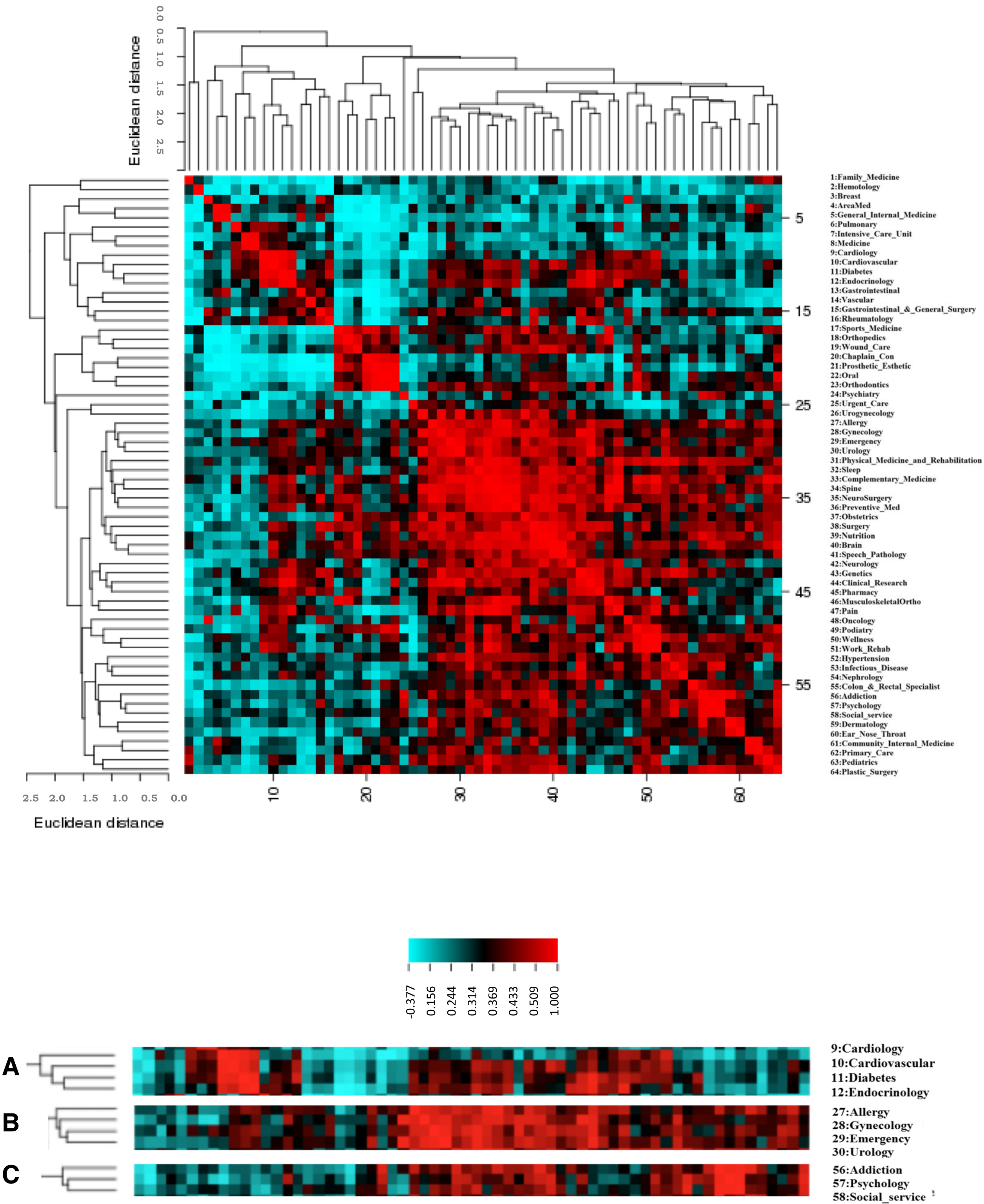

Fig. 3 Clustered image map (CIM) of settings based on Pearson correlation coefficients(X-axis and Y-axis both are settings). Red color indicates positive correlation among settings, and blue color indicates negative correlation among settings. Figure 3**a**, **b** and **c** are three enlarged clusters.

Table 1 Recall@k, Precision@k and F1@k (k = 2, 4, 6, 8, 10) for Pearson correlation coefficient

	@2	@4	@6	@8	@10
Recall	0.719	0.844	0.913	0.916	0.931
Precision	0.882	0.910	0.910	0.916	0.931
F1	0.790	0.875	0.911	0.916	0.931

similar, for example, Cardiology setting is similar to the Cardiovascular, Diabetes and Endocrinology settings (Fig. 3a). Allergy is similar to Gynecology, Emergency and Urology settings (Fig. 3b). Addiction is similar to Pychology and Social service settings (Fig. 3c).

Recall@k, Precision@k and F1@k were shown in Table 1. As k increases, the performance increases gradually. The reason that recall, precision and F1 are the same values when K equals 8 and 10 is FP equals FN when K increase to 8 and 10.

Discussion

During aggregating settings in Mayo Clinics, we have encountered the complexity in organization of the setting concept as stated in the study [29]. Due to the refined feature of practice settings at Mayo Clinic and for the purpose of simpler analysis, we have not totally aligned the extended setting values in Document Ontology (DO). First, we kept the settings that are similar but not exactly same, for example, Cardiology and Cardiovascular as separated settings. In contrast, in the proposed extensions to the DO [29] all settings are distinct. Second, we only used the extended setting values in DO in parallel, and have not studied settings in the hierarchy scenario [29]. For example, Emergency Setting is in parallel to Dermatology Setting in our study. While in the proposed extensions to the DO, Emergency Department is in parallel to Outpatient Setting that includes sub-level Clinic (Non-Acute) Settings, which embody the Dermatology Setting. Our mapping strategy kept features of clinical practices, and it could be used for future document hierarchy management.

In the clinical scenario, it is not easy for physicians from a specific setting to see the big picture with respect to problems most related to the setting. With the association between the problem list and practice setting revealed in our study, a prioritized and meaningful problem list above the irrelevant details could be generated, so as to help practitioners identify the most related problems from a succinct view. Our findings can predict practice setting based on problem list and providing a foundation for future document management. Furthermore, such findings also provide the premise for our next step toward automatic reformulation of problem lists as patients move from one practice setting to another, which would be a huge benefit. For example, when a patient is pursing help from the urology practice setting, his/her problems as the representative concepts associated with this setting, such as "bladder stones", or "prostatitis" could be generated and presented to the physicians. When the patient moves to other practice settings such as cardiovascular practice setting, physicians can easily find the most relevant problems, such as "atypical chest pain" or "coronary vasospasm".

From the practice setting level, highly associated settings which are unknown before can be revealed by using the similarity of problem lists. As shown in Fig. 3, Allergy is associated with Gynecology, Emergency and Urology settings. This finding will have implication in terms of health care for patients.

The reasons that we adopted LDA in our study instead of other methods include: 1) LDA is a unique bi-clustering approach with mixture models [55], considering both document-level and term level similarity. Other clustering methods such as k-means, can only cluster targets based on one similarity measurement. 2) LDA is also a robust generative Bayesian modeling approach, which specifically fits the big data analysis. The robustness is partially because LDA adopts conjugate distribution, such as Dirichlet and multinomial to build models. These features are unique in LDA which are not seen in many other unsupervised methods.

Conclusion

To our best knowledge, our study is the first attempting to discover the association between the problem list and hospital practice settings. The contributions of our method are multiple. First, the NLP techniques normalizing problems from various settings enabled LDA analysis. With our negation function in NLP method, this analysis would be more accurate, compared with other studies [56]. Second, Semantic Medline was used for enrichment analysis of concept pairs to help identify representative concepts for each setting before feeding into LDA model. Third, setting similarity was visualized providing the general view among various settings. Forth, our method realized good prediction for practice settings using the similarity of topics derived from unsupervised LDA model, with the advantage of potential semantic associations among problems in settings. In the future, we plan to investigate how to provide more tailored care by utilizing the association between problem list and practice setting revealed in this study.

Additional files

Additional file 1: Log likelihood values vs. Number of Topics. Note: The optimal number of topics is chosen when the maximum log-likelihoods are observed. This result includes a table showing the result of log likelihood method for choosing the optimal number of topics. (DOCX 100 kb)

Additional file 2: Frequency of concepts in randomly selected practice settings. This file includes a figure showing frequency information of top 10 concepts in each practice setting. (DOCX 13 kb)

Acknowledgements
None.

Funding
The work was supported by the National Institute of Health (NIH) grant R01LM011934, R01EB19403, R01LM11829, and U01TR02062. Publication costs are funded by U01TR02062.

Availability of data and materials
The EHR dataset are not publicly available due to the privacy of patients.

About this supplement
This article has been published as part of *BMC Medical Informatics and Decision Making Volume 19 Supplement 3, 2019: Selected articles from the first International Workshop on Health Natural Language Processing (HealthNLP 2018)*. The full contents of the supplement are available online at https://bmcmedinformdecismak.biomedcentral.com/articles/supplements/volume-19-supplement-3.

Author's contributions
All co-authors are justifiably credited with authorship, according to the authorship criteria. Final approval is given by each co-author. In detail: LW- design, development, data collection, analysis of data, interpretation of results, and drafting and revision of the manuscript; YW – design, analysis of data and revision of the manuscript; FS: analysis of data and revision of the manuscript; MRM- analysis of data; HL- conception, design, development, data collection, analysis of data, interpretation of results, critical revision of manuscript.

Ethics approval and consent to participate
This study was a retrospective study of existing records. The study and a waiver of informed consent were approved by Mayo Clinic Institutional Review Board in accordance with 45 CFR 46.116 (Approval #17–003030).

Consent for publication
Not applicable; the manuscript does not contain individual level of data.

Competing interests
The authors declare that they have no competing interests.

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

Published: 4 April 2019

References

1. Jha AK. Meaningful use of electronic health records: the road ahead. Jama. 2010;304(15):1709–10.
2. Medicare Cf, Services M: Medicare and Medicaid EHR Incentive Program: Meaningful Use Stage 1 Requirements Overview, 2010. In.; 2012.
3. Hsiao C-J, Hing E, Socey TC, Cai B. Electronic medical record/electronic health record systems of office-based physicians: United States, 2009 and preliminary 2010 state estimates. Natl Cent Health Stat. 2010:2001–11.
4. Henricks WH. "Meaningful use" of electronic health records and its relevance to laboratories and pathologists. J Pathol Inform. 2011;2.
5. Weed L. Medical records that guide and teach. N Engl J Med. 1968;278(11):652–7.
6. Fischetti L, Mon D, Ritter J, Rowlands D: Electronic health record–system functional model. Chapter Three: direct care functions 2007.
7. Wright A, Maloney FL, Feblowitz JC. Clinician attitudes toward and use of electronic problem lists: a thematic analysis. BMC Med Inform Decis Mak. 2011;11(1):36.
8. Hartung DM, Hunt J, Siemienczuk J, Miller H, Touchette DR. Clinical implications of an accurate problem list on heart failure treatment. J Gen Intern Med. 2005;20(2):143–7.
9. Franco M, Giussi BM, Otero C, Landoni M, Benitez S, Borbolla D, Luna D. Problem oriented medical record: characterizing the use of the problem list at hospital Italiano de Buenos Aires. Stud Health Technol Inform. 2014;216:877.
10. Pacheco JA, Thompson W, Kho A. Automatically detecting problem list omissions of type 2 diabetes cases using electronic medical records. In: AMIA Annual Symposium Proceedings. American medical informatics association; 2011. p. 1062. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3243294/.
11. Carpenter JD, Gorman PN. Using medication list--problem list mismatches as markers of potential error. In: Proceedings of the AMIA Symposium: 2002: American Medical Informatics Association; 2002: 106.
12. Meystre S, Haug PJ. Automation of a problem list using natural language processing. BMC Med Inform Decis Mak. 2005;5(1):30.
13. Meystre S, Haug PJ. Natural language processing to extract medical problems from electronic clinical documents: performance evaluation. J Biomed Inform. 2006;39(6):589–99.
14. Plazzotta F, Otero C, Luna D, de Quiros F. Natural language processing and inference rules as strategies for updating problem list in an electronic health record. Stud Health Technol Inform. 2012;192:1163.
15. Wright A, Pang J, Feblowitz JC, Maloney FL, Wilcox AR, McLoughlin KS, Ramelson H, Schneider L, Bates DW. Improving completeness of electronic problem lists through clinical decision support: a randomized, controlled trial. J Am Med Inform Assoc. 2012;19(4):555–61.
16. Zhou X, Zheng K, Ackerman M, Hanauer D. Cooperative documentation: the patient problem list as a nexus in electronic health records. In: Proceedings of the ACM 2012 conference on computer supported cooperative work. ACM; 2012. p. 911–20. http://hai.ics.uci.edu/papers/p911-zhou.pdf.
17. Blei DM. Probabilistic topic models. Commun ACM. 2012;55(4):77–84.
18. Weinstein JN, Myers TG, O'connor PM, Friend SH, Fornace AJ, Kohn KW, Fojo T, Bates SE, Rubinstein LV, Anderson NL. An information-intensive approach to the molecular pharmacology of cancer. Science. 1997;275(5298):343–9.
19. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. J Mach Learn Res. 2003;3:993–1022.
20. CIMMiner. http://discover.nci.nih.gov/cimminer/. Accessed 12 Apr 2018.
21. Semantic Medline. https://skr3.nlm.nih.gov/SemMed/. Accessed 25 Mar 2018.
22. Frazier P, Rossi-Mori A, Dolin RH, Alschuler L, Huff SM. The creation of an ontology of clinical document names. Stud Health Technol Inform. 2001;1:94–8.
23. Domain SM: Standardizing clinical document names using the HL7/LOINC document ontology and LOINC codes. 2010.
24. Hyun S, Shapiro JS, Melton G, Schlegel C, Stetson PD, Johnson SB, Bakken S. Iterative evaluation of the health level 7—logical observation identifiers names and codes clinical document ontology for representing clinical document names: a case report. J Am Med Inform Assoc. 2009;16(3):395–9.
25. Li L, Morrey CP, Baorto D. Cross-mapping clinical notes between hospitals: an application of the LOINC document ontology. In: AMIA annual symposium proceedings/AMIA symposium AMIA symposium. American medical informatics association. 2011;2011:777–83. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3243240/.
26. Hyun S, Bakken S. Toward the creation of an ontology for nursing document sections: mapping section headings to the LOINC semantic model. In: AMIA Annual Symposium Proceedings. American Medical informatics association; 2006. p. 364. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1839622/.
27. Shapiro JS, Bakken S, Hyun S, Melton GB, Schlegel C, Johnson SB. Document ontology: supporting narrative documents in electronic health records. In: AMIA. Citeseer: American medical informatics association; 2005. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1560738/.
28. McDonald C, Huff S, Deckard J, Holck K, Vreeman DJ. Logical observation identifiers names and codes (LOINC®) users' guide. Indianapolis: Regenstrief Institute; 2004. http://viw1.vetmed.vt.edu/Education/Documentation/LOINC/LOINCUserGuide200901.pdf.
29. Rajamani S, Chen ES, Wang Y, Melton GB. Extending the HL7/LOINC document ontology settings of care. In: AMIA Annual Symposium Proceedings: 2014: American medical informatics Association; 2014. p. 994. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4419877/.
30. Agrawal A, He Z, Perl Y, Wei D, Halper M, Elhanan G, Chen Y. The readiness of SNOMED problem list concepts for meaningful use of electronic health records. Artif Intell Med. 2013;58(2):73–80.

31. Liu H, Wagholikar K, Wu ST-I. Using SNOMED-CT to encode summary level data–a corpus analysis. AMIA Summits Transl Sci Proc 2012. 2012:30.
32. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004;32(suppl 1): D267–70.
33. Rhee SY, Wood V, Dolinski K, Draghici S. Use and misuse of the gene ontology annotations. Nat Rev Genet. 2008;9(7):509–15.
34. Winnenburg R, Shah NH. Generalized enrichment analysis improves the detection of adverse drug events from the biomedical literature. BMC bioinformatics. 2016;17(1):250.
35. Morota G, Beissinger TM, Peñagaricano F. MeSH-informed enrichment analysis and MeSH-guided semantic similarity among functional terms and gene products in chicken. G3: Genes| Genomes| Genetics. 2016;6(8): 2447–53.
36. Zhang Y, Tao C, Jiang G, Nair AA, Su J, Chute CG, Liu H. Network-based analysis reveals distinct association patterns in a semantic MEDLINE-based drug-disease-gene network. J Biomed Semantics. 2014;5(1):33.
37. L-w L, Long W, Saeed M, Mark R. Latent topic discovery of clinical concepts from hospital discharge summaries of a heterogeneous patient cohort. In: Engineering in Medicine and Biology Society (EMBC), 2014: 36th Annual International Conference of the IEEE. IEEE; 2014. p. 1773–6. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4894488/.
38. Yao L, Zhang Y, Wei B, Wang W, Zhang Y, Ren X, Bian Y. Discovering treatment pattern in traditional Chinese medicine clinical cases by exploiting supervised topic model and domain knowledge. J Biomed Inform. 2015;58:260–7.
39. Huang Z, Dong W, Duan H. A probabilistic topic model for clinical risk stratification from electronic health records. J Biomed Inform. 2015;58: 28–36.
40. Wang J, Liu P, She MF, Nahavandi S, Kouzani A. Biomedical time series clustering based on non-negative sparse coding and probabilistic topic model. Comput Methods Prog Biomed. 2013;111(3):629–41.
41. Howes C, Purver M, McCabe R. Investigating topic modelling for therapy dialogue analysis. In: Proceedings IWCS workshop on computational semantics in clinical text (CSCT). Association for computational linguistics. 2013;2013:7–16. http://www.aclweb.org/anthology/W13-0402.
42. Wang H, Ding Y, Tang J, Dong X, He B, Qiu J, Wild DJ. Finding complex biological relationships in recent PubMed articles using bio-LDA. PLoS One. 2011;6(3):e17243.
43. Flaherty P, Giaever G, Kumm J, Jordan MI, Arkin AP. A latent variable model for chemogenomic profiling. Bioinformatics. 2005;21(15):3286–93.
44. Chen X, He T, Hu X, An Y, Wu X. Inferring functional groups from microbial gene catalogue with probabilistic topic models. In: Bioinformatics and Biomedicine (BIBM), 2011 IEEE International Conference on: 2011: IEEE; 2011. p. 3–9. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6120400.
45. Hornik K, Grün B. Topicmodels: An R package for fitting topic models. J Stat Softw. 2011;40(13):1–30.
46. Arun R, Suresh V, Madhavan CV, Murthy MN. On finding the natural number of topics with latent dirichlet allocation: some observations. In: Pacific-Asia conference on knowledge discovery and data mining. Berlin: Springer; 2010. p. 391–402. https://link.springer.com/chapter/10.1007/978-3-642-13657-3_43.
47. Cao J, Xia T, Li J, Zhang Y, Tang S. A density-based method for adaptive LDA model selection. Neurocomputing. 2009;72(7–9):1775–81.
48. Deveaud R, SanJuan E, Bellot P. Accurate and effective latent concept modeling for ad hoc information retrieval. Doc Numérique. 2014;17(1):61–84.
49. Griffiths TL, Steyvers M. Finding scientific topics. Proc Natl Acad Sci. 2004; 101(suppl 1):5228–35.
50. Chen B, Chen X, Xing W. Twitter archeology of learning analytics and knowledge conferences. In: Proceedings of the fifth international conference on learning analytics and knowledge. New York: ACM; 2015. p. 340–9. https://dl.acm.org/citation.cfm?id=2723584.
51. Scott JG, Baldridge J. A recursive estimate for the predictive likelihood in a topic model. J Mach Learn Res. 2013;31:527–35.
52. Moslehi P, Adams B, Rilling J: Feature Location using Crowd-based Screencasts. 2018.
53. Asuncion A, Welling M, Smyth P, Teh YW. On smoothing and inference for topic models. In: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence. Arlington: AUAI Press; 2009. p. 27–34. https://dl.acm.org/citation.cfm?id=1795118.
54. Kumo. https://github.com/kennycason/kumo. Accessed 19 Feb 2018.
55. Shan H, Banerjee A. Bayesian co-clustering. In: Data Mining, 2008 ICDM'08 Eighth IEEE International Conference on. IEEE; 2008. p. 530–9. https://ieeexplore.ieee.org/document/4781148.
56. Hirsch JS, Tanenbaum JS, Gorman SL, Liu C, Schmitz E, Hashorva D, Ervits A, Vawdrey D, Sturm M, Elhadad N. HARVEST, a longitudinal patient record summarizer. J Am Med Inform Assoc. 2015;22(2):263–74.