Patient Clustering with Uncoded Text in Electronic Medical Records

Ricardo Henao[1], PhD, Jared Murray[1], PhDc, Geoffrey Ginsburg[1], MD, PhD, Lawrence Carin[1], PhD, Joseph E. Lucas[2], PhD
[1]Duke University, Durham, NC
[2]Quintiles, Durham, NC

Abstract

We propose a mixture model for text data designed to capture underlying structure in the history of present illness section of electronic medical records data. Additionally, we propose a method to induce bias that leads to more homogeneous sets of diagnoses for patients in each cluster. We apply our model to a collection of electronic records from an emergency department and compare our results to three other relevant models in order to assess performance. Results using standard metrics demonstrate that patient clusters from our model are more homogeneous when compared to others, and qualitative analyses suggest that our approach leads to interpretable patient sub-populations when applied to real data. Finally, we demonstrate an example of our patient clustering model to identify adverse drug events.

Introduction

Automating the creation of time evolving homogeneous clusters of patients is a powerful approach to the meaningful use of electronic medical records. If it is done successfully, the empirical distributions of important clinical parameters within each cluster of patients, such as diagnosis, medication and treatment options, can be used to address a wide array of important questions for individual and population level healthcare. (i) For any particular cluster, one might track outcomes for patients on different therapies, thereby identifying which therapy is most effective or has the least side effects. (ii) Time evolving clustering of patients allows adaptation to a potentially changing medical landscape. For example, if a new disease appears with a particular constellation of symptoms, over time the model should identify that disease by its symptoms and create a new cluster of patients who have that disease. (iii) By tracking incidence rates of each cluster independently, it is possible to identify disease outbreaks. (iv) Alternatively, tracking incidence by geographic location rather than through time may identify areas of hazardous environmental exposure. (v) The empirical distribution of diagnoses for patients in a particular subgroup is itself a differential diagnosis for a new patient belonging to that subgroup. Similar to diagnoses, empirical distributions of medications and outcomes lead to a menu of potential therapies and disease prognoses respectively.

Natural Language Processing: In any natural language there are a wide array of challenges that make automated computer analysis challenging. Ambiguities, synonyms, modifiers and other idiosyncrasies of human text and speech make the identification of concepts and the connections between them challenging. In medical text, where the author is also the physician responsible for ensuring proper care of many patients, the problem is compounded by competing job pressures. This leads to numerous mistakes in spelling and grammar, missing punctuation, and high levels of duplication from extensive copy and pasting.

Natural language processing (NLP) is a suite of methods – involving statistical modeling and artificial intelligence together with manually curated "knowledge bases" – that is designed to take messy, complicated text and convert it into a series of concepts, possibly embedded in an ontology. There are a number of commercial products designed specifically to do this with medical text such as MedLEE (1) and IBM's Watson. A similar, publically accessible, approach that is heavily reliant on manually curated databases of concepts is MetaMap (2). The general approach to NLP involves a multi-step process of data preprocessing, multiple parsing algorithms, regularization and encoding. The ultimate goal of NLP in this context is to map the unstructured text in a medical record to a set of pre-defined concepts (such as those contained in the UMLS Metathesaurus). The presence or absence of concepts in particular records can then be used in various statistical models or machine learning algorithms for specific tasks.

Wang et al. (3) utilize co-occurrence statistics to identify NLP derived concepts from discharge summaries that occur frequently with particular drugs. A similar approach was taken by Haerian et al. (4) to identify rhabdomyolysis and agranulocytosis caused by adverse drug events. In order to eliminate possible disease related causes of these symptoms, Haerian et al. performed manual curation based on chart reviews to eliminate NLP concepts that are more likely related to relevant diseases. Other uses of NLP include a comparison of clinical trials prescreening based on ICD9 code versus NLP concepts (5), a classifier of radiology reports designed to detect the

presence or absence of specific radiological features (6) and identification of surgical complications (7). All of these approaches utilize custom knowledge bases designed specifically for the study at hand, and success or failure is typically assessed based on agreement with a chart review.

There are a number of purely statistical approaches to working with medical records data. For example, the use of regression/change point models to detect infection disease outbreaks (8, 9). Classification models such as belief networks have been used for single or multiple disease diagnosis (10, 11). Dimensionality reduction and mixture models can be used for exploratory analysis and visualization to uncover adverse drug effects (12, 13). This work generally leaves unanswered the question of what data to use in fitting the proposed statistical models, though in some cases, NLP concepts would be appropriate.

Patient Similarity Metrics: There are numerous approaches to defining a distance between patients using NLP (14, 15). These approaches make use of both the concepts and ontologies produced by NLP. Distance between two patients is either proportional to the size of the overlap in concepts or a function of the ontological relationships between concepts assigned to the two patients. While a distance metric can, in theory, be used for the application we discuss in this paper, there is a substantial computational challenge to clustering on very large data sets. Distances must be computed for every pair of patients and the resulting matrix of distances must be manipulated. For even moderate population sizes this is intractable with standard approaches.

The use of our approach for clustering patients represents a potential synergy with NLP based similarity metrics. By computing NLP similarity only on patients in the same cluster (as determined by our model), one might only need to work with a block diagonal similarity matrix, thereby substantially decreasing the computational complexity of the NLP based clustering problem.

Our approach: In this paper we propose a statistical model that leads to the identification of homogeneous patient sub-populations based on text data in the initial history of present illness (HPI) and final diagnosis (DX) of a patient's electronic medical record. The model can be used to predict sub-population membership of a new patient based solely on HPI data. This not only gives us access to an *average* virtual HPI for each patient sub-population but a differential diagnosis in the form of a ranked list of possible diagnoses extracted from patients belonging to that subpopulation. This allows us to quantify the uncertainty of a diagnosis. We hypothesize that estimates of uncertainty will allow a system to know when to ask for help, and we expect that feature will be critical for successful clinical decision support tools in the future.

While we are using only a small part of the medical record, this work represents an initial step in the construction of an automated statistical model for clustering patients into homogeneous groups. We expect that improvements can be made through the inclusion of other information in the record such as medicines, demographics, labs, etc. In this study, we do not make use of natural language processing, nor do we generate models attempting to identify any specific disease or patient feature. However we believe that our work is synergistic with these approaches to the analysis of EMR data. Specifically, we are applying our model to the unprocessed text, but our model could as easily be applied to NLP concepts derived from that text. Additionally, a model of concept correlation could be used in future iterations of NLP software to help identify the correct choice of concept given the other concepts in a particular medical document.

Data

We are working with a database of 55,837 emergency department (ED) visit records collected during the first three quarters of 2009. Identifying information such as name, address, phone number, social security number and medical record number were removed before data processing. From each record, five fields were extracted: chief complaint (CC), initial history of present illness (HPI), diagnosis (DX), disposition and age. From all records, 5,616 were discarded due to lack of HPI or DX fields. HPIs consist of lists of words and DXs of lists of as many as 5 diagnoses per patient. The total number of terms in the HPI and DX dictionaries is 37,449 (7,148 present in more than 10 instances) and 8,246 (926), respectively. There are 836 CCs (100), 71 dispositions (40) and ages range from 1 to 103 years.

For this study, we will use a subset of the database. We consider 10,204 records (8,808 unique patients) corresponding to 11 common chief complaints. These are medical minor, cardiac symptoms, trauma complex, abdominal pain male, headache, neuro other, pain other, shortness of breath, laceration minor, infection local and seizure. The subset was selected to minimize compute times while maintaining a mixture of specific (seizure, headache) and non-specific (medical minor, pain other) chief complaints. We eliminate stop words, numbers and

terms occurring less than 10 times. The final subset contains a total of 725 words in the HPI dictionary and 323 terms in the DX dictionary.

Model

The Chinese restaurant process (16) (CRP) formally defines a distribution on the space of partitions of $\mathbb{N}$. For our purposes, the random process is described by two rules: (i) the first patient record analyzed always begins in the first cluster/component and (ii) the i-th patient starts a new cluster with probability proportional to parameter α or joins a current cluster with probability proportional to the number of patients already in it. The recently developed distance dependent CRP (17) (ddCRP), modified the second rule so the i-th patient joins a cluster with probability equal to a function of its *distance* to the other patients in that cluster. Inclusion of ancillary data in this distance function allows the topic model to be influenced by factors external to the documents. In our case, we can potentially utilize DX data to inform clustering of data from the HPI. However, direct application of ddCRP to analyze the plain text in the HPI requires that pairwise distances be set a-priori and computed from patient attributes other than HPI, which can become intractable for data sets with large numbers of patients. This is the same challenge faced by NLP patient similarity metrics (14, 15).

Our approach is different in the sense that distances are defined between patients and clusters rather than between pairs of patients, i.e. the i-th patient sits in an occupied cluster with probability proportional to its distance to that cluster. This approach allows our model to avoid the computation of distance between every pair of patients but still allows the flexibility of the ddCRP to modify cluster membership probabilities based on external attributes.

Let x_{im} be the m-th word (out of a total of M_i) in the DX for patient i. For the k-th cluster, let $d_{ik} = \prod_{m=1}^{M_i} \text{Discrete}(x_{im}|\boldsymbol{\psi}_k)$ be the distance between patient *i* and cluster *k*, where $\boldsymbol{\psi_k}$ is a probability vector over the space of words in the DX dictionary. We define *K* to be the number of non-empty patient clusters and $z_i \in 1 \ldots K$ to be the cluster assignment for patient i. The prediction rule in the CRP can be written as

$$z\_i \,|\boldsymbol{z}_{\setminus i}, \alpha \propto \alpha\delta_{k^*} + \sum_{k=1}^{K} n_k \;\delta_k \,,$$

where $\boldsymbol{z}_{\setminus i}$ is the vector of cluster assignments excluding patient i and k^* is a new cluster. The assignment z_i for patient i depends on the concentration parameter α and the number of patients n_k in cluster k. Note as well that n_k is a function of the assignments made to cluster k through $\boldsymbol{z}_{\setminus i}$. We label this approach *vanilla* CRP and we will compare our proposed method to this one in the results section of the paper. Alternatively, the prediction rule we propose is

$$z\_i \,|\boldsymbol{z}_{\setminus i}, \alpha, \boldsymbol{D} \propto \alpha\delta_{k^*} + \sum_{k=1}^{K} d_{ik}\,\delta_k \,,$$

where $\boldsymbol{D}$ is a matrix of distances d_{ik} between patients and clusters.

Conceptually, the CRP occupies components according to popularity, the ddCRP does it instead by looking at pair-wise similarities or links across observations whereas our modified ddCRP (mddCRP) occupies components with observations similar to those already assigned to it. Consequently, while mddCRP saves compute time by avoiding computing every pairwise difference between two observations, the distance matrix between patient and clusters changes with the assignment vector $\boldsymbol{z}$ and so must be updated during inference. It is easy to show that the traditional CRP is a special case of both ddCRP and mddCRP. In particular, for mddCRP it is enough to make $d_{ik} = n_k$ and for ddCRP we define the distance between any two observations to be 1.

Each patient record is composed of two lists of words, $\boldsymbol{w}_i$ and $\boldsymbol{x}_i$, defined to be the list of words in the HPI and the DX respectively for patient i. Both word vectors are assumed to be drawn from discrete distributions as follows

$$z_i \sim \text{mddCRP}(\alpha, \boldsymbol{D}), \qquad x_{im} \sim \text{Discrete}(\boldsymbol{\psi}_k), \qquad \boldsymbol{\psi}_k \sim \text{Dirichlet}(\gamma \mathbf{1}_M),$$
$$w_{im} \sim \text{Discrete}(\boldsymbol{\theta}_k), \qquad \boldsymbol{\theta}_k \sim \text{Dirichlet}(\beta \mathbf{1}_N),$$

where $\boldsymbol{x}_i$ and $\boldsymbol{w}_i$ have sizes M_i and N_i, respectively. The matrices of observed data are denoted by $\boldsymbol{W} = \{w_i\}$ and $\boldsymbol{X} = \{x_i\}$. Each patient cluster has two parameters $\boldsymbol{\psi}$ and $\boldsymbol{\theta}$, seen as word proportions from two different dictionaries of sizes M and N. For cluster k, $\boldsymbol{\psi}_k$ and $\boldsymbol{\theta}_k$ are interpreted as average virtual DX and HPI, respectively. The model is controlled by hyperparameters β, γ and α that specify the concentration of words in the modeling variable, distance attribute and the expected number of components, respectively.

Inference is performed with Markov chain Monte Carlo by iteratively sampling from the conditional posterior of the parameters of interest namely, z_i, $\boldsymbol{\psi}_k$, $\boldsymbol{\theta}_k$, β, γ and α. The first three parameters are sampled using Gibbs sampling whereas the last three use slice sampling with uniform bounded priors. In order to make predictions, we need to compute the conditional distribution of a new patient $\boldsymbol{w}^*$ (test set) given all training data. When diagnosis information is not available for test data, which is often the case, predictions are entirely based on the likelihood and posterior of HPIs as we assume the distribution over diagnoses to be flat in order to avoid favoring any component of the model. We have made inference and prediction procedures available together with processed data and Matlab source code as supplementary material at http://www.duke.edu/~rh137/mrecs.html.

Other techniques for grouping patients: We will compare the performance of our mddCRP model to three other clustering approaches. The first will be the use of *vanilla* CRP (vCRP) in which we fit a standard CRP model to a single document − HPI concatenated with DX − for each patient. The biggest drawback to this approach is that the average number of observed terms from HPI is much larger than that from the DX. This makes the model prone to produce components dominated by HPI features. The second "model" consists of clustering based entirely on chief complaint (naive). This is a very appealing straw man as it represents the ability of the triage nurse to diagnose the patient based solely on his/her brief initial interaction with the patient. For some chief complaints the consistency of final diagnosis is quite high. Additionally, there appear to be instances in which the nurse quite reasonably records that he/she doesn't know how to group the patient such as for the CC "medical minor". Finally, in order to try to take advantage of the relative accuracy of CC in some cases, we compare to a ddCRP model, which is fit independently to each of the different chief complaints. As we show in the next section, this model fails to collect some subsets of patients, which should be grouped because they present with different chief complaints.

A more direct approach to the task at hand is to treat it as a supervised problem in which components will be biased towards individual diagnosis discrimination. There are a number of models designed for this purpose (18, 19), however their use would require significant natural language processing of the records in order to remove synonyms. Additionally, such approaches are inflexible in the face of new types of diagnoses or new synonyms for diagnoses.

Results

We fit vCRP, ddCRP and mddCRP using a 2-fold cross validation scheme. For all models we used uniform bounded priors for α, β and γ with bounds $(10^{-3}, 10)$, $(10^{-2}, 10)$ and $(10^{-2}, 10)$, respectively. We verified empirically that further expanding these bounds did not produce noticeable changes in inference. Results are obtained after summarizing MCMC runs of the models. Each run consists of 300 posterior samples collected after 500 burin-in iterations. For naive no sampling is required as its components are obtained deterministically from chief complaint information. Traces of α, β and γ and marginal likelihoods were used for monitoring convergence. After inference we discard components containing less than 0.2% of the total number of observations. These *spurious* components are usually formed by outliers thus we regard them as not representative of the underlying structure of the data set. The mean number of resulting components for Naive, vCRP, ddCRP and mddCRP is 11, 16, 26, 17, respectively.

Technical assessment of homogeneity of patient clusters

Clustering of HPI data: One approach to quantifying the overall predictive power of a model is by means of its *perplexity*. This is a widely used performance metric from the natural language processing community (20) that is intended to quantify the goodness of fit of the model in held out samples. Let $\boldsymbol{w}_i^* \in \boldsymbol{W}^*$ to be the N_i^* words in the HPI for an observation from test set $\boldsymbol{W}^*$. Define $p(\boldsymbol{w}_i^*|\mathcal{M})$ to be the predictive distribution given a particular model, $\mathcal{M}$, that is being evaluated.

$$(\boldsymbol{W}^*|\mathcal{M}) = e^{-\sum_i \log p(\boldsymbol{w}_i^*|\mathcal{M})/\sum_i N_i^*},$$

Better models will in general assign larger probabilities, on average, to words in the HPI of test cases leading to lower perplexities. This is interpreted as the model being less confused in average by test cases. We computed perplexities for every posterior sample collected during inference for vCRP, ddCRP and mddCRP. Since Naive is deterministic, its perplexity is a single number: 35.4. Figure 1(a) shows perplexity boxplots computed for the posterior samples obtained during the 2-fold inference. We see clearly that mddCRP outperforms all the models being evaluated. This implies that our approach to incorporating diagnosis data has led to better clustering of HPI data even when compared to models that focus heavily on HPI data only.

Overlap in DX data: Let A_k be the set of all diagnoses for any patients associated with cluster k. We compute the overlap of component k with component j as $|A_k \cap A_j|/|A_k|$ where $|A|$ is the number of elements in the set A. This represents the amount of agreement of diagnoses between any two components of the model where 1 indicates that

$A_k \subseteq A_j$ and 0 indicates that there is no overlap in the set of diagnoses. Although we expect overlap, we prefer that those overlaps be small, as that indicates that the patient groupings are more homogeneous. Figure 1(b) shows overlap boxplots for all components. We see that mddCRP has in average the lowest overlap of the four models. As one might expect, vCRP produces overlaps spanning the whole [0,1] interval due to HPI terms dominating component formation in the model. The coverage we see for ddCRP is explained by its inability to merge components from different chief complaints, yet in average shows less overlap than Naive. Outliers (crosses) with 100% overlap are caused by small components with a small pool of diagnoses that are subsets of larger components. Naive does not show such extreme overlaps, however 8 out of 11 components in Naive have more than 60% overlap with the component labeled as "medical minor". Our model not only produces the smallest coverage but most of its components have less than 50% overlap indicating again that mddCRP has produced patient groups that are more homogeneous than those produced by the other approaches.

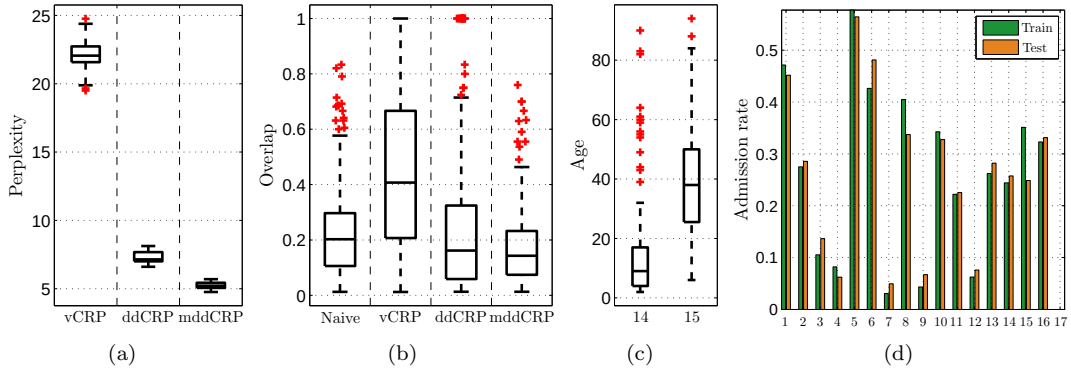

Figure 1. Performance measures and component comparison. (a) HPI perplexities, the reference is 35.4 (Naive). (b) DX based component overlap. Mean differences between mddCRP and the other models in (a, b) are significant at the 1% confidence level. (c) Age distribution for seizure components (14 and 15). (d) Admission rates for components of mddCRP computed for train/test data. Disposition information was not used to train the model.

Taken together, these two assessments of model performance demonstrate that mddCRP leads to a higher level of confidence in patient clustering as well as patient clusters with more homogeneous lists of diagnoses.

Qualitative analysis of patient clusters

We now focus on the results obtained from mddCRP. Table 1 shows the most frequent terms observed in each of the 17 components produced by mddCRP along with the percentage of observations assigned to each. One of the key features of our model is the ability to group synonymous terms without the need of prior knowledge about how diagnoses are related. Equivalent DXs such as chf/congestive heart failure (component 5), rabies vaccination/rabies vaccine (component 17), seizure febrile/febrile seizure (component 14) are collected in the same cluster without the need of a knowledge base of synonyms.

Table 1. Top HPI/DX terms from the mddCRP components. Sizes are proportions of total observations relative to the training set. First (grey) and second (white) lines in each cell correspond to HPI and DX terms, respectively.

Cluster	Size	HPI and Diagnosis terms
1	10.90%	pain, chest, sob, prior, denies, problem, worsens, diaphoresis, radiating, nausea, onset, past
		chest pain, unstable angina, shortness breath, palpitations, chest pain acute
2	10.50%	pain, abdominal, nausea, vomiting, denies, diarrhea, last, prior, past, problem, fever, abd
		abdominal pain, nausea vomiting, constipation, pancreatitis acute, diarrhea
3	9.60%	headache, pain, prior, nausea, denies, headaches, neck, past, problem, vision, head, last
		headache, migraine, hypertension, viral syndrome, nausea, dizziness, vomiting
4	8.50%	brought, immunizations, utd, fever, parents, mother, prior, cough, diarrhea, problem, pain, mom
		fever, uri acute, vomiting, cough, abdominal pain, constipation, viral syndrome, viral uri
5	7.30%	sob, pain, cough, chest, denies, past, last, prior, breath, fever, problem, shortness

		shortness breath, copd, pneumonia bacterial, chf, dyspnea, congestive heart failure
6	6.50%	pain, car, last, tetanus, mvc, loc, driver, utd, restrained, ems, neck, back
		mva, motor vehicle accident, pneumothorax closed traumatic, gunshot wound, headache
7	5.70%	laceration, utd, last, tetanus, head, loc, trauma, pain, brought, right, immunizations, hit
		laceration finger, laceration face, laceration scalp, laceration forehead
8	5.70%	pain, weakness, denies, numbness, headache, prior, symptoms, vision, past, last, left, right
		headache, tia, numbness, cva acute, paresthesia, hypertension, dizziness, syncope, vertigo
9	4.90%	pain, cough, throat, fever, sore, denies, mild, productive, symptoms, headache, chills, sick
		pharyngitis acute, viral syndrome, uri acute, cough, headache, fever, sinusitis acute
10	4.80%	pain, chest, denies, sob, prior, palpitations, past, felt, ems, feeling, last, episodes
		palpitations, chest pain, syncope, anxiety, atrial fibrillation rapid ventricular rate
11	4.40%	pain, swelling, prior, denies, problem, fever, left, fevers, right, leg, wound, redness
		cellulitis leg, cellulitis, wound check follow exam, wound infection surgical
12	4.00%	pain, back, denies, left, right, past, leg, numbness, prior, last, lower, trauma
		back pain, low back pain, leg pain, sciatica, knee pain, neck pain, fall accidental
13	3.90%	pain, denies, prior, chest, problem, sob, fevers, past, chills, abd, vomiting, nausea
		abdominal pain, uti, chest pain, headache, dehydration, constipation, back pain
14	3.30%	seizure, brought, mom, immunizations, utd, parents, prior, ems, seizures, last, mother, head
		seizure grand mal, seizure, febrile seizure, seizure febrile, epilepsy, fever, headache
15	3.30%	seizure, ems, past, episodes, denies, pain, seizures, prior, multiple, last, head, episode
		seizure grand mal, seizure, headache, syncope, altered mental status, alcohol withdrawal
16	3.20%	pain, fall, head, loc, fell, prior, denies, back, problem, utd, last, neck
		fall accidental, syncope, head injury unspecified, concussion
17	0.60%	dog, shot, exposure, well, complaints, last, utd, symptom, primary, control, bite, denies
		rabies vaccination, rabies vaccine, rabies exposure, wound check follow exam, uri acute

In addition to grouping synonyms, the model is capable of splitting components that at first glance seem homogeneous. While patients in both components 3 and 8 are given the diagnosis “headache” more than any other diagnosis, these two can be seen to have vastly different healthcare implications and outcomes. Component 3 is generally not life threatening, consisting of patients with migraines or sinus headaches while component 8 appears to be dominated by patients with hypertensive headaches or who are suspected of having a stroke. This is supported by the admission rates associated with the two components, shown in Figure 1(d). Even though admission rates are not explicitly included in the model, only around 10% of component 3 patients are admitted compared to approximately 35% of those falling in component 8.

The two seizure components (14 and 15) also seem to have different etiologies. From the top HPI terms, we see that 14 is loaded with terms that suggest younger patients. This fits with the diagnosis “febrile seizure” which is known to be more common in children. As shown in Figure 1(c) even though age is not incorporated in the model, there is a marked difference in the ages of patients assigned to the two components.

Pharmacovigilance

One commonly written about application of medical records analysis is the identification of adverse drug events. In order to assess the relevance of patient clusters, we reran our model on a larger patient cohort. We examined medication lists for each patient and tested for increased use of particular medications in each cluster (Chi-squared test, Bonferroni correction for multiple testing). The results of this analysis represent another confirmation of the relative purity of our patient clusters.

The drugs over-represented in an "allergic reaction" cluster include drugs often used to treat allergies or allergic symptoms such as Benadryl, Vitamin E, Hydroxyzine and Advil as well as drugs with well-known risks of allergic reaction such as Effexor XR, Celebrex and Estradiol. Similar results hold for the "eye pain" cluster, which included Cosopt and Sotalol (treatment for glaucoma), Prednisone (anti-inflammatory eye drop) and Erythromycin (antibiotic eye drop) as well as Zyprexa – a medicine for schizophrenia, which is known to cause eye pain in some patients.

The "sickle cell crisis" topic is particularly interesting, as examination of the associated medications offers insight into numerous aspects of the disease. The list included 5 different synonyms for folic acid, numerous strong pain medications, birth control (standard practice suggests that this is contraindicated), numerous anti-depressants, and Casodex. Casodex is an anti-androgen that is approved for use in the treatment of prostate cancer, but there is one case study in the literature (21) involving two patients that suggests it may prevent priapism in men with sickle cell anemia. All sickle cell patients on Casodex in our study were men. Incidentally, an Internet search for priapism and Casodex results in numerous international drug stores advertising this off label use of the drug and offering it without prescription.

Conclusion

We have presented a mixture model for exploratory analysis of ED visit data. We showed that our model has improved HPI perplexity and DX overlap when compared to related models and that it is possible to build a mixture model for patient data with meaningful components on relatively large patient populations. This is important because with constantly increasing data volume and evolving vocabulary, the computational cost of methods based on inter-patient distance is prohibitive.

We have shown compelling evidence that the text in the HPI section of the medical record contains significant information about patient health that can be modeled without significant preprocessing. However, there is obviously a tremendous amount of information that we have not utilized. Based only on the examples we have shown, it is clear that the incorporation of age, medications and admission rates will lead directly to more homogeneous patient clusters. Beyond this, there is almost certainly a vast amount of data available in both the current and historical medical records of patients. By collecting similar patients into sub-populations with methods that have been extended to incorporate this data, we will be able to identify features of those subpopulations, such as diagnosis and prognosis, that can help inform healthcare decisions on both an individual and a population level.

References

1. Friedman C. A broad-coverage natural language processing system. Proceedings / AMIA Annual Symposium AMIA Symposium. 2000:270-4. Epub 2000/11/18. PubMed PMID: 11079887; PubMed Central PMCID: PMC2243979.
2. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proceedings / AMIA Annual Symposium AMIA Symposium. 2001:17-21. Epub 2002/02/05. PubMed PMID: 11825149; PubMed Central PMCID: PMC2243666.
3. Wang X, Hripcsak G, Markatou M, Friedman C. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. Journal of the American Medical Informatics Association : JAMIA. 2009;16(3):328-37. Epub 2009/03/06. doi: 10.1197/jamia.M3028. PubMed PMID: 19261932; PubMed Central PMCID: PMC2732239.
4. Haerian K, Varn D, Vaidya S, Ena L, Chase HS, Friedman C. Detection of pharmacovigilance-related adverse events using electronic health records and automated methods. Clinical pharmacology and therapeutics. 2012;92(2):228-34. Epub 2012/06/21. doi: 10.1038/clpt.2012.54. PubMed PMID: 22713699.
5. Li L, Chase HS, Patel CO, Friedman C, Weng C. Comparing ICD9-encoded diagnoses and NLP-processed discharge summaries for clinical trials pre-screening: a case study. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2008:404-8. Epub 2008/11/13. PubMed PMID: 18999285; PubMed Central PMCID: PMC2656007.
6. Friedlin J, McDonald CJ. A natural language processing system to extract and code concepts relating to congestive heart failure from chest radiology reports. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2006:269-73. Epub 2007/01/24. PubMed PMID: 17238345; PubMed Central PMCID: PMC1839544.
7. Murff HJ, FitzHenry F, Matheny ME, Gentry N, Kotter KL, Crimin K, et al. Automated identification of postoperative complications within an electronic medical record using natural language processing. JAMA : the journal of the American Medical Association. 2011;306(8):848-55. Epub 2011/08/25. doi: 10.1001/jama.2011.1204. PubMed PMID: 21862746.

8. Sonesson C, Bock D. A review and discussion of prospective statistical surveillance in public health. J Roy Stat Soc a Sta. 2003;166:5-21. doi: Doi 10.1111/1467-985x.00256. PubMed PMID: ISI:000180772400002.
9. Buckeridge DL. Outbreak detection through automated surveillance: A review of the determinants of detection. J Biomed Inform. 2007;40(4):370-9. doi: DOI 10.1016/j.jbi.2006.09.003. PubMed PMID: ISI:000249262100002.
10. Heckerman D. A tractable inference algorithm for diagnosing multiple diseases. Proceedings of Uncertainty in Artificial Intellegence. 1989;5:163-71.
11. Miller RA. Computer-assisted diagnostic decision support: history, challenges, and possible paths forward. Advances in health sciences education : theory and practice. 2009;14 Suppl 1:89-106. Epub 2009/08/13. doi: 10.1007/s10459-009-9186-y. PubMed PMID: 19672686.
12. Bates DW, Evans RS, Murff H, Stetson PD, Pizziferri L, Hripcsak G. Detecting adverse events using information technology. Journal of the American Medical Informatics Association : JAMIA. 2003;10(2):115-28. Epub 2003/02/22. PubMed PMID: 12595401; PubMed Central PMCID: PMC150365.
13. Ruger JP, Richter CJ, Spitznagel EL, Lewis LM. Analysis of costs, length of stay, and utilization of emergency department services by frequent users: implications for health policy. Acad Emerg Med. 2004;11(12):1311-7. Epub 2004/12/04. doi: 10.1197/j.aem.2004.07.008. PubMed PMID: 15576522.
14. Melton GB, Parsons S, Morrison FP, Rothschild AS, Markatou M, Hripcsak G. Inter-patient distance metrics using SNOMED CT defining relationships. J Biomed Inform. 2006;39(6):697-705. Epub 2006/03/24. doi: 10.1016/j.jbi.2006.01.004. PubMed PMID: 16554186.
15. Cao H, Melton GB, Markatou M, Hripcsak G. Use abstracted patient-specific features to assist an information-theoretic measurement to assess similarity between medical cases. J Biomed Inform. 2008;41(6):882-8. Epub 2008/05/20. doi: 10.1016/j.jbi.2008.03.006. PubMed PMID: 18487093; PubMed Central PMCID: PMC2584163.
16. Pitman J. Combinatorial stochastic processes - Saint-Flour Summer School of Probabilities XXXII - 2002. Lect Notes Math. 2006;1875:1-+. PubMed PMID: ISI:000241856100001.
17. Blei DM, Frazier PI. Distance Dependent Chinese Restaurant Processes. J Mach Learn Res. 2011;12:2461-88. PubMed PMID: ISI:000298102200001.
18. Blei DM, McAuliffe J. Supervised Topic Models. 2008. In: Advances in Neural Information Processing Systems 20 [Internet]. MIT Press; [121-8].
19. Lacoste-Julien S, Sha F, Jordan MI. DiscLDA: Discriminative Learning for Dimensionality Reduction and Classification. In: Koller D, Schuurmans D, Bengio Y, Bottou L, editors. Advances in Neural Information Processing Systems: MIT Press; 2009. p. 897-904.
20. Jelinek F, Mercer RL, Bahl LR, Baker JK. Perplexity - a measure of the difficulty of speech recognition tasks. J Acoust Soc Am. 2977;62(1).
21. Dahm P, Rao DS, Donatucci CF. Antiandrogens in the treatment of priapism. Urology. 2002;59(1):138. Epub 2002/01/18. PubMed PMID: 11796309.