An Ensemble Topic Model for Sharing Healthcare Data and Predicting Disease Risk

Andrew K. Rider
Department of Computer Science & Engineering
Interdisciplinary Center for Network Science & Applications
University of Notre Dame
Notre Dame, IN, USA
arider1@nd.edu

Nitesh V. Chawla[*]
Department of Computer Science & Engineering
Interdisciplinary Center for Network Science & Applications
University of Notre Dame
Notre Dame, IN, USA
nvchawla@nd.edu

ABSTRACT

With the recent signing of the Affordable Care Act into law, the use of electronic medical data is set to become ubiquitous in the United States. This presents an unprecedented opportunity to use population health data for the benefit of patient-centered outcomes. However, there are two major hurdles to utilizing this wealth of data. First, medical data is not centrally located but is often divided across hospital systems, health exchanges, and physician practices. Second, sharing specific or identifiable information may not be allowed. Moreover, organizations may have a vested interest in keeping their data sets private as they may have been gathered and curated at great cost. We develop an approach to allow the sharing of beneficial information while staying within the bounds of data privacy. We show that the use of a probabilistic graphical model can facilitate effective transfer learning between distinct healthcare data sets by parameter sharing while simultaneously allowing us to construct a network for interpretation use by domain experts and the discovery of disease relationships. Our method utilizes aggregate information from distinct populations to improve the estimation of patient disease risk.

Categories and Subject Descriptors

J.3 [**Life and Medical Sciences**]: Health—*Medical information systems*; E.1 [**Data Structures**]: Graphs and networks

Keywords

Population Health; Patient-Centered Outcomes; Big Data

1. INTRODUCTION

The growing and mandated use of electronic medical records will allow scientists to unveil new discoveries about human health. An enormous quantity of healthcare data is created every year and there are vast amounts of past medical records that are being imported into electronic formats. A Center for Disease Control study estimates that patients in the United States made 1.2 billion visits or an average of 4.05 visits per patient to physicians' offices in 2007 [11]. The rate of visitation increased 11 percent since 1997.

Aspirin is one exemplary case of the utilization of available medical data. Multiple studies have found that Aspirin reduces the long term risk of colorectal cancer, the progression of cardiovascular disease and the likelihood of stroke [1, 2]. These studies relied on the wealth of data already available about Aspirin. Similarly, we hope to utilize existing electronic medical records (EMRs) to discover relationships between diseases and to improve disease risk prediction for patients.

While recently enacted healthcare laws mandate the use and utilization of EMRs, they do not specify how they should be stored or who should store and maintain the records in sufficient detail. At present this lack of centralization is impeding the meaningful use of this data. Each hospital and medical research organization may have their own data set, which they are compelled by law not to share due to privacy concerns. Regional data warehousing organizations have arisen to consolidate and store EMRs, but they are equally subject to restrictions on sharing the data. Additionally, EMRs have become a commodity, as the maintenance and security of the storage systems can be costly. Therefore, organizations may have incentive to protect their data sets.

Sharing complete EMRs would be the best means of promoting beneficial and meaningful use but there are obstacles to full disclosure of the data. However, the Health Insurance Portability and Accountability Act of 1996 stipulates that aggregated information can be shared freely [7]. We propose an approach that allows aggregated information to be shared between distinct organizations with EMR data in a way that increases the accuracy of computational prediction of disease risk. We utilize an ensemble approach to gain more predictive accuracy with little information. We posit that this approach is mutually beneficial to all organizations warehousing EMRs and maintains the privacy of the patients while protecting any potential interests in keeping valuable data sets private.

We further propose the use of learned parameters of topic models as an alternative approach to creating interpretable network models from EMR data. Networks have been an

[*]Corresponding Author

BCB '13, September 22 - 25, 2013, Washington, DC, USA
Copyright 2013 ACM 978-1-4503-2434-2/13/09 ...$15.00.

intuitive and useful approach to modeling complex data and presenting a representation that domain experts can understand. Perhaps the most closely related work in the healthcare domain utilizes collaborative filtering to create a disease-gene network based on some similarity criterion between diseases [4]. While network models can be very effective at identifying disease risk, many network approaches utilize different edge weighting methods, which may lead to different interpretations of the data [14]. Furthermore, many approaches to integrating distinct networks are computationally intractable. We view the network approach as a bottom-up construction of a relational model by inspecting individual health records. We propose a top-down topic modeling approach that begins with a partitioning function we wish to optimize on the data. By creating a topic model that explicitly measures disease co-occurrence we simultaneously learn the network that best models the data according to our criterion and partition it into meaningful groups with co-occurring diseases. The use of this approach simultaneously creates an interpretable model and allows easily computed solutions for combining information that creates a network. To this end we propose a novel extension to a well known probabilistic graphical model that optimizes the grouping of medical records on occurrence and co-occurrence of disease.

This is to our knowledge the first use of topic models to infer network structure and the first application to EMR data. However, topic models have been used to identify topics in medical documents and public health topics in Twitter [17, 10, 6]. Ensemble topic models have also been studied, although not in this context [15].

This study makes three contributions to the medical domain:

- A novel application of topic modeling to the analysis of disease risk.
- A novel topic modeling approach for the study of relational data.
- We support our proposition that distinct EMR warehousing organizations should share general information with evidence that it can improve the utility of disease risk prediction across individual data sets.

2. THE MODEL

Our approach extends the well studied Dirichlet Process Mixture Model (DPMM), which is depicted in Figure 1. DPMM is a non-parametric approach that learns an unspecified number of groups with distinct distributions over features.

2.1 The Dirichlet Process

The Dirichlet distribution is the multivariate generalization of the beta distribution. The Dirichlet process is an infinite dimensional generalization of the Dirichlet distribution. One formulation of the Dirichlet process is described in the stick-breaking process. Imagine that a stick is broken repeatedly such that the first section has a length dependent on the Beta distribution: $\beta_1' \sim Beta(1, \alpha)$. The remainder of the stick is broken in the same way such that $\beta_k = \beta_k' * \prod_{i=1}^{k-1}(1 - \beta_i')$. It has been shown that if $G \sim DP(\alpha_0, G_0)$,

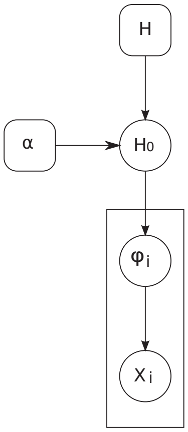

Figure 1: The Dirichlet Process Mixture Model. H is a base distribution from which weights are drawn as described in Equation 1. The mixing proportions of the components are specified by H_0. The parameters of the base distributions are specified by ϕ_i. X_i represents an observed instance.

$$G = \sum_{k=1}^{\infty} \beta_k \gamma_{\phi_k} \tag{1}$$

Where β_k are stick-breaking weights depending on the parameter α_0, γ_{ϕ_k} is an atom at ϕ_k, each representing an independent random variable [12]. Using this stick-breaking approach, any measure can be used to determine a set of discrete weights. DPs are often used to set priors for components of mixture models [13].

2.2 Dirichlet Process Mixture Model

We focus here on the use of multinomial base distributions as the multinomial is appropriate for the measurement of binary and count data.

The model is described by the hierarchical specification:

$$\begin{aligned} H_0 &\sim DP(H, \alpha) \\ \phi_i | H_0 &\sim H \\ x_i | \phi_i &\sim F(x_i, \phi_i) \end{aligned} \tag{2}$$

In the standard DPMM, F is the multinomial probability mass function. In the context of EMRs, the DPMM with a multinomial base distribution over disease occurrence optimizes for groups of patients that have received the same diagnoses.

2.3 Our approach

We propose an alternative formulation of a DPMM in which F is a function of both the multinomial over the diseases and a second multinomial over disease co-occurrence. This model allows us to construct a network representation of the data by utilizing co-occurrence counts explicitly. It is inspired partly by the effectiveness and generality of gaussian mixtures. A multivariate Gaussian is parameterized by a mean vector and a covariance matrix. The covariance matrix specifies not only the spread of values around the mean but the relationship between features. This is much more specific information than is captured by a multinomial.

However, inferring the parameters of a multivariate Gaussian can be much more complex than inferring the parameters of a multinomial. Furthermore, a multivariate gaussian is not appropriate for binary values as a Gaussian distribution only accurately models a set of binary values in the edge cases where all values are 1 or 0.

Our approach finds a balance between parsimony and specificness by placing equal weight on the co-occurrence of diseases and the presence of disease. We call our approach Co-occurrence Based Clustering (CBC) for its focus on explicitly learning the co-occurrence of diseases. In CBC, the multinomial over the diseases is analogous to the mean vector of a multivariate Gaussian. The multinomial over the disease co-occurrences is analogous to the covariance between each pair of diseases. This formulation allows CBC to capture co-occurrence explicitly while maintaining generality. This is essential as patients may be lacking multiple appropriate diagnoses The model is learned by Gibbs sampling in which the likelihood function gives equal weight to the two multinomials, as shown in Equation 3.

$$F(X_i, {X_i}', \phi, \phi') = \frac{\frac{[\sum_i^k X_i]!}{\prod_i^k X_i!} \prod_i^k \phi_i^{X_i} + \frac{[\sum_i^k X_i']!}{\prod_i^k X_i'!} \prod_i^k {\phi'}_i^{X_i'}}{2} \quad (3)$$

Where X is the matrix of diagnoses for all patients, ϕ is the matrix of disease occurrence parameters for each component, ϕ_i is the probability of observing a disease (alternatively $\phi_i = X_i / \sum X_i$), and X' and ϕ' are the analogous parameters for disease **co**-occurrence.

Relationships between diseases may not be apparent from examining their frequency separately. The use of a relational representation of the data —the co-occurrence of diseases— allows even simple models to take into account this more specific information.

While DPMM can be applied to the co-occurrence of diseases, the reliance on co-occurrence alone can undermine the generality of the model. If a patient has a disease that has not been diagnosed, then all 252 (in our data) potential co-occurrences will be missing, whereas in the flat occurrence representation, only a single value will be missing. Thus a little noise can have an overwhelming effect on the model. CBC is more tolerant to this source of noise by virtue of considering both co-occurrence and frequency.

We demonstrate these differences by comparing disease ranking results across three formulations of DPMMs: DPMM is a DPMM trained on disease occurrence data; COOC is a DPMM trained on co-occurrence data, and CBC is our model utilizing both the representations of data.

2.4 Markov Chain Monte Carlo inference

We utilize a version of Gibbs sampling with auxiliary parameters [8]. This approach allows us to sample the component membership of the model without having to integrate with respect to the prior distribution H. Algorithm 1 describes the steps in our sampler.

In our experiments we used the parameters $m = 1$ and $\alpha = .01$. The algorithm specifically describes the sampler for CBC, but the samplers for DPMM and COOC are the same with the exceptions that DPMM uses ϕ_c and y_i exclusively and COOC uses ${\phi'}_c$ and ${y'}_i$ exclusively.

2.5 Ensemble learning

The goal of ensemble learning in this context is to allow models to achieve performance increases through the use of data from distinct sources. Examples include data from different domains such as healthcare and genomics, data with different distributions such as healthcare data from different ethnic or socioeconomic groups, or even data with different feature spaces if for example there are no occurrences of a disease in one group that is present in another group. We utilize an approach that is common in transfer learning, known as parameter passing [9]. In this approach base models are learned on distinct data sets. The base models are then combined in a separate step by joining the parameters of the models. Wang et al. propose a similar approach in which different topic models are combined by running an additional clustering step on the component *labels* from base topic models [16].

Figure 2 outlines the process used to create ensembles. We build ensembles by training base models on each demographic data set. The ensemble step combines the occurrence parameters ϕ of the base models into a single matrix. A DPMM is trained on this matrix to form a ensemble-level model. Disease risk is assessed for an individual patient by first finding the component of this ensemble model that best fits their disease profile, then combining the parameters of every component from the base models whose parameters are in the ensemble-level component. The base-model parameters are averaged to form the parameters of a consensus model.

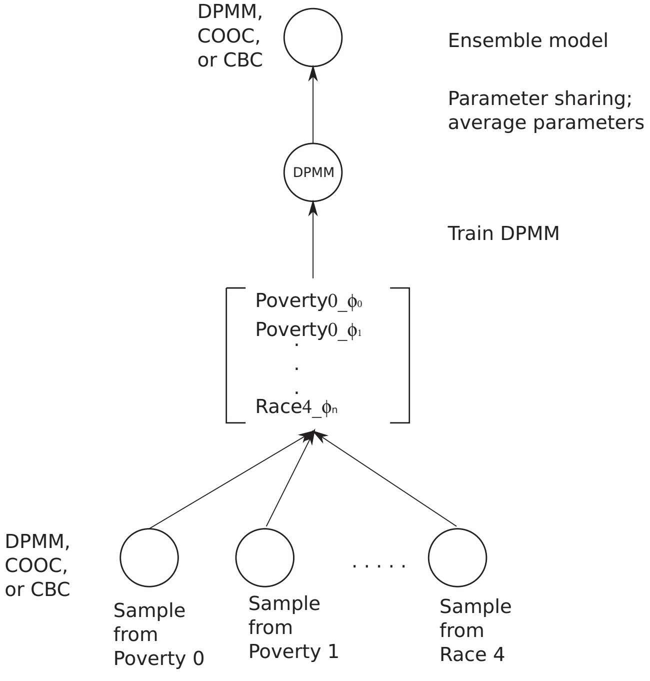

Figure 2: The ensemble takes base models, each consisting of the results from a single model trained on one data set, then combines the parameters of each component from the base mixture models into a single matrix.

3. DATA

Our deidentified and anonymous dataset contains data of 7,895,283 individuals with three or more diagnoses. The raw data set contains ICD-9-CM codes for describing the diag-

Algorithm 1: Gibbs sampler with auxiliary parameters

```
1: For $i = 1, ..., n$: Let $k^-$ be the number of components $c_j$ for $j \neq i$, and $h = k^- + m$, where $m$ is the number of auxiliary variables. If $c_i = c_j$ for some $j \neq i$, draw values independently from the base distribution $H$ for the parameters of components $\phi_c$ (and correspondingly ${\phi'}_c$) for which $k^- < c \leq h$. If $c_i \neq c_j$ for all $j \neq i$, let $c_i$ have the label $k^- + 1$, and draw values from H for those $\phi_c$ for which $k^- + 1 < c \leq h$. Draw a new value for $c_i$ from $1, ..., h$ with the following probabilities:

   $P(c_i = c | c_{-i}, y_i, {y'}_i, \phi_{-c}, {\phi'}_{-c}) = \begin{cases} b\frac{n_{-i,c}}{n-1+\alpha} F(y_i, {y'}_i, \phi_c, \phi'_c) & for\, 1 \leq c \leq k^- \\ b\frac{\alpha/m}{n-1=\alpha} F(y_i, {y'}_i, \phi_c, \phi'_c) & for\, k^- < c \leq h \end{cases}$   (4)

   where $y_i$ is an instance, ${y'}_i$ is the corresponding set of co-occurrences, $n_{-i,c}$ is the number of $c_j$ for $j \neq i$ that are equal to $c$, and $b$ is a normalizing constant. Remove $\phi_c$ that are not associated with at least one observation.
2: For all $c \in c_1, ..., c_n$: draw a new value from $\phi | y_i$ such that $c_i = c$.
```

noses that apply to each patient. ICD-9-CM codes exist in a hierarchy of disease that can complicate analysis [5]. Collapsed ICD-9-CM codes provide a mapping from specific diagnoses to general diagnoses. For example, ICD-9-CM codes "9843" and "9845" correspond to pneumonia from whooping cough and pneumonia from anthrax, respectively. Both can be described by the shortened ICD-9-CM code "984" or by the Clinical Classifications Software (CCS) code "122". CCS codes provide a standardized coding system based on the ICD-9 specification and is designed to be clinically meaningful and more useful for statistical analysis. Therefore, we utilize the CCS codes to provide a more general non-hierarchical classification of disease than ICD-9-CM codes [3].

We approach our goal of demonstrating the effectiveness of learning across distinct healthcare data sets by splitting the data into distinct populations by demographics based on poverty level, gender, race, and age. Table 1 shows that these groups tend to have similar numbers of disease diagnoses. The most significant difference appears to exist between the two poverty groups, in which the variance and kurtosis are strikingly different. This indicates that there is a wider range of number of diagnoses among these two groups.

Individual disease prevalence is much more strikingly different between these groups. The top 20 most common diseases in the original data set are listed in Table 2. A patient who is in demographic "poverty 1" is twice as likely to be diagnosed with a cognitive disorder as a patient on the other side of the poverty line. A patient of gender 0 is nearly three times as likely to suffer from genitourinary symptoms as a patient of gender 1. There are many additional differences between races and other demographics that demonstrate the distinctions between these populations. We were surprised to find that the disease prevalence in the 10% youngest patients in the data set was very similar to the disease prevalence in the 10% oldest patients. However, this may be explained by the fact that the distribution of patient age is strongly skewed towards the younger patients and that the data set consists entirely of patients with at least 65 years of age. Given the lack of interesting differences between the age groups, we focused on the poverty, gender, and race demographics for our analysis. Among the race demographics, only 8,075 patients were of Race 4, providing a relatively small sample. As such, Race 4 was not used for analysis.

4. EVALUATION

We trained DPMM, COOC, and CBC on fifty random samples of 4,500 instances from each poverty, race, and gender dataset. Models trained on a single sample from each demographic data set were combined as described in Section 2.5. Ensembles of CBC, DPMM, and COOC models were constructed in the same way. Gibbs sampling was carried out for 1000 iterations for each base model and for the ensemble models.

The accuracy of disease risk was measured by holding out a test set of instances of 500 patients from each demographic and calculating the likelihood of each disease given all but one of the observed diseases in the test instance. This was repeated by withholding each observed disease for every test instance. A test set of 500 instances with an average of 5 diseases per patient would result in 2500 individual rankings.

The ranking of the diseases was evaluated as in previous work by calculating the proportion of predicted disease rankings that were in the top ranks [5]. The lower ranks are more important as a medical professional reviewing a list of predicted diseases is much more likely to read predictions early in the list.

5. RESULTS

The log-likelihood plot in Figure 3 shows that the likelihood of the algorithms appear to be in stable states after 1000 iterations on the gender demographic.

The proportion of missing diseases that were ranked as disease risks for patients is shown in Figure 4. The proportions were determined by averaging across the ranking for patients in test sets from all demographics and all experiments. The ensemble of CBC models provides the best rankings in the highest ranks, with the most accurate predictions for 8 of the first 10 ranks, 18 of the first 20 ranks, 27 of the first 50 ranks. This approach ranks the missing disease from a patient's diagnosis in the first 10 listed diseases 47.5% of the time and the first 20 ranks 76.5% of the time. The ensemble of DPMM models performs best when considering ranks greater than 29. CBC is expected to perform better at lower ranks as the components in CBC utilize more specific co-occurrence data. All of the methods place nearly 100% of the missing diseases in the first 50 ranks. Notably, base-CBC performs next to worst, whereas the ensemble-CBC performs the best. This indicates that the base CBC models are diverse; they capture differences in the separate

Table 1: Descriptive statistics for the number of diseases patients suffer from in each demographic.

	All	Poverty 0	Poverty 1	Gender 0	Gender 1	Race 0	Race 1	Race 2	Race 3	Race 4	Race 5	Age 0	Age 9
# of patients	7895283	7050861	844421	3247168	4648114	83500	7050861	652158	79181	8075	18437	789528	789528
Maximum	66	66	63	66	57	47	65	51	51	39	43	66	51
Mean	9.25	9.12	10.36	9.18	9.30	8.77	8.21	8.69	8.28	7.12	7.93	8.88	8.85
Variance	20.87	19.97	26.98	20.31	21.25	23.00	20.47	24.55	22.04	14.43	20.21	20.33	19.99
Skewness	1.61	1.63	1.38	1.62	1.60	1.43	1.62	1.51	1.66	1.96	1.81	1.87	1.85
Kurtosis	3.31	3.42	2.34	3.34	3.28	2.43	3.36	2.75	3.60	5.31	4.35	4.71	4.52

Table 2: Percentage of patients with top 20 most prevalent disease by demographic.

	All	Poverty 0	Poverty 1	Gender 0	Gender 1	Race 0	Race 1	Race 2	Race 3	Race 4	Race 5	Age 0	Age 9
Essential Hypertension	43.78	43.88	42.92	39.10	47.04	33.34	36.28	46.10	36.88	37.32	39.34	45.93	46.77
Fluid and electrolyte disorders	39.48	38.18	50.30	34.19	43.18	38.09	31.64	39.29	35.85	29.91	30.69	33.92	34.14
Coronary atherosclerosis	38.69	38.99	36.16	43.39	35.40	34.57	37.07	29.97	31.21	31.45	35.77	39.91	39.45
Cardiac dysrhythmias	32.74	32.94	31.02	35.98	30.47	34.03	32.53	25.93	27.38	25.67	22.99	26.47	28.07
Congestive Heart Failure	27.84	27.09	34.09	27.50	28.08	31.77	25.27	26.11	25.01	17.15	21.12	21.91	21.32
Urinary tract infections	26.58	25.17	38.38	18.16	32.46	24.72	21.03	26.11	21.29	17.23	20.41	18.67	19.39
Bronchitis	25.42	25.13	27.83	31.68	21.05	25.81	25.48	18.73	19.85	16.27	20.91	28.31	27.42
Anemia	20.72	20.25	24.62	18.88	22.00	17.15	14.08	22.18	17.51	15.30	15.18	17.44	17.05
Diabetes w/o complication	18.80	18.31	22.88	18.74	18.84	14.15	13.76	20.57	19.74	17.36	23.29	23.66	22.29
Pneumonia	18.40	17.50	25.90	19.98	17.29	21.32	16.42	15.96	19.92	14.19	15.60	14.62	14.79
Surgical complications	16.67	17.22	12.15	19.70	14.56	13.66	16.50	12.63	15.61	15.62	14.02	19.58	20.37
Osteoarthritis	14.61	14.46	15.82	10.30	17.61	10.45	11.52	10.61	6.67	6.29	8.85	10.83	12.15
Bacterial infection	13.44	12.68	19.76	10.26	15.66	14.01	12.49	13.63	13.06	10.86	13.36	10.73	10.66
Heart valve disorders	12.41	12.70	09.97	11.73	12.88	12.68	12.29	10.33	10.23	9.79	8.77	10.01	10.54
Cerebrovascular disease	11.80	11.38	15.24	11.77	11.81	12.82	10.49	14.37	14.03	13.60	9.76	8.46	9.30
Pneumothorax	11.60	11.57	11.85	12.10	11.25	12.77	11.18	10.51	11.28	9.04	10.34	11.34	11.62
Genitourinary symptoms	11.51	11.56	11.08	17.99	6.98	12.03	11.31	11.36	11.51	11.41	12.24	9.34	10.10
Cystic fibrosis	11.07	11.14	10.54	10.85	11.23	9.52	10.17	7.95	10.67	10.41	7.76	9.64	10.48
Cognitive disorders	10.89	9.76	20.39	8.89	12.29	14.48	10.54	13.16	8.68	6.83	7.43	3.23	3.68
Gastrointestinal hemorrhage	10.69	10.49	12.33	11.07	10.42	8.63	7.33	8.21	9.60	9.18	6.54	8.72	8.99

demographics.

The nearest comparison to this study utilizes collaborative filtering on collapsed ICD-9 diagnoses to rank the likelihood of diagnoses in the last visit based on patients' medical history [5]. Where that method identifies 54.7% of future diagnoses in the top 20 ranks, our approach identifies 76.5% of held out diagnoses. While these methods utilize the same data, it is important to note that the approach of Davis et al. relies on temporal data and collapsed ICD-9 codes instead of CCS codes, making direct comparison problematic.

Different patients may have very different histories of diagnosis. The specific diagnoses that an individual patient has may be more or less predictive than others. Figure 5 shows the relationship between the number of diagnoses and the mean rank of diagnoses for individual patients. As expected, the variance in the accuracy of the model decreases sharply as the number of available diagnoses increases.

6. INTERPRETING THE MODEL

In addition to ranking individual patient disease risks, we are interested in creating a global model of disease relationships. Figure 6 shows a network constructed from one ensemble CBC model. Nodes represent diseases. Their groups (signified by color) are determined by the component in the ensemble that contains the most patients with the given diagnosis. Edge weight was determined by averaging disease co-occurrence across all components. Therefore, edges may tend to represent *global* co-occurrences rather than within component co-occurrences. The figure shows 64 distinct disease groups, determined by the number of patients in the base model components contributing to each ensemble component. Many of these groups contain diseases which are clearly similar. For example, the yellow group in the top row and fifth from the left, contains 9 cancer diagnoses. The remaining three are "gastritis and duodenitis," "intestinal infection," and perhaps oddly, "deficiency and other anemia." Other clusters appear less specific to the layman, but still contain common sense groups. For example, the larger red group, bottom left and three from the left, contains 9 pregnancy related diagnoses and 4 abdominal pain related diagnoses. Edges in the network represent the mean edge strength from across all components in the ensemble. We used the 99th percentile edges to form this network. These weights represent strong relationships that are not strong enough to determine component membership alone. Some of the strongest edges join two groups, the pink group third from the right in the middle row, and the red group third from the left in the third row. These join chemotherapy related issues, dizziness or vertigo, nervous system anomalies, unspecified circulatory disease, and unspecified eye disorders. A thorough analysis of this network requires medical expertise, however it is clear that there is meaningful structure to be investigated here. This model is provided with CCS designations as a Cytoscape file at http://www.cse.nd.edu/∼arider1/cbc_meta_meanedge.cys.

7. DISCUSSION

We set out with the goal to provide an approach that would allow and encourage EMR warehousing organizations and research centers to share EMR data for their mutual benefit and the benefit of patients. Our analysis demonstrates that the proposed use of aggregate data improves ranking across diverse patient populations. Therefore we strongly recommend that EMR warehousing organizations share this aggregate data both as an act of good will and as an act of self interest, as more available data will improve modeling on individual data sets.

We additionally sought to provide a means to create an

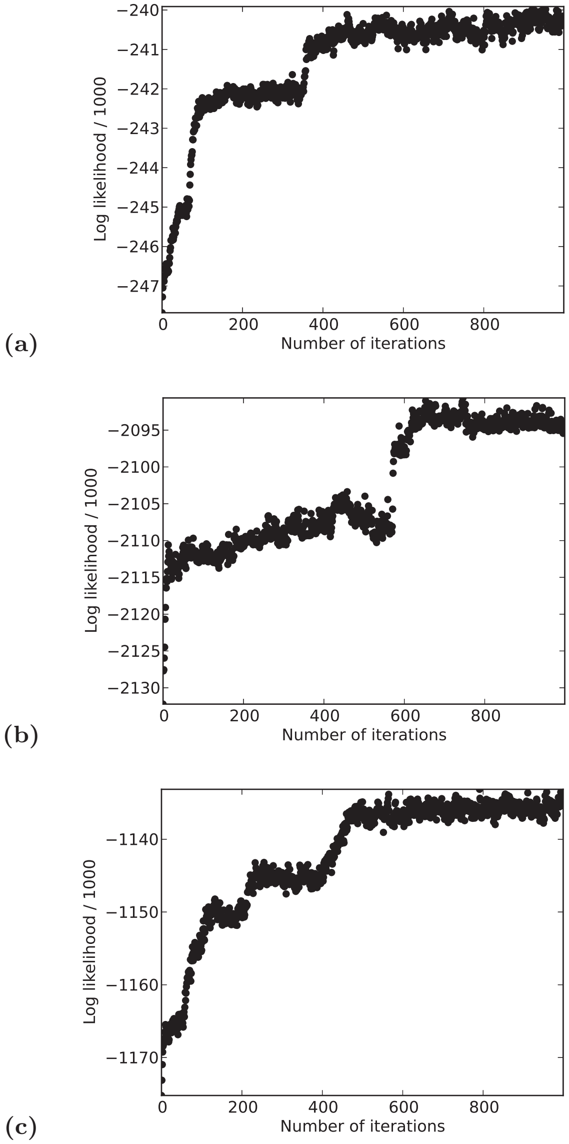

Figure 3: The log-likelihood of all three algorithms over 1000 iterations of Gibbs sampling. All log-likelihood values are divided by 1000 for readability. Panel **(a)** shows the log likelihood of DPMM. Panel **(a)** shows the log likelihood of DPMM. Panel **(b)** shows the log likelihood of DPMM using the co-occurrence data. Panel **(c)** shows the log likelihood of CBC.

interpretable model from disparate aggregate data. We proposed a method that explicitly utilizes co-occurence data to learn a network while simultaneously providing imroved disease risk predictions. We demonstrated that the network constructed contains comprehensible groupings of disease occurrence, based both on the component labels in our model and on the global mean edge weights used to construct the network. Although it can be exceedingly difficult to quantify the utility of a network model, the provided examples do indicate that this model may contain useful medical information.

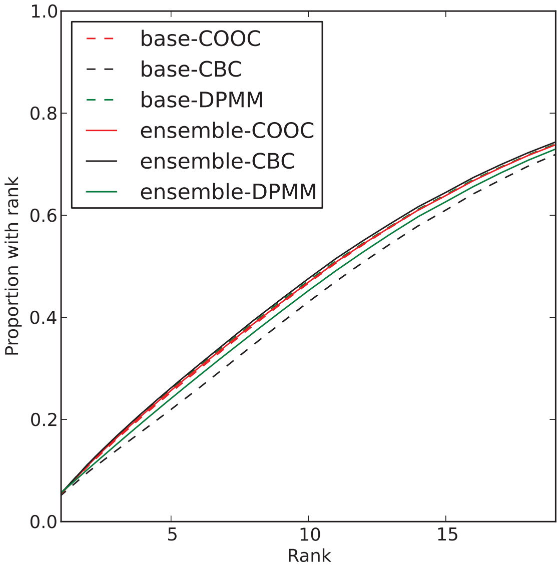

Figure 4: The proportion of held-out diseases given rank less than or equal to the value on the x-axis. Labels “base” and “ensemble” correspond to ranks given by the algorithms trained on a single demographic data set and ranks given by the ensemble across demographics.

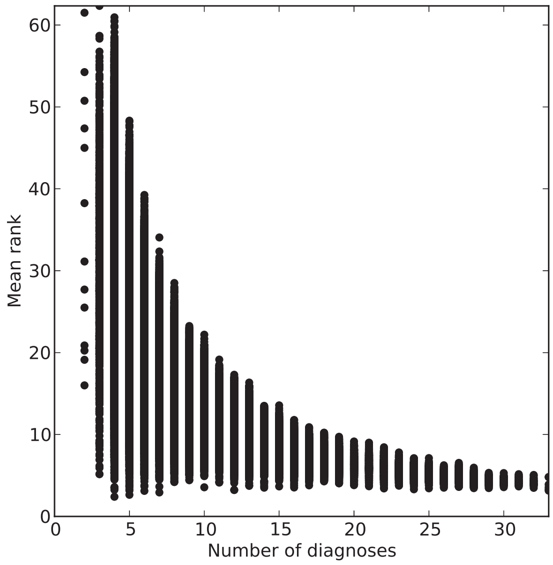

Figure 5: Mean rank for individual patient diagnoses versus the number of diagnoses available based on an ensemble of CBC models.

8. FUTURE WORK

The model described in this work extends DPMM to utilize specific co-occurrence data in addition to the normal occurrence data representation. While the approach does systematically improve results, it comes at a cost of increased computational complexity. However, the increased computational burden is distributed across numerous distinct computational resources in our proposed use case of sharing aggregate EMR data across distinct groups. The computational cost may be further reduced by utilizing a similar mixture modeling approach that relies on a conjugate prior distribution for the co-occurrence aspect of the model.

Figure 7 shows a spring-layout view of the same network. This figure shows that the clustering as determined by edge weight is also informative. The five of the six nodes in the group nodes on the rightmost side of the central cluster concern birth related diagnoses. The group of four nodes at the bottom most edge of the central cluster contain "OB-related trauma to perineum and vulva," "fetal distress and abnormal forces of labor," "Cardiac and circulatory congenital anomalies," and "acquired foot deformities." This layout additionally highlights two hubs, "disorders of lipid metabolism" and "coma; stupor; and brain damage."

We refrain from making a full enumeration of interesting clusters, but we have found that various edge weight based layouts provide additional clusters that seem to make sense. We encourage the reader to investigate these clusters by downloading the provided network.

9. REFERENCES

[1] A. M. Algra and P. M. Rothwell. Effects of regular aspirin on long-term cancer incidence and metastasis: a systematic comparison of evidence from observational studies versus randomised trials. *The lancet oncology*, 2012.

[2] C. Baigent, L. Blackwell, R. Collins, J. Emberson, J. Godwin, R. Peto, J. Buring, C. Hennekens, P. Kearney, T. Meade, et al. Aspirin in the primary and secondary prevention of vascular disease: collaborative meta-analysis of individual participant data from randomised trials. *Lancet*, 373(9678):1849, 2009.

[3] M. E. Cowen, D. J. Dusseau, B. G. Toth, C. Guisinger, M. W. Zodet, and Y. Shyr. Casemix adjustment of managed care claims data using the clinical classification for health policy research method. *Medical care*, 36(7):1108–1113, 1998.

[4] D. A. Davis and N. V. Chawla. Exploring and Exploiting Disease Interactions from Multi-Relational Gene and Phenotype Networks. *PLoS ONE*, 6(7):e22670, July 2011.

[5] D. A. Davis, N. V. Chawla, N. A. Christakis, and A.-L. Barabási. Time to care: a collaborative engine for practical disease prediction. *Data Mining and Knowledge Discovery*, 20(3):388–415, 2010.

[6] E. Diaz-Aviles, A. Stewart, E. Velasco, K. Denecke, and W. Nejdl. Epidemic intelligence for the crowd, by the crowd (full version). *arXiv preprint arXiv:1203.1378*, 2012.

[7] U. S. Government. Health insurance portability and accountability act. *45 CFR 164.514*, 1996.

[8] R. r. M. Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265, 2000.

[9] S. J. Pan and Q. Yang. A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, 22(10):1345–1359, 2010.

[10] M. J. Paul and M. Dredze. You are what you tweet: Analyzing twitter for public health. In *Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 2011)*, 2011.

[11] S. Schappert and E. Rechtsteiner. Ambulatory medical care utilization estimates for 2007. *Vital and Health Statistics. Series 13, Data from the National Health Survey*, (169):1, 2011.

[12] J. Sethuraman. A constructive definition of dirichlet priors. Technical report, DTIC Document, 1991.

[13] B. Shahbaba and R. Neal. Nonlinear models using dirichlet process mixtures. *The Journal of Machine Learning Research*, 10:1829–1850, 2009.

[14] R. Steuer, J. Kurths, C. O. Daub, J. Weise, and J. Selbig. The mutual information: Detecting and evaluating dependencies between variables. *Bioinformatics*, 18(suppl 2):S231–S240, Oct. 2002.

[15] S. Tang, Y.-T. Zheng, G. Cao, Y.-D. Zhang, and J.-T. Li. Ensemble learning with lda topic models for visual concept detection, multimedia - a multidisciplinary approach to complex issues.

[16] H. Wang, H. Shan, and A. Banerjee. Bayesian cluster ensembles. *Statistical Analysis and Data Mining*, 4(1):54–70, 2011.

[17] L. Yang, Q. Mei, K. Zheng, and D. A. Hanauer. Query log analysis of an electronic health record search engine. In *AMIA Annual Symposium Proceedings*, volume 2011, page 915. American Medical Informatics Association, 2011.

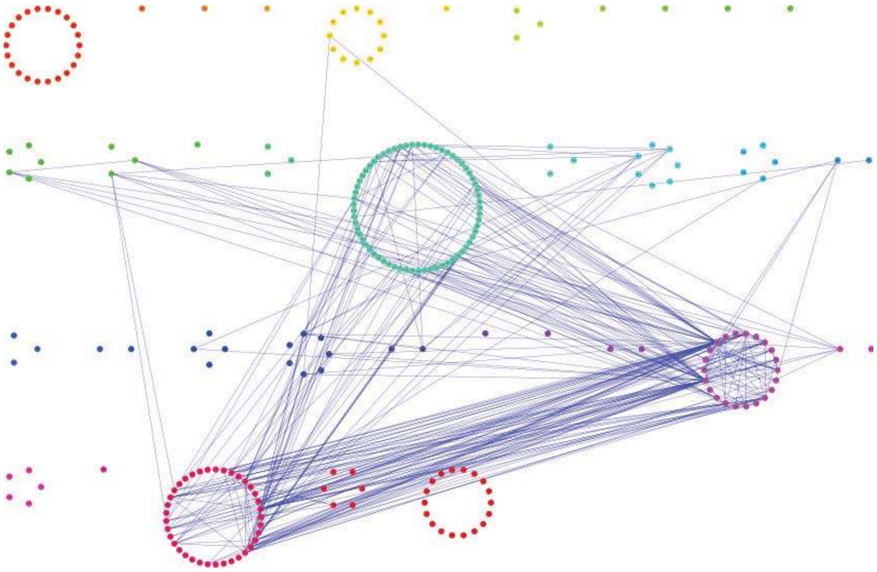

Figure 6: A network constructed from a single CBC ensemble. Nodes represent disease diagnoses and edges represent co-occurrences. Node groups are determined by the component in the model with the most diagnoses. Edge weight was determined by averaging disease co-occurrence across all components. Therefore, edges may tend to represent *global* co-occurrences rather than within component co-occurrences. Only edges in the 99th percentile weight category are shown.

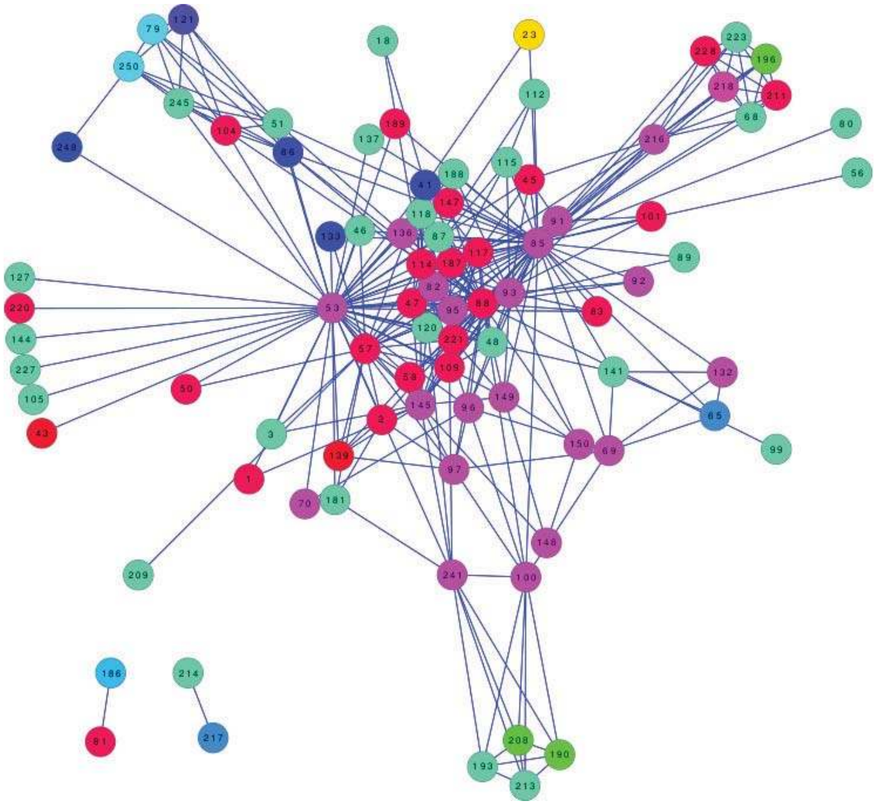

Figure 7: Another view of Figure 6 using the spring-layout based on edge weight. Edge weight was determined by averaging disease co-occurrence across all components and reflects global trends. This view reveals clusters and hubs in the network.