HEMnet: Integration of Electronic Medical Records with Molecular Interaction Networks and Domain Knowledge for Survival Analysis

Edward W Huang
University of Illinois at Urbana-Champaign
ewhuang3@illinois.edu

ABSTRACT
The continual growth of electronic medical record (EMR) databases has paved the way for many data mining applications, including the discovery of novel disease-drug associations and the prediction of patient survival rates. However, these tasks are hindered because EMRs are usually segmented or incomplete. EMR analysis is further limited by the overabundance of medical term synonyms and morphologies, which causes existing techniques to mismatch records containing semantically similar but lexically distinct terms. Current solutions fill in missing values with techniques such as mean imputation, which tends to introduce noise rather than reduce it. In this paper, we propose to simultaneously infer missing data and solve semantic mismatching in EMRs by incorporating molecular interaction networks and domain knowledge to build HEMnet, a heterogeneous medical information network. We project this network onto a low-dimensional space, then group entities in the network according to their relative distances in this new space. We evaluate the effectiveness of our method according to its ability to separate patients with dissimilar survival functions. We show that our method can obtain significant (p-value < 0.01) results for each cancer subtype in a lung cancer dataset, while the baseline methods cannot.

CCS CONCEPTS
• Information systems → Data analytics; • Applied computing → Health informatics;

KEYWORDS
Heterogeneous Information Networks, electronic medical records, survival analysis

1 INTRODUCTION
The Obama administration made electronic medical records (EMRs) a focus of the American Recovery and Reinvestment Act, which was a high-priority initiative that devoted significant resources to improving EMR adoption rate in healthcare practices across the United States [1]. In fact, EMR adoption has been estimated to eventually save more than $81 billion annually [14]. EMRs can also improve patient safety by alerting doctors to potentially harmful interactions that may result from prescriptions of multiple drugs [3]. Furthermore, with more widespread EMR databases, doctors can apply stronger statistical methods to accomplish previously infeasible tasks, such as relationship mining and clinical prediction of survival (CPS).

Relationship mining allows doctors to discover useful associations among entities in medical records, including novel usages of certain drugs and adverse drug events [13]. On the other hand, CPS enables the prediction of a new patient’s probability of survival based on his or her medical records. Successful prediction would help hospitals optimize resource allocation and treatment planning. For example, accurate survival estimates for terminally ill patients can prevent inappropriate therapies and avoid unnecessary toxicity [12].

Analyses involving EMRs, including both of the aforementioned applications, utilize patient features, which include test results, clinical notes, symptoms, diagnoses, and medical history. However, there are two main challenges that frequently appear with these features:

1) **Missing data:** many methods assume the availability of all features. However, this assumption typically does not hold for many EMR databases. One reason is that doctors do not perform all existing medical tests on all patients. Other reasons include incomplete medical records or inconsistent text data in the form of doctor notes. Mean imputation, the most common method to fill missing values, actually introduces noise rather than reduce it [25].

2) **Semantic mismatch:** patients with similar but distinct features may be judged to be dissimilar. In the traditional vector space model, in which each unique word in the vocabulary occupies a dimension, patient records require exact matches of features in order to be considered similar (Table 1). This problem of semantic mismatch has been addressed with methods such as word2vec, which trains similar representations for similar words [28]. However, this method has had limited success in the context of medical records (see Related Work for more details) [8].
Both of these challenges make it difficult for models to effectively group similar EMRs together. Fortunately, genetic and protein interactome databases are rapidly growing due to more advanced high-throughput experiments [35] and improved biocuration via text mining [15]. These interactions typically take the form of molecular interaction networks. We can analyze the topologies of these networks and extract meaningful patterns. This analysis stems from the “guilt by association” rule, which loosely states that associated or interacting proteins in the network are more likely to be functionally related [31]. We show that applying this rule to networks consisting of molecular interactions and domain knowledge can reveal implicit relationships among nodes with common neighbors. We further show that including features directly extracted from EMR databases can help solve the primary challenges of missing data and semantic mismatching.

Contributions:
We present the Heterogeneous Medical record network (HEMnet), which consists of information derived from medical records, molecular interaction networks, and domain knowledge. As far as we know, the HEMnet is the first to integrate molecular interaction information and EMRs into a single network. We show that efficient node representations trained from the HEMnet can be used to enrich EMRs. Furthermore, we show that the enriched EMRs can better group similar patients whose records may otherwise suffer from missing data and semantic mismatches. Though we perform experiments in the context of EMR enrichment, the HEMnet is a general model that can be applied to many different tasks, including survival prediction and drug-gene link prediction.

2 DATA DESCRIPTION
The National Data Center of Traditional Chinese Medicine in Beijing, China has been collecting traditional Chinese medicine (TCM) EMRs since 2007, integrating them into a clinical data warehouse [44]. The data warehouse stores over 300,000 clinical cases and will continue to acquire patient records from six hospitals in subsequent years. Each hospital in the program has received roughly 3,000,000 patient visits annually for a decade. In general, TCM data is a largely untapped resource, and provides an unprecedented opportunity to apply powerful mining algorithms to discover useful knowledge. Several interesting studies have already succeeded in detecting herb combinations for disease treatment [39], latent symptom phenotype regularities [41], and herb-symptom relationships [7]. Additionally, Li et al. showed that discovering knowledge from TCM data has significant clinical applications [23].

The data we use is curated from a hospital that also provides traditional Chinese medicine (TCM) services. In addition to the aforementioned benefits of TCM data, we choose this dataset because semantic mismatches are even more prevalent in the TCM field [16]. Thus, if our method can yield strong results on this dataset, it will also work well for EMR datasets in most other domains. All patients in the data were diagnosed with some type of non-small-cell lung carcinoma (NSCLC), including adenocarcinoma, squamous-cell carcinoma, adenosquamous carcinoma, and papillary adenocarcinoma. In addition to TCM herb prescriptions, patients received standard western medical treatments.

In our experiments, we only consider the first visit that a patient makes to the hospital. Overall, we selected 133 patients that have a high degree of missing data. For these patients, we have six categories of features: medical history, medical test results, prescribed herbs, prescribed drugs, symptoms, and syndromes. Here, syndromes denote “patterns” that are specific to TCM and have no direct western medical translation [9]. In total, there are 449 unique features. On average, each patient is missing 407 of the 449 possible features, providing a unique challenge of both missing data and semantic mismatch.

In addition, the database contains each patient’s survival information in the form of the number of months before the final event. In our data, a patient’s final event is either hospital discharge or death. The data is right-censored (i.e., if there is no recorded death event, then death occurs at an unknown time after hospital discharge).

3 HEMNET
We propose to create the heterogeneous medical record network (HEMnet) to address the challenges of EMR analysis. The basic idea is to leverage information from several external sources in order to supplement the knowledge contained in clinical data. We present its definition, the network embedding process, and how we use the resulting node representations to enrich the original medical record feature matrix in the following sections.

3.1 Definition of HEMnet
The HEMnet is a network that consists of nodes and edges associated with different types of information. Formally, we define the HEMnet as a graph \( G = (V, E, R) \), where \( V \) is the set of typed nodes (i.e., each node belongs to a specified type), \( E \) is the set of typed edges, and \( R \) is the set of edge types. An edge \( e \in E \) in the HEMnet is an ordered triplet \( e = \{u, v, r\} \), where \( u, v \in V \) are typed nodes and \( r \in R \) is the corresponding edge type.

The network in this study utilizes four distinct categories of edges. The first three categories are drawn from external databases, while the last category draws directly from the EMR database.

1. Protein-protein interaction network: this network is based on an external gene network of protein-encoding genes, called HumanNet [21]. For a functional linkage between two proteins \( p_1 \) and \( p_2 \), we create a node for \( p_1 \), a node for \( p_2 \), and an undirected edge \( \{p_1, p_2\} \) in the HEMnet. This is the molecular interaction network.
we have 133 patients and 449 features. Each row of A/f_terminating low-dimensional vector representations of nodes
133 sparse patient records, the external information is critical in
treating the categorical features.
as dosages, while metastasis sites are categorical features that may
ple, drugs may be continuous features because they are recorded
on the corresponding feature. For exam-
a patient record with binary, categorical, discrete, or continuous
values, depending on the feature. For example, a patient record
in the dictionary. This is domain knowledge.
(4) Electronic medical records: we directly add co-occurrence
edges from each medical record. For example, if a patient is
diagnosed with cancer type c and prescribed a drug d, then we
create a node c, a node d, and an undirected edge \{c, d\}. We
repeat this for all elements in each patient’s medical record.

All edges are unweighted and undirected. Because we only use
133 sparse patient records, the external information is critical in
discovering relationships among EMR entities that may otherwise
be completely hidden. Overall, \(|V| = 11,911, |E| = 379,715, \text{and}
|R| = 23\). It is important to note that there are only 449 actual
features in the EMRs, and that the remaining nodes are proteins
that are only used in the HEMnet.

3.2 Network Embedding
In this paper, we use the recently proposed embedding method
ProSNet to infer the relationships among different entities in the
HEMnet [37]. ProSNet was originally developed to discover
functionally similar proteins from a large biological network across
multiple species. It takes a heterogeneous network as input, and
then performs a novel dimensionality reduction algorithm on the
input network to optimize a low-dimensional vector representation
for each node. The vectors of two nodes will be co-localized in
the low-dimensional space if the nodes are close to each other in
the heterogeneous network. A key computational contribution is
that ProSNet obtains low-dimensional vectors through a fast online
learning algorithm instead of the batch learning algorithm used in
previous works [36]. In each iteration, ProSNet samples a path from
the heterogeneous network and optimizes the vectors based on this
path instead of all pairs of nodes. Therefore, it can easily scale to
large networks containing hundreds of thousands, or even millions,
of edges and nodes. This attribute of ProSNet makes it suitable for
the HEMnet. We choose our vectors to be of the recommended
number of 500 dimensions.

3.3 Feature Matrix Enrichment
After generating low-dimensional vector representations of nodes
in the HEMnet, we can tackle the problem of missing data and
semantic mismatches in the patient records.
The patient record matrix M is a 133 \times 449 feature matrix, since
we have 133 patients and 449 features. Each row of M consists of
a patient record with binary, categorical, discrete, or continuous
column values, depending on the corresponding feature. For exam-
ple, drugs may be continuous features because they are recorded
as dosages, while metastasis sites are categorical features that may
indicate multiple locations in the body. In this study, we binarize
the categorical features.

In order to impute missing features, we construct a similarity
matrix, S. S is a 449 \times 449 matrix such that \(S_{ij}\) is the absolute value
of the cosine similarity between feature i and feature j’s embedding
vectors. Furthermore, we set a threshold such that any entry \(S_{ij}\)
is set to 0 if \(S_{ij}\) is less than the threshold. We empirically set this
threshold to 0.3. After creating S, we can generate the enriched
patient record matrix \(M'\) with the following operation:

\[
M' = M \times S
\]

With this operation, we can simultaneously solve the problems
of incomplete data and semantic mismatches by filling in missing
values that are highly similar to existing features for each patient.
For example, a patient that has “halitosis” in his or her patient
record will receive a non-zero value for “bad breath”, as these two
symptoms are likely to have very similar embedding vectors.

4 EXPERIMENTAL DESIGN
Typical cancer dataset studies might cluster patients based on a
particular characteristic or set of characteristics (e.g., gene expres-
sion profiles) [4]. The hope is to identify features that discriminate
patients who develop metastases (or die) from those who remain
metastasis-free [38]. Following this line of reasoning, we cluster
patients into two groups. We can frame this as an unsupervised
classification task in which we use a patient’s medical record to
determine if he or she is likely to survive. Thus, a “good” clustering
will have one cluster with a significantly better survival rate than
the other.

We wish to evaluate the effectiveness of the HEMnet-based ex-
pansion of the original feature matrix in the context of clustering.
The expansion involves creating the heterogeneous medical record
network (HEMnet), training network embeddings, and then rebuild-
ing the feature matrix with similarity scores (Figure 1). Another
method, which we refer to as the baseline method, directly uses
the feature matrix generated from the patient records. Lastly, the
mean imputation method replaces each feature’s missing values
with the average of the available values. We wish to compare the
performances of these three feature matrices, all of which undergo
the same clustering processes. In the following sections, we dis-
cuss cancer subtypes, the clustering task, and then our evaluation
methods.

4.1 Cancer Subtypes
We address the fact that different cancer types are affected by dif-
ferent factors. For example, age has been shown to be negatively
related with survival rates for patients with glioblastoma multi-
forme [5]. On the other hand, younger breast cancer patients sta-
tistically have worse prognoses than older patients [30]. Therefore,
we first separate the patients in our database into two groups: one
of squamous-cell carcinoma (SCC) patients and the other of non-
squamous-cell non-small-cell lung carcinoma (non-SQ NSCLC) pa-
tients (Figure 2). There are 43 SCC patients and 90 non-SQ NSCLC
patients. We perform experiments and analysis on these two cancer
subtypes independently, but build the HEMnet with all available
records.
4.2 Clustering

For each method, we cluster on the corresponding patient record matrix. We choose to generate two clusters, for reasons stated previously. Note that the survival information is not included in the clustering process. We do not necessarily use all features during clustering, but rather combinations of feature categories. For example, we might cluster on feature matrices containing only drug features, drug and medical history features, symptom and herb features, etc. We found that clustering on symptom and medical history features (a total of 57 features), yielded the best performance.

In addition, we perform dimensionality reduction with principal component analysis (PCA) to dampen the impact of highly correlated features. From the reduced feature matrix, we compute a dissimilarity matrix using pairwise cosine distance, which is commonly used to calculate the dissimilarity between two documents [33]. We run k-means for two clusters on the resulting dissimilarity matrix.

4.3 Survival Analysis (Quantitative Evaluation)

In order to analyze the differences between the survival rates of two given clusters, we first compute the survival curves using the Kaplan-Meier estimator, one of the most frequently used methods in survival analysis [18]. After estimating the survival functions of both groups, we wish to determine whether they are significantly different. To this end, we compare the two curves with the log-rank test [26]. The log-rank test computes a \( \chi^2 \) statistic and a corresponding p-value to show if the two clusters have significantly different survival functions. We performed survival analysis with the R package \texttt{survival}\(^1\). Recall that we judge a clustering to be of high quality if one cluster possesses a significantly higher survival rate than the other.

When reporting the means of survival functions, we use a restricted mean. Since the last observation may not be a death, the survival curve estimates do not necessarily go to zero, which results in an undefined mean. Thus, we set the upper limit to be some constant \( u \), so that the reported “restricted” mean signifies the number of months out of the first \( u \) months that each group is expected to experience [32].

4.4 Feature Analysis (Qualitative Evaluation)

Throughout our experiments, we denote the cluster with longer survival \( c_{long} \) and the cluster with shorter survival \( c_{short} \). After identifying \( c_{long} \), we would like to discover the features that are responsible for its higher survival rate. Although we only use a subset of features during the clustering phase, we analyze all features that occur in the EMRs. This is because all features are used in the HEMnet to train embedding vectors, so they may indirectly impact the clustering features.

\(^1\)https://CRAN.R-project.org/package=survival
We discuss the results of the survival and feature analyses below where one set of samples comes from A/fur1 clustering squamous-cell carcinoma patients into two sets. The features with significant p-values (< 0.01) are candidate features for further exploration.

5 RESULTS AND DISCUSSION

We discuss the results of the survival and feature analyses below for each of the three methods. First, we discuss the SCC patients, then the non-SQ NSCLC patients.

5.1 The HEMnet substantially improves patient stratification on SCC patients

After clustering squamous-cell carcinoma patients into two sets using HEMnet-enriched feature matrices, c_long has 28 patients and c_short has 15 patients (Figure 3a). By the log-rank test, c_long’s survival function is significantly better than c_short’s with a $\chi^2$ statistic of 10.79 ($p$-value = 0.001020). We show the cluster summary statistics in Table 2, with $u = 33.3$.

In contrast, when using the baseline feature matrix, c_long has 25 patients and c_short has 18 patients (Figure 3c). Here, the survival functions are not significantly different at the 1% significance level, with a $\chi^2$ statistic of 6.214 ($p$-value = 0.01267).

| Table 2: SCC Patient Survival Time Statistics with HEMnet |
|-------------|-------------|-------------|-------------|-----------------|
|             | Cluster Size | Restricted Mean | Restricted Mean | Median |
| c_long      | 28          | 20.7          | 2.01          | 22.7  |
| c_short     | 15          | 11.0          | 1.60          | 11.0  |

When using feature matrices with mean imputation, c_long has 30 patients and c_short has 13 patients (Figure 3e). Here, the survival functions are significantly different with a $\chi^2$ statistic of 7.623 ($p$-value = 0.005760), but not as well-separated as with our proposed method.

By computing an unpaired t-test for each feature, we get a $p$-value denoting how different the feature’s values were in the c_long versus c_short (Table 3). We bolded and placed an asterisk after the features that were identified by our method but not by the baseline. In the baseline method, “umbrella poly pore” has a $p$-value of 0.01513, “sulphurweed” has a $p$-value of 0.01507, “sputum” has a $p$-value of 0.02150, “A. tataricus” has a $p$-value of 0.01352, and “cough” has a $p$-value of 0.1056. Our method missed one feature deemed significant by the baseline: “TNM staging system”, which describes the stage of the cancer’s progression [10]. Since there are many missing values in the original data, the baseline method fails to leverage a variety of features. However, the HEMnet does not have this limitation, and also achieves a more significant separation between the two clusters. The mean imputation method did not find any non-herb features to be significant.

Of all significant features, only “Karnofsky performance status” has a higher mean in c_long than in c_short. This is expected, as a higher KPS score indicates a relatively healthier cancer patient [19]. All other significant features have lower means in c_long than in c_short. This makes sense for symptom features, as patients with longer survival rates should be in better health. On the other hand, it might seem, at first, as if the higher values of herbs in c_short indicate their ineffectiveness. However, patients with more life-threatening conditions require higher dosages of treatments.

The baseline method did not find sputum and cough to be symptoms that are statistically different between c_long and c_short; while our method did. However, they are very common symptoms in lung cancer patients that tend to be more severe in patients with lower survival rates. Furthermore, the other three features that our method discovered are all herbs. We can interpret these features’ higher values in c_short as potential herb-symptom relationships. In the herb-symptom dictionary, sulphurweed is known to treat both cough and sputum, while umbrella polypore only treats urinary tract-related symptoms and A. tataricus simply does not appear. Despite this, a study showed that a naturally occurring compound derived from umbrella polypore mycelia induces apoptosis in human lung cancer cells [20]. Furthermore, a recent study showed that a polysaccharide isolated from A. tataricus inhibits the growth of cancer cells [42].

Our method was able to capture these meaningful herb-symptom relationships while the baseline methods could not. This is due to the fact that the HEMnet integrates external herb information in...
Figure 3: A comparison of survival functions for two patient clusters in SCC patients (subfigures (a), (c), and (e)) and non-SQ NSCLC patients (subfigures (b), (d), and (f)). Subfigures (a) and (b) use HEMnet-enriched feature matrices, subfigures (c) and (d) use the baseline feature matrices, and subfigures (e) and (f) use mean imputation on the baseline feature matrices.
of 0.021107, “Chinese cinnamon” had a $p$-value of 0.06895, and “A. asphodeloides” had a $p$-value of 0.03629. On the other hand, our method obtained $p$-values below 0.01 for each of these features. Atelectasis and shortness of breath are common symptoms among lung cancer patients, particularly those in more advanced stages.

Because our method found several treatments to be significant, this again affords us the opportunity to extract meaningful relationships. Pamidronate disodium injections are well-known treatments for patients with bone metastases [11], which is also a significant feature ($p$-value $= 1.294 \times 10^{-6}$). White peony roots have been shown to inhibit tumor growth in human non-small-cell lung cancer patients [17]. Lastly, a study has shown A. asphodeloides to have anti-tumor effects [2]. As with the SCC patients, our method was able to capture these meaningful herb-symptom relationships while the baseline methods could not. We again attribute this to the external herb relationship information integrated into the HEMnet.

### 5.3 KPS Feature Significance

Because “Karnofsky performance status” has the lowest $p$-value for both cancer types, we deemed it beneficial to further explore its significance. Performance status scores are assigned to cancer patients in attempts to quantify their overall health and well-being. It is therefore reasonable that KPS is a discriminative feature for patients of both cancer subtypes. Specifically, the Karnofsky score runs from 100 to 0, where 100 is relatively perfect health and 0 is death [19]. In this dataset, Karnofsky performance scores were assigned in standard intervals of 10.

To test whether KPS can accurately separate patients, we developed another experiment, classifying patients into $c_{\text{long}}$ if they had a Karnofsky score greater than 60, and the rest into $c_{\text{short}}$ (Figure 4). Although using only KPS can separate non-SQ NSCLC patients significantly ($p$-value $= 5.571 \times 10^{-4}$), we see that the survival differences are not very significant between the two clusters for SCC patients ($p$-value $= 0.3655$). Thus, we can conclude that KPS alone cannot accurately predict a patient’s survival rate.

### 5.4 Parameter Tuning

One of the parameters that requires tuning in our framework is the number of dimensions of each embedding vector, which was...
Table 5: Significant Non-SQ NSCLC Patient Features with HEMnet (p-value < 0.01). Features not found to be significant by the baseline method have asterisks. Feature means and standard deviations (SD) are shown.

<table>
<thead>
<tr>
<th>Feature Name</th>
<th>p-value</th>
<th>$c_{long}$ mean</th>
<th>$c_{long}$ SD</th>
<th>$c_{short}$ mean</th>
<th>$c_{short}$ SD</th>
<th>Feature Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Karnofsky performance status</td>
<td>5.570 × 10^{-8}</td>
<td>76.35</td>
<td>8.993</td>
<td>63.42</td>
<td>11.98</td>
<td>Medical Test</td>
</tr>
<tr>
<td>Pamidronate disodium injection*</td>
<td>2.196 × 10^{-7}</td>
<td>0.2115</td>
<td>0.8166</td>
<td>15.34</td>
<td>19.74</td>
<td>Drug</td>
</tr>
<tr>
<td>Bone metastasis</td>
<td>1.294 × 10^{-6}</td>
<td>0.03846</td>
<td>0.1923</td>
<td>0.4211</td>
<td>0.4977</td>
<td>Medical Test</td>
</tr>
<tr>
<td>Atelectasis*</td>
<td>1.250 × 10^{-3}</td>
<td>0.01923</td>
<td>0.1373</td>
<td>0.2105</td>
<td>0.4077</td>
<td>Symptom</td>
</tr>
<tr>
<td>TNM staging system</td>
<td>1.307 × 10^{-3}</td>
<td>7.346</td>
<td>2.472</td>
<td>6.832</td>
<td>0.6250</td>
<td>Medical Test</td>
</tr>
<tr>
<td>Distant metastasis</td>
<td>2.307 × 10^{-3}</td>
<td>0.3846</td>
<td>0.4865</td>
<td>0.6842</td>
<td>0.4648</td>
<td>Medical Test</td>
</tr>
<tr>
<td>White peony*</td>
<td>8.760 × 10^{-3}</td>
<td>1.115</td>
<td>2.584</td>
<td>2.789</td>
<td>3.894</td>
<td>Herb</td>
</tr>
<tr>
<td>Shortness of breath*</td>
<td>8.999 × 10^{-3}</td>
<td>0.6346</td>
<td>0.8555</td>
<td>1.105</td>
<td>0.9676</td>
<td>Symptom</td>
</tr>
<tr>
<td>Chinese cinnamon*</td>
<td>9.302 × 10^{-3}</td>
<td>0.03846</td>
<td>0.2747</td>
<td>0.4211</td>
<td>1.091</td>
<td>Herb</td>
</tr>
<tr>
<td>A. asphodeloides*</td>
<td>9.952 × 10^{-3}</td>
<td>0.05769</td>
<td>0.4120</td>
<td>1.158</td>
<td>3.273</td>
<td>Herb</td>
</tr>
</tbody>
</table>

Figure 4: A comparison of survival functions when separating patients only by their Karnofsky performance score.

6 RELATED WORK

Though our work is the first to utilize molecular interaction networks in conjunction with electronic medical records, it draws inspiration from many tasks and networks.

A previous work integrated genetics by linking EMR data to biobanked blood samples [27]. However, they use patient-matched DNA samples and patient genetic data rather than prior domain knowledge. Deep Patient is a previous work that also learns embedding vectors from EMRs [29]. However, they learn representations of patients rather than individual medical entities and do not use molecular information networks. Med2Vec is an algorithm that also learns efficient representations of medical concepts, but its joint optimization process relies on sequences of multiple visits [8], which is less applicable to inpatients. Word2vec, on which Med2Vec is based, provided the basis for much of the current research on learning efficient word embeddings [28]. However, word2vec can only be applied to patient records directly and cannot incorporate molecular interaction networks or domain knowledge.

Many related works have implemented data mining techniques on heterogeneous networks such as bibliographic networks [40, 43], gene-phenotype networks [22], and social media networks [24]. However, none of these studies focused on EMR tasks. A previous study performed similarity search within heterogeneous information networks [34]. It also exploited the idea of paths within a heterogeneous network, but instead used them for similarity computations rather than for vector optimization. Another previous work also trained embeddings on a network consisting of prior knowledge in order to enrich EMRs for clustering tasks [16]. However, they only used domain knowledge in the training network. Caballero and Akella also performed data imputation on EMR data [6]. However, they use an expectation-maximization method to fill out missing values rather than prior information and domain knowledge.
7 CONCLUSIONS AND FUTURE WORK
In this paper, we proposed to integrate EMRs with molecular interaction networks and domain knowledge using a heterogeneous medical record network, HEMnet, to solve two challenges in analyzing EMRs, i.e., missing data and semantic mismatch. By extracting knowledge from molecular interaction networks and domain knowledge and then combining them with medical records, we created a network that allows for the training of accurate embedding vectors. We showed how we can use these vectors to enrich EMR databases, and then evaluated their performance based on survival prediction. Our method is able to perform better than the baseline feature matrix with and without mean imputation by successfully splitting patients of both cancer subtypes (squamous-cell lung carcinoma and non-squamous-cell non-small-cell lung carcinoma). We showed this via quantitative evaluation by performing survival analysis on the resulting cluster pairs of each cancer subtype. Lastly, we verified the effects of the HEMnet with qualitative analysis, studying the features and their differences between healthy and unhealthy clusters.

For future work, we plan to implement tasks for clinical prediction of survival (CPS). Given the framework proposed in this paper, we can attempt to predict a new patient’s probability of survival by first grouping him or her into the most similar cluster, and then extrapolating the survival rate from the neighboring patient records (similar to the k-nearest neighbors algorithm). Furthermore, the HEMnet allows for even more sophisticated methods of feature discovery, including novel drug-drug interactions. This feature discovery would follow along the lines of our qualitative evaluation.

ACKNOWLEDGMENTS
This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1144245. This research was also supported by grant 1U54GM114838 awarded by NIGMS through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (www.bd2k.nih.gov).

Our work also received support from National Key Technology R&D Program grants 2013BAH02B01 and 2013BAH06F03, National S&T Major Project of China grant 2013BAI13B04, and Special Research Project of TCMS by State Administration of Traditional Chinese Medicine grant 201407001.

REFERENCES