2016 IEEE International Conference on Data Science and Advanced Analytics

MedCare: Leveraging Medication Similarity for Disease Prediction

Dipanwita Dasgupta
Dpt of Computer Science and Engineering and iCeNSA
University of Notre Dame
Notre Dame, Indiana
Email: ddasgupt@nd.edu

Nitesh V. Chawla
Dpt of Computer Science and Engineering and iCeNSA
University of Notre Dame
Notre Dame, Indiana
Email: nchawla@nd.edu

***Abstract*—The emergence of electronic health records (EHRs) has made medical history including past and current diseases, and prescribed medications easily available. This has facilitated development of personalized and population health care management systems. Contemporary disease prediction systems leverage data such as disease diagnoses codes to compute patients' similarity and predict the possible future disease risks of an individual. However, we posit that not all diseases (such as pre-existing conditions) may be represented in an EHR as a disease diagnosis code. It is likely that a patient is already taking a medication but does not have a corresponding disease in the EHR. To that end, we posit that the medication history can serve as a proxy for disease diagnoses, and ask the question whether medication and disease diagnoses combined together can improve the predictability of such systems. Building on our prior work in predicting disease risks (CARE), we develop two disease prediction systems: one using medication-based similarity (medCARE) and the other using both disease and medication-based similarity (combinedCARE). We show that combinedCARE provided a greater coverage and a higher average rank.**

I. Introduction

The Affordable Care Act (ACA) has led the digitization of healthcare, from mandating Electronic Health Records (EHRs) to quantifying and measuring patient outcomes [18], [3]. The EHRs have been leveraged extensively for prospective healthcare or predictive modeling of disease risks towards driving the vision of personalized and/or precision medicine [5], [4], [8]. A variety of contemporary methods or systems rely on disease diagnoses codes in the EHRs to compute patient similarity, learn predictive models, and generate predictions about possible diseases that an individual may be at risk for [7], [24], [11], [30], [12], [31], [11], [22].

However, there continue to be significant limitations in EHRs in their ability to truly represent a picture of a patient's current disease states. For instance, the EHR may only record the primary diagnosis of a patient during the visit or the chronic diseases / pre-existing conditions may not be recorded or a patient may be transferred from another healthcare system with pre-existing prescription records. Moreover, ICD-9-CM [1] codes can offer their own challenges [26] when used as part of predictive healthcare systems. Sometimes, errors in coding can be attributed to upcoding. Also, it may not be possible to verify the fidelity of the assignment of any particular code to a patient without repeated visits to a clinician [23]. Prescribed medications, on the other hand, target the underlying disease. A chronic disease patient may continue to use the prescribed medication, so the medications can be helpful in a more representative disease state view of the patient.

The above-mentioned scenarios and challenges posit important research questions: Is the reliance on disease diagnoses codes for predicting disease risks sufficient? Can the medication records serve as proxy for disease diagnoses? Can the medication records along with disease diagnoses codes be used to improve the predictability of disease risks for an individual?

To that end, we first demonstrate the utility of integrating disease diagnoses and medications using network based analysis and visualization [17], [31]. We use networks to establish the premise for developing a predictive model using an integrated dataset from disease diagnoses and medication records. We then detail our disease prediction methods that leverage both disease diagnoses and medication records, and demonstrate the improvement that comes from an integrated view. We implement our models by extending an existing disease prediction system, the Collaborative Assessment and Recommendation Engine (CARE) [10], [7], to include the medications in the calculation of disease risk likelihoods. We describe two adaptations of CARE that use this information: (1) medCARE, which only uses the medications for calculating the risk of a future disease, and (2) combinedCARE, that uses both medications and diseases for calculating this risk. This is the first work, to our knowledge, that uses medication records to drive the predictability of patient centered outcomes.

II. Related Work

A number of contemporary disease prediction systems leverage patient visit data defined using ICD-9-CM disease codes [31], [11] and/or online patient forum data [20], [21]. Various techniques like collaborative filtering using patient disease data and/or social network data have been employed for calculating patient similarity, which is used for predicting the risk for a particular disease for a particular patient [10], [20]. This patient similarity is also calculated using active learning or expert opinion to refine the similar patient neighborhood of a particular patient, constructed using counting cover method [29]. Association rules are generated using markov models and clustering [12], beta-binomial distribution and

978-1-5090-5206-6/16 $31.00 © 2016 IEEE
DOI 10.1109/DSAA.2016.90

regression models [24], association rule mining using Apriori algorithm [13], and association rule mining on clusters of patients generated using k-means algorithm with Jaccard similarity as the distance measure between the diseases of two patients [11]. Network-based approaches [31] and dirichlet mixture models [30] have also been used for predicting the probable disease risks. Many techniques including SVM-RFE feature selection method [32] have been used in predicting the risk for a single disease.

III. Data

We leverage two data sources to create a unified medication and disease diagnoses data. One data source contains de-identified 543,571 diagnoses records of 26,843 patients, spanning from 1999 to 2012. Each record contains a unique record identifier representing the visit date and time, patient identifier and list of diseases diagnosed for that visit. The diseases are represented using ICD-9-CM codes. Under this coding system, each disease is assigned a unique 5-digit long code. This coding system is hierarchical in nature, where the 5-digit code can be collapsed into 4-digit less specific code. The 4-digit code can further be collapsed to 3-digit code, representing the general condition [14]. For example, "Spinal stenosis, unspecified region" (ICD-9-CM 724.00) can be collapsed to "Spinal stenosis, other than cervical" (ICD-9-CM 724.0). ICD-9-CM 724.0 can be collapsed to a general code ICD-9-CM 724 ("Other and unspecified disorders of back"). As per the CARE paper [10], we collapse all the 5-digit and 4-digit codes into 3-digit codes for the analysis.

The medication data source contains 315,437 records of 11,230 patients. This database has both prescribed medications and procedures. There are 10,964 patients, which have medication records containing prescribed medications. Each record has a patient identifier and a unique encounter identifier. We, only, considered the records containing prescription drugs. The number of medication per patient record ranges from 1 to 38, with a median of 2.

The two databases are combined on basis of the encounter identifier. The intersection of these two databases is used for constructing the networks and experimental validation of the models proposed in this paper. We ignored all the entries, for which ICD-9-CM codes and the medication names were missing. The combined dataset has 4632 patients. There are 617 unique diseases and 790 unique medications in this combined dataset. The most frequently occurring diseases are hypertension (ICD-9-CM 401), followed by "Special investigations and examinations" (ICD-9-CM V72) and Diabetes Mellitus (ICD-9-CM 250). Acetaminophen-hydrocodone, albuterol and aspirin are the top three most frequently prescribed medications.

IV. Network Visualization

In this section, we demonstrate that the disease co-morbidities can be informed from medication histories, highlighting the importance of diseases, medications and a combination of both. This sets the context for leveraging medication and disease diagnoses information for personalized healthcare.

A. Network Construction

We constructed a disease-disease (DD) network (Figure 3(a)) where each node represents a disease and an edge represents the statistically significant co-occurrence of diseases in a patient's disease record. We calculated this statistical significance of co-occurrence using the method described in [9]. The thickness of the edges in the figure denote the frequency of occurrence of the two end nodes (disease-pair) in the dataset. There are 173 nodes (unique diseases) and 387 edges (disease co-morbidity) in this undirected, weighted network. The average clustering coefficient is 0.367 and the average shortest distance is 4.04.

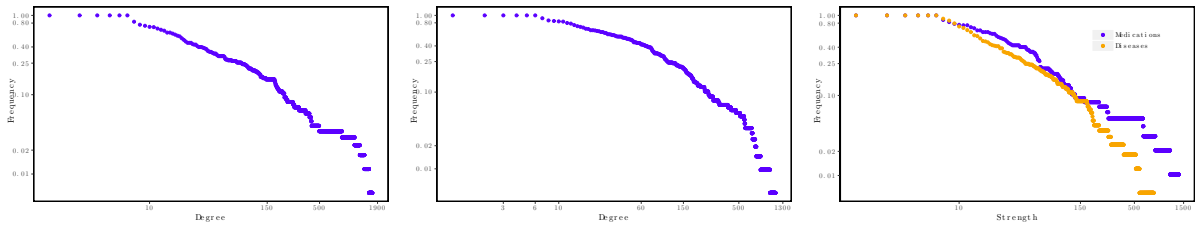
Using a similar network construction method, we built a medication-medication (MM) network (Figure 3(b)) with medications as nodes and edges representing the co-occurrences of medications in patients. We again relied on statistical significance to decide what relationships to keep and what to prune from this network. MM has 206 unique nodes (medications) and 932 unique edges. The average clustering coefficient is 0.40 and the average shortest distance is 3.00.

The integration of the disease-disease network (the DD network) and the medication-medication network (the MM network) can simultaneously provide insights on the side-effects of medications and co-morbidities of different diseases. Hence, we constructed a disease-medication network (DM network). It is a bipartite network with medications and diseases as nodes and an edge only connecting a medication and a disease. The edges are built using the same method used for DD or MM network. The DM network has 262 nodes and 464 unique edges. The average clustering coefficient [27] is 0.76 and the average shortest distance [28] is 5.16. We also calculated the degree distribution (Figure 1) and the shortest path distribution (Figure 2) for each of the networks.

Clearly, the MM network and DD network have differences, and the integrated DM network helps highlight the value that can be derived from a unified view. This is what we hope to leverage in the modeling constructs, as there is a value-add from including medications.

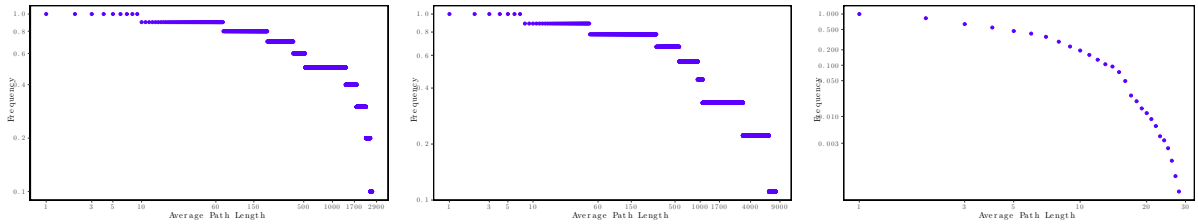
B. Network Analysis

The DD network has 7 connected components (Figure 3(a)). Upon analysis of these connected components, we found a strong association between hypertension (ICD-9-CM 401) and hyperlipidemia (ICD-9-CM 272) [indicated by the thickness of the edge connecting these diseases], implying a high comorbidity between them. In fact, studies have shown that there is strong correlation between hypertension and hyperlipidemia [2]. Further, we noticed hyperlipidemia to be connected with acquired hypothyroidism (ICD-9-CM 244) implying correlation between them [25]. Further, hypertension has been shown to be responsible for diabetes [16]. There is a strong edge between hypertension and diabetes in this network. In fact, the 2011 National Diabetes Fact Sheet showed that 67 percent of adults with diabetes have hypertension. Hypertension also shares



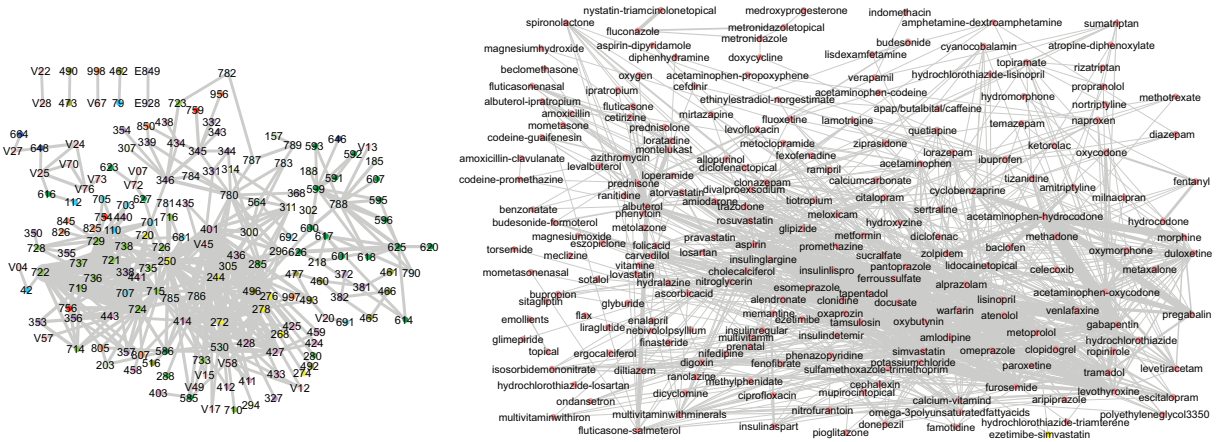
(a) Disease Degree (b) Medication Degree (c) Combined Degree

Fig. 1. Degree Distribution of the different networks



(a) Disease (b) Medication (c) Combined

Fig. 2. Shortest Path Distribution of the different networks



(a) Disease-Disease (b) Medication-Medication

Fig. 3. Homogeneous Networks: The different colors in the disease-disease network represent the different ICD-9-CM categories. The visualizations are done using Cytoscape v3.0.

an edge with coronary heart disease (ICD-9-CM 414). These association has been proven in the Framingham Heart Study [6]. Chronic kidney disease(CKD) (ICD-9-CM 585), belonging to a ICD-9-CM category (Diseases of The Genitourinary System) different from that of hypertension (Diseases of The Circulatory System), is connected with hypertension. Research has shown CKD to have a strong impact on risk of mortality for people having heart failure [15].

On the other hand, the MM network has 2 connected components with the largest connected component containing 204 nodes. Lisinopril, hydrochlorothiazide, amlodipine, metaprolol, and carvedilol are connected vertices in this network and are prescribed for hypertension. Some of the medications/nodes prescribed for hyperlipidemia are simvastatin, and atorvastatin. We observe a similar trend as seen in the DD network (Figure 3(a)) with connections existing between anti-hypertensive drugs, anti-hyperlipidemic drugs, and anti-diabetic drugs. For instance, the majority of the anti-hypertensive drugs (like amlodipine, and lisinopril) share an edge with simvastatin that treats high cholesterol (hyperlipidemia). Furthermore, the simvastatin node has levothyroxine (used in the treatment of hypothyroidism) as one of its neighbors in MM.

The largest connected component also contains connected nodes corresponding to medications for depression, anxiety, and mania disorders. For instance, clonazepam is given for the treatment of seizures, panic disorder and anxiety. Citalopram

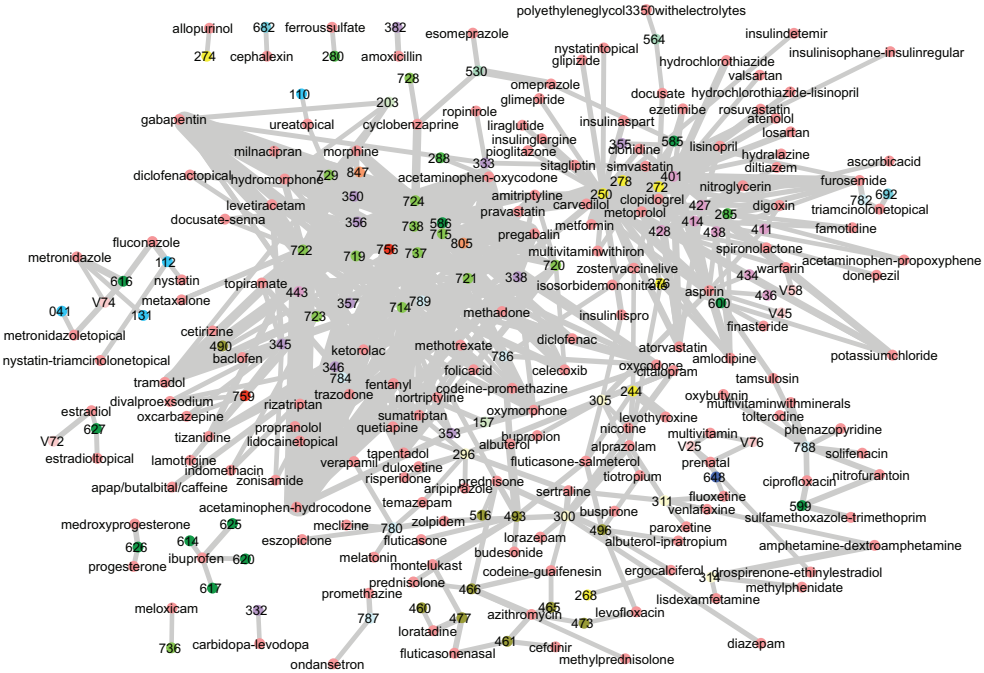


Fig. 4. Medication-Disease Network: The different colors represent the different ICD-9-CM categories and medications. The visualization is done using Cytoscape v3.0.

is a node in this component that is prescribed for treating depression [19]. Another node, quetiapine is prescribed for manic-depressive disorders like schizophrenia and bipolar disorder.

There are 10 connected components in the integrated DM network, with the largest one having 237 nodes (Figure 4). The most prevalent disease in our dataset is backache (ICD-9-CM 724.5 or node labelled 724). It shares connections/edges with medications like pregabalin, morphine, acetaminophen-oxycodone, tapentadol, celecoxib, cyclobenzaprine, acetaminophen-hydrocodone, oxymorphone, topiramate, naproxen, methadone, hydromorphone, tramadol, gabapentin, oxycodone, and lidocaine topical. It has the maximum number of edges with acetaminophen-hydrocodone and acetaminophen-oxycodone, representing the highest number of co-occurrences in our dataset.

This analysis clearly shows that there are differences among disease co-morbidities expressed through the different networks, and there is a potential for improved disease predictions by integrating this diversity among DD and MM networks.

V. Predicting Diseases

In this section, we first describe the disease prediction system, CARE, which is used as the basic framework [7]. We then propose two extensions to CARE: medCARE using medication-based similarity and combinedCARE using medication-based and disease-based similarity for calculating the similarity between two patients. First we briefly describe CARE [7], as medCARE and combinedCARE are dervatives of CARE.

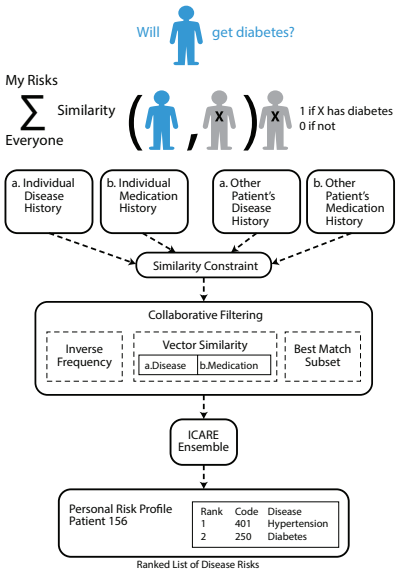


Fig. 5. Architectural representation of different models: CARE, medCARE and combinedCARE

A. CARE

CARE outputs a ranked list of diagnoses for any patient (test patient) based on his/her medical history of diseases by computing patient similarity using a collaborative filtering algorithm. Each of the diseases is provided a likelihood score, which affects its ranking.

Before we detail CARE, we briefly describe collaborative filtering in general terms, including the notation. Consider an active (or test) user a. Collaborative filtering method generates a prediction $p(a, j)$ on item *j* for the active user *a* based on the similarity between the active user *a* and training user *i*, who voted $v_{i,j}$ for item *j*. The entire training set of items is defined as I, and I_j is the subset of users who have voted for item *j*. The similarity score $w(a, i)$ between user *a* and a training user i is defined as:

$$w_d(a,i) = \sum_j \frac{f_j v_{a,j}}{\sqrt{\sum_{k \epsilon J_a} f_k^2 v_{a,k}^2}} \frac{f_j v_{a,j}}{\sqrt{\sum_{k \epsilon J_i} f_k^2 v_{i,k}^2}} \quad (1)$$

where:
I = Total number of users in the dataset
I_j = Total number of users in the dataset who have voted for item j
f_j = Inverse Frequency of item j calculated as $\log_{10} I/I_j$
$v_{a,j}$ = indicator function.
Equal to *1* if the training user has voted for j
0 otherwise

In our context of diseases and patients, an item represents the disease, and an user represents a patient. The vote $v_{i,j}$, in this case, can be interpreted whether the user *i* has been diagnosed with a disease *j*. The prediction score $p(a, j)$ is defined as:

$$p(a,j) = \overline{v}_j + k(1 - \overline{v}_j) \sum_{i \epsilon I_j} w(a,i) \quad (2)$$

where:
k = nornalizing constant
k = 1 / $\sum_{i \epsilon I} w(a,i)$

$p(a, j)$ is the prediction score that the patient *a* will be diagnosed with the disease *j*. The patient *a* has not been diagnosed with disease *j*. Thus, $p(a, j)$ denotes the probability of patient *a* diagnosed with disease *j* in the future. $\overline{v}_j$ represents the random expectation of the disease *j*. Alternatively, $\overline{v}_j$ can be interpreted as the baseline expectation of disease *j*, and is defined as I_j/I.

Any similarity function can be dominated by more commonly occurring diseases in a dataset. To mitigate this effect, CARE uses inverse frequency to bring up the contribution of the less frequent or relatively rare diseases. In addition, an ensemble of CARE models is constructed (ICARE). For each disease *j* that the patient *a* is diagnosed with, a training subset *c* of patients is generated with the ones who have been diagnosed with the disease *j*. CARE is applied to the training subset generated for each disease *j* that the patient *a* is diagnosed with, resulting in an ensemble of CARE models providing the likelihood score for disease *j*: $p(a, j)$. The maximal of all the $p(a, j)$ is taken as the prediction score for disease *j*. Thus, the prediction score for disease j can be defined as:

$$p(a,j) = \overline{v}_{j,c} + k(1 - \overline{v}_{j,c}) \sum_{i \epsilon I_{j,c}} w(a,i) \quad (3)$$

The baseline expectation $\overline{v}_{j,c}$ of the disease *j*, in this case, is the expectation of the disease within group *c*.

In this paper, we use the above described model as the base model for CARE. The architectural diagram for CARE is shown in Figure 5a.

Algorithm 1 Algorithm for medCARE

Initialize the score for each disease to baseline expectation of the disease in the population
for all disease j in the test patient a **do**
  $kappa\_m = 0$ ▷ Running Total for Medication Similarity
  for all patients i in the dataset who have disease j **do**
    best match function in CARE to get the weights*(vecsim)* and list of predictable diagnoses
    ▷ %comment: best match function time-sensitive CARE%
    $kappa\_m += vecsim$
    for all disease d in the predictable list **do**
      update the weight associated with d with *(vecsim)*
      update the number of training patients in this cluster
    end for
  end for
  for all disease d in the dataset **do**
    $weight\_m = weight\_m/\ kappa\_m$ ▷ Normalize the diagnosis-wise similarity
    $\overline{v}_{j,c}$ = baseline expectation of j in group c
    $score\_d = \overline{v}_{j,c} + weight\_m$ ▷ Revised prediction score for diagnosis d
    if (present score of d ¡ $score\_d$) **then**
      score of d = $score\_d$ ▷ ICARE
    end if
  end for
end for
Sort the diagnoses by score in descending order
The sorted list of diagnoses gives the predictions for test patient a

B. MedCARE

We calculate the similarity between patient *a* and training patient based on their medication history. The basic steps of the algorithm are akin to CARE, albeit the diseases are replaced by medications. The algorithm is mentioned in Algorithm 1.

C. combinedCARE

We showed that an integrated version of disease-medication network provides enriched information on disease-medication interactions than that obtained from disease or medication networks individually (Section IV). The integrated version also provides information on disease co-morbidities. These observations from the disease-medication interaction network led us to combine the two models (CARE and medCARE) to build combinedCARE. In this integrated model, we calculate the similarity between the patient *a* and the training patient *i* as the sum of the normalized vector similarity based on their

disease history as well as their medication history (w_{m_norm}). The revised predication score $p(a, j)$ is calculated using the following equation:

$$p\left(a,j\right)=\overline{v}_j+\left(1-\overline{v}_j\right)\sum_{i\epsilon I_j}\left(\frac{w_d\left(a,i\right)}{k_d}+w_{m\_norm}\left(a,i\right)\right) \tag{4}$$

$w_d\left(a,i\right)$ represents the similarity between the patient *a* and the training patient *i* based on their disease history. k_d is the normalizing constant for disease-wise similarity and is same as k, as defined in the CARE equation. The normalized medication-wise similarity is defined as the individual similarity divided by the sum of all the medication-wise similarities. The algorithm can be found in Algorithm 2.

Algorithm 2 Algorithm for combinedCARE

Initialize the score for each disease to baseline expectation of the disease in the population
for all disease j in the test patient a **do**
$kappa\_d = 0$ ▷ Running Total for Diagnosis Similarity
$kappa\_m = 0$ ▷ Running Total for Medication Similarity
for all patients i in the dataset who have disease j **do**
best match function in CARE to get the weights*(vecsim)* and list of predictable diagnoses
calculate medication similarity *(vecsim_m)* between i and a
▷ %comment: best match function time-sensitive CARE%
$kappa\_d += vecsim$
$kappa\_m += vecsim\_m$
for all disease d in the predictable list **do**
update the weight associated with d (both with *(vecsim)* and *(vecsim_m)*)
update the number of training patients in this cluster
end for
end for
for all disease d in the dataset **do**
$weight\_d = weight\_d/\ kappa\_d$
$weight\_m = weight\_m/\ kappa\_m$
$\overline{v}_{j,c}$ = baseline expectation of j in group c
$score\_d = \overline{v}_{j,c} + weight\_d + weight\_m$ ▷ Revised prediction score for diagnosis d
if (present score of d ¡ $score\_d$) **then**
score of d = $score\_d$ ▷ ICARE
end if
end for
end for
Sort the diagnoses by score in descending order

D. Evaluation

We evaluate the three models: CARE, medCARE and combinedCARE among themselves and also against a baseline method. The baseline method provides a ranking of diseases on the basis of random expectation $\overline{v}_j$ of diseases. Further, we use coverage and average rank for evaluating the performance of the proposed models [10]. We consider the Top 20, Top 50, Top 100 ranks for evaluating the models. The evaluation metrics are discussed in the following subsections.

1) Coverage: Coverage is defined as the percentage of diseases that are correctly predicted by the model [10]. A higher coverage is preferred as we want to cover as many of the future disease risks in our predictions.

2) Average Rank: Average or mean rank of diseases is used as a performance metric [10]. A lower average rank is preferred as low rank position is desirable. Lower rank position indicates higher risk. Moreover, a lower rank increases the probability of the disease being noticed and hence, prompt preventative measures for those diseases could be taken.

E. Experimental Setup

We used a leave-one-patient leave-one-visit out validation framework. For the patient *a*, we used the first visit as an input to the above discussed models to generate predictions on the next visit. The list of diseases or medications prescribed in this present visit is used to calculate the disease-wise or medication-wise similarity between the patient *a* and all the other patients present in the database. This similarity score is then used to generate the predictions. The predictions are verified against the subsequent visit of the patient *a*. A union of diseases or medications for all previous visits and the present visit is used as the disease or medication vector for the present iteration. For example, the disease or medication vector for visit i of the patient *a* is the combination of diseases diagnosed or medications prescribed in all the visits from 1 to *i*. The table I provides a presentation of the setup. For the sample given in this table (Table I), we consider that the patient has 3 visits. The present visit represents the set of diseases or medications of the patient considered for calculating the similarity. The testing set indicates the set of diseases used for validation of the predictions made by the models.

Visit #	Present Visit	Testing Set
1	Visit 1	Visit 2
2	Visit 1 ⋃ Visit 2	Visit 3

TABLE I
VALIDATION PROCESS: AN EXAMPLE FOR A PATIENT WITH 3 VISITS

We predict only the diseases, which the patient *a* has not been diagnosed with yet, and not repetitive chronic diseases, as it is of greater interest to identify what new diseases a patient might be at risk for. A patient can also be diagnosed with the same disease or prescribed with the same medication in multiple visits. However, we considered a unique set of diseases or medications for calculating the similarity as comparing the same disease multiple number of times is not useful. We do not consider the last visit (in chronological order) of a patient for training the models as there is no subsequent visit to validate the predictions. Similarly, we do not predict on the first visit as collaborative filtering or the calculation of similarity requires a set of diseases or medications to generate the prediction list.

For example, referring to the Table I, we do not predict for Visit 1, and do not train the models on the last or third visit.

All the testing patients are required to have a minimum of 3 diseases in a single visit or the union of diseases of all previous visits. That is, consider a patient with 4 visits. The patient is diagnosed with 1 disease in visit 1, 1 disease in visit 2, 2 diseases in visit 3, and 3 diseases in visit 4. We do not use visit 1 for calculating the similarity as the patient has only 1 disease. Hence, we do not generate predictions for testing against visit 2 and so on using visit 1. Similarly, the visit 2 is also not used as combining visit 1 and visit 2, the patient has less than three diseases. However, we would predict after visit 3 as combining visits 1, 2, and 3, results in four diseases for the patient and there is still the last visit 4 to make predictions on. Thus, in this case, the disease or medication set for calculating the similarity is the union of diseases or medications for the first 3 visits.

Of course, if a patient has three or more diseases in visit 1, then that visit in itself provides a disease vector for similarity computation and prediction on subsequent visits. And after visit 2 is observed, visit 2 diseases are added to visit 1 diseases for similarity computation and prediction on visit 3 onwards, and so on. We do not impose any restrictions on the number of medications prescribed because we found that the number of medications prescribed is usually more than that of the diseases diagnosed for a particular visit.

VI. Results

The evaluation metrics are first calculated on the basis of the correct matches or the rank of the diseases for all the test visits of a patient. Further, these values are averaged over the total number of patients in the database.

The Table II shows the experimental results for all the above-mentioned models. The best performance on the basis of a performance metric for each category of ranks is represented in bold. Although we evaluate for Top 20 ranks, Top 50, Top 100 and all; top 20, 50 and 100 ranks are of more interest as a clinician would be most likely to be interested in considering a smaller list of diseases for future risk identification. For this same reason, the performance in top 20 ranks is of greater interest.

The performance of all the models: CARE, medCARE and combinedCARE is better than that of our baseline method. When the entire list of predictions for diseases is considered, the coverage is same for all the models. This is because the list is exhaustive in terms of the diseases and hence, every disease is present in the list. The coverage for all predictions, ideally, should have been 100%. However, in our case, it is not 100% because we are not predicting on diseases that have been observed for a patient, even if that disease recurs. Irrespective, overall coverage is not a useful metric as there are over 15,000 unique ICD-9-CM codes, and a physician will not have the time or efficiency to review all possible likelihood scores. Instead, it is more important to identify the disease coverage in the top 20 predictions. That is, how many of a patient's future diseases are predicted in the top 20 disease rankings (based on disease likelihood scores). More than 60% of the diseases are covered by CARE, medCARE and combinedCARE in the Top 20 ranks. combinedCARE predicts more than 46% of the future diseases in the top 20 positions, in comparison to the baseline method, where that of medCARE is 37%.

We also achieved a considerable improvement in the average rank of diseases, which is indicative of the higher score assigned to each disease by CARE, medCARE and combinedCARE; in comparison to the baseline model. When compared to the baseline method in terms of average rank, there is about a 22% improvement for CARE and medCARE, and a 26% for combinedCARE.

Although medCARE and CARE have similar performance in terms of coverage, medCARE has a lower average rank compared to that of CARE. In fact, there is about 6% improvement in average rank for medCARE over CARE. This implies that medCARE assigns a higher prediction score to certain diseases, in comparison to that assigned by CARE to those diseases. Average rank depends on the exact position of the disease in the ranked list of predicted diseases. This position is determined by the prediction score assigned to each disease by each model. A higher prediction score corresponds to a lower rank. Hence, the lower average rank for medCARE indicates that it assigns a higher prediction score to some of the diseases (that are present on the validation list for iteration i.e. the set of diseases diagnosed in the subsequent visit) than that given by CARE. As mentioned earlier, ranking of a disease is very important as a medical professional is likely to pay more attention to the lower ranks. The prediction score for CARE depends on the similarity between the test patient and all the training patients, based on their diseases. On the other hand, this score for medCARE depends on their medications. Thus, the lower average rank of medCARE indicates a higher similarity score for the diseases, which, in turn, indicates a higher overlap between the test patient and the training patient in terms of medications than that with diseases.

Considering the top 20 ranks, combinedCARE has the best performance in terms of coverage, whereas medCARE has in terms of average rank. combinedCARE calculates the similarity between the test patient and a training patient as the sum of disease-wise and medication-wise similarity. The best coverage performance for combinedCARE indicates that there is a change in the prediction score, when compared to the baseline score (the random expectation of a disease), for a higher frequency of diseases for combinedCARE. However, this increase in the prediction score is not very high, indicated by a higher average rank. On the other hand, the increase in the prediction score for certain diseases are quite significant for medCARE. The higher increase in the score places a disease at a lower rank in the ranked list of predictable diseases; and hence, the lowest average rank for medCARE. Considering all the cases, both medCARE and combinedCARE outperform CARE, in terms of both coverage and average rank. Even though there is not a significant improvement in coverage, we notice a considerable improvement in the average rank. Medication similarity increases the score associated

TABLE II
EVALUATION METRICS FOR CARE, MEDCARE, COMBINEDCARE

	Baseline	CARE	combinedCARE	medCARE
Top 20				
Coverage	0.4493	0.6219	**0.6621**	0.6296
Average Rank	3.6730	3.5848	3.5193	**3.4329**
Top 50				
Coverage	0.7147	0.7806	**0.7957**	0.7835
Average Rank	12.5711	8.9308	**7.9543**	8.5835
Top 100				
Coverage	0.8297	0.8679	**0.8715**	0.8713
Average Rank	21.0193	15.0791	**13.4179**	14.2102
All				
Coverage	**0.9908**	**0.9908**	**0.9908**	**0.9908**
Average Rank	52.9883	41.2092	**39.1286**	41.1020

with each predicted diagnosis, implying the confidence in the prediction. The medication-wise similarity has no direct relationship with disease-wise similarity i.e. the medication-wise similarity can be greater than the corresponding disease-wise similarity. So, overall score for a particular disease can increase significantly, when the medications are considered for calculating the similarity between the test patient and a training patient. This increase places some of the diseases higher up in the ranked list, and hence, the decrease in the average rank. The coverage values are more or less the same because we are considering only those patients for training who share at least one diagnosis with the test patient. Coverage depends on the count of correctly predicted diseases. This count is more or less the same for all the models.

VII. DISCUSSIONS

We also analyzed the performance of the models for various ICD-9-CM categories. The disease categories are listed in Table III.

A. Performance Trends on ICD-9-CM Disease Categories

We analyzed the rank of a disease assigned by each of CARE, medCARE and combinedCARE. The analysis was done for top 20, 50 and 100 ranks. We further analyzed the ranking performance in the broad disease categories of ICD-9-CM codes in the Tabular Index of Diseases (Table III). For each disease of the test patient for all our testing iterations, we gave a score of 1 to the best performing model(s) i.e. the model(s) that gave the lowest rank to a particular disease. We aggregated the scores for each model for each disease category (mentioned in Table III) and for Top 20, Top 50 and Top 100 ranks.

For all the disease categories, it was observed that if the model performs best for Top 20, it performs well for Top 50 and Top 100. The same observation holds for rank of a disease. None of the models were able to predict E codes. E codes are, generally, used in conjunction with a primary ICD-9-CM code (numeric codes); and provides information on external causes of injury. They provide additional or specific information. Hence, the identification or prediction of E codes carries less significance than that of a primary ICD-9-CM code. combinedCARE was able to predict more than 50% of the diseases designed by V codes. The V codes are used to code supplementary information regarding health status like any procedure related to an existing diagnosis, family history of diseases and personal history of diseases. This very nature of V codes makes it very hard to predict.

CARE outperformed combinedCARE and medCARE in two categories of diseases: "Congenital Anomalies" (ICD-9-CM 740-759) and "Injury AND Poisoning" (ICD-9-CM 800-999). The most commonly occurring ICD-9-CM code in the "Congenital Anomalies" is "Other congenital musculoskeletal anomalies" (ICD-9-CM 756). This code corresponds to impairments in the musculoskeletal systems. These diseases can only be surgically treated, and hence, little or no long term medications can be prescribed [33]. One of the known disadvantages of collaborative filtering is that it does not perform well, in case of no or low frequency of items under consideration. Hence, little or no prescribed medications causes medCARE and combinedCARE to perform poorly on this category of diseases.

Medication-wise similarity influences the list of predicted diseases. combinedCARE and medCARE performs well in 6 categories each. combinedCARE performs well in the following categories:

- Endocrine, Nutritional and Metabolic Diseases, and Immunity Disorders (ICD-9-CM 240-279)
- Diseases of the nervous systems and the central organs (ICD-9-CM 320-389)
- Diseases of the circulatory system (ICD-9-CM 390-459)
- Complications of pregnancy, childbirth, and the puerperium (ICD-9-CM 630-679)
- Diseases of the musculoskeletal system and connective tissue (ICD-9-CM 710-739)
- V codes

medCARE performs well in the following categories:

- Diseases of the blood and blood-forming organs (ICD-9-CM 280-289)
- Mental disorders (ICD-9-CM 290-319)
- Diseases of the respiratory system (ICD-9-CM 460-519)
- Diseases of the digestive system (ICD-9-CM 520-579)
- Diseases of the genitourinary system (ICD-9-CM 580-629)
- Diseases of the skin and subcutaneous tissue (ICD-9-CM 680-709)

TABLE III
ICD-9-CM DISEASE CATEGORIES : TABULAR INDEX

ICD-9-CM Disease Categories	ICD-9-CM Code Ranges
Infectious And Parasitic Diseases	001-139
Neoplasms	140-239
Endocrine, Nutritional And Metabolic Diseases, And Immunity Disorders	240-279
Diseases Of The Blood And Blood-Forming Organs	280-289
Mental Disorders	290-319
Diseases Of The Nervous System And Sense Organs	320-389
Diseases Of The Circulatory System	390-459
Diseases Of The Respiratory System	460-519
Diseases Of The Digestive System	520-579
Diseases Of The Genitourinary System	580-629
Complications Of Pregnancy, Childbirth, And The Puerperium	630-679
Diseases Of The Skin And Subcutaneous Tissue	680-709
Diseases Of The Musculoskeletal System And Connective Tissue	710-739
Congenital Anomalies	740-759
Certain Conditions Originating In The Perinatal Period	760-779
Symptoms, Signs, And Ill-Defined Conditions	780-799
Injury And Poisoning	800-999
Supplementary Classification Of Factors Influencing Health Status And Contact With Health Services	V01-V89
Supplementary Classification Of External Causes Of Injury And Poisoning	E800-E999

- Symptoms, signs, and ill-defined conditions (ICD-9-CM 780-799)

B. Performance Breakdown for ICD-9-CM Codes (Diseases)

There are multiple occasions where any two of the methods perform well in predicting a disease and assigning the same rank, higher than that assigned by the leave-out model. "Special investigations and examinations" (ICD-9-CM V72) was the most frequently ranked disease by CARE and combinedCARE simultaneously. On other hand, the same disease "Other and unspecified disorders of back" (ICD-9-CM 724) was ranked the same and the highest number of times by the pairs: CARE and combinedCARE; and medCARE and combinedCARE. Further, we observed that the two models that performed well in the Top 20, also, performed well in Top 50 and Top 100, as well as the exhaustive list of diseases predicted. Thus, the performance in assigning a rank to a disease for the test patient among Top 20 ranks is a strong indicator of how the models will perform when considering higher values of K for Top K positions.

For cases where only a single model performed best, the top 20 rank performance determined the performance in top 50, top 100 ranks as well as overall. CARE performed best for "Other and unspecified disorders of back" (ICD-9-CM 724); medCARE for "Symptoms involving respiratory system and other chest symptoms" (ICD-9-CM 786), combinedCARE for "Intervertebral disc disorders" (ICD-9-CM 722). There were a number of instances when all the models calculated the same rank for a disease. Some of the same ranked diseases are "Hypertension" (ICD-9-CM 401), "Hyperlipidemia" (ICD-9-CM 272), "Asthma" (ICD-9-CM 345), and "Diabetes mellitus" (ICD-9-CM 250). These diseases are chronic in nature and are some of the highly prevalent diseases in our dataset.

VIII. CONCLUSIONS

The main goal of this paper was to see whether incorporating medication history in addition to disease diagnoses can improve the predictability of diseases, and further enhance the performance of disease prediction systems for personalied and population healthcare management. The models, proposed in this paper, utilize the basic framework of an existing disease prediction system (CARE) that calculates disease-based similarity for generating the prediction score of a disease for a patient. The only difference was the use of prescribed medication history in calculating the similarity between a test patient and a training patient. We found that medication-based similarity gave a better performance than that compared to that of disease-based similarity in predicting future disease risks. We also found that the combination of medication-based and disease-based similarity outperform the individual component models. Our results show that medications provide additional utility in exposing underlying disease conditions, and improve the predictability of future diseases. These future disease risks, as identified by the models discussed in this paper, could act as early warning indicators for taking preventative measures, and help develop a personalized disease management and wellness plan for an individual.

ACKNOWLEDGMENTS

This research is supported in part by NSF grants IIS-1447795 and BCS-1229450.

REFERENCES

[1] International classification of diseases, 9th revision, clinical modification (ICD-9-CM). *NC for Health Statistics*, 2007.
[2] R. P. Ames. Hyperlipidemia in hypertension: causes and prevention. *American Heart Journal*, 122(4):1219–1224, 1991.
[3] ARRA. *ARRA*, 2009.
[4] M. A. Barrett, O. Humblet, R. A. Hiatt, and N. E. Adler. Big data and disease prevention: From quantified self to quantified communities. *Big Data*, 1(3):168–175, 2013.
[5] A. Belle, M. A. Kon, and K. Najarian. Biomedical informatics for computer-aided decision support systems: a survey. *The Scientific World Journal*, 2013, 2013.
[6] W. Castelli. Epidemiology of coronary heart disease: the framingham study. *The American Journal of Medicine*, 76(2):4–12, 1984.

[7] N. V. Chawla and D. A. Davis. Bringing big data to personalized healthcare: A patient-centered framework. *Journal of General Internal Medicine*, 28(3):660–665, 2013.

[8] K. L. Chen and H. Lee. The impact of big data on the healthcare information systems. *ICHITA-2013 TRANSACTIONS*, page 43, 2013.

[9] D. A. Davis and N. V. Chawla. Exploring and exploiting disease interactions from multi-relational gene and phenotype networks. *PloS one*, 6(7):e22670, 2011.

[10] 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.

[11] F. Folino and C. Pizzuti. A comorbidity-based recommendation engine for disease prediction. In *Computer-Based Medical Systems (CBMS), 2010 IEEE 23rd International Symposium on*, pages 6–12. IEEE, 2010.

[12] F. Folino and C. Pizzuti. Combining markov models and association analysis for disease prediction. In *Information Technology in Bio-and Medical Informatics*, pages 39–52. Springer, 2011.

[13] F. Folino, C. Pizzuti, and M. Ventura. A comorbidity network approach to predict disease risk. In *Information Technology in Bio-and Medical Informatics, ITBAM 2010*, pages 102–109. Springer, 2010.

[14] C. for Disease Control and Prevention. International classification of diseases - 9, abbreviated titles. [online], 2016. http://wonder.cdc.gov/wonder/sci_data/codes/icd9/type_txt/icd9abb.asp.

[15] A. S. Go, J. Yang, L. M. Ackerson, K. Lepper, S. Robbins, B. M. Massie, and M. G. Shlipak. Hemoglobin level, chronic kidney disease, and the risks of death and hospitalization in adults with chronic heart failure the anemia in chronic heart failure: outcomes and resource utilization (anchor) study. *Circulation*, 113(23):2713–2723, 2006.

[16] E. Grossman and F. Messerli. Hypertension and diabetes. 2008.

[17] C. A. Hidalgo, N. Blumm, A.-L. Barabási, and N. A. Christakis. A dynamic network approach for the study of human phenotypes. *PLoS Comput Biol*, 5(4):e1000353, 2009.

[18] HiTECH. *HiTECH*, 2009.

[19] J. Hyttel. Citalopram-pharmacological profile of a specific serotonin uptake inhibitor with antidepressant activity. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 6(3):277–295, 1982.

[20] X. Ji, S. Chun, and J. Geller. A collaborative filtering approach to assess individual condition risk based on patients' social network data. In *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 639–640. ACM, 2014.

[21] X. Ji, S. A. Chun, J. Geller, and V. Oria. Collaborative and trajectory prediction models of medical conditions by mining patients' social data. In *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on*, pages 695–700. IEEE, 2015.

[22] R. Kost, B. Littenberg, and E. S. Chen. Exploring generalized association rule mining for disease co-occurrences. In *AMIA Annual Symposium Proceedings*, volume 2012, page 1284. American Medical Informatics Association, 2012.

[23] I. Kurbasic, H. Pandza, I. Masic, S. Huseinagic, S. Tandir, F. Alicajic, and S. Toromanovic. The advantages and limitations of international classification of diseases, injuries and causes of death from aspect of existing health care system of bosnia and herzegovina. *Acta Informatica Medica*, 16(3):159, 2008.

[24] T. H. McCormick, C. Rudin, D. Madigan, et al. Bayesian hierarchical rule modeling for predicting medical conditions. *The Annals of Applied Statistics*, 6(2):652–668, 2012.

[25] T. O'Brien, S. F. Dinneen, P. C. O'Brien, and P. J. Palumbo. Hyperlipidemia in patients with primary and secondary hypothyroidism. In *Mayo Clinic Proceedings*, volume 68, pages 860–866. Elsevier, 1993.

[26] K. J. O'malley, K. F. Cook, M. D. Price, K. R. Wildes, J. F. Hurdle, and C. M. Ashton. Measuring diagnoses: Icd code accuracy. *Health Services Research*, 40(5p2):1620–1639, 2005.

[27] T. Opsahl. Triadic closure in two-mode networks: Redefining the global and local clustering coefficients. *Social Networks*, 35(2):159–167, 2013.

[28] T. Opsahl, F. Agneessens, and J. Skvoretz. Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*, 32(3):245–251, 2010.

[29] B. Qian, X. Wang, N. Cao, H. Li, and Y.-G. Jiang. A relative similarity based method for interactive patient risk prediction. *Data Mining and Knowledge Discovery*, 29(4):1070–1093, 2015.

[30] A. K. Rider and N. V. Chawla. An ensemble topic model for sharing healthcare data and predicting disease risk. In *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*, page 333. ACM, 2013.

[31] K. Steinhaeuser and N. V. Chawla. A network-based approach to understanding and predicting diseases. In *Social computing and behavioral modeling*, pages 1–8. Springer, 2009.

[32] G. Stiglic, I. Pernek, P. Kokol, and Z. Obradovic. Disease prediction based on prior knowledge. In *Proceedings of the ACM SIGKDD Workshop on Health Informatics, in Conjunction with 18th SIGKDD Conference on Knowledge Discovery and Data Mining*, 2012.

[33] WHO. Congenital anomalies. [online], 2015. http://www.who.int/mediacentre/factsheets/fs370/en/.