RESEARCH-ARTICLE

Predicting individual disease risk based on medical history

DARCY A DAVIS, University of Notre Dame, Notre Dame, IN, United States

NITESH V. CHAWLA, University of Notre Dame, Notre Dame, IN, United States

NICHOLAS BLUMM, Northeastern University, Boston, MA, United States

NICHOLAS A CHRISTAKIS, Harvard Medical School, Boston, MA, United States

ALBERT LÁSZLÓ BARABÁSI, Northeastern University, Boston, MA, United States

Published: 26 October 2008

CIKM08: Conference on Information and Knowledge Management
October 26 - 30, 2008
California, Napa Valley, USA

Conference Sponsors:
SIGIR
SIGWEB

CIKM '08: Proceedings of the 17th ACM conference on Information and knowledge management (October 2008)
https://doi.org/10.1145/1458082.1458185
ISBN: 9781595939913

Predicting Individual Disease Risk Based on Medical History

Darcy A. Davis
University of Notre Dame
ddavis4@nd.edu

Nitesh V. Chawla*
University of Notre Dame
nchawla@nd.edu

Nicholas Blumm
Northeastern University
nblumm@ccs.neu.edu

Nicholas Christakis
Harvard Medical School
christak@hcp.med.harvard.edu

Albert-László Barabási
Northeastern University
alb@neu.edu

ABSTRACT

The monumental cost of health care, especially for chronic disease treatment, is quickly becoming unmanageable. This crisis has motivated the drive towards preventative medicine, where the primary concern is recognizing disease risk and taking action at the earliest signs. However, universal testing is neither time nor cost efficient. We propose CARE, a Collaborative Assessment and Recommendation Engine, which relies only on a patient's medical history using ICD-9-CM codes in order to predict future diseases risks. CARE uses collaborative filtering to predict each patient's greatest disease risks based on their own medical history and that of similar patients. We also describe an Iterative version, ICARE, which incorporates ensemble concepts for improved performance. These novel systems require no specialized information and provide predictions for medical conditions of all kinds in a single run. We present experimental results on a Medicare dataset, demonstrating that CARE and ICARE perform well at capturing future disease risks.

Categories and Subject Descriptors

I.2.1 [**Artificial Intelligence**]: Applications and Expert Systems—*Medicine and science*; I.5.1 [**Pattern Recognition**]: Models—*statistical*; J.3 [**Computer Applications**]: Life and Medical Sciences—*Medical information systems*

General Terms

Algorithms, Experimentation

Keywords

collaborative filtering, disease risk prediction, ensemble, prospective health care

*Corresponding Author

CIKM'08, October 26–30, 2008, Napa Valley, California, USA.

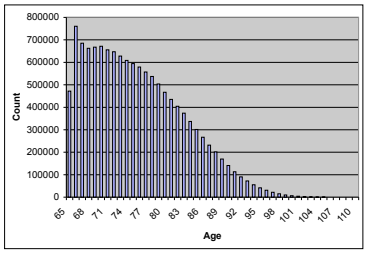
1. INTRODUCTION

Medical care and research are literally the most vital part of science for humans, as none of us are immune to physical ailments and biological deterioration. Annual health care expenditure in the U.S. alone is an overwhelming sum, with a strong majority of this money used for chronic disease treatment. Experts expect the burden on the system to continually increase in coming years. The rapidly increasing medical concerns of the baby boomer generation is one major factor stressing the health care system. A CDC study estimates that 880.5 million visits were made to physician offices, about 3.1 visits per patient, in 2001 [3]. Since 1992, the average age increased to 45 years, and the visit rate for persons 45 years of age and over increased by 17% from 407.3 to 478.2 visits per 100 persons [3].

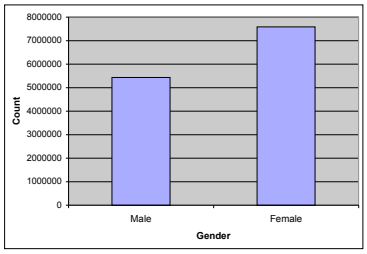
Health care, thus, needs to become more proactive than reactive in recognizing the onset of disease and risk. However, the combinatorial problem generated by the different disease factors and the previous medical history of a patient is so complex that no single health care professional can fully comprehend it all. Currently, physicians can use family and health history and physical examination to approximate the risk of a patient, guiding laboratory tests to further assess the patient's stage of health. However, these sporadic and qualitative 'risk assessments' generally focus on only a few diseases and are limited by a particular doctor's experience, memory, and time. Therefore, current medical care is reactive, stepping in once the symptoms of a disease have emerged, rather than proactive, treating or eliminating a disease at the earliest signs.

Today the prevailing model of prospective heath care is firmly based on the genome revolution. Indeed, technologies ranging from linkage equilibrium and candidate gene association studies to genome wide associations have provided an extensive list of disease-gene associations, offering us detailed information on mutations, SNPs, and the associated likelihood of developing specific disease phenotypes [4]. The underlying hypothesis behind this line of research is that once we catalogue all disease-related mutations, we will be able to predict the susceptibility of each individual to future diseases using various molecular biomarkers, ushering us into an era of predictive medicine. Yet, these rapid advances have also unraveled the limitations of the genome based approaches [12].

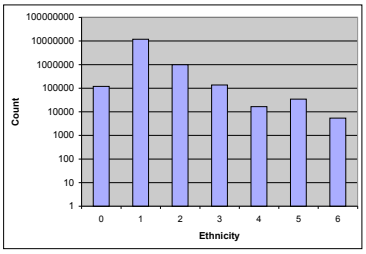
Given the weak signals that most disease associated SNPs or mutations offer, it is increasingly clear that the promise



(a) Patients by age.



(b) Patients by gender.



(c) Patients by Race (Logplot).

Figure 1: Data Statistics.

of the genome based approaches may not be realized soon. Does this mean that prospective approaches to health care will have to wait until the genomic approaches sufficiently mature? Our aim here is to show that phenotype and disease history based approaches offer the promise of rapid advances towards disease prediction. Recent literature further justifies the move towards a more prospective medical care system [9, 16].

Related research has largely focused on computer-aided medical prediction systems. One very widely used system is the Apache III [17], a prognostic scoring system for predicting inpatient mortality. Apache uses a combination of acute physiological measurements, age, and chronic health status. There are also a number of systems developed for predicting risk of individual diseases, such as specific heart conditions [5], hepatitis [15], Alzheimer's disease [11], etc.

Our approach is distinctly different in that we are trying to build a *general* predictive system which can utilize a less constrained feature space, i.e. taking into account all available demographics and previous medical history. Moreover, we rely primarily on ICD-9-CM (International Classification of Diseases, 9th revision, Clinical Modification) insurance codes (see Section 1) for making predictions to account for the previous medical history, rather than specialized test results.

1.1 Contribution

This research seeks to aid the development of a predictive system by examining the use of medical history to examine information about disease correlations and inexpensively assess risk. An effective proactive approach requires an understanding of disease interdependencies and how they translate into a patient's future. Due to common genetic, molecular, environmental, and lifestyle-based individual risk factors, most diseases do not occur in isolation [1, 4, 13]. Shared risk and environmental factors have similar consequences, prompting the co-occurrence of related diseases in the same patient. Therefore, a patient diagnosed for a combination of diseases and exposed to specific environmental, lifestyle and genetic risk factors may be at considerable risk of developing several other genetically and environmentally related diseases.

How can we exploit such interconnections and generate predictions about the future diseases a patient may develop? The underlying thesis of our work is to generate a patient's prognosis based on the experiences of other similar patients. We attempt to build on these relationships across millions of patients to effectively determine a prediction for a patient. Our goal is to provide every patient with an personalized answer to the question: *What are my disease risks?*

We approach this problem using collaborative filtering methodology. Collaborative filtering is designed to predict the preferences of one person (active user) based on the preferences of other similar persons (users). The technique is based on the intuitive assumption that people will enjoy the same items as their similar peers, or more specifically, having some common preferences is a strong predictor of additional common preferences. Predictions are based on datasets consisting of many user profiles, each containing information about the individual user's preferences. This has made a significant impact on marketing strategies. We draw an analogy between marketing and medical prediction. Each user is a patient whose profile is a vector of diagnosed diseases. Using collaborative filtering, we can generate predictions on other diseases based on a set of other similar patients. However, the ratings in our case are binary – a patient either has a disease (1) or does not have a disease (0). There is no ordinal set of ratings as is typically observed in movie or music data. Another difference is that the users choose to rate movies and music, while the diseases are not a patient choice, per se.

Key contributions in this work include the following:

1. A novel application of collaborative filtering in the medical domain for advancing the field of prospective medicine. To our knowledge, collaborative filtering has not been used for disease prediction. Unlike other disease prediction software, we present a general system which makes predictions on all types of diseases and medical conditions. Another novelty in our work is the use of ICD-9-CM codes [6] data for building our collaborative filtering based predictive models. We do not require any other information such as lab tests, etc., which can be expensive.

2. The collaborative filtering employed, while building upon prior work, incorporates new elements of clustering, significance testing, and ensemble methods within the CARE framework.

3. A case study is provided as a real-world example of potential benefits of CARE.

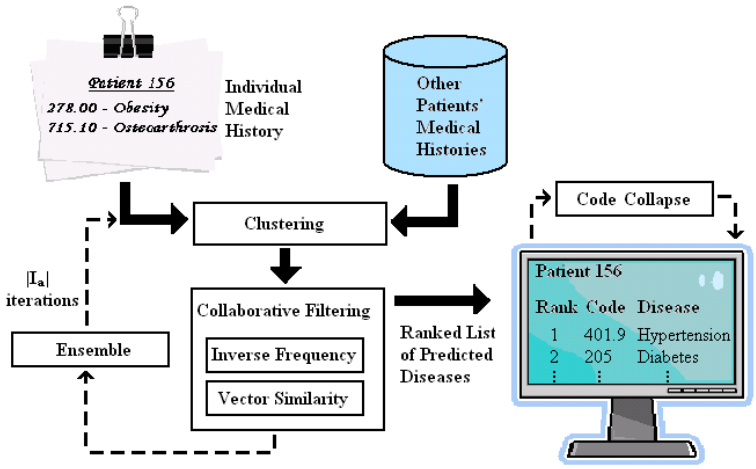


Figure 2: A high-level overview of the CARE system

Disease	Prevalence
unspecified essential hypertension	33.64%
coronary atherosclerosis	21.16%
congestive heart failure	18.16%
urinary tract infection	16.67%
chronic airway obstruction	14.69%
atrial fibrillation	14.03%
volume depletion	11.90%
hypopotassemia	11.34%
diabetes uncomplicated type II	10.47%
pneumonia, organism unspecified	9.35%
angina, unstable	8.72%
hyposmolality and/or hyponatremia	8.47%
unspecified anemia	8.38%
acute posthemorrhagic anemia	8.14%
unspecified angina pectoris	7.90%
hyperplasia of prostate	6.54%
other spec cardiac dysrhythmias	5.61%
osteoarthros uns gen/loc uns site	5.20%
unspecified hypothyroidism	5.14%
unspec chronic ischemic hrt disease	5.13%

Table 1: The 20 most prevalent diseases.

2. DATA

Our entire database comprises the Medicare records of 13,039,018 elderly patients in the United States with a total of 32,341,348 hospital visits. Such Medicare records are highly complete and accurate, and they are frequently used for epidemiological and demographic research [10, 14]. Our data is completely anonymized; that is we have no mean to identify the patient or the hospital the patient visited. The input for our methods consists of each patient's diagnosis history, provided per inpatient visit. Each data record represents a hospital visit, represented by a patient ID and a list of up to ten diagnosis codes, as defined by the *International Classification of Diseases, Ninth Revision, Clinical Modification* (ICD-9-CM). The International Statistical Classification of Diseases and Related Health Problems provides codes to classify disease and a wide variety of signs, symptoms, abnormal findings, social circumstances, and external causes of injury or disease. It is published by the World Health Organization. Each disease or health condition is given a unique code, and can be up to 5 digits long. However, the 5 digit codes can be collapsed for some disease to fewer characters for identifying a family of diseases. In the Medicare data, the first code is the principal diagnosis, followed by any secondary diagnoses made during the same visit. A sample patient medical history is shown in Table 2; each line represents one hospital visit.

Patient ID	Vector of ICD-9-CM Disease Codes
9142409	40291 57420 5301 5533 2780
9142409	29624 4019 2768 2780
9142409	2967
9142409	25090 7906 E9331 20300
9142409	25090 E9331 20300 4019
9142409	3101 20300 25001

Table 2: A sample patient medical history

The number of visits per patient ranges from 1 to 155, with a median of 2. Also, though up to ten diagnosis codes are permitted, the average is only 4.32 per visit. There are a total of 18,207 unique disease codes expressed in the database. However, only 169 diseases occur at 1% or more in the population (across visits for patients). Table 1.1 shows the 20 most prevalent diseases in our database.

Demographic data was also available and was used to examine CARE's predictive power. Figure 1 shows the distribution of patients across the demographics of age, gender, and ethnicity. The different races are coded as 0 through 6 in our database. In spite of being a relatively homogeneous database (all senior citizens), there is still significant diversity in the age, gender, and ethnicity distribution. Since these factors are known to influence certain medical conditions, we use them to partition our database and perform experiments on the demographic-specific subsets.

3. THE CARE METHODOLOGY

3.1 System Overview

Before detailing the individual components, a high-level preview of the entire CARE framework is provided in Figure 2. The dotted lines represent optional methods. As

shown, both testing and training data enter the system at the same time. We form a cluster of relevant patients based on the 'known diseases' of the testing patient. Collaborative filtering is performed on the resulting cluster, generating predictions for the future visits of the testing patient. Each component is further defined in the subsequent sections. In the case of ICARE, this process is performed multiple times for each patient, with each iteration using a different basis for clustering. These different clusterings are combined to form an ensemble. The output after CARE and ICARE is a ranked list of diseases in the subsequent visits of the testing patient, ranked in order from the highest risk score to the lowest. If desired, the output can be easily collapsed into less specific groups of medical conditions due to the hierarchical nature of the disease codes.

3.2 Vector Similarity

Our collaborative filtering technique is derived from the vector similarity algorithm presented by [2]. Traditionally, collaborative filtering is used to make a prediction $p(a, j)$ of the likelihood of user a, the active user (testing), on item j based on the similarity between user a and every member of the set I_j who have previously rated that item. The similarity $w(a, i)$ between users a and i is calculated by vector similarity; that is,

$$w(a,i) = \sum_j \frac{v_{a,j}}{\sqrt{\sum_{k \in J_a} v_{a,k}^2}} \frac{v_{i,j}}{\sqrt{\sum_{k \in J_i} v_{i,k}^2}}. \quad (1)$$

J_i is the set of items rated by user i. The prediction and similarity weight takes into account the average vote $\overline{v}_i$ of each user to account for personal differences. A normalizing constant κ is added so that the sum of weights is equal to 1, constraining the prediction within the range of possible votes (in this case, 0 and 1). Thus, the general collaborative filtering equation is:

$$p(a,j) = \overline{v}_a + \kappa \sum_{i \in I_j} w(a,i)(\overline{v}_{i,j} - \overline{v}_i). \quad (2)$$

However, this equation will not suffice for the proposed application in the medical domain. The user in this case is a patient and the items are diseases. Each patient i either has ($v_{i,j} = 1$) or does not have (no vote) disease j. Since every vote is 1, it is easy to see that every $\overline{v}$ term will be one, and the algorithm then predicts that every user has every disease with a likelihood of 1, an obvious error. The proposed changes modify the general equation to incorporate binary the diagnoses and remove the effect of the range of ratings. The general equation 2 is also modified to be dependent on the average number of occurrences, or random expectation, of each disease. This average is referred to as the baseline prediction about the disease, $\overline{v}_j$. Thus, the likelihood of the active patient a on disease j can be expressed as follows:

$$p(a,j) = \overline{v}_j + \kappa(1 - \overline{v}_j) \sum_{i \in I_j} w(a,i) \quad (3)$$

with the normalizing constant

$$\kappa = \frac{1}{\sum_{i \in I} w(a,i)}$$

Intuitively, the equation treats the random expectation $\overline{v}_j$ as the baseline expectation of each patient having disease j, and adds additional risk based on similarity to other patients with disease j.

3.3 Inverse Frequency

We further extended Equation 1 to include inverse frequency (IF), which gives lower weights to very common diseases in the training set, based on the intuition that sharing a rare disease has more impact on similarity than sharing a common disease. For instance, individuals sharing a rare genetic disease are assumed to be more similar than two patients with general hypertension. Furthermore, two patients with the same disease are considered more similar if they share a specific type of complication. This is particularly influential in our medical database. There can be many medical diagnoses shared between patients but most important contributions arise from uncommon connections. The inverse frequency of disease j is defined as

$$f_j = log \frac{n}{n_j} \quad (4)$$

where n is the number of patients in the training set, and n_j is the number of patients who have j. This is incorporated into the similarity weighting equation by multiplying each disease vote by the corresponding IF factor. The resulting equation for $w(a, i)$ is

$$w(a,i) = \sum_j \frac{f_j v_{a,j}}{\sqrt{\sum_{k \in J_a} f_k^2 v_{a,k}^2}} \frac{f_j v_{i,j}}{\sqrt{\sum_{k \in J_i} f_k^2 v_{i,k}^2}}. \quad (5)$$

No changes to the general equation are needed. All of the experimental results discussed were found using this method, which we call inverse frequency vector similarity (**IFVS**).

3.4 Clustering

We cluster patients on the basis of shared diseases. Before each application of collaborative filtering, clustering is applied to the training set to discover connected components of patients. This serves to remove the influence of patients who have little or no similarity with the testing patient for whom predictions are being made. This is determined by the number of diseases which the patients have in common. In the most basic case, patients are removed only if they have no diseases in common with the active patient. It can be seen in Equation 5 above that these patients have a weight of zero and do not contribute to the prediction scores. Thus, removing these patients does not result in loss of information, but effectively reduces the runtime of the algorithm. In practice, we cluster such that all patients in the training set have two or more diseases in common with the known diagnoses of the active patient.

Introducing the constraint that clustering patients in the training set must have at least two common diseases with the active (testing) patient enforces stronger similarities for all patients influencing the predictions. Essentially, we build a network of patients that are connected by at least two diseases and then perform collaborative filtering in this network. In theory, this helps to avoid the noise resulting from common diseases that introduce a very high number of weak influences. The clustering provides an additional benefit by reducing the number of diseases predicted on, which both simplifies and improves the collaborative filtering results. This effect will be further discussed in the next chapter. It is important to note that the frequency of diseases is different within the cluster than the overall occurrence in the

entire dataset. We will refer the global v_j as the 'baseline' of disease j and the new v_{cj} after clustering as the 'cluster baseline' of disease j. In all experiments where clustering is employed, the cluster baseline is used in the IFVS equation.

3.5 ICARE with Ensembles

Even with the double-overlap clustering method combined with IFVS, we still observed that common diseases can dominate the effect of collaborative filtering since they account for the majority of the patients in the cluster. Ideally, we want to capture the effect of each individual disease with minimal noise from other diseases, but without the loss of information due to removing them. To meet this goal, we developed an iterative version of CARE using ensembles of individual-disease clusters. Specifically, for each disease j developed by the test patient a, collaborative filtering is applied only to the cluster of training patients with disease j. As before, the collaborative filtering scores build onto the cluster baselines. Each component of the ensemble is a round of collaborative filtering on an individual disease cluster, and it follows that the number of components is equal to the number of unique diseases which patient a has had. Within each component, the collaborative filtering still uses the entire past disease vector of patient a. Thus, each disease has a chance at making a strong impact individually, but all disease interactions are preserved. The ensembles are combined by taking the maximum prediction score for each disease, that is

$$\max_{c \in C} \left(\overline{v}_{j,c} + \kappa(1 - \overline{v}_{j,c}) \sum_{i \in I_{j,c}} w(a,i) \right) \tag{6}$$

where C is the set of clusters c. We choose the maximum since diseases are generally not protective against each other, with few exceptions. In other words, additional unrelated diseases do not lessen the probability of developing a disease. Such ensembles can be easily run in a distributed fashion as each cluster evaluation is independent of the other.

In order to reduce the number of predictions and the runtime of the ensembles, we only predict on diseases for which the cluster baseline is significantly higher than the population baseline. That is, if the population baseline is larger than the cluster baseline, then the disease being predicted on does not have a good set of predictive diseases in the cluster. We determine the significance of a disease in the cluster using a difference of proportions test. This statistical test determines whether the difference between two sample proportions taken from different populations is significant. The null hypothesis is always that the two proportions are equivalent, and the alternative hypothesis is that they are not equivalent. A z score is then found using the equation

$$z = \frac{p_1 - p_2}{S_{p_1 - p_2}}. \tag{7}$$

Here, $p_1 - p_2$ is the difference between the sample proportions and S is the associated standard error determined by the equation

$$S_{p_1 - p_2} = \sqrt{\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}} \tag{8}$$

where p is the weighted average of p_1 and p_2, while n_1 and n_2 are the respective sizes of the samples. In our formulation, p_1 is the cluster baseline, p_2 is the population baseline, n_1 is the number of patients in the cluster, and n_2 is the number of training patients. We use a 95% confidence interval.

3.6 ICD-9-CM Code Collapse

In some cases, it is desirable for all 4 or 5-digit ICD-9-CM codes to be collapsed into more general 3-digit codes, which represent small groups of related or similar diseases. In general, these groups are not based on comorbidity; they are often comprised of specific forms or complications of the same disease or injury. The grouping is based entirely on the structure of the ICD-9-CM coding scheme. For example, the ICD-9-CM code of 426 corresponds to *Conduction disorders*. The specific version of 426.0 corresponds to *Atrioventricular block, third degree*; this can be further specified as (426.11) *Atrioventricular block, first degree*; (426.12) *Atrioventricular block, Mobitz II* and (426.13) *Atrioventricular block, Wenckebach's*.

Such 4 or 5 digit codes can be truncated to 3 digits either before (pre-collapse) or after (post-collapse) applying collaborative filtering. In the first case, collaborative filtering is applied to vectors of already shortened codes. This significantly reduces the number of diseases being predicted, consequently reducing the runtime. However, pre-collapsing results in loss of all information provided by the more detailed codes, since only one composite prediction is made for each 3-digit disease group. When post-collapsing, the collaborative filtering is run normally on the original codes, and the results are merged after completion. The 3-digit code group adopts the highest prediction score given to one of the members. That is, the likelihood of having a general disease is equal to the highest likelihood of having some specific instance of the disease.

Post-collapsing can be done in a hierarchical manner, so that the detailed results provided by specific ICD-9-CM codes are preserved. Collapsing the ICD-9-CM codes is beneficial in multiple ways. In the case of pre-collapsing, algorithm efficiency is improved. In both cases, the reduced number of diseases predictions makes the results simpler to evaluate and interpret. Also, collapsing reduces the negative effects of assuming that all undiagnosed diseases are not present. For example, a high score for diabetes will be evaluated as a successful prediction of diabetes with a specific complication. Without collapsing, the relationship between the two diabetes codes could not be directly considered, and the rareness of the complication could cause the diabetes diagnosis to be overlooked or highly underrated. This is particularly relevant since Medicare data does not reliably capture complications [14]It is important to note that post-collapsing the codes does not change the performance of collaborative filtering; this method primarily serves to make evaluation of the performance more accurate, giving the medical practitioner the choice to conduct further tests to identify the specific nature of the disease.

4. EVALUATION

CARE and ICARE generate predictions only on 'future' visits of a patient in the testing set based on the medical history provided; that is, we only want to evaluate performance on diseases which happen on a later date than those that the collaborative filtering algorithm was given (akin to leave-one-out testing). For this reason, the collaborative filtering algorithm is given information about the active user one visit at a time, and performance is measured only in

terms of those diseases which occur in the following visits. The reported metrics are averaged across all future visits predictions across all testing patients.

It is difficult to determine whether an individual prediction is successful or not, since setting a threshold on the prediction score is unreasonable in this domain. The highest risk scores for one patient might be relatively low for another patient with more obvious concerns. We determine performance based on the overall list of predictions, ranked in order from the most likely to the least likely. Specifically, the diseases are given a rank k in order from highest prediction score p to the lowest, with the highest score having $k = 0$. Note that a baseline ranking can also be determined by ordering the diseases by their prevalence in the overall population. The performance measures on the baseline ranking serve as a benchmark for experiments; that is, a good collaborative filtering method should produce a significantly better ranking than one based solely on knowledge of disease prevalence. The baseline ranking is essentially the best guess for a patient for which no further information is known, or alternately, who is assumed to be equally similar to everyone in the database. Since our data is from a targeted group (senior citizens), the likelihoods of diseases are more meaningful than in a general database. We use three metrics to assess the baseline ranking and the prediction lists generated by CARE and ICARE.

The first performance metric is list *coverage*. A method's coverage is defined as the percentage of diseases for which a prediction is made and ranked. This is necessary since test patients occasionally express diseases which never occur in the training set, and significance testing can cause some diseases to be dropped from consideration. Obviously we wish to capture as many future diseases as possible, so high coverage is preferred. The *average rank* of future diseases is also used as an evaluation metric, since it is desirable for future diseases to have low rank positions. Ideally, the diseases which a patient actually develops should be near the top of the list, where they are most likely to be noticed and used.

The last metric is also based on this concept. Referred to as *half-life accuracy* [8], this metric is intended to measure the expected utility of the ranked list [7]. Based on the rank k, $p(k)$ is defined as the probability that a user reading the list would consider the disease in position k before stopping. The scenario is, given a long list, a user would start with the highest risk diseases, but will not read the entire list due to lack of time or further interest. Thus, $p(k)$ is an exponentially decaying function defined

$$p(k) = 2^{-k/a} \tag{9}$$

where a is a user-defined constant that determines the speed of decay. For our experiments, we use $a = 5$. The utility of the list is then

$$\text{utility} = \sum_k p(k)\delta_k \tag{10}$$

, where $\delta_k = 1$ for future diseases, and $\delta_k = 0$ otherwise. Intuitively, this means that utility is entirely based on how highly future diseases are ranked. The half-life accuracy is then defined as the average over all test patients i of the expected utility of the ranked list of predictions for i divided by the utility of a perfect ranking for i, where all future diagnoses are in the highest possible rank positions. That is,

$$\text{accuracy} = \frac{100}{N}\sum_{i=1}^{N}\frac{\sum_{k=0}^{R_i-1}\delta_{ik}p(k)}{\sum_{k=0}^{M_i-1}p(k)} \tag{11}$$

where N is the number of test users, R_i is the number of items that are predicted on for user i, and M_i is the number of diseases in R_i such that $\delta_{ik} = 1$. The denominator of the half-life accuracy measure is a per-user normalization, which takes into account the varying number of patient diagnoses.

As implied above, a doctor may not have time or interest for looking at the entire list of predictions, which can contain thousands of prediction scores in the worst case. A more attainable goal would be to consider only the top 20 or top 100 predictions. In addition to overall performance, we also consider the coverage, average rank, and half-life accuracy within those ranges. The performance on the top 20 or top 100 ranks is a much stronger measure of realistic usefulness than the overall results. Coverage is particularly important in these limited ranges. A doctor could conceivably consider all diseases on a list of 20, making actual rank less meaningful. However, each additional 'correct' prediction on the list could have a substantial impact. There is some tradeoff between average rank and coverage, since higher coverage captures less obvious diseases with lower rank.

ID Visit 1		Round 1: ID Visit 1
ID Visit 2	→	Round 2: ID Visit 1∪Visit 2
ID Visit 3		Round 3: ID Visit 1∪Visit 2∪Visit 3

Figure 3: Example of how patient visits are processed by the IFVS algorithm. ID refers to a patient ID.

5. EXPERIMENTS

For our experiments, we selected patients that had at least five visits recorded in our database to allow for sufficient patient history for both training and evaluation. We then randomly created equal sized training and testing sets; each testing patient was further evaluated using leave-one-visit-out validation. We first present and compare the performance of the baseline ranking, CARE, and ICARE. Then, we show the impact of collapsing of ICD-9-CM codes on the predictive performance. Finally, we analyze the effect of demographic-based segmentation.

5.1 CARE Performance

The predictions were generated on the future visits of a patient. Since the order of disease occurrence is necessary for making meaningful predictions, the testing set was left in the original format, with each visit as a separate record. Both CARE and ICARE make one round of predictions for each visit, adding the diagnoses of the next visit in each successive round while retaining all diagnoses from previous visits. The idea is that on round i, the algorithm 'knows' all diagnoses up through visit i, and is evaluated on ability to predict diagnoses which occur in visits $i + 1$ and on. Figure 3 provides a pictorial explanation of this process.

Table 3 presents the analysis. The baseline method corresponds to a list of the diseases ranked in order from highest baseline prevalence to lowest. As mentioned earlier, results on the top 100 and top 20 ranks are more meaningful, since

Comparison of Methods			
	Baseline	CARE	ICARE
Top 20			
Coverage	.283	.344	.413
Average Rank	7.504	7.819	5.771
Half-Life Accuracy	30.574	30.255	47.663
Top 100			
Coverage	.552	.606	.607
Average Rank	30.082	26.734	20.400
Half-Life Accuracy	31.115	30.759	50.346
All			
Coverage	.994	.940	.775
Average Rank	266.523	177.495	81.345
Half-Life Accuracy	31.115	30.759	50.346

Table 3: Evaluation of performance of CARE and ICARE compared with the baseline ranking.

Post-Collapsed Results			
	Baseline	CARE	ICARE
Top 20			
Coverage	.374	.405	.488
Average Rank	8.731	7.347	4.783
Half-Life Accuracy	35.396	34.786	47.930
Top 100			
Coverage	.774	.712	.678
Average Rank	41.443	25.998	17.592
Half-Life Accuracy	36.181	34.993	48.274
All			
Coverage	.999	.944	.781
Average Rank	153.008	101.750	48.060
Half-Life Accuracy	36.181	34.993	48.274

Table 4: Evaluation after post-collapsing ICD-9-CM codes.

a medical practitioner or other user is unlikely to consider a very large portion of the list. CARE shows significantly better performance than baseline across the board overall and in the top 100 ranks. In the top 20 ranks, CARE covers 6% more diseases than the baseline method with minimal impact on the average rank.

ICARE shows very substantial improvement over both the baseline and CARE in all cases. This method captures about 13% more of the future diseases than the baseline method in the top 20 rankings alone, while the average rank of 5.77 suggests that most of these captured diseases are in the first few positions on the list. It is particularly powerful that both average rank and coverage improve simultaneously, since there is some tradeoff between the two metrics. The most impressive result is that ICARE predicts more than 41% of all future diseases in the top 20 ranks, a list of a manageable size for use by a doctor or other medical professional.

It merits explanation that the half-life accuracy overall and in the top 100 are the same, although actually not identical at higher precision. This happens because of the way half-life accuracy is defined, where the utility decreases as a future disease moves down the list. The exponential decay is such that information beyond the top 100 ranks has minimal impact on the half-life accuracy. By modifying the a value defined in 4 to slow the decay, these accuracies could be forced to diverge. Regardless, it seems unreasonable that a medical professional would seriously consider the list beyond 100 diseases, making the equal utility realistic.

5.1.1 Collapsing ICD-9-CM codes

After the initial results, we also performed experiments using both the pre- and post-collapsing methods described earlier for condensing the disease codes. Our initial experiments showed the two methods tend to perform very similarly. Since the hierarchical nature of post-collapsing preserves all information, we determined it to be the better method and present results from post-collapsing. We believe that post-collapsing is a more amenable method to the eventual use of the system — it still provides a medical practitioner a choice to retrieve the complete resolution of ICD-9-CM code. If we pre-collapse the ICD-9-CM codes and then run CARE, the full resolution is lost. The results of our methods on the 3-digit ICD-9-CM codes are shown in Table 4.

The relative performance of the three methods shows very similar trends to those in Table 3. This is reasonable, since post-collapsing is actually a post-processing step applied to the earlier results. Post-collapsing results in an improvement in ranking and coverage across the board. ICARE shows a slight dip in the half-life accuracy measure. We believe this arises because of multiple high-ranking diseases collapsing to a common code, eliminating the dominance in the top ranks. Since, the half-life accuracy metric is strongly dependent on the highest ranks, the decrease occurs.

These results from collapsing of ICD-9-CM codes are very encouraging, with nearly 49% of future disease 'families' among the top 20 predictions. Still, it is an important distinction that the collapsed results are not necessarily better than the original 5-digit results if measured in terms of granularity. They are a more condensed but less detailed version of exactly the same results. However, this list could conceivably be used to present a medical practitioner with a greater breadth of predictions in the same concise format. The details could then be selectively considered, based on the hierarchy preserved by the post-collapsing method.

5.2 Demographic Experiments

In addition to overall performance, we experimented to see if demographic information can be used to improve the predictive performance of the CARE framework. It is well known that many biological, social, and environmental factors can influence health, so we hypothesized whether using more homogenous data sets would be beneficial. The demographics explored were age, gender, and race. The specific categories include both genders, 5 racial groups, and 7 age groups spanning 5 years each. We partitioned the training and testing sets based on the considered demographic categories. New experiments were run on each partition of the testing set using the corresponding training set. We present only the results on the top 20 ranks after using ICARE, as this method was shown in the previous section to be consistently superior and the highest ranks are most meaningful.

The demographic categories, prevalence statistics, and experimental results are shown in Table 5. We point out that we did not consider races 4, 5, and 6, since they had low prevalence rates of 0.13%, 0.26%, and 0.04%, respectively. For comparison, we also present the results on each demographic segment when the original data was used in the en-

Category		Original Training Data		Demographic Training Data	
Age	% Prevalence	Average Rank	Coverage	Average Rank	Coverage
65-69	19.59	5.951	0.401	5.301	.398
70-74	21.71	5.804	0.411	5.820	.411
75-79	23.33	4.912	0.406	5.831	.392
80-84	18.67	5.741	0.420	5.782	.392
85-89	11.51	5.609	0.419	5.754	.373
90-94	4.27	5.627	0.438	5.797	.353
95+	0.92	5.214	0.449	5.640	.423
Gender	% Prevalence	Average Rank	Coverage	Average Rank	Coverage
Male	40.12	5.841	0.408	5.834	.397
Female	59.88	5.724	0.415	5.743	.424
Race	% Prevalence	Average Rank	Coverage	Average Rank	Coverage
Race0	1.22	5.903	0.417	5.734	.413
Race1	89.22	5.753	0.411	5.893	.382
Race2	8.14	5.885	0.428	5.886	.434
Race3	1.09	6.112	0.385	6.179	.389

Table 5: Performance on different demographics using ICARE. For comparison, we include the results of using the aggregate population with no demographic specific modeling and prediction (Original Training Data) and the demographic specific segments (Demographic Training Data).

tirety. These are displayed under 'Original Training Data' in Table 5. There is clearly a marginal difference in the relative performance of the global method across the partitions. The older population and females generally have a higher coverage. Race 2 performs relatively better than the other races, while race 3 is the lowest across all the segments.

Application of ICARE after the demographic segmentation resulted in dip in performance, interestingly. The homogeneity of the group offered by demographic split does not seem to add any improvement to the performance of ICARE. We believe this is due to the "cluster ensemble" effect of ICARE that intrinsically identifies more similar groups.

Since each cluster in ICARE is simply a group of all patients expressing a disease, the demographic distribution for that disease is expressed within the cluster. For example, if it clusters on prostate cancer, this eliminates all the female patients and already generates a more homogeneous cluster of patients, reflective of the demographic of males. However, if one clusters on heart disease and the patient is a woman, the cluster could still carry prostrate cancer as the clustering constraint was heart disease and no gender information was implied. But we conjecture that because we are clustering on each individual disease in the history of the patient, some of the demographic distribution will begin to emerge with ICARE. Also, while the actual prevalence of a disease may vary within different demographic groups, the relative ordering of the diseases tends to be fairly similar across the categories, especially for common diseases. This means that a disease usually will have fairly similar baseline rank despite the demographic. Since we evaluate on relative rankings rather than thresholds, the actual prevalence percentage matters less than relative baseline ranking.

6. CASE STUDY

To demonstrate the CARE process proposed and applied in this work, we present case studies which place the algorithm results in the context of real patients. We look at the ranked list of disease predictions generated for a cancer patient after each of 3 subsequent hospital visits. This study was done using ICARE, which is demonstrated to be our best method. The patient diagnoses and top 10 predictions are provided in Figure 6.

We point the reader to the most prevalent diseases in Table 1.1. These are relevant to the case study since they pose the greatest challenge for other future diseases to overcome. It is worth noting that many of these diseases have been linked with one another in other medical studies. In fact, 4 out of the 10 are forms of heart disease, which has known links with hypertension and diabetes. This only serves to increase their influence.

Figure 4(a) shows the actual diseases developed by the patient. It is evident that we are dealing with a cancer patient. The first visit has the initial diagnosis of esophagal cancer, which spreads into secondary malignancies in the following visits. Since cancer is not a quickly treated disease, the original diagnoses recur in later visits. Since predicting these diseases is not interesting, we don't include them in the top 10 lists. In the final visit, the diagnoses diversify to include hypertension, regional enteritis, and a mineral deficiency. Figure 4(b) shows the results after applying ICARE to the first visit. Even from the first visit, we are able to predict the two locations of cancer spread with rank 3 and 4.

Figure 4(c) shows the prediction after the second visit is observed. Upon adding an additional form of cancer in the second visit, we see little change except for a slight reordering of the list. The space left after removing liver cancer was filled by urinary tract infection. This is a good example of prevalent diseases overtaking others once they make it through the significance test. Despite the fact that hypertension is the most prevalent disease in the database, we are not able to predict the occurrence in visit 3. This does not necessarily imply a mistake. Hypertension did not appear anywhere on the prediction list for visits 1 and 2. Considering the significance testing, this implies that it is not strongly connected to the cancers and thus should not be predictable. A similar argument applies for the enteritis. The disorder of mineral metabolism does appear in the rankings after the first two visits, at 71 and 83 respectively.

Visit 1
cancer, esophagus
chemotherapy

Visit 2
cancer, esophagus
chemotherapy
secondary cancer, liver

Visit 3
cancer, esophagus
chemotherapy
secondary cancer, lymph nodes
secondary cancer, liver
hypertension
regional enteritis
disorder of mineral metabolism

(a) Patient Diagnoses

1	cancer, bronchus/lung
2	anemia
3	secondary cancer, lymph nodes
4	secondary cancer, liver
5	pneumonia
6	secondary cancer, unspecified
7	disorder of fluid/elecrolyte balance
8	pleurisy
9	cancer, ovary
10	disease of white blood cells

(b) ICARE Prediction After Visit 1

1	cancer, bronchus/lung
2	anemia
3	urinary tract infection
4	secondary cancer, lymph nodes
5	disorder of fluid/elecrolyte balance
6	pneumonia
7	secondary cancer, unspecified
8	pleurisy
9	cancer, ovary
10	disease of white blood cells

(c) ICARE Prediction After Visit 2

1	disorder of fluid/elecrolyte balance
2	unspecified anemia
3	cancer, breast
4	congestive heart failure
5	cardiac dysrhythmias
6	acute ischemic heart disease
7	other digestive system complications
8	cancer, bronchus/lung
9	urinary tract infection
10	pneumonia

(d) ICARE Prediction After Visit 3

Figure 4: Case Study

This acknowledges a significant link to the disease,placing it still within the top 100 but not among the strongest concerns. The predictions in Figure 4(d) cannot be validated, since we only have ground truth up to visit 3. Nevertheless, these predictions are interesting because they exemplify list reaction when a patient has more than one type of condition. Two of the predictions are still cancers. The list now has a digestive condition, attributable to the enteritis. However, the strong links associated with hypertension are by far the dominant effect in this final list; that is, the heart conditions become the strongly predicted diseases after this visit.

From this case study, we can see that ICARE is able to make reasonable and intuitive predictions. When multiple unrelated conditions are introduced simultaneously, the list is able to diversify. In the case of this conflict, the more common or heavily linked condition is dominant, securing a higher percentage of the ideal rank positions.

7. CONCLUSIONS

The goal of our work was to come up with a system that can assist a medical practitioner in decision making. If a sampling of future diagnoses can be provided to a practitioner, appropriate medical tests can be ordered sooner and lifestyle adjustments can be adopted by the patient proactively. This will not only result in improving the quality of life for the patient, but also in reducing the health care costs. To that end, we proposed CARE, a collaborative recommendation engine for prospective and proactive healthcare. CARE relied solely on the ICD disease codes, which are a standard across insurance and medicare databases. This exploitation of ICD codes by CARE allows for a seamless integration with a variety of electronic healthcare systems that use or will embrace the standard of ICD. Also, as the medical community moves toward comprehensive electronic records, CARE becomes increasingly relevant.

ICARE's use of ensembles clearly demonstrated that isolating significant relationships and controlling high-prevalence diseases is essential for making better predictions. The impressive future disease coverage of ICARE represents more accurate early warnings for thousands of diseases, some even years in advance. In its most conservative use, the rank lists can provide reminders for conditions that busy doctors may have overlooked. Applied to full potential, the CARE framework can be used to explore broader disease histories, suggest previously unconsidered concerns, and facilitate discussion about early testing and prevention

Incorporating demographic information did not positively influence the predictive power of ICARE, in general. While additional work is necessary before making a judgement about the potential usefulness of this information, the current results suggest that a randomly sampled training set reflecting the distribution of the entire population is sufficient for testing patients of most demographic groups.

7.1 Future Work

Our development and evaluation of CARE has shown that collaborative filtering is a strong and viable approach to disease prediction. However, there are still many interesting avenues for future work. While In this paper, CARE is limited to ICD-9-CM data, the underlying collaborative framework has no such limitation. While it is an advantage that our system doesn't *require* test results or special information, such advanced results when available can add further value. CARE could exploit this information through similarity metrics that are appropriately modified for more complex representations of medical history. It is part of our future work to incorporate such prognostic and diagnostic information about patients.

The current implementation of CARE captures an aspect the temporal information available in the data. The experimental setup limits prediction to future disease, an obvious necessity. However, other temporal data is implicit, such as the length of time between visits or the absolute order or disease occurrence in each patient. Developing methods to harness this information could lead to more precise predictions and even estimated time of disease onset.

We are also incorporating clinical use by collaborating with medical professionals. A longer term study with explicit testing (where reasonable) and monitoring for predicted conditions would be the gold standard.

Acknowledgments

Center of Research Computing at the University of Notre Dame for the high performance and distributed computing resources for this work. This work was partially supported by the Arthur J. Schmitt Fellowship at Notre Dame.

8. REFERENCES

[1] A.-L. Barabasi. Network medicine — from obesity to the dieaseasome. *New England Journal of Medicine*, 357:404–407, 2007.

[2] J. S. Breese, D. Heckerman, and C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. Technical Report MSR-TR-98-12, Microsoft Research, May 1998.

[3] D. K. Cherry, C. W. Burt, and D. Woodwell. A national ambulatory medical care survey: 2001 summary. *Advance Data*, 337:1–16, 2001.

[4] W. T. C. Consortium. A national ambulatory medical care survey: 2001 summary. *Nature*, 447:661–678, 2007.

[5] O. Cordón, F. Herrera, J. de la Montaña, A. Sánchez, and P. Villar. A prediction system for cardiovascularity diseases using genetic fuzzy rule-based systems. In *Proceedings of the 8th Ibero-American Conference on AI*, pages 381–391. Springer Berlin, 2002.

[6] N. C. for Health Statistics. International Classification of Diseases, Ninth Revision, Clinical Modification (icd-9-cm), 2007. `http://www.cdc.gov/nchs/about/otheract/icd9/`.

[7] D. Heckerman, D. M. Chickering, C. Meek, R. Rounthwaite, and C. Kadie. Dependency networks for inference, collaborative filtering, and data visualization. Technical Report MSR-TR-2000-16, Microsoft Research, February 2001.

[8] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, 22:5–53, 2004.

[9] J. Langheier and R. Snyderman. Prospective medicine: The role of genomics in personalized health planning. *Pharmacogenomics*, pages 1–8, 2004.

[10] D. S. Lauderdale, S. E. Furner, T. P. Miles, and J. Goldberg. Epidemiologic uses of medicare data. *Epidemiologic Reviews*, 15:319–27, 1993.

[11] Y. Liu, L. Teverovskiy, O. Lopez, H. Aizenstein, C. Meltzer, and J. Becker. Discovery of biomarkers for alzheimer's disease prediction from structural mr images. In *2007 IEEE International Symposium on Biomedical Imaging*, April 2007.

[12] J. Loscalzo. Association studies in an era of too much information - clinical analysis of new biomarker andgenetic data. *Circulation*, 116(17):1866–1870, 2007.

[13] J. Loscalzo, I. Kohane, and A.-L. Barabasi. Human disease classification in the postgenomic era. *Molecular Systems Biology*, 2007.

[14] J. B. Mitchell, T. Bubolz, J. E. Paul, C. I. Pashos, J. J. Escarce, L. H. Muhlbaier, J. M. Wiesman, W. W. Young, R. S. Epstein, and J. C. Javitt. Using medicare claims for outcomes research. *Medical Care*, 32:38–51, 1994.

[15] F. Piscaglia, A. Cucchetti, A. Orlandini, E. Sagrini, A. Gianstefani, C. Crespi, G. Pelosi, M. Valli, L. Sacchelli, C. Ferrari, and L. Bolondi. Prediction of significant fibrosis in chronic hepatitis c patients by artificial neural network analysis of clinical factors. volume 39, March 2007.

[16] R. Snyderman and R. S. Williams. Prospective medicine: The next health care transformation. *Future Medicine*, 2003.

[17] D. T. Wong and W. A. Knaus. Predicting outcome in critical care: the current status of the apache prognostic scoring system. *Canadian Journal of Anesthesia*, 38:374–383, 1991.