Data Min Knowl Disc (2010) 20:388–415
DOI 10.1007/s10618-009-0156-z

Time to CARE: a collaborative engine for practical disease prediction

Darcy A. Davis · Nitesh V. Chawla · Nicholas A. Christakis · Albert-László Barabási

Received: 22 January 2009 / Accepted: 8 October 2009 / Published online: 25 November 2009

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 patient's medical history using ICD-9-CM codes in order to predict future disease risks. CARE uses collaborative filtering methods 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. Also, we apply time-sensitive modifications which make the CARE framework practical for realistic long-term use. 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 large

Responsible editor: R. Bharat Rao and Romer Rosales.

D. A. Davis · N. V. Chawla (✉)
Department of Computer Science and Engineering, Interdisciplinary Center for Network Science and Applications (iCeNSA), University of Notre Dame, Notre Dame, Indiana, USA
e-mail: nchawla@nd.edu; nchawla@cse.nd.edu

D. A. Davis
e-mail: ddavis4@nd.edu

N. A. Christakis
Harvard Medical School, Boston, MA, USA
e-mail: christak@hcp.med.harvard.edu

A.-L. Barabási
Northeastern University, Boston, MA, USA
e-mail: alb@neu.edu

Medicare dataset, demonstrating that CARE and ICARE perform well at capturing future disease risks.

Keywords Collaborative filtering · Prospective medicine · Disease prediction · Electronic healthcare record

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. A Center for Disease Control and Prevention (CDC) study estimates that 880.5 million visits were made to physician offices, about 3.1 visits per patient, in 2001 (Cherry et al. 2001). 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.

Research has shown many conditions to have recognizable indicators before onset or preventable risk factors. From these discoveries comes the idea of prospective medicine, aimed at determining and minimizing individual risk, as well as actively addressing conditions at the earliest indication. In theory, these practices reduce the number of conditions needing treatment and improve the effectiveness of necessary interventions. 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 (Consortium 2007). 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 (Loscalzo 2007). Given the weak signals that most disease associated SNPs or mutations offer, it is increasingly clear that the promise 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.

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 the common genetic, molecular, environmental, and lifestyle-based individual risk factors, most diseases do not occur in isolation (Barabasi 2007; Consortium 2007; Loscalzo et al. 2007). 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 a 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. Our goal is to provide every patient with a 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.

Key contributions in this work are listed below. Earlier work on the first two contributions can also be found in (Davis et al. 2008a,b).

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. Our system uses only ICD-9-CM (International classification of Diseases) codes (NC for Health Statistics 2007) to make predictions, which are a common standard for insurance and medical databases. We do not require any other information such as lab tests, etc., which can be expensive.

2. The collaborative filtering method employed, while building upon prior work, incorporates new elements of significance testing and ensemble methods within the CARE framework.
3. A time-sensitive system which uses a best sub-vector matching concept to exploit the known ordering of disease diagnosis. The time-sensitive improvements to our framework make it applicable to long-term, diverse data such as public health records. They also help to automatically differentiate and correctly deal with chronic versus non-chronic diseases.
4. Analysis of performance trends dependent on the amount of data known and the length of time between diagnoses. This information provides guidelines for efficient use in a practical setting.
5. Case studies are provided as a real-world example of the potential benefits of CARE.

2 Related work

Our related work includes the larger body of research on collaborative filtering, studies from the medical community which further support the need for preventative medicine, and various interdisciplinary efforts which previously led to computer-aided medical prediction systems. While most of these systems are only loosely comparable to CARE, they are representative of the same goals. We are not aware of any work which is directly comparable to CARE.

As mentioned in the introduction, collaborative filtering is a data mining technique that makes predictions about an active user based on information about other similar users. The usual method is to find the other users that are most similar to the active user, and generate predictions based on their preferences. The first automated collaborative filtering systems were GroupLens (Konstan et al. 1997; Resnick et al. 1994) and Ringo (Shardanand and Maes 1995), which recommended internet news articles and music, respectively. These systems are part of the larger class of memory-based algorithms, which make predictions using the entire user database. This is typically accomplished by calculating a weight of similarity between the active user and all others, and the active user's opinion is determined by the weighted average of the others' opinions. In many cases, only a limited number of 'nearest neighbors' are included in the calculation. The most common similarity metrics are the Pearson correlation coefficient (Resnick et al. 1994) and vector similarity (Salton and McGill 1983; Breese et al. 1998). Memory-based algorithms are simple, easily updated, and generally produce good results. These advantages come at the cost of high resource consumption, since the entire database must be retained and used. While correlation is usually cited as the superior method, other results have shown vector similarity to perform equally well or better (Grcar et al. 2005). The second widely-used class of collaborative filtering algorithms are model-based, where predictions are generated by a model of user preferences which was preconstructed on the user database. The model-based algorithms are faster and more scalable, in general. However, model building tends to be expensive, leading to inflexibility for introducing new data. The quality of predictions for model-based methods widely vary (Si and Jin 2003). Well known model-based methods include

Bayesian clustering or models (Breese et al. 1998), Personality Diagnosis (Pennock and Horvitz 1999), Singular Value Decomposition (Goldberg et al. 2000; Paterek 2007), and the Aspect model (Hofmann 2004; Hofmann and Puzicha 1999). There are also many content-based recommender systems. There is not an appropriate and available source of disease 'content', so these are of little relevance to our problem.

Early treatment (Coyle and Hartung 2002), screening (Institute 2007), lifestyle change (Hunt et al. 1995), and other interventions (Edelman 2006; Koertge et al. 2003) are common themes in modern medical research, where early intervention is shown repeatedly to improve disease outcome and quality of life. Nonetheless, these proactive treatments are far from the norm in our largely reactive health care system. In Snyderman and Williams (2003) provide an outstanding overview of the flaws of the current system and potential benefits of a prospective health care system. They suggest that data mining is a "central feature" of prospective health care. Glasgow (2001) support the feasibility of the preventive approach. They state that much of the chronic disease burden can be prevented, and further posit that existing management strategies can also be used to advance prevention.

Many proponents of prospective medicine emphasize genomic studies and other breakthrough research in human biology. It is undeniable that genomic research is rapidly advancing (Consortium 2007) and holds great promise for medicine. Unfortunately, applicability to the general public is still very limited (Loscalzo 2007). Similarly, in Weston and Hood (2004) express excitement with advancements in systems biology and proteomics, but acknowledge that we still need to learn how to realistically translate discoveries into health benefits. Also, they recognize that there are still "enormous challenges" to overcome. Though low-tech in comparison, CARE demonstrates that existing data and technology can provide immediate advancement toward prospective medicine.

Beyond the basic analogy in 1.1, collaborative filtering seems well suited to disease prediction due to the known collaborative nature of diseases. A wide variety of studies on disease comorbidity, i.e., the simultaneous occurrence of two or more distinct diseases, have shown that multiple risk factors cannot reliably be considered in isolation (Starfield et al. 2003). Co-occurring factors can have a synergystic effect, leading to unexpectedly high risk (Loscalzo 2007; Kannel et al. 1961). In van den Akker et al. (1998), mention that the incidence of comorbid diseases is increasing. They state that statistical clustering of comorbid diseases was surprisingly strong, even among young subjects. This results implies likely interaction between many of the coinciding diseases.

Many different computer-aided methods have been developed for medical prediction. Most of these systems are designed to make predictions about a single disease or class of diseases. Usually, the predictions are generated from some combination of basic data such as demographic information and physical description with addition condition-specific test results or family history. One well-known system is Apache III (Wong and Knaus 1991), a prognostic scoring system for predicting inpatient mortality. Apache uses a combination of acute physiological measurements, age, and chronic health status. A wide variety of systems have been developed for predicting risk of individual diseases or complications, such as specific heart conditions (Cordn et al. 2002), Alzheimer's disease (Liu et al. 2007), and cancer (Mould 2003). While data

mining has been widely used to explore medical problem, collaborative filtering has not been used. An exception is Kahn (2005), which discusses the use of collaborative filtering into the CHORUS system for efficiently locating relevant radiology information. CHORUS is essentially a text-classification program and is not comparable to our work.

3 Data

Our database comprises the Medicare records of 13,039,018 elderly patients in the United States with a total of 32,341,348 hospital visits. The data was originally compiled from raw claims data for beneficiaries who were at least 65 years old as of January 1993 (Christakis and Allison 2006). Such Medicare records are highly complete and accurate, and they are frequently used for epidemiological and demographic research (Lauderdale et al. 1993; Mitchell et al. 1994).

The input for our methods consists of each patient's diagnosis history, provided per inpatient visit. Each data record consists of a hospital visit, represented by a patient ID and a list of up to ten diagnosis codes per visit, as defined by the *International Classification of Diseases, Ninth Revision,Clinical Modification*. The International Statistical Classification of Diseases and Related Health Problems (ICD) provides codes to classify diseases 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 character long. ICD-9 codes are hierarchical in nature, so the 5 characters codes can be collapsed to fewer characters identifying a small family of related medical conditions. For instance, code 40201 is a specific code for malignant hypertensive heart disease with heart failure. This code can be collapsed to 4020, non-specific malignant malignant hypertensive heart disease, or further to 402, the family of all hypertensive heart disease (non-specific).

A sample patient medical history is shown in Table 1; each line represents one hospital visit. The first code for any visit is the principal diagnosis, followed by any secondary diagnoses made during the same visit. Demographic data was available, but our experiments showed little effect on predictive power (Davis et al. 2008b). The length of time between patient visits was also included, which we harnessed to develop a time-sensitive version of our system.

In our Medicare database, 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

Table 1 A sample patient medical history

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 The 20 most prevalent diseases

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 specified cardiac dysrhythmias	5.61
Osteoarthrosis unspec gen/loc unspec site	5.20
Unspecified hypothyroidism	5.14
Unspecified chronic ischemic heart disease	5.13

is only 4.32 per visit, making this dataset very sparse. 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 2 shows the 20 most prevalent diseases in our database.

4 The CARE methodology

4.1 System overview

Before detailing the individual components, a high-level preview of the entire CARE framework is provided in Fig. 1. The dotted lines represent optional methods. The testing patient (denoted as a) is the individual for whom we are making predictions based on the histories of the training patients (denoted as I, with each individual denoted as $i \in I$). Thus, the individual medical history in Fig. 1 is the testing patient while other patients' medical histories is the set of training patients. All patients are represented by their medical history in the format shown in Table 1. The training set is constrained to patients with at least two diseases in common with the testing patient, prior to the application of collaborative filtering. This results in a group of patients similar to the individual testing set patient a. Collaborative filtering is performed on the resulting group, generating predictions for the future visits of the testing patient. In the case of ICARE, this process is performed multiple times for each patient, with

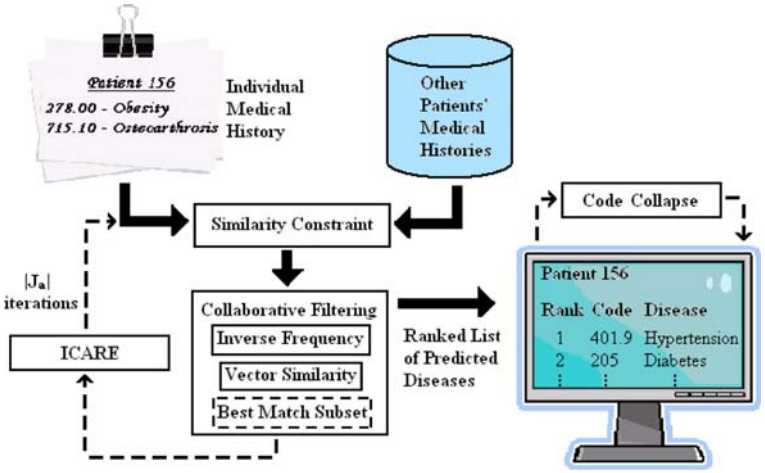

Fig. 1 A high-level overview of the CARE system

each iteration creating a different training patient group based on an individual disease. These multiple resulting predictions are combined to form an ensemble. The output after CARE and ICARE is a ranked list of diseases for the subsequent visits of the testing patient, ranked in order from the highest risk score to the lowest. If desired, these lists can be collapsed into a shorter, less specific version by grouping medical conditions according to the hierarchical nature of the disease codes. Each component is further defined in the subsequent sections.

4.2 Vector similarity

Our collaborative filtering technique is derived from the vector similarity algorithm presented by Breese et al. (1998). Traditionally, collaborative filtering is used to make a prediction $p(a, j)$ on user a, the active user (testing), for item j based on the similarity between user a and every other user i who has previously given a vote $v_{i,j}$ for that item. The entire training set of users is defined as I, and I_j is the subset of users who have voted on j. The similarity $w(a, i)$ between users a and i is calculated by vector similarity; that is,

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

J_i is the set of items rated by user i. The prediction score 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. Thus, the general collaborative filtering equation is:

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

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 1, and the algorithm then predicts that every user has every disease with a score of 1, an obvious error. Our proposed changes modify the general equation to incorporate binary diagnoses and remove the effect of the range of ratings. The modified general equation is also dependent on the random expectation of each disease, referred to as $\overline{v}_j$. Specifically,

$$\overline{v}_j = \frac{|I_j|}{|I|} \tag{3}$$

Thus, the prediction score for the active patient a on disease j is now expressed as follows:

$$p(a, j) = \overline{v}_j + \kappa_a (1 - \overline{v}_j) \sum_{i \in I_j} w(a, i) \tag{4}$$

with the normalizing constant

$$\kappa_a = \frac{1}{\sum_{i \in I} w(a, i)} \tag{5}$$

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.

4.3 Inverse frequency

We further extended Eq. 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 the most important contributions arise from uncommon connections. The inverse frequency of disease j is defined as

$$f_j = \log \frac{n}{n_j} \tag{6}$$

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

$$w(a,i) = \sum_j \frac{f_j v_{a,j}}{\sqrt{\sum_{k \in J_a} f_k^2 v_{a,k}^2}} \frac{f_j v_{i,j}}{\sqrt{\sum_{k \in J_i} f_k^2 v_{i,k}^2}}. \tag{7}$$

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**).

4.4 Grouping of training patients

Before each application of collaborative filtering, a group of relevant training patients is determined based on the number of diagnoses in common with the testing patient. This serves to remove the influence of patients who have little or no similarity with the patient for whom predictions are being made. Training patients with no diseases in common with the active patient have a similarity weight of 0 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 for CARE, we include all patients with two or more diseases in common with the active patient. This constraint enforces stronger similarities for all patients influencing the predictions. In theory, this helps to avoid the noise resulting from similarity on a single common disease, which can introduce a very high number of weak influences. Restricting the training set 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 section.

It is important to note that the random expectation of diseases is different within the group than the overall occurrence in the entire dataset. We defined the global expectation as $\overline{v}_j$ and similarly, we refer to the expectation with a group c as $\overline{v}_{j,c}$. In all experiments, the random expectation within the relevant group is used in Eq. 4.

4.5 ICARE with ensembles

Even with the restricted training set combined with IFVS, we still observed that common diseases can overwhelm less common diagnoses 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 (Dietterich 2000) of CARE runs on individual-disease training groups.

Specifically, for each disease j developed by the test patient a, collaborative filtering is applied separately to the group of training patients with disease j. Since each group is specific to a disease, we do not enforce the two-in-common constraint described in Sect. 4.4. As before, the collaborative filtering uses the within-group

random expectation $\overline{v}_{j,c}$. Thus, each member of the ensemble is a round of collaborative filtering on an individual disease group, and it follows that the number of members is equal to the number of unique diseases developed by patient a prior to the visit for which the predictions are being generated. While a single visit is used to define the ensemble member, 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

$$\max_{c \in G} \left(\overline{v}_{j,c} + \kappa (1 - \overline{v}_{j,c}) \sum_{i \in I_{j,c}} w(a, i) \right) \tag{8}$$

where G is set of ensemble members or disease groups, conditioned on the individual diseases of each testing patient. We choose the maximum since diseases are generally not protective against each other, with few exceptions. In other words, having additional diseases does not lessen the probability of developing a disease.

In order to reduce the number of predictions and the runtime of the ensembles, we only predict on diseases for which the $\overline{v}_{j,c}$ is significantly higher than $\overline{v}_j$. That is, if a disease's prevalence is higher in the entire population than the focused group, we do not generate a prediction on that disease since the group does not show a strong influence on its occurrence. We determine the significance of a disease 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

$$z = \frac{p_1 - p_2}{S_{p_1 - p_2}}. \tag{9}$$

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

$$S_{p_1 - p_2} = \sqrt{\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}} \tag{10}$$

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, $p_1 = \overline{v}_{j,c}$, $p_2 = \overline{v}_j$, $n_1 = |c|$, and $n_2 = |I|$. We use a 95% confidence interval.

4.6 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 coding scheme. For example, the ICD-9 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 (pre-collapsing), 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 codes are preserved. Collapsing the ICD-9 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 (Mitchell et al. 1994). 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.

5 Time-sensitive CARE

CARE and ICARE do not take the order of or length between disease diagnoses into account when generating vector similarity among patients. However, a patient should be considered more similar to another if their shared diseases follow a similar temporal pattern, as well. Similarly, matching with two diseases which occurred many years apart may not be relevant. For this reason, we modify our methods to incorporate the length of time between medical events (in our case, hospital visits). In addition to more realistic similarity weights, using temporal information allows our framework to extend to broad, general datasets with more complete medical history. A limitation of our dataset is that the disease onset can only be identified with the hospital admission, which might not accurately reflect the time the disease was developed. Our dataset

is also limited to a scope of 4 years, another disadvantage. Nevertheless, the goal of our system is to have the capability to incorporate temporality, which has distinct advantages.

As explained in 4.2, CARE determines the similarity of the active patient a and a training patient i as the vector similarity between the disease vector of a and the entire disease vector of i. The prediction score $p(a, j)$ for every disease j in the training vector will be weighted by this similarity. This implementation is blind to the order of disease occurrence in the training patient; a common disease between the active patient and training visit 5 will increase prediction scores for diseases which occurred in training visit 1. This captures correlation, but it misses any causality effects or natural ordering of disease occurrence. We are only interested in predicting the future, so an overlapping disease should ideally only increase prediction scores for diseases occurring in later training visits. However, this is too simplistic. In most cases, a and i will have multiple overlaps in different visits, and considering them individually would lose complex or synergistic effects.

Algorithm 1 Pseudocode for finding the best match subset of training visits

Algorithm best_match(a,$visits$)

```
maxsofar = 0
maxstart = 0
maxend = 0
currentmax = 0
currentstart = 0
for all m in visits do
  if w(a, sub_{m,m}) ≥ w(a, sub_{currentstart,m}) then
    currentmax = w(a, sub_{m,m})
    currentstart = m
  else
    currentmax = w(a, sub_{currentstart,m})
  end if
  if currentmax ≥ maxsofar then
    maxsofar = currentmax
    maxstart = currentstart
    maxend = m
  end if
end for
Return maxsofar, maxstart, maxend
```

Our method is a compromise. First, we find the subset of consecutive training visits of i with the best vector match to the active patient a. We define $sub_{s,z}$ to be the consecutive set of visits from visit s to visit z. For training patient i with n visits, the best match $best(a, i)$ to active patient a is $sub_{s,z}$ such that

$$\max(w(a, sub_{s,z})), \quad 1 \leq s \leq z \leq n \tag{11}$$

Similar to other maximum subsequence problems, we can find $best(a, i)$ in linear time. Our pseudocode is shown in Algorithm 1. While our algorithm is heuristic, it

will be accurate for nearly all sequences that would realistically occur. As the algorithm scans, it tests whether the current visit yields a higher vector similarity in conjunction with the preceding best sequence or standing alone. If the current visit v performs better without the preceding set of visits, then those visits will not be beneficial to any sequence containing v. Conversely, if similarity is higher when v is combined with earlier visits, then those earlier visits will continue to be beneficial in any subset containing v. The concept is similar, though not identical, to fraction multiplication; note that for any two vectors, $0 < w(a, i) < 1$. Starting with the largest fraction will always be better, regardless of future values.

Intuitively, $best(a, i)$ is the time period when training patient i was having the most similar medical experience to the active patient a, and the visits immediately following should have the most relevant information to the prognosis of a. Thus, we modify the general equation so the "best match" vector similarity only adds prediction weight for diseases which occur in visits after the "best match" time frame. Assuming that $best(a, i) = sub_{s,z}$ and i has n visits, then Z_j is the set of patients i such that $j \in sub_{z+1,n}$. The time-sensitive general equation is then

$$p(a, j) = \overline{v}_{j,c} + \kappa(1 - \overline{v}_{j,c}) \frac{\sum_{i \in Z_j} w(a, best(a, i))}{\sum_{i \in I} w(a, best(a, i))} \tag{12}$$

If the "best match" includes the last visit for i, we assume that i provides no knowledge about the future of a.

Finding a best match subset of visits incorporates the ordering of diseases and resolves the problem of "predicting the past", while still preserving multiple-disease interactions. Additionally, this strategy makes the CARE framework feasible for long-term, diverse data, such as public health records. Over a lifetime, people may go through many different medical experiences and phases, and very few people will have the same experience over a period of many years. However, similarity within a short window may be very strong. The best match is able to isolate the most relevant time periods without all of the noise generated by the rest of the medical record. A simple cutoff mechanism could easily be used to limit the breadth of the training patients' 'future' influencing the predictions, as well.

6 Evaluation

CARE and ICARE generate predictions only on 'future' visits of a patient 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. 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. For each round of collaborative filtering, each disease j is assigned an actual value $A(a, j)$ which describes when the active patient a is diagnosed with j.

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 $k = 1$.

We compare all the methods against the following *baseline* method. A baseline ranking for each testing patient a is determined by ordering the diseases by their random expectation $\overline{v}_{j,c}$ within the group c of relevant training patients formed around a. The performance measures on the baseline ranking serve as a benchmark for experiments. This baseline ranking determines the patient-specific risk based only on the training patients with whom they share diagnoses, but without the benefits of collaborative filtering. 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* (Herlocker et al. 2004), this metric is intended to measure the expected utility of the ranked list (Heckerman et al. 2001). 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

$$p(k) = 2^{-k/a} \tag{13}$$

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

$$\text{Utility} = \sum_{k} p(k)\delta_k \tag{14}$$

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 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,

$$\text{Accuracy} = \frac{100}{N} \sum_{i=1}^{N} \frac{\sum_{k=0}^{R_i - 1} \delta_{ik} p(k)}{\sum_{k=0}^{M_i - 1} p(k)} \tag{15}$$

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 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 accuracy of 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 trade-off between average rank and coverage, since higher coverage captures less obvious diseases with lower rank.

7 Experiments

In this section, we evaluate the predictive performance of CARE and ICARE. We also show results after applying the time-sensitive modifications described in 5 as well as the pre- and post-collapsing methods described in 4.6. 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 2 provides a pictorial explanation of this process. All testing patients were required to have at least five visits. We used a 2-fold cross-validation scheme for the experiments. For all methods, individual predictions are independent, making it easy to run experiments in a distributed fashion. Also, static calculations such as determining individual disease groups, random expectations, and inverse frequencies are preprocessed to avoid repetition.

Table 3 displays the experimental results. The metrics are also applied to the baseline ranking, which is a list of the diseases ranked in order from highest baseline

ID Visit 1		Round 1: ID Visit 1
ID Visit 2	$\rightarrow$	Round 2: ID Visit 1$\cup$Visit 2
ID Visit 3		Round 3: ID Visit 1$\cup$Visit 2$\cup$Visit 3

Fig. 2 An example of how patient visits are processed by the IFVS algorithm. ID refers to a patient ID

Table 3 Evaluation of performance of CARE, ICARE, and time-sensitive ICARE compared with the baseline ranking

	Comparison of methods			
	Baseline	CARE	ICARE	Time ICARE
Top 20				
Coverage	.321	.344	.412	.385
Average rank	7.326	7.819	5.755	6.806
Half-life accuracy	30.574	30.255	49.274	41.238
Top 100				
Coverage	.585	.606	.605	.594
Average rank	27.766	26.734	20.299	22.024
Half-life accuracy	31.115	30.759	49.645	41.683
All				
Coverage	.986	.940	.773	.770
Average rank	229.572	177.495	81.191	90.317
Half-life accuracy	31.115	30.759	49.645	41.683

prevalence to lowest. As mentioned earlier, results on the top 100 and top 20 ranks are more meaningful, since a medical practitioner or other user is unlikely to consider a very large portion of the list. CARE shows better performance than baseline across the board overall and in the top 100 ranks. In the top 20 ranks, CARE covers 2% 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 9% more of the future diseases than the baseline method in the top 20 rankings alone, while the average rank of 5.755 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.

We show the effect of the time-sensitive modifications only as applied to ICARE, the clearly superior method. Time-sensitive ICARE shows a small loss of performance, but still outperforms CARE and the baseline significantly. This loss is easily explainable, since ICARE has a stronger bias for ranking chronic diseases, which are very prevalent among senior citizens. Since the ensemble method looks at each diagnosis individually, any repeat diseases will create a group around themselves with $\overline{v}_{j,c} = 1$. This leads to a perfect ranking of chronic diseases. The time-sensitive method does not carry that bias, since only visits occurring after the "best match" affect $\overline{v}_{j,c}$, which may or may not contain a repeat code. The time-sensitive version will predict chronic diseases based on their likelihood of repeat visits rather than assuming 100% likelihood. In the Medicare data, the dominance of chronic diseases causes ICARE's assumption to be beneficial, but would likely be less influential in a more general setting. In fact, we find that the time-sensitive version performs slightly better when considering non-repeat (not previously diagnosed to the active patient) diseases. Specifically, we see

19.9% coverage in the top 20 ranks versus 16.9% with unmodified ICARE. Finally, it is worth noting that the time-sensitive methods are necessary for computational efficiency and noise control when applied to a long-term general database. We posit that a minor drop in performance is acceptable in light of these more practical concerns.

It merits explanation that the 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 accuracy. By modifying the α value defined in Sect. 6 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.

We post-collapsed the full code results as described in Sect. 4. The performance of ICARE on resulting the 3-digit ICD-9 codes are shown in the "Postcollapse" column of Table 4. The same trend is seen when applied to CARE or time-sensitive experiments. Post-collapsing results in an improvement in ranking and coverage across the board. There is a slight dip in the accuracy measure. We believe this arises because of multiple high-ranking diseases collapsing to a common code, eliminating the dominance in the top ranks. These results from collapsing of ICD-9 codes are very encouraging, with more than 51% 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. 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.

Table 4 Effect of post-collapsing and pre-collapsing on ICARE

	Comparison of ICD-9 collapsing methods		
	ICARE	Postcollapse	Precollapse
Top 20			
Coverage	.412	.513	.547
Average rank	5.755	5.668	5.820
Half-life accuracy	49.274	58.731	59.472
Top 100			
Coverage	.605	.722	.785
Average rank	20.299	18.101	18.612
Half-life accuracy	49.6455	58.731	59.924
All			
Coverage	.773	.779	.833
Average rank	81.191	29.742	26.471
Half-life accuracy	49.6455	58.731	59.924

We also run ICARE experiments on pre-collapsed data. The results here are particularly impressive, with a coverage of nearly 55% of future disease groups in the top 20 ranks. The performance measures achieved here are substantially higher than the post-collapsed method. Still, the utility of pre-collapsing depends on the situation and the goals of the doctor. Unlike the post-collapsed results, detailed information is permanently lost. Depending on the family of conditions, the difference between the 5-digit members ranges may be minimal or very crucial, requiring different paths of response. In general, we believe that post-collapsing is more amenable to the eventual use of the system; it still provides a medical practitioner a choice to retrieve the complete resolution of ICD-9 codes.

8 Performance trends

We also are interested in how performance changes with respect to the amount of data known about the testing patient. This analysis provides insight into optimal deployment of such a system in a practical setting. It provides guidelines for the minimum amount of information needed for meaningful (better than baseline) results and a threshold for good results without overcomplicated computation. Specifically, we look at the number of visits known by CARE about the testing patient (Fig. 3a), the total number of unique diseases known about the patient (Fig. 3b), and the length of time in days between the patient's last known visit and the following unknown, and thus predictable, visit (Fig. 3c). We look at the coverage within the top 20 ranks, which we believe is our most practical measure of performance.

The visit and diseases trends show that performance continually increases as more information is known about the patient. The results suggest that ICARE on a single visit is sufficient to outperform the baseline, though 3b shows that the visit should have at least 3 diseases. The benefit of additional diseases flattens around 25 unique diagnoses. The data for patients with more than 35 diseases is too sparse for further conclusions, but can be expected to continue in a flat line near 57% coverage.

Unsurprisingly, as the length of time since the last visit increases, a modest drop in performance can be observed. The intuition here is that older diagnoses are less relevant to immediate concerns, on average. Despite the downward trend, ICARE still outperforms the baseline after gaps of more than 2–3 years. A more long-term study of this effect would be interesting, but we are limited by the scope of our data.

Since we have mentioned multiple times the dominance of common diseases, we examine our method's ability to control this effect. To do this, we look at the distribution of disease prevalence of the patients' actual future diseases compared to the predictions from the baseline and ICARE. This analysis is shown in Table 5. The first column is the percent prevalence of a disease in the patient population, which is equivalent to the random expectation $\overline{v}_j$. The second column shows what percentage of the actual diagnoses fall within each prevalence range. We see that while there are a few diseases which are very common, there are many different uncommon diseases which account for most of the actual diagnoses. In fact, 53% of actual diagnoses are diseases which are less than 5% prevalent in the whole patient population. The final

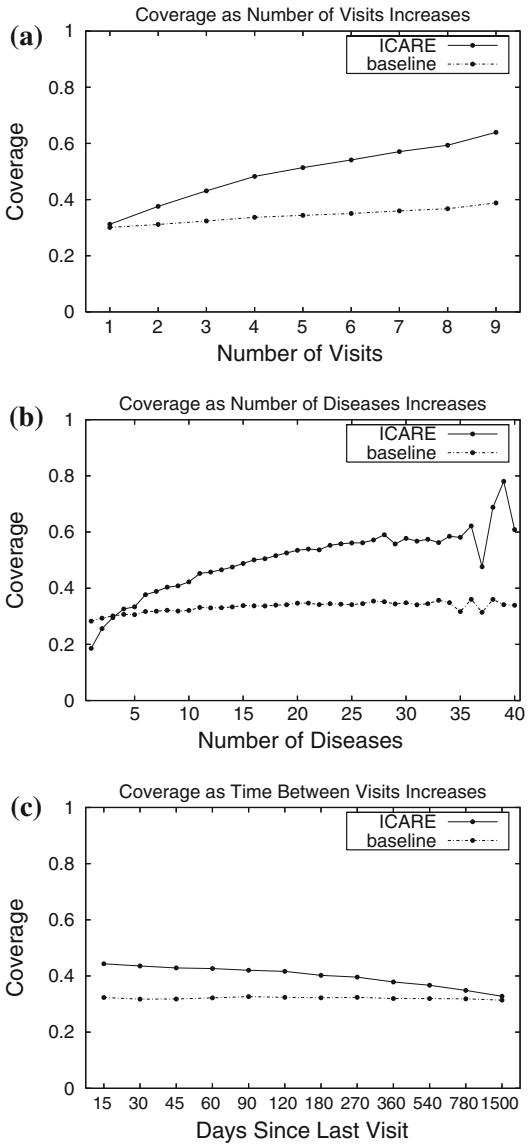

Fig. 3 Coverage trends with respect to known testing patient data

two columns shows the percentage of the top 20 predictions which fall within each prevalence range; the third column uses the baseline ranking and the fourth column uses ICARE. The baseline results clearly show that extra controls are needed to avoid skew toward common diseases. Also, the results show that ICARE does well at limiting very common diseases to a realistic percentage of the strongest predictions. Finally, we note that most of the top predictions by ICARE are low-prevalence diseases, which is also the case in reality.

Table 5 Comparison of distribution of diseases prevalence between the actual patient diagnoses, the top 20 ranked diseases of the baseline method, and the top 20 ranked diseases of ICARE

Disease prevalence trends in actual diagnoses vs. top 20 rankings			
% Disease prevalence	Actual diagnoses (%)	Baseline top 20 (%)	ICARE top 20 (%)
0–5	53	1	24
5–10	20	6	22
10–15	5	29	10
15–20	3	13	8
20–25	4	18	10
25–30	4	13	8
30–35	2	4	5
35–40	0	0	0
40–45	4	8	8
45–50	2	4	2
50–55	2	4	3

9 Case studies

To further demonstrate our work, we present case studies which place the algorithm results in the context of real patients. First, we look at the ranked list of disease predictions generated for a cancer patient after each subsequent hospital visit. This is a demonstration of the intended usage of the process. The case study shows the evolution of the prediction lists as new information is added. Also, a real example of actual versus predicted diseases facilitates an intuitive understanding of the method's strengths and limitations. Next, we look at patients with an unusually high prediction score for a disease. These case studies provide interesting insight into the important factors and correlations which influence disease risk and how they translate to a risk score. All of the case studies are done using ICARE, which is demonstrated to be our best method.

9.1 Case study of ranked list

We applied ICARE to a patient with three hospital visits. The patient's actual diagnoses and top 10 predictions after each visit are provided in Fig. 4. These results are based on post-collapsed disease codes to avoid unnecessary complication.

We would like to point the reader to the list of most prevalent diseases in Table 2. 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.

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)

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)

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)

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)

Fig. 4 Case study: ranked prediction of cancer patient. **a** Patient diagnoses, **b** ICARE prediction after visit 1, **c** ICARE prediction after visit 2, **d** ICARE prediction after visit 3

Figure 4a 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 reoccur 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 4b 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 4c 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. This acknowledges a significant link to the disease,placing it still within the top 100 but not among the strongest concerns.

The predictions in Fig. 4d 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.

9.2 Case study of highest score

In this section, we provide examples of disease-specific case studies where we are interested in the highest 'scoring' patients for a single condition. For this kind of case study, it is interesting to explore the statistical relationship between each disease in a patient's medical history and the disease that is being predicted. This provides an intuition as to how much effect each expressed disease had on the resulting prediction. We define two metrics describing the relationship between the expressed disease (E) and the target disease (T) being predicted. $E \rightarrow T$ is the percentage of patients with the expressed disease who also have the target disease. Conversely, $T \rightarrow E$ is the percentage of patients with the target disease who also have the expressed disease. The first metric, $E \rightarrow T$, is the stronger explanatory factor, since ICARE forms a training patient group for each expressed disease. The second metric has a less direct impact.

We will look at two patients who have an unusually strong prediction for diabetes, our target disease. Patient 1 is diagnosed with diabetes in the sixth hospital visit.

Table 6 Case study: patient 1—does develop diabetes

	T → E	E → T
Patient 1—visit 1–9/1990		
Syncope and collapse	0.049	0.175
Toxic diffuse goiter no crisis	0.001	0.154
Cellulitis and abscess leg except foot	0.035	0.288
Unspec peripheral vascular disease	0.071	0.340
Urinary tract infection	0.221	0.236
Patient 1—visit 2–7/1992		
Contusion of thigh	0	0
Multiple involv of mitral and aortic valves	0.002	0.125
Cerebral atherosclerosis	0.0162	0.308
Awaiting admission to adeq facilities ELS	0.003	0.444
Patient 1—visit 3–7/1992		
Unspec nonpsycht mntl disorder, brain damage	0.012	0.207
Patient 1—visit 4–2/1993		
Urinary tract infection	0.221	0.236
Hyperosmolality and/or hypernatremia	0.013	0.292
Hyperpotassemia	0.033	0.296
Pneumonia, organism unspecified	0.103	0.194
Unspecified pleural effusion	0.0528	0.211
Hypopotassemia	0.093	0.176
Unspecified anemia	0.087	0.181
Patient 1—visit 5–3/1993		
Gastrostomy status	0.002	0.3

Patient 2 is never diagnosed with diabetes. We analyze the expressed diseases and the corresponding T → E and E → T measurements for the first 5 visits for each patient. The month and year of each hospital visit are also included. The case study for patient 1 is in Table 6 and patient 2 is shown in Table 7.

From the beginning, patient 1 expresses diseases which are highly co-occurrent with diabetes. This trend continues through several years and the patient is diagnosed with uncomplicated diabetes mellitus in 4/1993. Amazingly, all diagnoses for this patient have E → T ≥ 0.125. This means that at least 12.5% of similar patients had diabetes, regardless of which diagnosis the training group was based on. Referring back to the Table 2, we see that the population baseline is only 10.47%. Even from the first visit, 2.5 years before diabetes is officially listed as a diagnosis, we expect a minimum 34% risk. Also, the fact that all expressed diseases co-occur rather strongly with diabetes results in an unusually low amount of noise from unrelated medical conditions. Overall, it is unsurprising that this patient was easily recognized as at high-risk for diabetes. Another interesting observation is that, with the exception of urinary tract infections, most of the E → T are fairly low. This means that while these conditions are strong predictors of diabetes, diabetes is not a very strong predictor of them.

Table 7 Case study: patient 2—does not develop diabetes

	T → E	E → T
Patient 2—visit 1–6/1992		
Unspecified cerebral artery occlusion	0.058	0.25
Acute bronchiolitis	0.001	0.5
Patient 2—visit 2–7/1992		
Care involving other spec rehab process	0.009	0.165
Cerebral thrombosis	0.012	0.3
Unspecified cardiovascular disease	0.056	0.241
Bronchitis, not spec. if acute/chronic	0.007	0.159
Urinary tract infection	0.221	0.236
Spondylof uns site w/o myelpath	0.006	0.170
Osteoarthros unspec gen/loc low leg	0.017	0.122
Patient 2—visit 3–5/1993		
Congestive heart failure	0.300	0.266
Hypopotassemia	0.093	0.176
Unspecified asthma	0.028	0.201
Other and unspecified angina pectoris	0.133	0.239
Coronary atherosclerosis	0.283	0.237
Unspecified essential hypertension	0.412	0.225
Spondylof uns site without myelopathy	0.006	0.170
Osteoarthros unspec gen/loc low leg	0.017	0.122
Patient 2—visit 4–7/1993		
Congestive heart failure	0.300	0.266
Generalized osteoarthrosis unspec site	0	0
Patient 2—visit 5–7/1993		
Congestive heart failure	0.300	0.266

Similar to the first patient, patient 2 has many strongly diabetes-linked diseases and a limited amount of noise. There is one completely unconnected general osteoarthrosis diagnosis. However, this code is only expressed by 10 patients in the dataset, so the effect is minimal. From the first visit, this patient had an expected 50% risk of developing the disease. This patient obviously suffered from advanced heart disease, which is known to link with type 2 diabetes. In contrast to patient 1, this table shows multiple conditions for which both E → T and T → E are exceptionally high. Again, a high prediction score for diabetes is unsurprising. We contend that monitoring, such as routine blood tests or educational intervention, for such a patient would be justified.

10 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. The result of our effort was CARE, a collaborative recommendation engine for prospective and proactive healthcare. CARE relied solely on ICD-9 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. By making our framework more time sensitive, we reap multiple practical benefits. The time-sensitive approach is able to differentiate between chronic disease and lone occurrence. Also, it makes the CARE framework feasible for large, diverse datasets spanning many years, such as the comprehensive records mentioned above. In its most conservative use, the rank lists can provide reminders that busy doctors may have overlooked. Applied to full potential, the CARE framework can be used explore a broader disease histories, suggest previously unconsidered concerns, and facilitating discussion about early testing and prevention.

11 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.

In this paper, CARE is limited to ICD-9 data and temporal data, but the underlying collaborative framework has no such limitation. While it is an advantage that our system doesn't *require* test results or special information, it would be naive to ignore these advanced results when they are available. CARE could exploit this information through similarity metrics which are appropriately modified for more complex representations of medical history. Such generalizations will allow CARE to advance in parallel with the field of medicine. Also, due to limited availability, we only explore one dataset. Further experiments should be done on datasets with varying degrees of diversity to determine the best uses for the system. Additional collaborative filtering methods can also be explored.

Using the temporal data exploited by our time sensitive approach, CARE could be extended to predict the time of expected disease diagnosis in addition to the likelihood of occurrence. Such a mechanism is not well suited to our data, since inpatient visits are fairly sporadic and may include diagnoses which do not relate to the timing of the hospitalization. However, in a database providing a more complete medical picture, this functionality could be an additional guide for scheduling of future checkups, screening, and tests.

Finally, the real utility of CARE cannot be determined without clinical testing. Doctors are the best judge of the utility of this system. Use by medical experts can also provide better insight into needed improvements. A long term study with explicit testing (where reasonable) and monitoring for predicted conditions would be the gold standard.

Acknowledgments The work was supported in part by the Arthur J. Schmitt Foundation.

References

Barabasi A-L (2007) Network medicine—from obesity to the dieaseasome. N Engl J Med 357:404–407

Breese JS, Heckerman D, Kadie C (1998) Empirical analysis of predictive algorithms for collaborative filtering. Technical Report MSR-TR-98-12, Microsoft Research, May

Burges C (1998) A tutorial on support vector machines for pattern recognition. Data Min Knowl Disc 2(2):121–167

Cherry DK, Burt CW, Woodwell D (2001) A national ambulatory medical care survey: 2001 summary. Adv Data 337:1–16

Christakis NA, Allison PD (2006) Mortality after the hospitalization of a spouse. N Engl J Med 354(7):719–730

Cordn O, Herrera F, de la Montaб1a J, Sdfnchez A, Villar P (2002) A prediction system for cardiovascularity diseases using genetic fuzzy rule-based systems. In: Proceedings of the 8th Ibero–American Conference on AI, pp 381–391. Springer, Berlin

Coyle P, Hartung H-P (2002) Use of interferon beta in multiple sclerosis: rationale for early treatment and evidence of dose- and frequency-dependent effects on clinical response. Multiple Scler 8(1):2–9

Davis D, Chawla NV, Blumm N, Christakis N, Barabasi A-L (2008a) Care for your future: prospective disease prediction using collaborative filtering. In: Proceedings of the KDD 2008 workshop on mining medical data

Davis D, Chawla NV, Blumm N, Christakis N, Barabasi A-L (2008b) Predicting individual disease risk based on medical history. In: Proceedings of the ACM conference on information and knowledge management

Dietterich TG (2000) Ensemble methods in machine learning. In: Proceedings of the first international workshop on multiple classifier systems, pp 1–15, June

Edelman D et al (2006) A multidimensional integrative medicine intervention to improve cardiovascular risk. J Gen Intern Med 21(7):728–734

Glasgow RE et al (2001) Does the chronic care model serve also as a template for improving prevention? Milbank Q 79(4):579–612

Goldberg K, Roeder T, Gupta D, Perkins C (2000) Eigentaste: a constant time collaborative filtering algorithm. Technical report, University of California, Berkley, August

Grcar M, Fortuna B, Mladenic D (2005) Knn versus svm in the collaborative filtering framework. In: WebKDD August

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

Herlocker JL, Konstan JA, Terveen LG, Riedl JT (2004) Evaluating collaborative filtering recommender systems. ACM Trans Inf Sys 22:5–53

Hofmann T (2004) Latent semantic models for collaborative filtering. ACM Trans Inf Sys 22:89–115

Hofmann T, Puzicha J (1999) Latent class models for collaborative filtering. In: Proceedings of the 16th international joint conference on artificial intelligence, pp 688–693

Hunt J, Kristal A, White E, Lynch J, Fries E (1995) Physician recommendations for dietary change: their prevalence and impact in a population-based sample. Am J Public Health 85:722–726

Kahn CE Jr (2005) Collaborative filtering to improve navigation of large radiology knowledge resources. J Digit Imaging 18(2):131–137

Kannel WB, Dawber TR, Kagan A, Revotskie N, Stokes J III (1961) Factors of risk in the development of coronary heart disease: six-year follow-up experience: The Framingham Study. Ann Intern Med 55: 33–50

Koertge J et al (2003) Improvement in medical risk factors and quality of life in women and men with coronary heart disease in the Multicenter Lifestyle Demonstration Project. Am J Cardiol 91(11):1316–1322

Konstan JA, Miller BN, Maltz D, Herlocker JL, Gordon LR, Riedl J (1997) Grouplens: applying collaborative filtering to usenet news. Commun ACM 40:77–87

Lauderdale DS, Furner SE, Miles TP, Goldberg J (1993) Epidemiologic uses of medicare data. Epidemiol Rev 15:319–327

Liu Y, Teverovskiy L, Lopez O, Aizenstein H, Meltzer C, Becker J (2007) Discovery of biomarkers for alzheimer's disease prediction from structural mr images. In: 2007 IEEE international symposium on biomedical imaging, April

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

Loscalzo J, Kohane I, Barabasi A-L (2007) Human disease classification in the postgenomic era. Mol Syst Biol

Mitchell JB, Bubolz T, Paul JE, Pashos CI, Escarce JJ, Muhlbaier LH, Wiesman JM, Young WW, Epstein RS, Javitt JC (1994) Using medicare claims for outcomes research. Medical Care 32:38–51

Mould R (2003) Prediction of long-term survival rates of cancer patients. Lancet 361:262

NC for Health Statistics (2007) International Classification of Diseases, 9th Revision, Clinical modification (icd-9-cm). http://www.cdc.gov/nchs/about/otheract/icd9/abticd9.htm

NC Institute (2007) Cancer trends progress report—2007 update

Paterek A (2007) Improving regularized singular value decomposition for collaborative filtering. In: KDD-Cup August

Pennock DM, Horvitz E (1999) Collaborative filtering by personality diagnosis: a hybrid memory- and model-based approach. In: Proceedings of the IJCAI workshop on machine learning for information filtering

Resnick P, Iancovou N, Sushak M, Bergstrom P, Riedl J (1994) Grouplens: an open architecture for collaborative filtering of netnews. In: Proceedings of the ACM conference on computer supported cooperative, pp 175–186

Salton G, McGill M (1983) Introduction to modern information retrieval. McGraw-Hill, New York

Shardanand U, Maes P (1995) Social information filtering: algorithms for automating "word of mouth". In: Proceedings of the computer human interaction, pp 210–217, May

Si L, Jin R (2003) Flexible mixture model for collaborative filtering. In: Proceedings of ICML

Snyderman R, Williams RS (2003) Prospective medicine: the next health care transformation. Future Med

Starfield B, Lemke KW, Bernhardt T, Foldes SS, Forrest CB, Weiner JP (2003) Comorbidity: implications for the the importance of primary care in case management. Ann Fam Med 1:8–14

van den Akker M, Buntinx F, Metsemakers JF, Roos S, Knottnerus JA (1998) Multimorbidity in general practice: prevalence, incidence, and determinants of co-occuring chronic and recurrent diseases. J Clin Epidemiol 51:367–375

Weston AD, Hood L (2004) Systems biology, proteomics, and the future of health care: toward predictive, preventative, and personalized medicine. J Proteome Res 3(2):179–196

Wong DT, Knaus WA (1991) Predicting outcome in critical care: the current status of the apache prognostic scoring system. Can J Anesth 38:374–383

WTC Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661–678