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
The Author(s) 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...
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 health 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