Darcy A. Davis  
Advisor: Nitesh V. Chawla  
University of Notre Dame

Introduction
Faced with unsustainable costs and enormous amounts of under-utilized data, health care needs more efficient practices, research, and tools to harness the benefits of data. These methods should create a feedback loop where computational tools guide and facilitate research, leading to improved biological knowledge and clinical standards, which in turn should generate better data. In order to facilitate the necessary changes, better tools are needed for assessing risk and optimizing treatments, which further require better understanding of disease interdependencies, genetic influence, and translation into a patient’s future. We propose network-centric data mining approaches for benefit in multiple stages of this feedback loop: for better understanding of disease mechanisms and for development novel clinical tools for personalized and prospective medicine.

Network Analysis of NICU Team Structure
Patients in the neonatal intensive care unit (NICU) are at relatively high risk for preventable medical harm. Long and complex stays, with up to 300 handoffs, leave infants at risk. Maintaining an well functioning nursing team is a daunting challenge. Clinical studies have shown organizational characteristics of care to be predictors of performance, and networks have been used to study performance and fault tolerance with respect to team structure in related domains. We demonstrate of the use of network analysis to answer the question:

How does the structure of the nursing team affect the quality of care?

Methods
Using EHR data, we construct individual patient networks representing handoffs (edges) that occurred between nurses (nodes) who cared for the patient.

Results
High MeRCI scores were significantly correlated with poor scores for family satisfaction with care.

Evaluation
Our strongest evaluation metric is coverage, the percentage of a patient’s actual future diseases which a prediction is made and ranked. For evaluation, we usually limit the prediction list to the top 20 highest risk scores, a practical size for consideration by a medical professional.

Results
Our best method, the ensemble-based ICARE, captures 41% and 45% of all future diagnoses in the top 20 ranks for the Medicare and outpatient data, respectively.

The CARE Framework
The central component of CARE is collaborative filtering. In the same way that services like Netflix predict that users will enjoy. CARE predicts diseases that patients will develop based on the diagnoses of patients who had a similar medical history. It is very likely that other patients among millions have experienced genetic and environmental risk factors that closely mirror our own.

Data and Methods
Our data comprises of Medicare records of more than 13 million elderly patients. Each data record consists of a hospital visit, represented by up to ten ICD-9-CM diagnosis codes. We also have outpatient data for 800,000 patients within a large regional health system, also in ICD-9-CM format, spanning all age ranges.

We use collaborative filtering with vector similarity weighting. We also incorporate inverse frequency, meaning rare diseases influence similarity more strongly. Our iterative version, ICARE, uses collaborative filtering rounds to isolate significant correlations and control common diseases. We also determine patient similarity based on the subset of consecutive visits from a training patients record that best matches the active patient (for whom predictions are being made). This allows the algorithm to use only the most relevant portion of each medical record, which also reduces noise and complexity. For each patient, the system outputs a ranked list of diagnoses from the highest risk score to the lowest.