PheMap: An NLP-based resource for high-throughput phenotyping within electronic health records

Neil S. Zheng1, Qiping Feng2,3, V. Eric Kerchberger1,2, Juan Zhao1, Todd L. Edwards2,4, Nancy J. Cox1,2, C. Michael Stein2,3, Dan M. Roden1,2,3, Joshua C. Denny1,2, Wei-Qi Wei1

1Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN; 2Department of Medicine, Vanderbilt University Medical Center, Nashville, TN; 3Division of Clinical Pharmacology, Vanderbilt University Medical Center, Nashville, TN; 4Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN, USA

Abstract
Electronic health records (EHRs) are valuable resources for medical research. However, developing algorithms to extract disease phenotypes from EHRs can be difficult and time-consuming. We propose PheMap, a natural language processing (NLP) based resource that assigns a “phenotype score” to each patient using their EHRs. In an initial validation, PheMap distinguished cases and controls for type 2 diabetes mellitus (T2DM), demonstrating its potential for high-throughput phenotyping.

Introduction
Phenotyping algorithms and electronic health records (EHRs) have facilitated medical research1. However, phenotyping algorithms can be difficult to develop, often requiring particular domain knowledge and several iterations of manual chart review to achieve reasonable accuracy2. To overcome this expensive and lengthy development process, we applied natural language processing (NLP) to five publicly available resources to create PheMap, a knowledgebase that streamlines the phenotyping process within EHRs.

Methods
PheMap is a knowledgebase of medical concepts with quantified relationships to phenotypes. We collected articles describing diagnoses, symptoms and treatments for diseases of interest (phenotypes) from five online medical resources: Mayo Clinic, MedlinePlus, MedicineNet, WikiDoc, and Wikipedia. We combined the articles related to each disease phenotype into a single “phenotype document.” We used KnowledgeMap Concept Indexer, our locally developed NLP pipeline, to identify Unified Medical Language System concepts in each phenotype document3. To estimate the importance of a concept to a phenotype, we computed and assigned the term frequency–inverse document frequency (TFIDF) score to each concept.

To calculate a phenotype score, we search a patient’s EHRs for the top weighted concepts from PheMap for the phenotype of interest, including observations, billing codes, laboratory test orders, procedure orders, medication prescriptions, and sum across the weighted scores of uniquely identified concepts. To validate the utility of PheMap, we calculated the type 2 diabetes mellitus (T2DM) phenotype scores for 131,347 patients from Vanderbilt’s deidentified EHR data using the top 100 weighted concepts for T2DM. We compared the phenotype score distributions between T2DM cases and controls defined by clinician-validated T2DM algorithm from the eMERGE network4.

Results
PheMap contains quantified concepts for 942 unique disease phenotypes. We were able to incorporate information from two or more online resources for 875 (93%) of the disease phenotypes.

Using the eMERGE T2DM algorithm, we identified 18,313 patients as cases and 38,233 patients as controls from a total number of 131,347 patients. There were 74,806 patients that did not fulfill either criteria. Other than T2DM diagnosis codes, the PheMap for T2DM also highlighted concepts such as ‘blood glucose measurements’, ‘hyperglycemia’, ‘metformin’ and ‘thirst’. The average PheMap phenotype score (standard deviation) was 54.23 (16.92) for eMERGE cases, 5.26 (5.50) for eMERGE controls, and 17.31 (17.69) for eMERGE unclassified patients.
The t-test for T2DM phenotype scores between eMERGE cases and eMERGE controls yielded a p-value < $2.2 \times 10^{-16}$. From manual chart review, we observed that many “low-score” eMERGE cases lacked a T2DM diagnosis but were taking T2DM medication for other reasons. There were only a few “high-score” eMERGE controls, but all of them had some sort of blood sugar disorder.

Figure 1. PheMap phenotype score distributions for eMERGE T2DM classification

Discussion

In our initial validation, PheMap effectively differentiated T2DM cases from controls. PheMap assigns a markedly higher average phenotype score for eMERGE-defined cases than controls. We observed that many of the eMERGE unclassified patients with high phenotype scores were diagnosed with prediabetes or blood sugar disorders, suggesting that the phenotype score may capture information about disease progression. Although PheMap does not assign a case-control label, the significant difference between the distributions of the phenotype score between T2DM cases and controls may allow researchers to use the score as a feature to identify case and control groups. Moreover, the PheMap phenotype score could be used as a continuous variable in future T2DM studies as it may reflect the continuum of disease progression.

PheMap significantly streamlines the phenotyping process with EHRs. We can quickly calculate a phenotype score for all 942 unique phenotypes. Moreover, PheMap is easily portable since the knowledgebase of can be mapped to local EHR structures.

There are some limitations to our current approach. PheMap does not currently incorporate clinical note parsing or evaluate whether a given lab measurement is normal or abnormal. PheMap also relies on online medical resources, which may not include some specific phenotypes or concepts. However, in both of these instances, PheMap and its dictionary of concepts offer a good starting point for researchers to develop enhanced approaches. We plan to continue improving our approach and validating against additional eMERGE phenotyping algorithms.

In conclusion, the PheMap approach has the potential for high-throughput phenotyping that can accelerate the pace of impactful clinical or translational research.

References