Sequencing EHR for Disease Subtyping

Po-Hsiang Chiu, PhD[a], Ning Shang, PhD[a], Chunhua Weng, PhD[a]
[a]Department of Biomedical Informatics, Columbia University, New York, NY

Introduction

Complex diseases are in general not single entities but can be classified into several subtypes. However, disease subtyping is, by and large, qualitative in etiological and pathophysiological differences on a disease-specific basis (e.g. type I vs. type II diabetes) but does not have computationally quantifiable definitions generalizable from one disease to another beyond predefined clinical coding systems. Inspired by *life language* processing[1], this study introduces the notion of *EHR sequencing* in which medical coding sequences (MCSs), comprising temporally ordered medical codes, are generated as a patient representation that facilitates the unraveling of characteristic diagnosis progressions and treatment pathways as *temporal phenotypes* useful for delineating disease subgroups. In synergy with the MCS representation is a pipeline that encapsulates interrelated predictive analytics ranging from sequence pattern recognitions within MCSs to the vectorization of MCSs using neural linguistic approach[2]. These analytics can be used to identify similar patients sharing coherent sequence properties as archetypal examples linked to a subgroup. Using a validated eMERGE phenotyping algorithm[3], we annotated a patient cohort diagnosed with chronic kidney disease (CKD), wherein severity stages are determined, to demonstrate the utility of EHR sequencing.

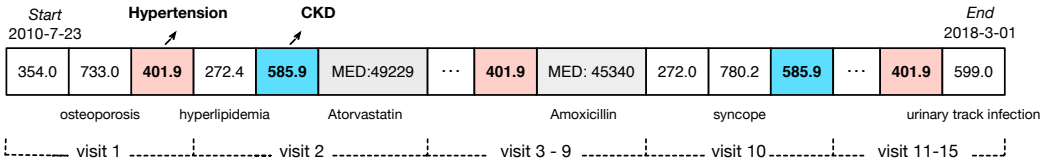

Figure 1. An example MCS with a characteristic alternating pattern between hypertension (401) and CKD (585) intermixed with other comorbid conditions such as disorders of lipoid metabolism (272).

Method

EHR sequencing, as the name suggests, begins by transforming the codified EHR data from clinical data repository (CDR) to a sequential format (i.e. MCS) according to a predefined protocol that delineates the scope of the medical concepts of interest and their ordering in the sequence. As shown in Figure 1, this study focuses on the MCS (one per patient) structured in a temporal order of diagnoses and medications and implicitly grouped in units of clinical visits, each with a distinct timestamp. We postulate that a given disease cohort (e.g. CKD) and its subgroups (e.g. severity stages) can be characterized collectively by patients exhibiting similar diagnosis progressions (e.g. sequence of ICD-9 codes) and treatment pathways (i.e. sequence of medications). Using a CKD cohort as an example, we define *MCS chromosome* as the complete set of MCSs that express a CKD stage, drawing a conceptual analogy to genetic sequences. Each chromosome references *MCS genes* as coding segments within MCSs (potentially disjointed) that functionally determine the stage; by assumption, different stages are characterized by different genes that phenotypically express their associated severity degrees. In light of the MCS abstraction, one can stratify a disease cohort such as CKD in at least two different ways: i) establish differential clinical properties through variations of MCSs across CKD stages and ii) identify MCS genes highly associated with each CKD stage. In particular, we use *paragraph vector*[2] to encode each MCS and evaluate the predictability of CKD stages by formulating a multiclass classification problem using the severity stages as the class labels. In addition, we use Longest Common Subsequences (LCSs) to represent MCS genes, which are obtained by subsampling pairs of MCSs from within each CKD stage, deriving pairwise LCSs (of specified lengths) via dynamic programming, and subsequently selecting the most enriched LCS patterns within each stage.

Experimental Settings and Results

We assess the utility of EHR sequencing by examining i) the predictability of CKD stages through the vectorized MCSs and ii) the difference in enriched LCS patterns across various stages. Using paragraph

vectors of dimension 100 and window length 5, every patient-specific MCS instance is embedded into a 100-D vector space, for which an extension of ROC to multiclass domain is used to evaluate their predictive performance through micro-averaging. The eMERGE-annotated CKD cohort, curated from ODHSI database, consists of 2833 cases and 5 stage-specific labels ranging from stage 1 to stage 5 (n=89, 630, 422, 84, and 778 respectively) representing diminishing kidney functions, plus the control group (n=830). Running the random forest classifier (with 500 trees and Gini impurity as split quality measure) yields, as shown in Figure 2, stage-wise AUC estimates between 0.67 (stage 4) and 0.96 (stage 5). In Table 1, we illustrate example characteristic LCSs as temporal phenotypes that distinguish various CKD stages while leaving out intermixing medication segments for the simplicity of clinical interpretation. Prominent shifts in LCS patterns are observed in end stage CKD compared to those that frequently occur in stage 1-4.

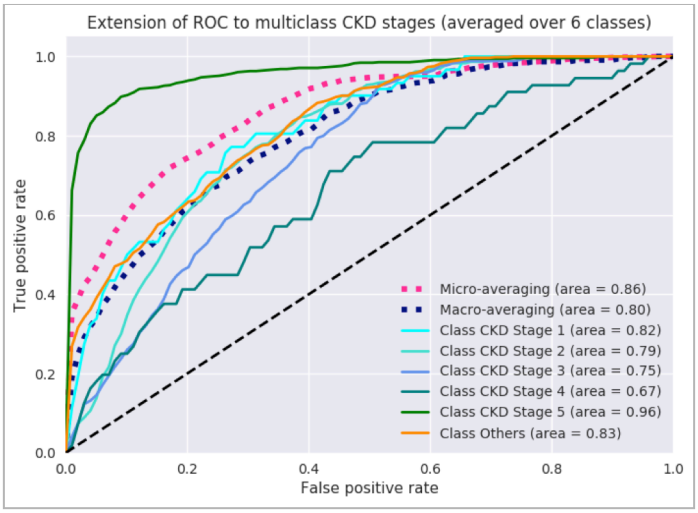

Figure 2. ROC curves of the random forest classifier trained on MCS-derived document vectors predicting CKD stages; performance evaluated via micro-averaging AUCs using 5-fold cross validation.

Table 1. LCSs of high-frequency occurrences as potential MCS genes expressing different CKD stages.

Severity	**LCS Examples** (Length=5)	**Freq.**	**Interpretations**
Stage 1-2	(1) 272.0 401.9 272.4 401.9 272.0 (2) 250.00 401.9 250.00 401.9 250.00	95/719 89/719	Both (1) and (2) represents interleaving patterns between hypertension (401) and comorbid conditions such as diabetes (250), corresponding to pattern (2) that occurs in 89 out of 719 MCSs.
Stage 3-4	(3) 401.9 **599.0** 401.9 272.0 401.9 (4) **786.50** 401.9 272.4 401.9 272.4	64/506 62/506	In addition to (1) and (2), diagnoses involving disorders of urinary track (599) and chest symptoms (786) are observed more often in cases with moderate kidney damages.
Stage 5	(5) V42.0 V58.69 V42.0 V58.69 V42.0 (6) 996.81 V42.0 V58.69 V42.0 V58.69 (7) **585.6 V42.0** V58.69 V42.0 V58.69	585/778 534/778 491/778	Cases with severe loss of kidney functions exhibit noticeable shifts in LCS patterns: i) 585.6+V42.0 signifies end stage renal disease with a kidney transplant status ii) 996+V42 indicates complications with the kidney and iii) long-term medication (V58.69) is another common indicator.

Acknowledgment

This work was funded by CTSA grant 1U54 TR001633-01 and U01 HG008680.

References

1. Asgari E, Mofrad MRK. Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics. PLoS ONE 2015;10(11).
2. Le Q, Mikolov T. Distributed Representations of Sentences and Documents. Int Conf Mach Learn - ICML 2014;32:1188–1196.
3. Shang N, Drawz PE, Kiryluk K, et al. Electronic Health Records-based Computable Phenotype for CKD Diagnosis and Staging. J Am Soc Nephrol 28, 2017: 42.