Developing Phenotypes from Electronic Health Records for Chronic Disease Surveillance

Sarah Conderino[1], Justin Feldman[1], Tom Carton[2], Lorna Thorpe[1]

[1]Population Health, New York University Langone Medical Center, New York, New York, United States, [2]Louisiana Public Health Institute, New Orleans, Louisiana, United States

Objective

To utilize clinical data in Electronic Health Records (EHRs) to develop chronic disease phenotypes appropriate for conducting population health surveillance.

Introduction

Chronic diseases, including hypertension, type 2 diabetes mellitus (diabetes), obesity, and hyperlipidemia, are some of the leading causes of morbidity and mortality in the United States. Monitoring disease prevalence guides public health programs and policies that help prevent this burden. EHRs can supplement traditional sources of chronic disease surveillance, such as health surveys and administrative claims datasets, by offering near real-time data, large sample sizes, and a rich source of clinical data. However, few studies have provided clear, consistent EHR phenotypes that were developed to inform population health surveillance.

Methods

Retrospective EHR data were obtained for patients seen at New York University Langone Health in 2017 (n=1,397,446). To better estimate chronic disease burden among New York City (NYC) adults, the patient population was limited to NYC residents aged 20 or older, who were seen in the ambulatory primary care setting (n=153,653). Rule-based algorithms for identifying patients with hypertension, statin-eligibility, diabetes, and obesity were developed based on a combination of diagnostic codes, lab results or vitals, and relevant prescriptions. We compared the performance of our metric definitions to selected phenotypes from the literature using percent agreement and Cohen's kappa. Patients with discordant disease classifications between the two sets of definitions were analyzed through natural language processing (NLP) on the patients' 2017 medical notes using a support vector machine model. Statin-eligibility is a novel phenotype and therefore did not have a comparable definition in the literature. Sensitivity analyses were conducted to determine how disease burden changed under alternative rules for each metric.

Results

Of 153,653 adult ambulatory care patients in 2017, an estimated 53.7% had hypertension, 12.4% had diabetes, 27.8% were obese, and 30.0% were statin- eligible under our proposed definitions. The estimated prevalence of hypertension increased from 28.1% to 53.7% when diagnostic codes were supplemented with blood pressure measurements and anti-hypertensive medications, while the estimated prevalence of diabetes increased less than one percentage point with inclusion of diabetes-related medications and elevated A1C measurements. There was high agreement between our obesity (94.5% agreement, k=0.86) and diabetes (96.2% agreement, k=0.81) definitions and selected definitions from the literature and moderate agreement between the hypertension definitions (74.8% agreement, k=0.41). NLP classification of discordant cases had greater alignment with the classification results of our definitions for both hypertension (78.0% agreement) and diabetes (71.2% agreement) but did not show strong agreement with either obesity algorithm. Sensitivity analyses did not have large impacts on prevalence estimates for any of the indicators, with all estimates within two percentage points of the final algorithms.

Conclusions

Our proposed rule-based phenotypes using prescriptions, labs, and vitals improved ascertainment of conditions beyond diagnostic codes and were robust to modifications per sensitivity analyses. Results from our algorithms were highly consistent with standard phenotypes from the literature and may improve case capture for surveillance purposes. These algorithms can be replicated across diverse EHR networks and can be weighted to generate population prevalence estimates.