Understanding Chronic Kidney Disease: Leveraging Clinical Records for Phenotype Research

Ellen L Palmer, PhD1,2, Dana Crawford, PhD1,2, John O’Toole, MD1,3, John Sedor, MD1,3, William Bush, PhD1,2

1Case Western Reserve University, Cleveland, OH; 2Cleveland Institute for Computational Biology, Cleveland, OH; 3Cleveland Clinic, Cleveland, OH

Introduction

Chronic Kidney Disease (CKD), defined as having a eGFR of 15-60, affects approximately 13% of the US population1. The majority of these cases are likely due to Diabetes Mellitus (DM) and/or hypertension (HT), and while many people with CKD will not face any complications from their condition1,2, people with DM are at an increased risk of End Stage Renal Disease (ESRD)3. Other clinical risk factors strongly associated with CKD are cardiovascular disease (CVD), structural renal tract disease, hereditary kidney disease/family history, and systemic lupus4. These latter conditions are often treated in conjunction with nephrologists, ensuring that patients are receiving appropriate monitoring to maintain renal function.

However, given that DM, and HT are common conditions, precision medicine approaches to differentiate between individuals at high risk of progression to ESRD and individuals with stable CKD is essential for efficient monitoring and improved patient outcomes5. Previous work developed a clinical CKD registry for tracking patient outcomes1,5, but the scale and scope of information captured was mostly limited to billing codes. In this work, we aim to augment an existing code-based registry1 with additional clinical details derived from EHR measures to construct a higher-resolution research registry for examining the impact of CKD causes, its comorbid conditions, and factors that influence ESRD trajectories over time.

Methods

We accessed complete medical records for 203,344 patients seen at the Cleveland Clinic. These patients were abstracted from an existing clinical CKD registry1, kidney transplant registry, and a chronic renal insufficiency cohort. Our primary aim was to robustly identify CKD, DM, and HT in individuals with clinically identified kidney disease.

Two methods of identifying the primary phenotypes were used: ICD billing codes, and laboratory test results. ICD-9 code inclusion and exclusion criteria were developed from existing clinical registries1,5 and converted to ICD-10; both ICD-9 and ICD-10 were utilized in order to identify all pertinent records. A1c and eGFR values were extracted from laboratory reports and utilized to confirm a diagnosis of DM and CKD respectively. Blood pressure was extracted from visit vitals to confirm a diagnosis of HT. Vectors of eGFR values per patient over time were retained for estimating slopes and clinical trajectories.

To create a high-resolution research registry, we identified individuals who had 1) at least one ICD code for the phenotypes of interest (CKD, DM, HT) and 2) two confirmatory laboratory/vital measures at least 90 days apart. Records which had documented CKD via either a mix of a single ICD code and single blood measure or only contained one data source were retained as a low-resolution validation set for future clinical application testing.

We are in the process of abstracting medications lists in conjunction with these two other data sources. Medications for DM and/or HT documented twice at least 90 days apart are considered strong evidence for actual phenotype presence. The number and quality of records available from all pairwise measures (medication/lab, medication/ICD, lab/ICD) as well as the number and quality of records when all three sources are required will be evaluated compared to a gold standard manual review of 30 records.

Results

Out of 18,580,162 ICD codes for 203,344 individuals, 18,240 ICD codes for 13,464 individuals met the eMERGE CKD algorithm exclusionary criteria5. These individuals were not considered for further analyses.

We identified 286,023 ICD billing codes related to CKD for 82,721 individuals. We obtained 5,144,227 eGFR measures <60 on 150,922 individuals. Given the criteria of one ICD code and at least two blood measures >90 days apart, we were able to label 65,777 records as being high-resolution for CKD documentation.
For DM, we identified 985,849 ICD billing codes for 52,840 individuals. We obtained 532,824 A1c measures ≥ 6.5 for 60,971 individuals. Given the criteria of one ICD code and at least two blood measures >90 days apart, we were able to identify 44,126 records as having high resolution DM documentation.

For HT, we identified 1,461,110 billing codes for 96,574 individuals. We obtained 7,084,818 blood pressure measures that met the stage 2 HT criteria (≥140 systolic or ≥90 diastolic) for 177,272 individuals. Given the criteria of two ICD codes and at least two vital measures >90 days apart, we were able to label 14,331 records as having high resolution HT documentation.

Overall, we classified 65,777 (34.6%) of 189,880 patients with documented kidney dysfunction as having high resolution records. For these 65,777 records, we also had high-resolution documentation of DM 1,568 (2.4%), HT 33,725 (51.3%) and both DM and HT 23,793 (36.2%); only 6,691 high-resolution records. Additionally, we classified 102,089 (53.8%) records as low resolution for CKD reporting; 16,944 contained ICD codes only, and 85,145 records contained lab measures only. In the low resolution CKD records, some individuals had high quality reporting for DM and HT. Finally, we were unable to verify CKD status on 22,014 (11.6%) records.

Table 1. Counts of high-quality CKD records by major known risk factors for 65,777 patients

<table>
<thead>
<tr>
<th></th>
<th>No HT</th>
<th>HT</th>
</tr>
</thead>
<tbody>
<tr>
<td>No DM</td>
<td>6,691 (10.2%)</td>
<td>33,725 (51.3%)</td>
</tr>
<tr>
<td>DM</td>
<td>1,568 (2.4%)</td>
<td>23,793 (36.4%)</td>
</tr>
</tbody>
</table>

Discussion

Development of a high resolution registry for CKD requires additional data beyond clinical codes. Additional validation of code-based conditions through laboratory values, requiring longitudinal follow-up for individuals, and usage of multiple data sources to confirm the presence of a phenotype is necessary to ensure that those included in the registry have high resolution records useful for addressing clinically relevant questions.

Understanding data missingness is equally important for research applications. We retained records that fail to meet the full requirement (verification of the primary phenotype from two data sources on at least two dates >90 days apart). These records can be utilized later on in the research process for 1) determining if there is differential missingness between those with high resolution records and those without and 2) assessing if any novel phenotypic patterns identified in the high resolution set can be found in those with incomplete records. Since this lower-resolution set is the more typical state of records, validation within the lower resolution cohort will provide more accurate information on the likely impact of precision medicine algorithms in a clinical setting.

The research registry we report on here has been developed with OMOP data formatting considerations in mind so that future work can include validating primary and secondary findings in other clinical datasets. We are in the process of integrating medications lists as a third source of phenotype data for DM and HT, assessing the concordance between our registry classification and expert reviewers, and data mining for relevant novel clinical patterns that contribute to ESRD.

References