Limitations of ICD-10-CM for Quantifying the Burden of Rare Diseases in Health Care Systems

Rachel L. Richesson¹, Gary J. Koplik¹, Matthew R. Tribby¹, Isaac Lavine¹, Joseph E. Lucas¹, Olivier Bodenreider², Kin Wah Fung²

¹Duke University, Durham, NC, ²National Library of Medicine, Bethesda, MD

Abstract
There are more than 6,000 recognized rare diseases, and the true prevalence for most is unknown. This student project used existing data, standardized coding systems, and published cross-terminology maps to quantify the number of rare diseases treated in the Duke University Health System and to estimate the number of patients affected by these conditions. We queried our health system over a 5-year period for all patients with at least one ICD-10-CM code associated with a rare disease. Face validation of our results indicated that all of the 40 most prevalent ICD-10-CM rare disease codes in our sample are likely used for common or unknown conditions as well as the rare condition. This work illustrates the known limitations of the ICD-10-CM classification scheme for identifying rare conditions. Additional data types are necessary to identify and quantify the collective burden of rare diseases in healthcare settings.

Introduction
Rare diseases are defined in the U.S. as conditions that affect less than 200,000 Americans. There are over 6,000 named rare diseases, although the exact number varies by source.[1-4] The National Organization for Rare Disorders estimates that up to 30 million (or 1 in 10) Americans have a rare disease,[3] but the true prevalence of a rare disease is difficult to measure. Electronic health record (EHR) data can facilitate the identification of rare disease patients for population surveillance, treatment, and research. The objective of this project was to estimate the number and prevalence of different rare diseases in our health system by examining diagnoses encoded using the International Classification of Diseases (ICD) version 10-CM and to estimate the burden with respect to number of visits.

Because most rare diseases do not have unique ICD-10-CM codes, identifying patients with rare conditions from data routinely captured through EHRs is challenging. Previous work by terminology experts at the National Library of Medicine (NLM) has shown that only 1,386 (21%) of 6,519 rare diseases have ICD-10-CM codes, and more than a quarter of those codes are non-specific, meaning that they map to more than one condition (e.g., Q82.8 Other specified congenital malformations of skin includes 22 rare diseases).[5] Despite the known limitations of ICD-10-CM for identifying rare diseases, we believed that it was an important and logical first step to begin to estimate the number of different rare conditions treated in one large academic medical center.

Methods
This work was conducted as part of the Duke University Data+ summer research program that allows Duke students to work collaboratively to explore data-driven approaches to interdisciplinary challenges. In May 2017, we queried the Duke University Health System (DUHS) for all patients with at least one visit during a 5-year period (May 1, 2012 through April 30, 2017). To ensure anonymity, we requested limited demographics (gender, age in years) and limited the date of visit to month and year only (with a unique visit ID). A statistician (JL) extracted the data from the Epic® Clarity® database. Students completed CITI Human Subjects training and had supervised access to the data on a Virtual Machine. The data set contained all visits and associated diagnoses, encoded in ICD (versions 9 and 10-CM). R software was used for statistical computing and graphics.

A master list of ICD-10-CM codes that correspond to various rare diseases was constructed by Drs. Fung and Bodenreider at the NLM by mapping the latest list of rare diseases recognized by the U.S. Office of Rare Diseases Research (ORDR, part of The National Center for Advancing Translational Sciences (NCATS) at the National Institutes of Health (NIH)[6]) to the Unified Medical Language System (UMLS) and the published maps between SNOMED CT and ICD, as described in [5]. The list of 6,627 rare disease names recognized by ORDR in 2017 was mapped to the UMLS, using exact and normalized string matches, followed by semantic group validation (with
restriction to the Semantic Group ‘Disorders’). Through the UMLS concept structure, matches to SNOMED CT, ICD-9-CM and ICD-10-CM codes were identified. For rare diseases with a UMLS-identified SNOMED CT code match but no direct ICD match, we used publicly available maps from SNOMED CT to ICD-9-CM and ICD-10-CM to identify matches other than direct equivalent matches, following our methodology in [5]. This method generated mappings for a total of 2,913 ORDR rare diseases (1,218 direct mappings, 1,632 indirect mappings) to ICD-10-CM, including some one-to-many and many-to-one maps.

The 2,913 ICD-10-CM rare disease codes were used to query the 5-year DUHS dataset for unique patients with at least one instance of rare disease code at any visit in the 5-year time period. The frequencies for the ICD-10-CM rare disease codes were generated and reviewed by a clinical domain expert (pediatrician, geneticist, rare disease specialist) for interpretation.

The same expert then reviewed codes from ICD-10-CM Chapter 10 (Metabolic Disorders, E70 - E88) that were associated with ORDR rare disease names in order to identify a set of ICD-10-CM codes that were precise enough to identify patients with a given rare disease (high recall) and only those patients who truly have that rare disease (high precision). Our expert has experience managing and documenting these rare disorders as part of her work as physician and translational scientist for genetic conditions. She reviewed 429 relationships of rare disease names to corresponding ICD-10 CM Chapter 10 codes. Of the 429 rare diseases - ICD-10-CM mappings, 119 ICD-10-CM codes were identified by the expert as specific (i.e., used uniquely for a rare disease) enough to be used in a query for that rare disease. Considering only these 119 codes as rare, we partitioned a subset of only patients with at least one ICD-10-CM Chapter 10 code into rare and non-rare patients. We ran frequencies for all of these codes, and experts reviewed the numbers of records retrieved. Because we had no gold standard for rare disease diagnosis and no resources for clinical validation of our results, we did not calculate recall, precision, specificity, sensitivity, or PPV.

Results

Our initial query of the DUHS data set yielded approximately 1,245,243 unique patients, of which 580,952 included at least one visit diagnosis coded with one of 3,166 ICD-10-CM rare disease codes. The counts, ICD-10-CM code, and corresponding ORDR rare disease names for the ten most frequent ICD-10-CM rare disease codes by prevalence are presented in Table 1.
For face validity of our rare disease query, a clinical domain expert reviewed a report of the most common ICD-10-CM rare diseases that we found and their corresponding frequencies. The expert opined that none of the 40 most prevalent ICD-10-CM rare disease codes were specific enough to identify rare diseases. In other words, the expert verified that the ICD-10-CM codes with the highest frequencies in our dataset could very plausibly be used in practice for common conditions as well. For example, our highest frequency code (ICD-10-CM code E11.9, Type 2 diabetes mellitus without complications, n=78,370) was in our list of ICD-10 codes associated with a rare disease because the ORDR rare disease Maturity-Onset Diabetes of the Young, type 2 maps to the same ICD-10-CM code as the very prevalent diagnosis of Type 2 Diabetes. Similarly, the ICD-10-CM code E55.9 (Vitamin Deficiency, Unspecified) corresponds to the rare disease Rickets, but clearly could be used for a range of symptoms or findings suggesting any suspected vitamin deficiency as well.

In our next iteration, we tried to validate a set of ICD-10-CM codes for rare diseases with the help of an expert who reviewed the codes in Chapter 10 Metabolic Diseases from the perspective of a physician that uses those codes in practice. Of 119 ICD-10-CM codes that our expert found potentially specific for rare metabolic diseases based on her experience as a physician treating and documenting these disorders, 79 were in our data set. The resulting data contained 150,024 patients, of which 2,062 had at least one rare metabolic code as defined by our expert. The five most common rare metabolic diseases using the expert-vetted rare disease ICD-10-CM codes, along with the associated rare disease names and synonyms recognized by the ORDR, are presented in Table 2.
Table 2. Counts of unique patients for the five most frequent ICD-10-CM metabolic rare disease codes (validated by clinical expert), with corresponding rare disease names recognized by the NCATS Office of Rare Diseases Research (ORDR), DUHS 2012-2017.

<table>
<thead>
<tr>
<th>Count</th>
<th>ICD-10-CM Code</th>
<th>ICD-10-CM Name</th>
<th>ORDR Rare Disease Names that Map to this ICD Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>431</td>
<td>E74.02</td>
<td>Pompe disease</td>
<td>Adrenomyopathy; AMD; MAD; Mandibuloacral dysplasia; Glycogen storage disease type 2; Acid maltase deficiency disease; Alpha-1,4-glucosidase deficiency; Cardiomegaly glycogenica diffusa; Deficiency of alpha-glucosidase; Deficiency of lysosomal alpha-glucosidase; GSD II; Pompe disease; Age-related macular degeneration; Age-related maculopathy; AMD; ARM</td>
</tr>
<tr>
<td>299</td>
<td>E72.20</td>
<td>Disorder of urea cycle metabolism, unspecified</td>
<td>Urea cycle disorders; UCD</td>
</tr>
<tr>
<td>201</td>
<td>E72.11</td>
<td>Homocystinuria</td>
<td>Homocystinuria due to CBS deficiency; CBS deficiency; Cystathionine beta-synthase deficiency; Homocystinuria due to cystathionine beta-synthase deficiency; Homocystinuria</td>
</tr>
<tr>
<td>175</td>
<td>E78.3</td>
<td>Hyperchylomicronemia</td>
<td>Hyperlipoproteinemia type 5; Hyperchylomicronemia late onset; Hyperchylomicronemia with hyperprebetalipoproteinemia, familial; Hyperlipemia combined fat and carbohydrate-induced; Hyperlipemia mixed; Hyperlipemia type V; Hyperlipoproteinemia type V; Mixed hyperlipemia; Type V hyperlipoproteinemia; Chylomicron retention disease; CMRD; Hypobetalipoproteinemia with accumulation of apolipoprotein b-like protein in intestinal cells; Lipid transport defect of intestine; Familial lipoprotein lipase deficiency; Burger-Gnzt syndrome; Endogenous hypertriglyceridermia; Familial fat-induced hypertriglyceridermia; Familial hyperchylomicronemia; Familial LPL deficiency; Lipase D deficiency; LIPD deficiency; Lipoprotein lipase deficiency, familial; LPL deficiency; Type I hyperlipoproteinemia</td>
</tr>
<tr>
<td>97</td>
<td>E85.8</td>
<td>Other amyloidosis</td>
<td>AL amyloidosis; Amyloidosis AL; Amyloidosis primary systemic; Light chain amyloidosis; Primary AL amyloidosis; Primary amyloidosis (Formerly); Primary systemic AL amyloidosis; Primary systemic amyloidosis; Systemic AL amyloidosis; Amyloidosis AA; AA Amyloidosis; Amyloid A amyloidosis</td>
</tr>
</tbody>
</table>

Discussion

The aim for this project was to develop an approach for quantifying the number and prevalence of rare diseases that can be replicated in other health systems. The ICD-10-CM coding system for medical diagnoses, mandated by the Department of Health and Human Services, is one of the few coding systems standardized nationally across health providers. Therefore, the use of ICD-10-CM codes was a logical strategy to understand the prevalence of different rare diseases while ensuring the generalizability of the query and approach. Rare diseases might also be noted in patient problem lists using SNOMED CT, but there are many known limitations with problem lists with respect to accuracy, completeness, and timeliness.[7, 8] Rare diseases could also be documented in free text reports, but at this time, most organizations do not have the capability or resources to examine free text data on a large scale.

Our investigation has several significant limitations. First, the ICD-10-CM classification system was not designed to identify rare conditions. In fact, as a global classification of all diseases, it has many categories of “other” disorders specifically designed to capture rare conditions rather than enumerate thousands of conditions with codes that might never be used in most settings. Consequently, the precision of our initial query (for all ICD-10-CM codes associated with rare diseases) was clearly quite low as determined by clinical experts simply by looking at the ICD-10-CM term name and the count of records returned. Second, we only validated our query results by whether the counts seemed reasonable to our expert. A true validation of our query would include the verification of diagnoses by a clinical expert for all or a random sample of the actual patient records retrieved. Unfortunately, the validation of patient records was beyond the scope and resources of this summer student research project.

All health systems were required to transition from ICD-9-CM to ICD-10 CM in 2015. However, our health system used a consistent interface terminology with mappings to both ICD versions before and after October 1, 2015, so we have ICD-10-CM codes available for all years of our dataset. It should be noted that any translation of ICD-9-CM to ICD-10-CM most certainly introduces noise or loss of information. Further, the use of an interface or intermediary terminology brings uncertainty about which terms are presented to clinicians on the EHR screen how those are mapped to different ICD versions by the software provider.

Despite the limitations of ICD-10-CM and our methods, we believe our experience provides a valuable first step in the unsolved problem of estimating the burden of rare diseases in any health system. Because so many ICD-10-CM codes are not precise for rare conditions, future research into rare disease prevalence will need to leverage other data such as labs, medications, genetic tests, and the information recorded in clinical notes to better identify rare disease
cohorts. For example, we could use NLM tools to link RxNorm medication codes to the sulfonylureas drug class and identify which of the 78,370 patients coded with ICD-10-CM E11.9 (Type 2 diabetes mellitus without complications) were prescribed with sulfonylurea drugs, which are used in the rare MODY-type 2 but not generic type 2 diabetes mellitus. Regardless of the data sources, organizations should embrace a strategy to use previously validated rare disease code sets if possible and to validate locally as well.

Although we failed to meet our project objectives to quantify the number and impact of rare diseases at DUHS, we successfully exposed the significant challenges in generating a cohort of rare disease patients from health system data, and we believe that we are the first to attempt this type of query. The challenges we faced underscore the need for analysts to understand the limitations of the coding systems that are used in structured healthcare data as well as the need for domain knowledge in both the forming and interpretation of results. Future health services researchers and analysts should be wary of using any kind of rare disease ICD code list.

Our work and experience speaks to the immediate yet unaddressed need for explicit, validated, reproducible, and most importantly computable (EHR-based) phenotype definitions for rare diseases. Computable phenotypes (or query definitions) for rare diseases could be used in national research networks to generate national prevalence estimates for various rare diseases, but unfortunately are not included in existing phenotype inventories such as PheKB.[9] As researchers develop and validate rare disease query definitions using different types of data commonly found in EHRs, they can share their experience on portals such as PheKB and enable others to replicate their work and identify rare diseases in other health care organizations.

Acknowledgements
Matthew Tribby (BS Computer Science anticipated 2019) and Gary Koplik (MS Economics and Computer Science anticipated 2018) conducted this work as part of the Duke Data+ program, a paid summer internship experience representing Duke’s commitment to data science. The authors wish to thank Paul Bendich and Ashlee Valente for coordination and support of the Duke Data+ program. The work described used cross terminology mappings generated by Olivier Bodenreider and Kin Wah Fung of the National Library of Medicine. We are especially grateful to Priya Kishnani for her hours of assistance reviewing ICD-10 codes and suggestions for moving forward.

References