Finding That Needle in the Haystack: Computable Phenotypes

Dorey Glenn and Keisha L. Gibson
School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina

JASN 30: 2279–2280, 2019.
doi: https://doi.org/10.1681/ASN.2019101037

In a 1532 publication, Sir Thomas More—an English lawyer and social philosopher—likened one's ability to find a line in St. Austin's writing to finding a needle in a meadow—later interpreted as a "needle in the haystack."[1] The isolation experienced by patients with rare diseases can often make them feel like lost needles in the meadow of our large complex health care systems. As of 2018, there were 327.2 million people with the potential to have data stored across 5534 registered hospitals in the United States alone along with over 1100 different electronic health record vendors.[2] The frustration experienced by those eager to connect individuals with rare diseases to clinical care and research opportunities is magnified by the fact that our electronic medical record systems at times fall short of "meaningful use" and do many things to the exception of connecting to other systems and providing efficient communication. The effect that this complicated meadow has on clinical care has been well described, but the barriers that this poses to patient recruitment into clinical and other research trials are less addressed in published literature.

Recruitment of patients for clinical trials is an area ready for innovation. At a time when there is measurable interest and support from industry sponsoring clinical trials in the glomerular disease space, connecting the right patients to the right trials as efficiently as possible is of great import. Harnessing data from electronic health records (EHRs) to streamline and personalize clinical trial recruitment using computable phenotypes has the potential to decrease the time and money associated with clinical trial setup. Indeed, global spending on clinical trial technology and services is expected to reach $69 billion per year by 2025.[3] EHR-powered analytic tools are already commercialized for the purposes of protocol feasibility, cohort analysis, and study site identification.

Computable phenotypes represent the building blocks with which we can interact with "big" health care data for the purposes of advancing patient care and clinical outcomes. Comparative effectiveness research, quality improvement initiatives, clinical trial recruitment, or even the possibility of conducting high-quality, multicenter, clinical trials within an EHR also exist. Operationalizing these activities using EHR data depends on having validated tools to define patient characteristics, diseases, or clinical events of interest. Numerous computable phenotypes have been developed, validated, and made available publicly, including for CKD, treatment resistant hypertension, type 2 diabetes, steroid-induced osteonecrosis, and many others.[4] Most recently, Norton *et al.*[5] published an article in the September 2019 issue of *CJASN* detailing their development of a pragmatic computable phenotype to identify patients likely to have CKD.

In this issue of *JASN*, Denberg *et al.*[6] report their newly developed computable phenotype for glomerular disease using a rule-based algorithm derived from 561 outpatient nephrology clinic encounters at the Children's Hospital of Philadelphia. Multiple data elements were evaluated for use in the computable phenotype, including demographic information, diagnostic codes, laboratory data (urine protein to creatinine ratio), medication prescriptions (*i.e.*, angiotensin-converting enzyme inhibitor), and kidney biopsy reports. The authors show that a phenotype using a combination of health care encounters, diagnosis codes, and kidney biopsy procedure codes could be used to identify patients with glomerular disease who have a positive predictive value of 92% and a negative predictive value of 100%. The computable phenotype was then validated using structured chart reviews at eight additional institutions within the PEDSnet collaborative (a large, national community of hospitals and healthcare organizations, researchers and clinicians, and patients and families), totally 800 patients with glomerular disease and 798 patients with nonglomerular disease.[6]

Several strengths of the study should be noted, including its rigorous development and validation across multiple collaborating children's hospitals. The authors report that the majority of urine protein assessments were on the basis of urine dipstick measurements, and they were excluded due to concerns that dipsticks might not be able to discriminate glomerular from nonglomerular proteinuria.

Published online ahead of print. Publication date available at www.jasn.org.

Correspondence: Dr. Keisha L. Gibson, Division of Nephrology and Hypertension, UNC Kidney Center, University of North Carolina at Chapel Hill, 7054 Burnett Womack, Chapel Hill, NC 27599-7155. Email: kgibson@med.unc.edu

Future work might test this hypothesis given the potential application of this phenotype to identify clinically significant subgroups of patients with glomerular disease or particular disease states, such as disease remission or relapse.

The recent boom of tools to query EHRs is a welcome development given the many potential applications, such as facilitating quality initiatives in our growing era of value-based care medicine and efficiently identifying outcomes or adverse events in clinical trials/studies to name a few. Any new technology involving protected health information requires a delicate balance between patient privacy and data privacy. With the enormity of data points and complexity of the phenotypes, collaboration between physicians, computer scientists, and informaticians is paramount. Lastly, clinical phenotypes can change over time and are dynamic, and therefore, the methods to update these computable phenotypes must also be dynamic and maintain the capability of being validated efficiently.

DISCLOSURES

Dr. Gibson reports honorarium received as a member of the Reata Inc., CKD Advisory Board unrelated to this editorial work.

Dr. Gibson and Dr. Glenn are supported in part by the Cure GN Study: National Institutes of Health grant NIDDK-4-UM1-DK100867-04. Dr. Gibson is also a co-investigator for the Neptune Study (NIDDK 1U45 DK083912), and has research support from Retrophin, Inc and Aurinia Inc. Funders of both authors had no role in the editorial content.

REFERENCES

1. Gordon WM: A needle in a meadow: A missing reference in the More-Frith controversy. *Moreana* 52: 19–22, 1976
2. 2019 AHA Hospital Statistics; Publisher: Health forum, American Hospital Association Affiliates; January 2019 edition. Chicago, IL; United States.
3. May M: Clinical trial costs go under the microscope [published online ahead of print March 6, 2019]. *Nat Med* doi:10.1038/d41591-019-00008-7
4. Kirby JC, Speltz P, Rasmussen LV, Basford M, Gottesman O, Peissig PL, et al.: PheKB: A catalog and workflow for creating electronic phenotype algorithms for transportability. *J Am Med Inform Assoc* 23: 1046–1052, 2016
5. Norton JM, Ali K, Jurkovitz CT, Kiryluk K, Park M, Kawamoto K, et al.: Development and validation of a pragmatic electronic phenotype for CKD. *Clin J Am Soc Nephrol* 14: 1306–1314, 2019
6. Denberg M, Razzaghi H, Bailey LC, Sorrano D, Pollack A, Dharnidharka V,et al.: Using electronic health record data to rapidly identify children with glomerular disease for clinical research. *J Am Soc Nephrol* 30: 2427–2435, 2019

See related article, "Using Electronic Health Record Data to Rapidly Identify Children with Glomerular Disease for Clinical Research," on pages 2427–2435.