Computable Phenotypes: Standardized Ways to Classify People Using Electronic Health Record Data

by Lilia Verchinina, PhD; Lisa Ferguson, MSI; Allen Flynn, PharmD; Michelle Wichorek, PhD; and Dorene Markel, MS, MHSA

Abstract

Computable phenotypes (CPs) are an increasingly important structured and reproducible method of using electronic health record data to classify people. CPs have the potential to provide important benefits to health information management (HIM) professionals in their everyday work. A CP is a precise algorithm, including inclusion and exclusion criteria, that can be used to identify a cohort of patients with a specific set of observable and measurable traits. With the use of CPs, a series of technical steps can be taken to automatically identify people with specific traits, such as people with a particular disease or condition. CPs were first used outside of the HIM domain for clinical trials and network-based research. Because CPs are becoming more easily shareable, they have the potential to be used by HIM professionals to help improve coding, reporting, management, sharing, and reuse of clinical information.

Keywords: computable phenotype, electronic health record (EHR), cohort, patient data

Introduction

Health information management (HIM) professionals are stewards of health information. They have expertise in the use of information from electronic health records (EHRs) to meet business objectives and to solve problems for clinicians, caregivers, patients, and families. As different methods of using patient data from EHRs arise, HIM professionals strive to understand and support them. Here, we discuss the use of a type of specification incorporating EHR data called a computable phenotype (CP). A CP is a precise, shareable, reproducible, and documented method for using EHR data to categorize people for a variety of purposes. CPs are currently being used in biomedical research. Because CPs specify how to identify cohorts of individuals using EHR data, they can be useful in the work of HIM professionals.

To better explain CPs, we begin by defining what a phenotype is, and then we discuss the definition of the term *computable phenotype*. According to the glossary at genome.gov, a phenotype is a composite of an individual person's "observable traits, such as their height, eye color, and blood type."[1] The prefix *pheno-* means "showing." A phenotype is something that can be observed. The suffix *-type* suggests a typology, or classification scheme. Hence, a phenotype is a way of classifying organisms, including people, based on characteristics they have that are observable and measurable.

Building on the definition of *phenotype* above, the term *computable phenotype* refers to a shareable and reproducible algorithm precisely defining a condition, disease, complex patient characteristic, or clinical event using only data processed by a computer, principally EHR data.[2] As a shareable and reproducible algorithm, a CP is a tool that can be used to identify patient cohorts with a specific medical condition of interest associated with a specific set of observable and measurable traits. Further, because they are shareable and provide formal specifications of diseases and events, CPs can serve as standards, allowing patient cohorts to be compared and combined more easily than they are today.

Although this definition of a CP is a good start, the definition is not fully settled. Robinson asserts that a CP is a standardized method for capturing phenotypic manifestations of disease.[3] Richesson et al. suggest that CPs are intended to be based *solely* on data that can be processed by a computer, that is, EHR data.[4] For our purpose here, we will consider CPs to be formally specified, written algorithms that detail a list of EHR data and serve to define inclusion and exclusion criteria for a specific disease cohort.

We believe that CPs are growing in importance beyond their original use to establish patient cohorts for clinical trials with EHR data. We anticipate that HIM professionals will soon be likely to encounter CPs in their work because CPs can also be used to determine treatment eligibility and to assess the quality of record coding, among other possible clinical, research, public health, and business uses.

A Look at Computable Phenotypes in General

To further explain CPs and how they are used, we outline a three-step process for generating patient cohorts using a CP. On the basis of a literature review[5–9] and our experience using CPs to inform EHR data queries for research,[10–13] we have developed a process diagram to show how CPs have generally been used to classify patients (see Figure 1).

The process described in Figure 1 begins on the far left with a CP, or an algorithm that precisely describes a condition, a disease, or a set of traits. Like all CPs, this CP provides a standard algorithm for including and excluding individuals in a cohort based on criteria documented in EHR records.

To implement a CP locally, the user must transform the CP algorithm into a database query. The query is created by mapping the inclusion and exclusion criteria documented in the CP to a defined set of data elements and logical expressions that are specific to the EHR system, clinical data repository, or other data source being used.[14] This query is then run against the source database to generate a patient cohort representing those individuals that meet the CP algorithm. Because different source databases have different schemas, CPs cannot include actual database queries. Instead, the role of the CP is to provide a standard algorithm, with inclusion and exclusion criteria, so that HIM professionals, database administrators, clinicians, and others can construct similar database queries that fit their own source databases.

Hence, as a specified algorithm, the CP accurately, but only generally, communicates a set of standard inclusion and exclusion criteria. Its criteria define a condition or cohort so that the results of local database queries are directly comparable across time and place. For this reason, CPs are particularly helpful for research.

Because source databases typically include many tables of data, queries that correspond to CPs often involve complex “join” functions that combine data from many tables in a source database according to the criteria included in a CP. The complexity of this work increases further when data from multiple sources are queried. In such cases, patient data must first be merged from the various sources with the use of unique patient identifiers. Such unique identifiers have been described in the HIM literature previously.[15] This paper reports on work that has taken place over the past several years to develop and share CPs in ways that support improved standardization of patient cohort development and clinical event definitions across diverse sites for a variety of purposes.

Types of Data for Which CPs Can Be Used

Within source databases, which contain data extracted from EHRs, several types of data are particularly well suited for identification with CPs. These types of data include vital signs, prescribed medications, coded diagnoses and procedures, and clinical notes. They may also include International Classification of Diseases (ICD) codes and Current Procedural Terminology (CPT) codes. What makes CPs different from the locally developed queries that HIM professionals currently use is that CPs have the potential to be widely reproduced and to be shared as standards that formally specify patient cohorts and clinical events.

Laboratory data are often used in CPs. For instance, to select a cohort and generate a list of patients with Type 1 diabetes and not Type 2 diabetes, the results of blood tests may be included in the CP to improve the accuracy of its criteria for Type 1 diabetes.

When medication data are used in a CP, sometimes pharmacy dispensing data can be helpful to infer whether a prescribed medication is actually being taken by a patient.

Data from patient surveys and screening questionnaires may also be included in CPs. For example, the American Diabetes Association recommends screening youth who have difficulty achieving treatment goals about their mental health, including screening for depression and coping skills.[16] Providers collect these screening data. Once these data are stored in EHR databases, they become observable and measurable traits that can be included in CPs and queried to identify patient cohorts and classify individual patients.

Finally, organizational and provider-specific information are additional types of EHR data that can be used to identify and classify patients.[17, 18] These types of data help to distinguish patients by medical services received[19] and by the characteristics of the providers who provide the medical services.[20]

Common Uses of CPs

CPs have traditionally been used in domains outside of HIM. In clinical research, especially, investigators regularly use CPs to identify a specific population, or cohort, of patients they wish to study. With the use of CPs, cohorts are selected systematically according to a specified set of criteria that can be checked using EHR data.

The effectiveness of using CPs for cohort identification in research is facilitated by the widespread sharing and dissemination of CPs. Sharing CPs can accelerate their refinement and validation by multiple healthcare organizations. The Phenotype KnowledgeBase, or PheKB, is an online collaborative environment that provides tools for editing and improving CPs, as well as a place for storing and sharing them publicly.[21, 22]

CPs have become increasingly relevant in this era of network-based clinical research. An example of network-based research is the Electronic Medical Records and Genomics (eMERGE) national network, which is organized and funded by the National Human Genome Research Institute. The eMERGE network brings together data from multiple institutions, combining genetic research with EHR data to support research on the genetic determinants of disease. A significant aspect of the eMERGE network includes developing CPs for multiple genetic diseases, including cancer, epilepsy, chronic kidney disease, and hearing loss.[23, 24]

CPs are also used by a national research network called PCORnet. Funded in 2014 by the Patient Centered Outcomes Research Institute, PCORnet comprises 13 smaller regional Clinical Data Research Networks, 20 disease-specific People-Powered Research Networks, and two Health Plan Research Networks.[25] At the national level, PCORnet has developed CPs to identify clinical obesity and other medical conditions. PCORnet members must demonstrate their ability to use common CPs for network-wide cohort identification in support of patient-centered research.[26]

Two Examples of CPs

Sickle Cell Disease

To provide an example of an actual CP, here we describe a CP for the identification of patients with sickle cell disease. To develop this CP, Michalik et al. conducted a retrospective study using EHR data from the Children’s Hospital of Wisconsin.[27]

The sickle cell disease CP has just two criteria:

1. ICD-9 diagnosis codes for sickle cell disease or "other sickle cell disease" in the patient's EHR medical problem or diagnosis lists; and
2. a documented history of two outpatient visits, at least 30 days apart, or one hospitalization, related to sickle cell disease.

The researchers ran queries using the sickle cell disease CP against their institutional research data warehouse, which contains EHR data, to generate a qualifying list of patients. The resulting list of patients demonstrated a 99.4 percent positive predictive value (PPV) for confirmed sickle cell disease, indicating that the CP is very accurate for the identification of this population.[28]

To further validate these results, the researchers asked a neighboring health system to also run the sickle cell disease CP. In this second validation, 415 of 433 patients were confirmed to meet the inclusion criteria, resulting in a PPV of 95.8 percent. This relatively high PPV, in combination with the first PPV, signaled that the sickle cell disease CP could be more widely adopted and used to identify patients with sickle cell disease.[29] Its developers subsequently made the sickle cell disease CP available to others by depositing it in the online PheKB database.[30]

Peripheral Arterial Disease

An example of a more complex CP is one developed by the Mayo Clinic to identify patients with peripheral arterial disease (PAD).[31] The PAD CP consists of five complex criteria, each involving a variety of data domains. The first criterion is definitive for identifying a patient with PAD. At least two of the second through fifth criteria can also establish a definitive case of PAD. The five criteria that constitute the PAD CP are as follows:

1. Ankle brachial index below 0.9 OR ankle systolic blood pressure greater than 255 mmHg as a result of nonatherosclerotic causes of PAD.
2. A diagnosis code for PAD from one of three ICD-9 code families is found in the EHR.
3. One of these three subcriteria:
    a. One of the ICD-9-CM/CPT-4 codes for lower extremity artery angiography plus one concurrent code for noncoronary vessel stents.
    b. One of the ICD-9-CM/CPT-4 codes for lower extremity artery surgical and percutaneous vascular interventions excluding the cases when one of the codes for alternate reasons for this surgery are present.
    c. One of the ICD-9-CM/CPT-4 codes for lower extremity amputation excluding the cases when one of the codes for nonvascular amputation is also present.
4. One of these two subcriteria:
    a. A lower extremity arteries phrase is found through natural language processing (NLP).
    b. An occlusive arterial disease phrase is found through NLP.
5. One of two medications for claudication are prescribed and in use by the individual.

This PAD CP is being used in a research project to identify patients with PAD.

How and Why CPs Are Relevant in HIM

CPs may be relevant for HIM professionals for several reasons. The first reason is the most obvious. As the use of CPs shifts from clinical research into treatment, payment, and operations, HIM professionals will likely be asked to work with CPs to standardize their everyday efforts. Because CPs provide common and precise definitions of diseases, CPs could become part of the general tool kit that all HIM professionals use for coding, reporting, managing, and sharing clinical information.

In addition, for the purpose of improving coding quality, CPs have the potential to automate comparisons between what is coded and what the CPs would suggest could be coded. By automatically applying CPs, software applications can become capable of highlighting gaps between what is documented, what is coded, and the precise definition of a disease or condition in a CP. The use of tools such as CPs to help standardize and thereby potentially improve the quality of coding has obvious financial implications.

We anticipate that CPs are likely to be shared by many organizations in the future. If this sharing occurs and CPs begin to form a standard set of specifications for patient cohorts and clinical events, then we would expect HIM professionals to have new opportunities to help define CPs and to collaborate with database and information technology experts to implement and test CPs in HIM practice. This work could result in significant efficiency gains if CPs can be developed once and used multiple times by many organizations.

Furthermore, in an age of genomics, disease classification is expected to continue to become more complex.[32–35] Hence, CPs may eventually play a role by directly guiding the coding process as more diseases come to have clinically relevant genetic variants identified.

CPs are also relevant for HIM record searching and sampling procedures. CPs can be used to discover and sample records to identify precise populations of individuals who meet the criteria expressed in a CP.

Finally, CPs can support the HIM professional goal of minimum necessary data sharing. With the use of CPs, it becomes possible to precisely explain why an individual's documented observations either include or exclude the individual as a person with a specific disease. For this reason, CPs offer predefined and limited data that may often be the minimum necessary data to share for the purpose of documenting or evaluating a specific condition.

Conclusion

CPs are standardized, shareable, and reproducible algorithms that precisely define a condition or disease. They can be used to guide the development of database queries that result in cohorts of patients who either do or do not exhibit certain observable and measurable traits. Although CPs have typically been used to support research using EHR data, they are likely to become tools that HIM professionals will use in their everyday work. HIM professionals now have the opportunity to begin evaluating how CPs can support record coding, record sampling, appropriate EHR data sharing, and other aspects of their professional work. Expansion of the use of CPs could further advance the field of HIM and increase the value of EHR data in many aspects of treatment, payment, and operations as well as medical/clinical research.

Lilia Verchinina, PhD, is a data analyst at the University of Michigan Brehm Center for Diabetes Research in Ann Arbor, MI.

Lisa Ferguson, MSI, is a program manager at the University of Michigan Medical School Department of Learning Health Sciences in Ann Arbor, MI.

Allen Flynn, PharmD, is a research analyst and technology lead at the University of Michigan Medical School Department of Learning Health Sciences in Ann Arbor, MI.

Michelle Wichorek, PhD, is a project manager at the University of Michigan Brehm Center for Diabetes Research in Ann Arbor, MI.

Dorene Markel, MS, MHSA, is the director of the University of Michigan Medical School Department of Learning Health Sciences in Ann Arbor, MI.

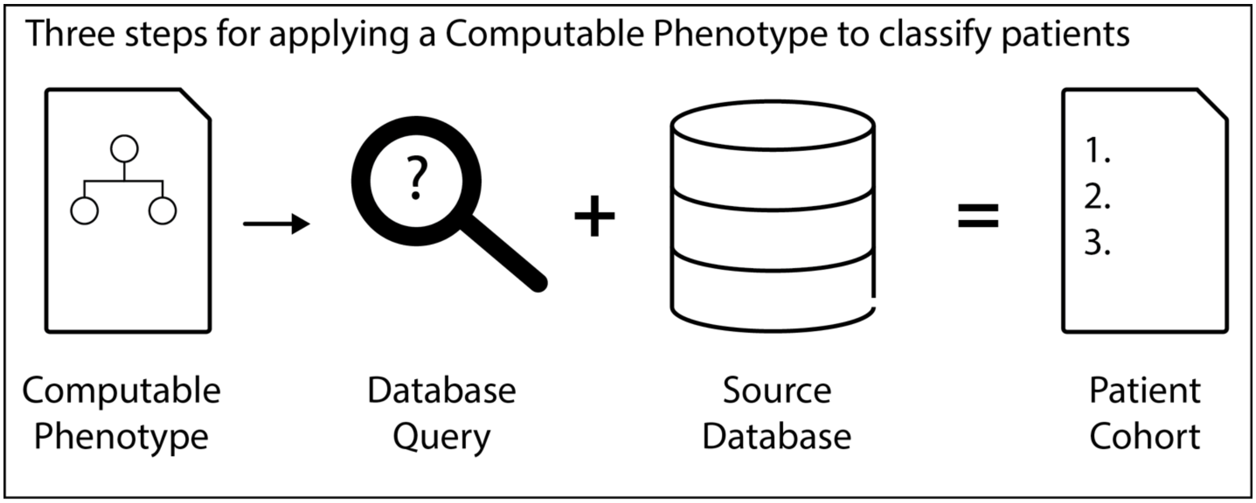
Notes

1. National Human Genome Research Institute. “Phenotype.” Available at https://www.genome.gov/glossary/index.cfm?id=152 (accessed January 18, 2018).
2. Richesson, R. L., et al. “Electronic Health Records Based Phenotyping in Next-Generation Clinical Trials: A Perspective from the NIH Health Care Systems Collaboratory.” *Journal of the American Medical Informatics Association* 20, no. e2 (2013): e226–e231.
3. Robinson, P. N. “Deep Phenotyping for Precision Medicine.” *Human Mutation* 33, no. 5 (2012): 777–80.
4. Richesson, R. L., et al. “Electronic Health Records-based Phenotyping.” Rethinking Clinical Trials: A Living Textbook of Pragmatic Clinical Trials. 2014. Available at https://sites.duke.edu/rethinkingclinicaltrials/ehr-phenotyping/ (accessed May 13, 2018).
5. Richesson, R. L., M. M. Smerek, and C. Blake Cameron. “A Framework to Support the Sharing and Reuse of Computable Phenotype Definitions across Health Care Delivery and Clinical Research Applications.” *EGEMS* 4, no. 3 (2016): 1232.
6. Newton, K. M., et al. “Validation of Electronic Medical Record-based Phenotyping Algorithms: Results and Lessons Learned from the eMERGE Network.” *Journal of the American Medical Informatics Association* 20, no. e1 (2013): e147–e154.
7. Mo, H., et al. “Desiderata for Computable Representations of Electronic Health Records-driven Phenotype Algorithms.” *Journal of the American Medical Informatics Association* 22, no. 6 (2015): 1220–30.
8. Denny, J. C. “Chapter 13: Mining Electronic Health Records in the Genomics Era.” *PLoS Computational Biology* 8, no. 12 (2012): e1002823.
9. Jensen, P. B., L. J. Jensen, and S. Brunak. “Mining Electronic Health Records: Towards Better Research Applications and Clinical Care.” *Nature Reviews Genetics* 13, no. 6 (2012): 395–405.
10. Goodrich, D. E., et al. “Sex Differences in Weight Loss among Veterans with Serious Mental Illness: Observational Study of a National Weight Management Program.” *Women’s Health Issues* 26, no. 4 (2016): 410–19.
11. Littman, A. J., et al. “National Evaluation of Obesity Screening and Treatment among Veterans with and without Mental Health Disorders.” *General Hospital Psychiatry* 37, no. 1 (2015): 7–13.
12. Janney, C. A., et al. “The Influence of Sleep Disordered Breathing on Weight Loss in a National Weight Management Program.” *Sleep* 39, no. 1 (2016): 59–65.
13. Farmer, M. M., et al. “Depression Quality of Care: Measuring Quality over Time Using VA Electronic Medical Record Data.” *Journal of General Internal Medicine* 31, no. 1 (2016): 36–45.
14. Richesson, R., et al. “Electronic Health Records-based Phenotyping.”
15. Godlove, T., and A. W. Ball. “Patient Matching within a Health Information Exchange.” *Perspectives in Health Information Management* 12 (Spring 2015).
16. Silverstein, J., et al. “Care of Children and Adolescents with Type 1 Diabetes: A Statement of the American Diabetes Association.” *Diabetes Care* 28, no. 1 (2005): 186–212.
17. Harrold, L. R., T. S. Field, and J. H. Gurwitz. “Knowledge, Patterns of Care, and Outcomes of Care for Generalists and Specialists.” *Journal of General Internal Medicine* 14, no. 8 (1999): 499–511.
18. Krauss, J. C., et al. “Is the Problem List in the Eye of the Beholder? An Exploration of Consistency across Physicians.” *Journal of the American Medical Informatics Association* 23, no. 5 (2016): 859–65.

19. Harrold, L. R., T. S. Field, and J. H. Gurwitz. “Knowledge, Patterns of Care, and Outcomes of Care for Generalists and Specialists.”
20. Krauss, J. C., et al. “Is the Problem List in the Eye of the Beholder? An Exploration of Consistency across Physicians.”
21. Kirby, J. C., et al. “PheKB: A Catalog and Workflow for Creating Electronic Phenotype Algorithms for Transportability.” *Journal of the American Medical Informatics Association* 23, no. 6 (2016): 1046–52.
22. PheKB. “What Is the Phenotype KnowledgeBase?” Available at https://phekb.org/ (accessed January 18, 2018).
23. McCarty, C. A., et al. “The eMERGE Network: A Consortium of Biorepositories Linked to Electronic Medical Records Data for Conducting Genomic Studies.” *BMC Medical Genomics* 4 (2011): 13.
24. eMerge. “Phenotyping: Cohort Discovery Using EHR Data.” Available at https://emerge.mc.vanderbilt.edu/phenotyping-cohort-discovery-using-ehr-data/ (accessed January 18, 2018).
25. PCORnet. “About PCORnet.” Available at http://pcornet.org/about-pcornet/ (accessed January 18, 2018).
26. PCORnet. “Data Network Request.” Available at http://pcornet.org/data-network-request/ (accessed January 18, 2018).
27. Michalik, D. E., B. W. Taylor, and J. A. Panepinto. “Identification and Validation of a Sickle Cell Disease Cohort within Electronic Health Records.” *Academic Pediatrics* 17, no. 3 (2017): 283–87.
28. PheKB. *Computable Phenotype for Identification of Patients with Sickle Cell Disease*. Available at https://phekb.org/sites/phenotype/files/Computable Phenotype Description.pdf (accessed January 18, 2018.
29. Michalik, D. E., B. W. Taylor, and J. A. Panepinto. “Identification and Validation of a Sickle Cell Disease Cohort within Electronic Health Records.”
30. PheKB. “Phenotype 615: Sickle Cell Disease.” Available at https://phekb.org/phenotype/sickle-cell-disease-0 (accessed January 18, 2018).
31. PheKB. “Peripheral Arterial Disease – 2012.” Available at https://phekb.org/phenotype/16 (accessed May 14, 2018).
32. Jones, D. T., et al. “Dissecting the Genomic Complexity Underlying Medulloblastoma.” *Nature* 488, no. 7409 (2012): 100–105.
33. Choi, M., et al. “Genetic Diagnosis by Whole Exome Capture and Massively Parallel DNA Sequencing.” *Proceedings of the National Academy of Sciences of the United States of America* 106, no. 45 (2009): 19096–101.
34. Curtis, C., et al. “The Genomic and Transcriptomic Architecture of 2,000 Breast Tumours Reveals Novel Subgroups.” *Nature* 486, no. 7403 (2012): 346–52.
35. Hinoue, T., et al. “Genome-Scale Analysis of Aberrant DNA Methylation in Colorectal Cancer.” *Genome Research* 22, no. 2 (2012): 271–82.

Figure 1

Using Computable Phenotypes to Generate Patient Lists


Three steps for applying a Computable Phenotype to classify patients
?
+
=
1.
2.
3.
Computable Phenotype
Database Query
Source Database
Patient Cohort