A USER'S GUIDE TO COMPUTABLE PHENOTYPES

By

C. Blake Cameron, M.D., M.B.I.

TABLE OF CONTENTS

Abstract

Computable phenotypes are re-usable computerized search queries that detect specific clinical events or diseases using electronic health record data. Compared to manual chart review, computable phenotypes extract large scale information from the medical record with greater speed and lower overall cost. Public phenotype repositories are being developed to encourage sharing and re-use of definitions. Hundreds of phenotypes definitions have now proliferated, often overlapping and differing in subtle ways. No consumer tools currently exist to help prospective users evaluate and select the most appropriate definition among multiple options. The purpose of this project is to provide a practical framework that will help physicians, clinical researchers and informaticians evaluate published phenotype algorithms for re-use in various purposes. The framework is divided into three phases, aligned with expected user roles: overall assessment, clinical validation, and technical review. Evaluation templates are provided in the appendix.

Chapter 1: Introduction

“Show me a list of everyone within our health system who has chronic kidney disease.”

“Find all the patients hospitalized with heart attacks in the last 30 days.”

Queries like these undergird healthcare quality improvement programs, pragmatic clinical trials, genomic studies, outcomes research, and virtually every activity where data, information and knowledge could be added to the health system to improve health care, as in the learning health system.[1, 2] In this era of widespread electronic health record (EHR) use, one would expect that even the most basic EHR system could execute these searches. Yet few EHRs currently possess the capability out-of-the-box to *accurately* identify patient cohorts or clinical events using large-scale queries.[3] Among those systems that are capable of providing valid results, they can do so for a handful of conditions or events, and only then because of hundreds of hours of programming and rigorous statistical evaluation invested in custom query development.[4]

Richesson and colleagues define computable phenotypes as “a definition of a condition, disease, characteristic or clinical event that is based solely on data that can be processed by a computer.”[5] Phenotypes form the basis of re-usable EHR search queries that can be used to identify patient populations and establish registries for additional analysis.[5, 6] The terminology of phenotyping traces its origins to genomic research, where automated, high-throughput methods were needed to identify patients with and without conditions of interest.[7] Usage of computable phenotypes extends beyond genomic research. Computable phenotypes are employed for multiple purposes related to health care operations, public health and biomedical research. Since long before the genomic era, organizations such as National Quality Forum,[8] National Committee for Quality Assurance[9] and Centers for Medicare & Medicaid Services[10] have overseen the creation and implementation of computable phenotypes that define clinical quality measures. Distributed research networks such as PCORnet[11] and the NIH Collaboratory[3] conduct observational and comparative effectiveness research that relies on common data models and computable phenotypes to identify patient cohorts and clinical events. The FDA employs computable phenotypes to conduct drug safety surveillance in its Mini-Sentinel program.[12] Cleveland Clinic applied a chronic kidney disease phenotype to construct a patient registry to target patients for enrollment in a patient navigator program and for other quality improvement activities.[13-15] Other efforts applying phenotypes include eMERGE,[7] the NIH Precision Medicine Initiative,[16] and the Million Veteran Program.[17] Compared to manual chart review, computable phenotype applications extract information from the medical record with far greater speed and lower overall cost.

Historically, each extant phenotype definition represented the independent efforts of a single entity. Phenotypes were developed and validated to meet an immediate need, such as for recruiting patients in a research study, measuring quality, or for performing large-scale genome wide association studies. Anecdotally, organizations rarely shared these definitions outside their own walls, hindering transparency, reproducibility and scalability of research.[5] And many disincentives and barriers to sharing currently exist.[5] As a result, hundreds of potentially overlapping phenotype definitions have now proliferated. In large or complex organizations, there might conceivably be multiple definitions per condition within the organization. Definitions that appear similar on their face may differ in subtle ways, such that they yield vastly different results when applied to the same data set. For example, a comparison of seven phenotype definitions for prevalent type 2 diabetes mellitus found that the cohorts they identified were largely non-overlapping.[18] These varying definitions are not necessarily incorrect, but rather may reflect fitness for use for differing purposes. The most appropriate definition for a quality improvement program that narrowly targets individuals with uncontrolled diabetes will differ by necessity from a definition aimed at broadly identifying diabetics for genetic analysis of disease subtypes. Electronic health record data quality issues are pervasive[19] and may limit the broad applicability of current phenotype definitions.[20] A group of researchers using computable phenotypes to investigate diabetes-related complications found that they could not estimate the prevalence of those complications due to inconsistencies in clinical documentation and varying definitions in EHRs.[21] These findings underscore the fact that phenotype implementation is complex and performance depends to a great extent on the quality of the underlying data.[7]

Public phenotype repositories are increasingly being promoted to encourage sharing, re-use, and iterative improvement of definitions. It is believed that standardization and dissemination of definitions will facilitate analytical transparency, promote use of common data models, increase quality and consistency, and minimize duplication of effort.[5] PheKB.org is the largest and best-known example of such a repository. (See the "How do I locate existing phenotype definitions?" section for further information.) While these repositories may improve the dissemination of phenotypes, they will require tools that help prospective users evaluate existing phenotype definitions and select the most appropriate definition among multiple options. To my knowledge, no such tools currently exist.

The purpose of this document is to provide a practical user's guide to computable phenotypes that will help physicians and informaticists evaluate published phenotype algorithms for re-use in various purposes. Phenotype development for clinical research usage is a new science. There is little high quality evidence that can define best practices. This guide is based on the consensus opinion and experience of experts in the field of phenotype development and implementation.

Chapter 2: Methods

This user's guide to computable phenotypes was developed using a consensus-building approach that obtained iterative feedback from a number of multi-disciplinary experts in the field of computable phenotype development and validation. These experts included clinical researchers, physicians, and practicing informaticists in biomedical informatics departments at several academic medical centers. Experts were recruited through established professional connections with the author (CBC) and mentors (RR, DD).

The author conducted unstructured interviews with several practicing informaticists to understand current approaches to selecting computable phenotypes for re-use and lessons learned.   Content areas for interviews included ideal phenotype definition characteristics, sources for obtaining existing phenotypes, considerations in building a *de novo* definition versus re-use, and evaluation of phenotype definitions and documentation.

Based on these interviews and review of relevant published literature, a draft user's guide was written and distributed electronically to phenotyping experts for review. Reviewer comments were incorporated in an iterative feedback process, leading to a final consensus document.

Chapter 3: Results

A total of 12 experts reviewed and contributed to this guide over the course of four months. These experts represent belong to various disciplines (practicing physicians, informaticists, statisticians, and data scientists) and participate actively in phenotype-based research networks (such as NIH Collaboratory, eMERGE, PCORnet), although they were not officially representing those networks. The user checklist and supporting information were iteratively revised and sent to reviewers until no additional feedback was received. A formal evaluation of the checklist and user guide is being planned.

Below I present the recommendations and guidance that we developed for the user guide, which is provides further background, resources and a framework for reviewing and comparing phenotype definitions.

What makes a good phenotype?

Phenotypes define the collections of concepts and logical elements that will be used to support various analyses. It is important that they can be understood, implemented, reported and shared. A "good" phenotype is one that is explicit, reproducible, reliable, and valid for its intended use.[6] *Explicit* means that the documentation is sufficiently detailed and unambiguous such that the phenotype can be implemented with high fidelity. *Reproducible* means that the phenotype will achieve the same results with repeated implementations (assuming the same underlying data). *Reliable* means that the phenotype will return the same results with repeated executions. *Valid* means that the phenotype search measures the intended clinical concept. The caveat of "fit for its intended use" is critically important. Many phenotypes were not designed for re-use, and the authors could not have anticipated all the possible implementation challenges. Therefore, a phenotype that is explicit, reproducible, reliable and valid for a particular use in a particular context is not necessarily generalizable to other uses or contexts.

"Garbage In = Garbage Out." A high quality phenotype, if provided inaccurate or faulty data, will almost always return faulty results. Phenotype algorithms may include logic routines that resolve minor internal inconsistencies in input data. However, these quality checks are limited in power and are not exhaustive. The output errors resulting from poor quality source data may be subtle, such that only subject matter experts identify inconsistencies, or even undetectable in the absence of rigorous evaluation.

Ensuring the validity of underlying clinical EHR data is a prerequisite to applying phenotypes. A detailed discussion of data quality assurance falls beyond the scope of this user's guide. Healthcare organizations collect clinical data through electronic health records primarily for the purposes of medical billing and patient care activities, rather than rigorous research. EHR data are prone to poor quality and biases. According to Weiskopf, domains of quality include completeness, correctness, concordance, plausibility and currency.[22] EHR data about a patient are often incomplete due to lack of information exchange among healthcare organizations and due to variable capture of data elements during routine operations. EHR data are often incorrect.[23] A recent study conducted within the Veterans Health Administration found at least one error in 84% of progress notes and an average of 7.8 documentation errors per patient.[23] The presence of diagnosis codes may suggest that a patient suffers from a particular illness, when in fact the code was selected to justify billing for a diagnostic test used to rule out that condition. Furthermore, the data contained in various healthcare information systems may conflict without a clear method to resolve inconsistencies. For these reasons, phenotype validation is necessary – even when the logic is deemed sound – in order to ensure that the algorithm performs satisfactorily. If the phenotype's output is valid, then the underlying (input) data quality can be assumed to be acceptable.

How do I locate existing phenotype definitions?

Definitions can be obtained from a variety of sources, and have varying degrees of specificity and validation. Phenotype repositories such as PheKB.org contain phenotype definitions, documentation and information about validated performance characteristics. Quality measures[8, 10] generally include two phenotypes each: a denominator defining an eligible population and a numerator defining the event or process of interest, and may include validation details.

Phenotypes can also be derived from medical professional society guidelines, which often include structured clinical definitions of disease that can be mapped to common EHR data elements.[13, 24] Phenotype definitions may be difficult to locate in the peer-reviewed literature. Journals seldom provide authors sufficient space for full documentation of phenotype definitions and implementation details. Due to the nascent terminology surrounding clinical phenotypes, phenotype definitions may be described in the medical literature using various terms that require multiple searches and subject matter expertise to locate.[25] The NIH Health Systems Research Collaboratory Phenotypes, Data Standards, and Data Quality Core maintains a list of sources for existing phenotypes and suggested search terms for locating phenotype definitions in the medical literature (a phenotype for phenotypes, if you will).[25] Use of standardized terminology resources such as the Unified Medical Language System (UMLS) and UMLS Terminology Services can help match phenotype definitions to clinical concepts. The process of locating phenotype definitions will likely be simplified as phenotype repositories (e.g., PheKB, PhenX) and authoring tools mature.[26]

Deciding whether to "build or buy"

Whether applied in healthcare operations or clinical research, computable phenotypes are almost always deployed in response to a particular information need. The decision to build a phenotype algorithm from scratch, re-use an existing algorithm, or modify an existing algorithm should start with a thoughtful requirements analysis, evaluation of existing resources, and consideration of relevant tradeoffs.

Re-using an existing algorithm can have several advantages. Re-use helps establish a *de facto* standard that enables scalable use within and across organizations. The network benefits of standardization cannot be emphasized enough; they almost always outweigh technical or performance shortcomings. For example, phenotype standardization for quality measurement facilitates comparison of clinical performance against peer organizations. In research, phenotype standardization allows large-scale enrollment and analysis of subjects in multicenter research programs or networks. Re-use tends to save resources because others have already absorbed the high upfront costs of development and validation.

Building a phenotype from scratch may be useful or mandatory in certain scenarios:

1. No suitable phenotype exists. Although phenotypes now exist for hundreds of common health conditions and clinical events, these phenotypes represent only a tiny sliver of a nearly infinite universe of possibilities. For undeveloped conditions or events, there is no choice but to develop an algorithm from scratch. (Note that, if a custom-built phenotype is then shared in a phenotype library, it may become the de facto standard for that condition or event in the future.)
2. Unique local circumstances. Phenotype algorithms make assumptions about data structure and the delivery of medical care that may not be universal. If local circumstances deviate substantially from these assumptions, re-use will not be feasible. For example, a phenotype that

relies upon deprecated coding vocabularies (such as ICD-9-CM) will not be usable in modern clinical environments or overseas where different coding systems are employed. Or, for example, a phenotype that performs natural language processing of radiology reports to identify cerebral revascularization procedures using keywords may prove inaccurate at centers where those procedures are performed by neurologists rather than radiologists, who may use different vocabulary to describe similar findings. (Use of "anchors," or atomic keywords identified by subject matter experts that are indicative of certain clinical characteristics, can improve the efficiency and accuracy of natural language processing in these circumstances.[27])

Creating a new definition is resource intensive and contributes to the problem of overlapping definitions. Extending or modifying an existing phenotype is a compromise that reduces development effort compared to developing a new phenotype, but still allows adaptation to local needs. For example, the clinical logic of a phenotype may be sound, but the terminology set may be outdated (such as the case with ICD-9-CM). In this case, remapping codes to ICD-10-CM using Medicare general equivalence mappings would require relatively little development effort.[28] However, the change could have unintended and unforeseeable implications that affect diagnostic accuracy.[29] To the extent that phenotype repositories support sharing of user-contributed extensions in the future, "crowdsourcing" of these modifications may further reduce development resource requirements.[5]

Both approaches – extension and build from scratch – require verification and validation to ensure that the algorithm delivers satisfactory performance. Verification involves evaluation and testing of the algorithm to ensure that it was built as intended, according to specification. Validation requires analysis of the output to ensure that the underlying clinical concept is correctly measured. These activities require coordinated effort from clinical subject matter experts (usually physicians), informaticists (including programmers and data analysts), and statisticians. Whenever possible, it is preferable to re-use an existing phenotype to limit the extent of validation required.

How do I evaluate phenotype definitions for re-use?

There are myriad uses for phenotypes. The features determining quality or suitability for re-use differ across applications, and the profile of strengths and weaknesses depends on the particular use case. This document provides a generalizable framework for reviewing and comparing existing phenotype definitions for local implementation. The framework guides potential users through a detailed assessment of the strengths and weaknesses of a given definition in context of their intended purpose.

The framework is divided into three phases: an overall evaluation of fit and purpose, a review of clinical validity, and an analysis of the technical feasibility of implementing a given phenotype definition. The evaluation process cannot be distilled to a simple rule-based algorithm. Rather, each section presents a detailed, but non-exhaustive, list of considerations to help the reviewer evaluate and compare phenotype definitions relevant to a particular purpose.

What follows is a step-by-step guide to applying the assessment framework to a particular definition. A suggested template for use during review and for documenting responses is presented in Appendix 1. After considering relevant strengths and weaknesses, each section can be assigned a grade indicating a summative assessment of overall suitability, as shown in table 1.

Table 1. Phenotype Evaluation Rubric

Grade	Notes
A	No major weaknesses. Few minor weaknesses
B	One major weakness or several minor weaknesses
C	More than one major weakness
D	Multiple major and minor weaknesses, phenotype not valid or implementation clearly infeasible

Anticipated reviewer roles

Proper evaluation of phenotype definitions requires competencies in clinical medicine, data architecture and standards, and statistical reasoning. It is rare that a single individual possesses all of these competencies and sufficient experience to evaluate a phenotype definition properly. In most cases, the evaluation process requires a small team comprising (at a minimum) a physician or clinical subject matter expert and an informaticist. The content areas for each anticipated role are denoted in the table below.

Table 2. Anticipated Reviewer Roles

	Administrator	Physician	Clinical Researcher	Informaticist or Data Analyst
Phase 1 – Overall Evaluation	X	X	X	X
Phase 2- Clinical Validity Assessment		X	X	
Step 3 – Technical Feasibility Assessment			X	X

Review Phase 1: Overall Evaluation – Who, What, Where, When, Why?

The first step in evaluating a candidate phenotype is to address the 5 W's: "who?", "what?", "where?", "when?" and "why?". The answers to these simple questions provide important clues to the suitability of a particular phenotype for re-use.

What is the name of the phenotype? The title establishes initial relevance; the greater the specificity, the better. If you wish to identify adult type 2 diabetics, a phenotype titled "Diabetes Mellitus Type 2 in Adults" will likely be a better fit than one titled "Diabetes" (which could potentially include undesired conditions such as diabetes insipidus, gestational diabetes, or type 1 diabetes). It is also important to understand the type of event or condition being identified. A taxonomy of phenotypes is shown in Table 3.

Table 3. Phenotype Classification.
[Table adapted from Shelley Rusincovitch, Duke Clinical Research Institute, 2015. Used with permission.]

Phenotype Classification	Description
Prevalent Disease	Does the patient have a given condition within the observation period? This type of phenotype identifies the

	presence of a condition within an individual patient, but does not identify the date of onset or resolution. Most disease-based phenotypes belong to this category.
Incident Disease	When did the patient acquire the disease? This phenotype attempts to identify the onset of a particular condition. Classification is dependent upon being able to pinpoint precisely the onset of disease and/or duration, which is challenging due to *incompleteness* in EHR data. Phenotypes in this category may be few and far between, because the "onset date" of a condition is highly dependent on healthcare utilization. For example, type 2 diabetes physiology may exist silently in a patient for a decade before clinical detection.[30]
Health Care Event or Utilization	Did a particular event (e.g., hospitalization, cardiac catheterization) occur? These can often be identified via administrative records (CPT procedure codes) rather than diagnostic criteria. As such, these phenotypes will tend to be highly specific (e.g., hemodialysis, coronary artery bypass grafting), but may not be directly attributable to a clinical condition.
Atomic / "Anchor" Traits	These simple phenotypes may draw from a variety of clinical criteria to describe a discrete or continuous patient trait or clinical event, and can be used to assemble higher-order phenotype definitions.[27] For example: Is the patient male or female? What is the patient's race? What is the patient's height, weight or BMI? What is the patient's average red cell distribution width (RDW) over the last 6 months?
Risk	What is the probability that the patient will develop a given condition or experience a given clinical event? For example: the Kidney Failure Risk Equation uses clinical laboratory variables to estimate the likelihood that an individual with chronic kidney disease will progress to permanent kidney failure.[31, 32]

Who was/were the author(s) of the phenotype? Phenotypes authored by a group or consortium may be more likely to consider a broader spectrum of relevant factors, approaches and limitations. What are the authors' affiliations? Phenotypes authored by academicians for research purposes may possibly undergo more rigorous validation, but validation could also reflect a clinical context unique to academic medical centers.

Who is using the phenotype? Stakeholder endorsements, peer-review, and widespread use each signal strong confidence in the methodology and suggest portability across settings. Widespread usage also establishes a de facto standard that enables comparison and benchmarking across sites.

Where was the phenotype developed? The burden and causes of disease vary dramatically within the United States and internationally. For example, the causes of anemia are fundamentally different in various parts of the world. Among phenotype definitions that rely on medication usage, variations in local and regional prescribing practices may impact the performance of the phenotype. Similarly, if a phenotype is deployed across international lines, the set of medications approved by national regulatory bodies may differ. Terminologies may differ across countries; for example, the United States uses a heavily modified version of the World Health Organization ICD-10 coding system that is not directly comparable to the ICD-10 coding set used by the rest of the world. Laboratory assays and measurements units may differ across sites. The location of phenotype development provides important clues to ultimate validity and technical feasibility of re-implementation.

When was the phenotype developed and last updated? The definition likely reflects the standard of care at the time of development. Medical practice evolves quickly, resulting in significant changes in the defining features of diseases and their treatments. Without updating, phenotypes eventually become outdated. For example, a phenotype definition for hypertension may include logic that establishes blood pressure thresholds. The thresholds defining hypertension change frequently as new guidelines emerge from the Joint National Committee.[33] Similarly, a phenotype definition for diabetes mellitus may include logic that considers the usage of blood glucose lowering medications to determine the presence of the disease. Failure to include newer therapies (such as sodium/glucose contransporter 2 inhibitors) reduces the sensitivity of the algorithm. Phenotypes reflect coding systems in place at the time of development (e.g., ICD-9-CM to ICD-10-CM transition in the United States in 2015). Therefore, the phenotype definition should clearly state the date it was developed, validated, last updated, and ideally a version number.

Why was the phenotype developed? Or, what was the original application? Potential reasons include quality measurement and reporting, epidemiologic research, and clinical trial enrollment. Each intended use has a different set of tradeoffs. For example, trial enrollment may maximize sensitivity for broad catchment of potential subjects, whereas QI aims for representative (specific) cases of a given condition, excluding outliers or marginal cases. When repurposing a phenotype for a different type of application, extra care should be taken to ensure that the phenotype remains valid for the new use. These trade-offs are discussed in more detail in the next section.

Descriptive Information & Overall Evaluation	
Name	□What is the name of the phenotype? □What condition or event does it identify?

Author(s)	□Who developed the phenotype? □What are their affiliations?
Authorship Date and Version	□When was the phenotype originally developed? □When was it last updated or revised (if applicable)? What version is it?
Type of Event or Condition	□Prevalent chronic disease □Incident chronic disease □Acute/transient disease or event prevalence □Acute/transient disease or event incidence □Procedural event □Patient trait
Original Application	□Epidemiologic research? □Clinical trial enrollment? □Genomic research? □Quality and practice improvement? □Regulatory or quality reporting? □Other?
Tradeoffs	□What trade-offs may have been made for use in the original application? □Are those trade-offs optimal for my intended use?
Dissemination and Acceptance	□What organizations have endorsed the phenotype? (e.g., CMS for quality measures) □Who is using it? □What peer-reviewed publications depend on it?

Review Phase 2: Clinical Diagnostic Evaluation

If properly implemented, is this algorithm valid in my patient population for my intended purpose? A computable phenotype algorithm is analogous to a laboratory test: some operation is performed on a specimen and a result returns. In the case of a laboratory test, the specimen is human tissue, whereas for a computable phenotype, the "test" is performed on a patient's medical record.   Like laboratory tests, immense effort goes into developing, validating and operationalizing computable phenotypes, and the results may only be useful when certain conditions are met. Evaluating the validity of a phenotype definition requires expertise in the clinical subject and in statistical reasoning.

The first step is to compare the population on which the algorithm was derived to your own patient population. How are they similar and how are they different? Demographic factors such as age, race, ethnicity, gender, health insurance and socioeconomic status influence the prevalence of conditions or certain events. Factors specific to the setting also influence the prevalence and severity of conditions and treatment approaches. For example, a phenotype designed to detect hospital admissions for heart failure that was developed and validated using a patient population at an urban academic medical center may have limited applicability to a rural community hospital where the structure and intensity of care differ. The academic center could have an intensive outpatient care unit that treats patients that would have otherwise required hospitalization at other centers. Phenotype definitions often rely on severity thresholds (e.g., hemoglobin A1c >8%) or care intensity (e.g., at least three occurrences of a test or diagnosis code in two years) that may be less sensitive in primary and secondary care settings. A strong candidate definition will have been developed and validated on a similar population in a similar setting for a similar purpose. Caution should be used when the population, setting, and purpose are dissimilar. The degree of concordance required for validity may vary depending on the intended use case.

The second step is to evaluate the criteria by which patients were included or excluded from the phenotype definition and validation process. Do those criteria appear reasonable clinically? Are they consistent with the intended reuse? For example, a phenotype definition for diabetes mellitus that explicitly excludes type 1 diabetics would likely not be appropriate for use with a quality improvement program aimed at reducing diabetic foot infections, which complicate both types of the disease. These inclusion and exclusion criteria may be applied at various points in the logic of the phenotype algorithm. If not specified clearly in the documentation, it may be necessary to look "under the hood" at the logic to understand which patients or events are included or excluded. Note that the purpose of this step is to determine the eligible patient population, not to evaluate the clinical soundness of the underlying decision logic. Clinical face validity is helpful, but not always necessary. High-performing phenotypes may group or omit clinical elements in ways that appear counterintuitive to clinical experts. For example, a recently validated risk phenotype for uncontrolled hypertension found that consideration of historical blood pressure measurements did not improve performance.[34] Therefore, unless a phenotype directly implements a reference definition of a disease, its validity should be determined by empirical comparison against a diagnostic gold standard. A strong phenotype definition must detect the desired condition or clinical events without being overly broad or narrow.

The third step is to examine the phenotype's validity in its original application. All phenotypes must be validated prior to use in a production environment. Validation ensures that the phenotype detects the intended clinical concept by adjudicating the algorithm's output against a reference (gold) standard or through a controlled process, such as expert review or by comparing its ability to predict health outcomes. Validation can be divided conceptually into low-level and high-level phases.

Low-level validation ensures concordance at the interface between the underlying data structure and the atomic data elements incorporated in the phenotype. For example, a phenotype that requires laboratory measurements is designed with certain assumptions about the laboratory test as well as data types, value sets and units. Low-level validation confirms that atomic data elements match the intended clinical concept and are provided in the appropriate data structure. Phenotype documentation should provide data dictionaries that provide detailed specifications of the required data elements. For example, a laboratory test for “creatinine” may come from blood or from urine specimens. The former is used to estimate kidney function, whereas the latter is not. Clinical information systems may report laboratory results numerically as continuous or as ranges, or discretely as categories. Phenotype documentation should provide data dictionaries that provide detailed specifications of the required data elements, to which the underlying data model must be mapped. Ideally, documentation would provide anticipatory guidance for commonly encountered remapping tasks. Low-level validation generally requires manual inspection of the underlying data and verification with the electronic health record. Low-level validation is implementation-specific and must be performed whenever re-using a phenotype definition.

High-level validation ensures clinical concordance between the phenotype and the condition or event being measured, and is generalizable across information systems. The high-level validation process should be publicly reported. Key elements of evaluation include the choice of reference standard, the breadth of validation, blinding, and the presence or absence peer-review. The reference standard should be appropriate for the condition being evaluated. Manual adjudication by one or more clinical experts is frequently used as the gold standard and usually appropriate if blinded. However, other phenotypes or computable definitions could conceivably be used as the reference standard – for example, when testing a simplified definition against a previously validated phenotype algorithm. Phenotype definitions that have been validated at multiple sites should be considered more generalizable than those that have been validated at only a single site. Peer reviewed publication of the definition provides some assurance that the validation methodology is sound.

High-level validation should yield formal reporting of performance characteristics. The most commonly reported characteristics are positive and negative predictive values, sensitivity and specificity. Among patients identified by a phenotype algorithm as having a condition, the positive predictive value is the proportion that actually have the condition. Negative predictive value is the inverse. Sensitivity (also known as recall) indicates the proportion of patients with a given condition that are properly detected by the algorithm. Specificity (also known as precision) indicates the proportion of patients lacking a given condition that are properly rejected by the algorithm. Positive and negative predictive values are the most useful for determining the real-world accuracy of the algorithm because the measures take into account the prevalence of the condition. (All other things being equal, the positive predictive value worsens as the prevalence of the condition decreases.) Because disease prevalence can vary widely across sites, the PPV and NPV reported in validation studies may not be generalizable to other settings. Ideally, validation studies should report sensitivity and specificity, from which site-specific PPV and NPV can be derived using local prevalence figures. Determining sensitivity and specificity requires adjudicating the reference standard among (at least a sample of) all patients. A very large sample may be required to calculate specificity with reasonable confidence for conditions with low prevalence. For that reason, resource constraints may prevent precise estimation of sensitivity and specificity during the validation process. A strong phenotype would have undergone a rigorous, multi-site validation including peer-review and have reported all relevant performance characteristics including sensitivity and specificity.

Sensitivity, specificity, PPV and NPV often exist in tension. Tradeoffs that improve sensitivity and NPV usually worsen specificity and PPV. Phenotypes with high PPV or high specificity are most useful for definitively ruling in a condition. All other things being equal, patients with a positive result are likely to have the condition. For example, a high-specificity performance profile could be useful for identifying a focused subset of at-risk patients for care management services. Conversely, phenotypes with high NPV or high sensitivity are most useful for ruling out a condition. Patients with a negative result are unlikely to have the condition. This profile, for example, might be useful narrowing down a list of candidate subjects for a study of a rare disease. A strong phenotype performs well with high

(>90%) sensitivity and specificity or with a profile that is appropriate for the intended use.

Validation (Is the algorithm valid in my population for my intended purpose?)	
Derivation	□On what population was the algorithm derived? Are the population and setting similar to my patient population and setting?
Gold standard	□Was there a gold standard against which the algorithm was validated? □Is the gold standard an appropriate choice for the condition or event?
Validation	□Was validation performed on a separate cohort (same-site) in a blinded fashion? □Was validation conducted at another site? □Was validation performed at multiple sites? □Has validation undergone peer-review?
Performance characteristics	□What is the sensitivity? Specificity? Positive predictive value? Negative predictive value? □Is that performance profile satisfactory for my intended purpose?

Review Phase 3: Technical Evaluation

The final portion of the review is an assessment of the technical feasibility for implementation. This review requires an in-depth assessment of phenotype documentation and implementation requirements. For reasons of efficiency, this review should be conducted once the phenotypes have been narrowed to a small number of clinically-appropriate candidates.

The first step involves evaluation of the documentation quality. Written documentation should include descriptive information (discussed in the overall review section) and provide a clear, unambiguous description of the algorithm and supporting details. The algorithm must be described in a verbal, graphical or pseudocode representation that is sufficient for a programmer or informaticist to reproduce with fidelity. Ideally, the logic should be encoded in a structured and computable format, such as in Clinical Quality Language,[35] Quality Data Model, or similar,[36] with source code available for review. The specifications must include a detailed data dictionary that specifies the data element name, written description, data type, value sets, and dependencies on other standards. The documentation should indicate best practices and any caveats experienced during implementation and low-level validation at other sites. Phenotype authoring tools and common data models that enable machine-readable phenotype definitions are under development. If provided in machine-readable format, the software requirements and all other dependencies should be clearly specified.

The second step addresses feasibility. Phenotype definitions may rely on some combination of demographic data, diagnosis and procedure codes, medication or pharmacy data, orders, structured clinical observations (such as vital signs), lab results, unstructured text, genetic data, and patient reported (survey) data. Required information types or processing techniques may not be available at all sites. For example, most electronic health record repositories do not currently contain genetic or patient reported survey data. Natural language processing capabilities are relatively uncommon despite widespread adoption outside of healthcare. It is not feasible to implement a phenotype that requires information that does not exist or cannot be accessed. The phenotype definition should specify which data elements are required versus optional, and the acceptable degree of "missing-ness" for each.

The third step looks at concordance between the organization's data model and the phenotype's input requirements to understand the scope of resources required for implementation. Some degree of data transformation will certainly be required to map an electronic health record data model to the phenotype definition's requirements, even if the high level logic is entirely sound. Mapping element names and data types is relatively straightforward and usually involves no significant loss of information. However, challenges frequently arise when mapping source data to required value sets, particularly when the source data does not adhere to a common standard and has less granularity than the phenotype demands. For example, a source system may report race as "White," "Black" and "Other," whereas a phenotype definition may use a more exhaustive list. Similar difficulties may arise when a source system reports a result as a range (e.g., "30-300" or ">14%") and the phenotype expects a numeric result. Various imputation approaches may be required to translate between value sets. Ideally, phenotypes should be mapped to standardized clinical terminologies (such as SNOMED CT®, ICD-10-CM, and RxNorm) and common data models (CDMs). Examples of CDMs include those used by Observational Health Data Sciences and Informatics (OHDSI) Observational Medical Outcomes Project (OMOP)[37] and PCORNet,[38] which together possess datasets encompassing more than 600 million patients across 11 countries. The OMOP CDM includes a collection of software tools that facilitate data element mapping, cohort selection, and data quality assessment. Dependencies on outside standards, data models, and other phenotypes should be clearly specified in the documentation, including the version number of the required standard. Implementation of phenotypes that require deprecated standards (such as ICD-9-CM) or proprietary standards may demand so many resources as to not be feasible.

Technical Review of Documentation and Implementation Feasibility	
Human-readable	□Is there a description of the meta-data including name, authorship, date/versioning, and intended purpose? □Is there a verbal, graphical or pseudocode representation of the algorithm and data dictionary sufficient to reproduce? □Are all appropriate dependencies and value sets described or referenced? □Are best practices or caveats indicated?
Machine-readable	□Is the phenotype algorithm provided in a machine-interpretable format? □Are data dictionaries and value sets provided in a machine-interpretable format?
Data Elements and Modalities Required	□Demographics? □Diagnosis codes? □Procedure codes? □Pharmacy/Medications? □Orders? □Structured clinical observations (e.g. vital signs)? □Lab results? □Unstructured text / natural language processing? □Patient reported data (survey responses) □Genetic data (biobank repositories)
Value sets	□Internal value sets (unique to the phenotype) □External value sets (mapped to a standard)
Relationship to other phenotypes	□Other phenotypes embedded or required? □Are those phenotypes available?

Relationship to other standards	□Terminologies (e.g., ICD-10, SNOMED dependencies) □Other

Chapter 4: Conclusion

Computable phenotype definitions enable identification of patient cohorts or clinical events using electronic health records. Numerous phenotype definitions have proliferated, often for the same condition. Platforms for sharing definitions are under development and will soon become available. To my knowledge, currently no tools exist that help users evaluate the suitability of a particular definition or compare definitions for re-use. This “users’ guide to computable phenotypes” provides a starting framework and evaluation tools for physicians, clinical researchers and informaticists to evaluate the clinical validity and technical feasibility of re-using an existing phenotype definition for a particular purpose, and lays the groundwork for future empirical research in this area.

References

1. Greene SM, Reid RJ, Larson EB. Implementing the learning health system: From concept to action. Ann Intern Med. 2012;157(3):207-10.

2. Weber GM, Murphy SN, McMurry AJ, Macfadden D, Nigrin DJ, Churchill S, et al. The shared health research information network (shrine): A prototype federated query tool for clinical data repositories. J Am Med Inform Assoc. 2009;16(5):624-30.

3. Richesson RL, Hammond WE, Nahm M, Wixted D, Simon GE, Robinson JG, et al. Electronic health records based phenotyping in next-generation clinical trials: A perspective from the nih health care systems collaboratory. J Am Med Inform Assoc. 2013;20(e2):e226-31.

4. Horvath MM, Winfield S, Evans S, Slopek S, Shang H, Ferranti J. The deduce guided query tool: Providing simplified access to clinical data for research and quality improvement. J Biomed Inform. 2011;44(2):266-76.

5. Richesson RL, Smerek MM, Cameron CB. A framework to support the sharing and re-use of computable phenotype definitions across health care delivery and clinical research applications. In press. 2015.

6. Richesson R, Smerek M. Electronic health records-based phenotyping Durham, NC: Duke University Medical Center; 2014 [Nov 19, 2014]. Available from: http://sites.duke.edu/rethinkingclinicaltrials/ehr-phenotyping/.

7. Newton KM, Peissig PL, Kho AN, Bielinski SJ, Berg RL, Choudhary V, et al. Validation of electronic medical record-based phenotyping algorithms: Results and lessons learned from the emerge network. J Am Med Inform Assoc. 2013;20(e1):e147-54.

8. Maintenance of nqf-endorsed® performance measures National Quality Forum; [May 10, 2016]. Available from: http://www.qualityforum.org/Measuring_Performance/Endorsed_Performance_Measures_Maintenance.aspx.

9. Hedis measures: National Committee for Quality Assurance; [May 10, 2016]. Available from: http://www.ncqa.org/hedis-quality-measurement/hedis-measures.

10. E-clinical quality measures library: Centers for Medicare and Medicaid Services; [May 10, 2016]. Available from: https://www.cms.gov/regulations-and-guidance/legislation/ehrincentiveprograms/ecqm_library.html.

11. Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS. Launching pcornet, a national patient-centered clinical research network. J Am Med Inform Assoc. 2014;21(4):578-82.

12. Curtis LH, Weiner MG, Boudreau DM, Cooper WO, Daniel GW, Nair VP, et al. Design considerations, architecture, and use of the mini-sentinel distributed data system. Pharmacoepidemiol Drug Saf. 2012;21 Suppl 1:23-31.

13. Navaneethan SD, Jolly SE, Schold JD, Arrigain S, Saupe W, Sharp J, et al. Development and validation of an electronic health record-based chronic kidney disease registry. Clin J Am Soc Nephrol. 2011;6(1):40-9.

14. Jolly SE, Navaneethan SD, Schold JD, Arrigain S, Konig V, Burrucker YK, et al. Development of a chronic kidney disease patient navigator program. BMC Nephrol. 2015;16:69.

15. Jolly SE, Navaneethan SD, Schold JD, Arrigain S, Sharp JW, Jain AK, et al. Chronic kidney disease in an electronic health record problem list: Quality of care, esrd, and mortality. Am J Nephrol. 2014;39(4):288-96.
16. Tenenbaum JD, Avillach P, Benham-Hutchins M, Breitenstein MK, Crowgey EL, Hoffman MA, et al. An informatics research agenda to support precision medicine: Seven key areas. J Am Med Inform Assoc. 2016.
17. Gaziano JM, Concato J, Brophy M, Fiore L, Pyarajan S, Breeling J, et al. Million veteran program: A mega-biobank to study genetic influences on health and disease. J Clin Epidemiol. 2016;70:214-23.
18. Richesson RL, Rusincovitch SA, Wixted D, Batch BC, Feinglos MN, Miranda ML, et al. A comparison of phenotype definitions for diabetes mellitus. J Am Med Inform Assoc. 2013;20(e2):e319-26.
19. Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PR, Bernstam EV, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. Med Care. 2013;51(8 Suppl 3):S30-7.
20. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc. 2013;20(1):117-21.
21. Gregg EW, Li Y, Wang J, Burrows NR, Ali MK, Rolka D, et al. Changes in diabetes-related complications in the united states, 1990-2010. N Engl J Med. 2014;370(16):1514-23.
22. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research. J Am Med Inform Assoc. 2013;20(1):144-51.
23. Bowman S. Impact of electronic health record systems on information integrity: Quality and safety implications. Perspect Health Inf Manag. 2013;10:1c.
24. Kdigo 2012 clinical practice guideline for the evaluation and management of chronic kidney disease.
25. Richesson RL, Smerek MM, Rusincovitch SA, Heath A. Ehr-based phenotyping tools: National Institutes of Health Systems Research Collaboratory; 2014 [Mar 3, 2016]. Available from: https://sites.duke.edu/rethinkingclinicaltrials/tools-for-research/tools-ehr-phenotyping/.
26. Xu J, Rasmussen LV, Shaw PL, Jiang G, Kiefer RC, Mo H, et al. Review and evaluation of electronic health records-driven phenotype algorithm authoring tools for clinical and translational research. J Am Med Inform Assoc. 2015;22(6):1251-60.
27. Halpern Y, Horng S, Choi Y, Sontag D. Electronic medical record phenotyping using the anchor and learn framework. J Am Med Inform Assoc. 2016.
28. Cimino JJ, Remennick L. Adapting a clinical data repository to icd-10-cm through the use of a terminology repository. AMIA Annu Symp Proc. 2014;2014:405-13.
29. Turer RW, Zuckowsky TD, Causey HJ, Rosenbloom ST. Icd-10-cm crosswalks in the primary care setting: Assessing reliability of the gems and reimbursement mappings. J Am Med Inform Assoc. 2015;22(2):417-25.
30. American Diabetes A. Diagnosis and classification of diabetes mellitus. Diabetes Care. 2014;37 Suppl 1:S81-90.

31. Tangri N, Grams ME, Levey AS, Coresh J, Appel LJ, Astor BC, et al. Multinational assessment of accuracy of equations for predicting risk of kidney failure: A meta-analysis. JAMA. 2016;315(2):164-74.
32. Tangri N, Stevens LA, Griffith J, Tighiouart H, Djurdjev O, Naimark D, et al. A predictive model for progression of chronic kidney disease to kidney failure. Jama. 2011;305(15):1553-9.
33. James PA, Oparil S, Carter BL, Cushman WC, Dennison-Himmelfarb C, Handler J, et al. 2014 evidence-based guideline for the management of high blood pressure in adults: Report from the panel members appointed to the eighth joint national committee (jnc 8). JAMA. 2014;311(5):507-20.
34. Sun J, McNaughton CD, Zhang P, Perer A, Gkoulalas-Divanis A, Denny JC, et al. Predicting changes in hypertension control using electronic health records from a chronic disease management program. J Am Med Inform Assoc. 2014;21(2):337-44.
35. Hl7 standard: Clinical quality language specification, release 1: Health Level 7; [May 30, 2016]. Available from: http://www.hl7.org/implement/standards/product_brief.cfm?product_id=400.
36. Mo H, Thompson WK, Rasmussen LV, Pacheco JA, Jiang G, Kiefer R, et al. Desiderata for computable representations of electronic health records-driven phenotype algorithms. J Am Med Inform Assoc. 2015;22(6):1220-30.
37. OHDSI. Omop common data model [June 1, 2016].
38. PCORnet. Pcornet common data model (cdm) [June 1, 2016]. Available from: http://www.pcornet.org/pcornet-common-data-model/.

Acknowledgements

I appreciate the thoughtful feedback and assistance provided by Rachel Richesson (Duke), David Dorr (OHSU), and advisor Ebony Boulware (Duke).

I would also like to thank J.D., A.H., J.K., B.N., J.P., L.R., S.R., M.S., M.S., and J.S. for reviewing early drafts and contributing their expertise.

This work was made possible by financial support from the Duke Division of Nephrology, Duke Institute for Health Innovation, and the National Institutes of Health (5T32DK007731).

Appendix

Phenotype Evaluation Template

Reviewer Name: ____

Reviewer Role: □Clinician □Researcher □Informaticist □Other _________

Review Date: _______________

Phenotype: ____

____

Review Phase 1: Overall Evaluation

□Check if not reviewed

Strengths:

-
-
-

Major Weaknesses

□Check if none

-
-
-

Minor Weaknesses

□Check if none

-
-
-

Overall Assessment A B C D

(circle one)

Review Phase 2: Clinical Validity

□Check if not reviewed

Strengths:	•			
	•			
	•			
Major Weaknesses	•			
	•			
□Check if none	•			
Minor Weaknesses	•			
	•			
□Check if none	•			
Overall Assessment	A	B	C	D
(circle one)				

Review Phase 3: Technical Assessment

□Check if not reviewed

Strengths:	•			
	•			
	•			
Major Weaknesses	•			
□Check if none	•			
	•			
Minor Weaknesses	•			
□Check if none	•			
	•			
Overall Assessment	A	B	C	D
(circle one)				

Phenotype Comparison Template

Phenotype Def.	1:	2:	3:
Overall Evaluation			
Clinical Validity			
Technical Assessment			
Comments			