Characterizing Variability of EHR-Driven Phenotype Definitions

Pascal S. Brandt[1*], Abel Kho[2], Yuan Luo[2], Jennifer A. Pacheco[2], Theresa L. Walunas[2], Hakon Hakonarson[3], George Hripcsak[4], Cong Liu[4], Ning Shang[4], Chunhua Weng[4], Nephi Walton[5], David S. Carrell[6], Paul K. Crane[7], Eric Larson[8], Christopher G. Chute[9], Iftikhar Kullo[10], Robert Carroll[11], Josh Denny[12], Andrea Ramirez[11], Wei-Qi Wei[13], Jyoti Pathak[14], Laura K. Wiley[15], Rachel Richesson[16], Justin B. Starren[2], Luke V. Rasmussen[2]

[1]Department of Biomedical and Medical Education, University of Washington, Seattle, WA, [2]Northwestern University Feinberg School of Medicine, Chicago, IL, [3]Center for Applied Genomics, Children's Hospital of Philadelphia, Philadelphia, PA, [4]Department of Biomedical Informatics, Columbia University, New York, NY, [5]Intermountain Precision Genomics, Intermountain Healthcare, St George, UT, [6]Kaiser Permanente Washington Health Research Institute, Seattle, WA, [7]Department of Medicine, University of Washington, Seattle, WA, [8]Institute for Learning and Brain Sciences, University of Washington, Seattle, WA, [9]Schools of Medicine, Public Health, and Nursing, Johns Hopkins University, Baltimore, MD, [10]Department of Medicine, Mayo Clinic, Rochester, MN, [11]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, [12]All of Us Research Program, National Institutes of Health, Bethesda, MD, [13]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, [14]Weill Cornell Medicine, New York, NY, [15]Department of Biomedical Informatics, University of Colorado Anschutz Medical Campus, Aurora, CO, [16]Department of Learning Health Sciences, University of Michigan Medical School, Ann Arbor, MI

*Corresponding author:

Pascal S. Brandt
Department of Biomedical Informatics and Medical Education
University of Washington
Box 358047
Seattle, WA 98195

Keywords: FHIR, CQL, EHR-driven phenotyping, cohort identification

ABSTRACT

Objective

Analyze a publicly available sample of rule-based phenotype definitions to characterize and evaluate the types of logical constructs used.

Materials & Methods

A sample of 33 phenotype definitions used in research and published to the Phenotype KnowledgeBase (PheKB), that are represented using Fast Healthcare Interoperability Resources (FHIR) and Clinical Quality Language (CQL) was analyzed using automated analysis of the computable representation of the CQL libraries.

Results

Most of the phenotype definitions include narrative descriptions and flowcharts, while few provide pseudocode or executable artifacts. Most use 4 or fewer medical terminologies. The number of codes used ranges from 5 to 6865, and value sets from 1 to 19. We found the most common expressions used were literal, data, and logical expressions. Aggregate and arithmetic expressions are the least common. Expression depth ranges from 4 to 27.

Discussion

Despite the range of conditions, we found that all of the phenotype definitions consisted of logical criteria, representing both clinical and operational logic, and tabular data, consisting of codes from standard terminologies and keywords for natural language processing. The total number and variety of expressions is low, which may be to simplify implementation, or authors may limit complexity due to data availability constraints.

Conclusion

The phenotypes analyzed show significant variation in specific logical, arithmetic and other operators, but are all composed of the same high-level components, namely tabular data and logical expressions. A standard representation for phenotype definitions should support these formats and be modular to support localization and shared logic.

BACKGROUND AND SIGNIFICANCE

The generation of biomedical knowledge from electronic health record (EHR) data requires establishing cohorts of patients meeting certain criteria.[Banda2018] This cohort identification process is referred to as *EHR-driven phenotyping* and the sets of inclusion and exclusion criteria are known as *phenotype definitions*, or just *phenotypes* (the term used here for brevity). Although here we focus on the use within a research context, phenotypes support a wide variety of purposes beyond research, including quality improvement, clinical decision support, population health management, and public health. While often developed and executed at a single institution, research networks such as the National Patient-Centered Clinical Research Network (PCORnet),[Fleurence2014] the electronic Medical Records and Genomics (eMERGE) Network, [McCarty2011, Gottesman2013, Zouk2019] and the Observational Health Data Sciences and Informatics (OHDSI) program [Hripcsak2015] have run distributed studies to pool their results to improve statistical power and cohort diversity, leveraging shared phenotype definitions to meet these goals.[Ahmad2020, Hripcsak2019, Burn2020]

Beyond coordinated research networks, the re-use of phenotype definitions can save time across organizations and create many efficiencies.[Richesson2016] But this requires that potential users be able to search, retrieve, and assess existing definitions. At present, there is no widely used platform for sharing existing phenotype definitions, although multiple phenotype libraries or inventories have been created within research consortia. Historically, phenotypes have been shared between sites via repositories in the form of narrative descriptions,[Kirby2016, Denaxas2019] sometimes accompanied by flowcharts or pseudocode. Lists of codes from common terminologies like the International Classification of Diseases version 9 (ICD-9) are usually also included, but in most cases directly computable artifacts, such as SQL scripts or programming code, are not. Despite the potential gains of phenotype re-use, this current method of phenotype distribution has proven to be a major limiting factor in scaling up biomedical knowledge generation,[Pathak2013a, Newton2013, Shivade2014a, Adekkanattu2020] since implementing sites must individually interpret narrative descriptions to produce queries that can extract patient cohorts from local data sources. This process is time-consuming, may be operator dependent and error prone, compounded by the fact that narrative descriptions can be ambiguous or difficult to interpret.[Yu2021]

However, at present there is no universally accepted standard for representing the attributes and specification for computable phenotype definitions, although several have been proposed and evaluated [Peterson2014a, Pathak2013, Mo2015b, Mo2016, Jiang2017, Chapman2020a, Hong2019, Brandt2020]. The lack of a common representation for computable phenotypes has hindered the analysis and comparison of different phenotype algorithms. For example, it is not possible to do an equal technical comparison of a phenotype implemented in SQL against a local data warehouse and one done in Python against the OMOP CDM. Previous analyses have been performed in a few domains including an early review of 11 narrative phenotype descriptions;[Conway2011] clinical quality measures (CQMs);[Dorr2011] and high-level categorizations of the elements that make up clinical trial inclusion and exclusion criteria.[VanSpall2007, Ross2010a] Comparisons of multiple phenotype definitions have also been done,[Richesson2013] but have focused on a single disease or condition. However, no work to date has evaluated phenotype composition across diseases using a consistent computable representation.

The goal of this study is to characterize EHR-based phenotype definitions using a consistent computable representation, and generate new knowledge about the types and variability of constructs that must be accommodated in a system that is capable of formally representing a rich and diverse set of phenotype algorithms.

Materials & Methods

Data Set

We used a set of 33 phenotype definitions that were represented using FHIR and CQL. Full details about the creation of the phenotype definitions are explained elsewhere,[Brandt2021] but briefly, the data set is comprised of phenotype algorithms that utilized structured data, were marked as "Final" in PheKB, and were used in a published research study. These phenotype algorithms were then translated into FHIR (selected for its growth and use in healthcare and research contexts) and CQL (selected for its use within eCQMs and ability to represent phenotype algorithms), and validated using manual review and automated testing. The dataset is open source and available on GitHub[1].

Data Analysis

Metadata Analysis

PheKB allows phenotype submitters to voluntarily annotate their phenotype algorithm with metadata. We extracted and reviewed the metadata available on the PheKB page for each phenotype in JSON format, but this metadata was not used in our analysis due to issues with irrelevant, incomplete or outdated information. Instead, supplemental metadata was manually assembled by two authors (PSB, LVR) who independently conducted a review of PheKB for each phenotype in the dataset and curated relevant dimensions, emergent patterns, and characteristics. We categorized the artifacts provided with each phenotype definition (e.g., flowcharts) and whether or not the definition for controls, subtypes, or suspected cases is provided. We also provided a "Type" categorization to capture the intent of the phenotype, and note whether or not the phenotype used tabular data, and how these data are provided. In most cases tabular data refers to lists of codes from standard terminologies, but may also include lists of keywords or medication names. Following this manual review, the two reviewers met to discuss their findings and resolved any discrepancies.

Phenotype Definition Analysis

To evaluate the phenotype definition logic, we conducted an automated analysis of the Expression Logical Model (ELM) representation of each CQL library. An ELM representation is an instance of an Abstract Syntax Tree (AST),[Parr2009] which acts as a machine-readable representation of a complete program and is used to evaluate or execute the program. ASTs can be used in program translation, as has been shown for CQL,[Brandt2020] or for program analysis, as we demonstrate here. We evaluated the ELM for each phenotype by making use of the Visitor Pattern,[Gamma1995] which is a mechanism for inspecting each node of tree-like data structures and executing custom code in the context of each node. Our Java implementation used an interface provided by the reference implementation of the CQL translator,[2] which

[1] https://github.com/PheMA/phekb-phenotypes
[2] https://github.com/PheMA/elm-utils

is the same one used by the CQL engine during program execution. Our evaluator program calculated a number of measures about a given CQL library, including how many value sets are referenced, how many Boolean, temporal, and aggregate operators are used, and how these operators are combined. We also determined the total number of expressions, how many data types are used, and how many unique data queries are performed. After a preliminary manual review of the phenotype definitions, we identified 11 dimensions along which to evaluate each phenotype, shown in Table 1. We note that the library used did not include NLP implementations, and so NLP-related metrics were not considered in our analysis.

Category	Description	Examples
Aggregate	Operations that calculate single values from collections	**`sum`**, **`count`** or **`mean`**
Arithmetic	Mathematical operations	+, -, or *
Collection	Operations on collections of data like sets and lists	**`first`**, **`exists`** or **`union`**
Comparison	Numeric or date comparisons	**`>`** or **`=`**
Conditional	Branching logic	**`if`** or **`case`**
Data	Data retrieval and filtering operations	FHIR resource querying and filtering by value set
Expressions	Total number of expressions as well as their depth	Total expression count, where clause expression depth
Literals	Explicit values, codes and quantities	**`23`**, **`5`** months or **`0.5 mg/dL`**
Logical	Boolean logical operators	**`and`** or **`not`**
Temporal	Operators relating to dates and times	**`before`**, **`starts`** or **`overlaps`**
Terminology	Number of value sets used and the number of individual codes	Value sets per phenotype, codes per code system

Table 1. Phenotype definition analysis dimensions.

Results

Metadata

Table 2 provides metadata extracted by manually reviewing each phenotype definition. Our manually determined types align with the self-reported PheKB types, but we introduce a new type with the label "Valid Data". This indicates that the phenotype is trying to identify patients without disqualifying data. For example, the *Height* phenotype identifies individuals who have a valid height measurement and do not have any conditions that may impact height. We also introduce the "Treatment / Therapy" type, which identifies individuals who have had a specific treatment, for example, *Bone Scan Utilization*. Finally, we describe the approach outlined for NLP implementation (if applicable).

About two thirds of the definitions provide a narrative description (n=20) and flowchart (n=19), while only about one third (n=12) provide pseudocode. While tabular data is provided by all but three phenotypes, only 4 provide this data in a computable format. Computable artifacts in the form of KNIME workflows are provided for 5 phenotypes, and we found that these workflows require users to prepare their data in a specified custom format before execution.

Most phenotypes (n=20) provide control definitions, 8 provide phenotype subtype definitions, and 4 include a category for suspected cases. About half (n=16) of the phenotypes provide a list of covariates to be collected. All but 5 phenotypes rely on some form of NLP, with 17 providing a list of keywords, 8 providing regular expressions, and 6 providing a list of medication names.

Terminologies

Figures 2 and 3 provide histograms related to codes, code systems and value. The figures were generated using automated analysis of the ELM representation of the phenotype definitions and the value sets in FHIR format. Most phenotypes (n=28) use four or fewer code systems, and almost all (n=30) use ICD-9 codes, driven in part by the prevalence of phenotype algorithms in PheKB developed prior to or shortly after the adoption of ICD-10 in the United States. RxNorm (n=21), LOINC® (n=17), and CPT (n=16) are the next most commonly used. Five code systems (AMT, dm+d, BDPM, CIEL, and MedDRA) are each only used by a single phenotype.

About half of all codes (n=7020) are ICD-9 codes, and about a quarter (n=4112) are ICD-10 codes. CPT (n=1221) and RxNorm (n=699) are the next most common. The total number of codes used varies from 5 (*Warfarin Dose/Response*) to 6865 (*Developmental Language Disorder*), with a median of 147 (mean: 509.2, std: 1206.3). The total number of value sets used ranges from 1 (*Lipids and Sickle Cell Disease*) to 19 (*Resistant Hypertension*), with a median of 5 (mean: 6, std: 4.4).

Name	Type	Narrative	Flowchart	Pseudocode	Tabular	Executable	Suspected	Controls	Subtypes	Covariates	NLP
Asthma Response to Inhaled Steroids	Drug Response	✓			NC						regex
Atrial Fibrillation	Disease			✓	NC		✓	✓	✓		keywords; regex
Autism	Disease	✓	✓		NC			✓	✓	✓	DSM-IV criteria
Benign Prostatic Hyperplasia	Disease		✓	✓	NC	KNIME		✓		✓	keywords
Bone Scan Utilization	Treatment / Therapy	✓	✓		NC			✓			keywords
Cardiac Conduction	Trait	✓			NC					✓	keywords; negation; uncertainty
Cataracts	Disease		✓	✓	NC			✓	✓	✓	MedLEE concepts; negation; regex
Clopidogrel Poor Metabolizers	Drug Response		✓		NC			✓	✓		keywords
Crohn's Disease	Disease			✓	NC		✓	✓	✓		keywords; regex
Developmental Language Disorder	Disease	✓	✓		CSV	KNIME			✓		
Digital Rectal Exam	Treatment / Therapy	✓	✓		NC			✓			documentation is obtained from clinical notes
Drug Induced Liver Injury	Drug Response	✓	✓		NC			✓		✓	medication names; "diagnosis mentioned"
Familial Hypercholesterolemia	Disease	✓	✓	✓	XLSX			✓		✓	detailed pseudocode
Height	Healthy / Valid Data		✓		NC	KNIME				✓	keywords
Herpes Zoster	Disease		✓	✓	XLSX			✓		✓	
High-Density Lipoproteins	Trait		✓	✓	NC						keywords
Hypothyroidism	Disease	✓			NC			✓		✓	keywords
Lipids	Healthy / Valid Data	✓	✓	✓	NC	KNIME				✓	
Multimodal Analgesia	Treatment / Therapy	✓	✓		NC			✓			medication names
Multiple Sclerosis	Disease			✓	NC		✓	✓			keywords; regex
Peripheral Arterial Disease	Disease	✓			NC		✓			✓	keywords
Red Blood Cell Indices	Healthy / Valid Data	✓	✓		NC					✓	medication names
Resistant hypertension	Drug Response	✓			NC			✓		✓	medication names
Rheumatoid Arthritis	Disease			✓	NC		✓	✓			keywords; regex
Sickle Cell Disease	Disease	✓			NC						
Statins and MACE	Drug Response	✓	✓		NC			✓	✓	✓	medication names; keywords
Steroid Induced Osteonecrosis	Drug Response		✓		NC			✓			keywords
Systemic Lupus	Disease	✓									keywords
Type 2 Diabetes	Disease			✓	NC			✓			keywords; regex
Type 2 Diabetes Mellitus	Disease	✓	✓	✓	XLSX	KNIME; SQL fragments		✓		✓	medication names; regex
Urinary Incontinence	Treatment / Therapy	✓									executable Python code
Warfarin Dose/Response	Drug Response		✓						✓		keywords
White Blood Cell Indices	Healthy / Valid Data	✓			NC					✓	

Table 2. Metadata manually extracted from PheKB.

NC – Non-computable (e.g., Word or PDF), **XLSX** – Microsoft Excel file, **CSV** – Comma Separated Value file, **KNIME** – KoNstanz Information MinEr files, **SQL** – Structured Query Language

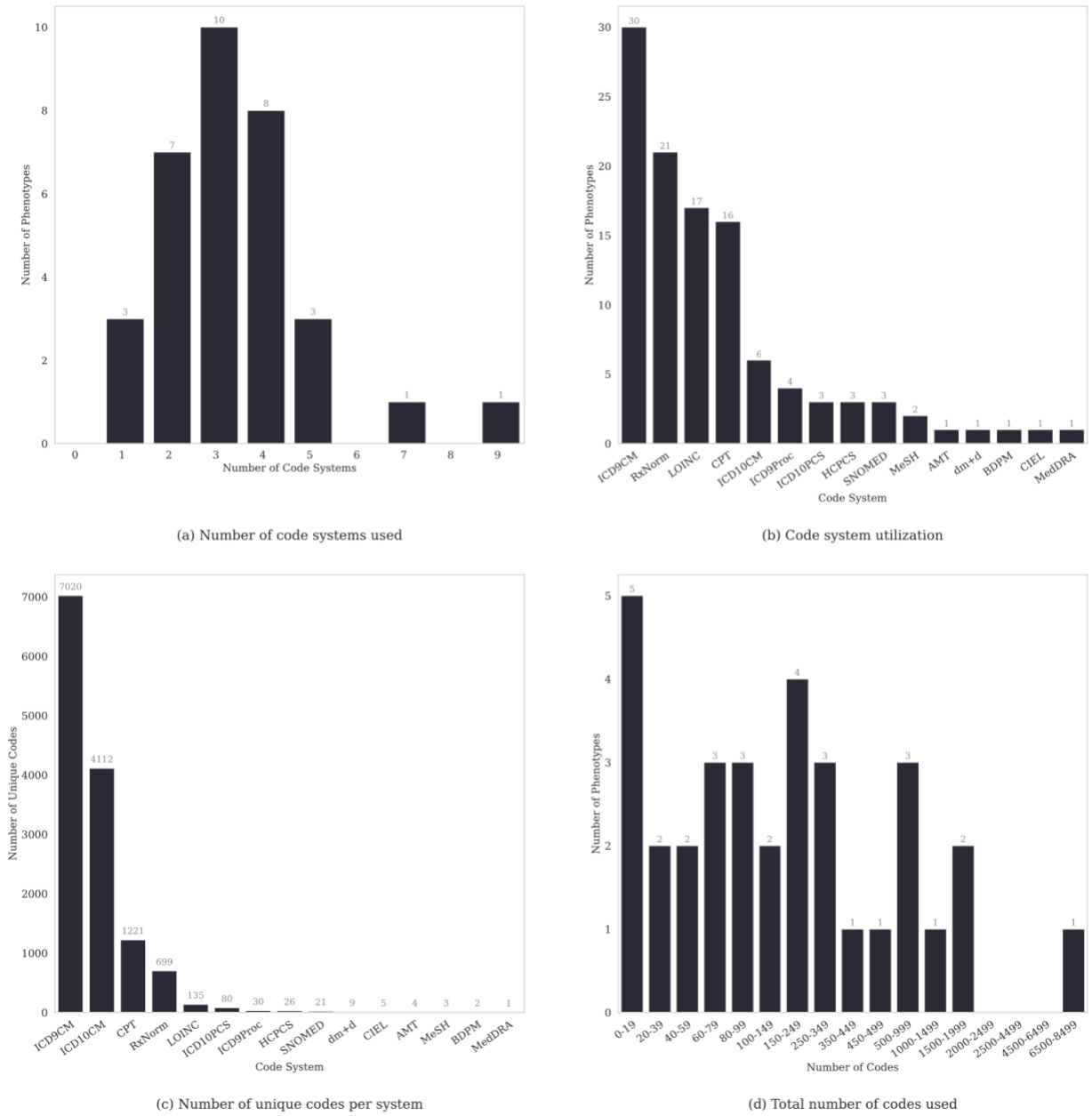


Figure 2. Histograms of code and code system usage.

ICD9CM – International Classification of Diseases, Ninth Revision, Clinical Modification; **ICD10CM** – International Classification of Diseases, Tenth Revision, Clinical Modification; **LOINC** - Logical Observation Identifiers Names and Codes; **CPT** - Current Procedural Terminology; **ICD9Proc** – International Classification of Diseases, Ninth Revision, Procedures; ; **ICD10PCS** – International Classification of Diseases, Tenth Revision, Procedure Coding System; **HCPCS** - Healthcare Common Procedure Coding System; **SNOMED** – Systematized Nomenclature of Medicine; **MeSH** - Medical Subject Headings; **AMT** - Australian Medicines Terminology; **dm+d** - Dictionary of Medicines and Devices; **BDPM** - Public Database of Medications; **CIEL** - Columbia International eHealth Laboratory; **MedDRA** - Medical Dictionary for Regulatory Activities

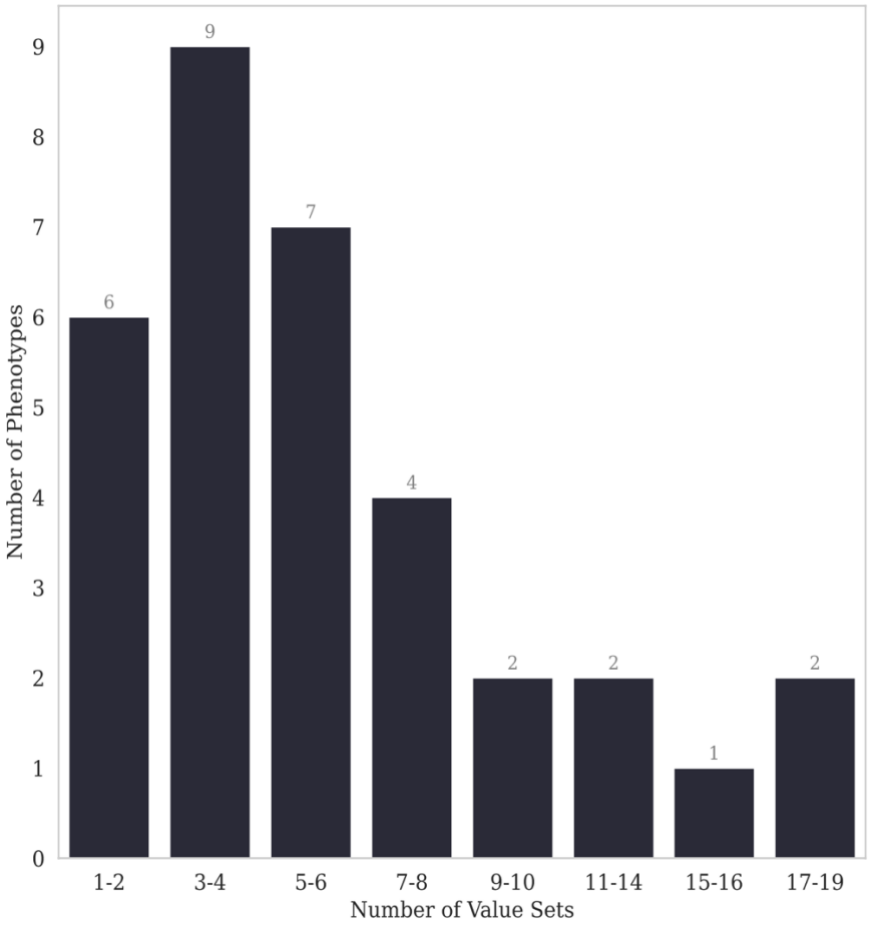


Figure 3. Histogram of value sets used.

Logical Expressions

Logical expressions were analyzed by programmatically examining the ELM representation of each phenotype definition. Figure 4 illustrates the total number of expressions used in each phenotype broken down by CQL expression category, enabling easy comparison between phenotypes and demonstrating the range of variability among the definitions. Figure 5 provides a histogram of the individual expressions within each category and Figure 6 illustrates the data types utilized by literal expressions.

The most widely used expression categories are literal (n=767), data (n=229), logical (n=341), and collection (n=292), with the latter three categories used by every phenotype. The least commonly used expression categories are aggregate (n=30) and arithmetic (n=80). The total number of expressions used ranges from 6 (*Autism*) to 228 (*Familial Hypercholesterolemia*), with a median of 59 (mean: 69.7, std: 53.3).

Only two types of aggregate expressions were used, with **`count`** (n=28) being the most common. The **`exists`** (n=191) expression is the most common collection expression used, with **`equal`** (n=109) and **`if`** (n=141) the most common comparison and conditional expressions respectively. The **`retrieve`** (n=228) expression was the most common non-literal expression overall, while the **`aggregate`** data expression was used only once. The **`and`** (n=170) expression was the most common logical operator, being used about twice as many times as **`not`** (n=83) and **`or`** (n=88). The **`start`** (n=30), which extracts the start date or time from an interval, is the most common temporal operator.

Figure 6 shows the frequency of data types for literal expressions. Of these, terminology literals are the most commonly used types, with the **`code`**, **`code system`**, and **`concept`** types each occurring in 27 or more phenotypes. The next most common are primitive types like **`Boolean`** (20) and **`Integer`** (24). In total, 18 phenotypes make use of a **`Quantity`** data type (a scalar value with a unit).

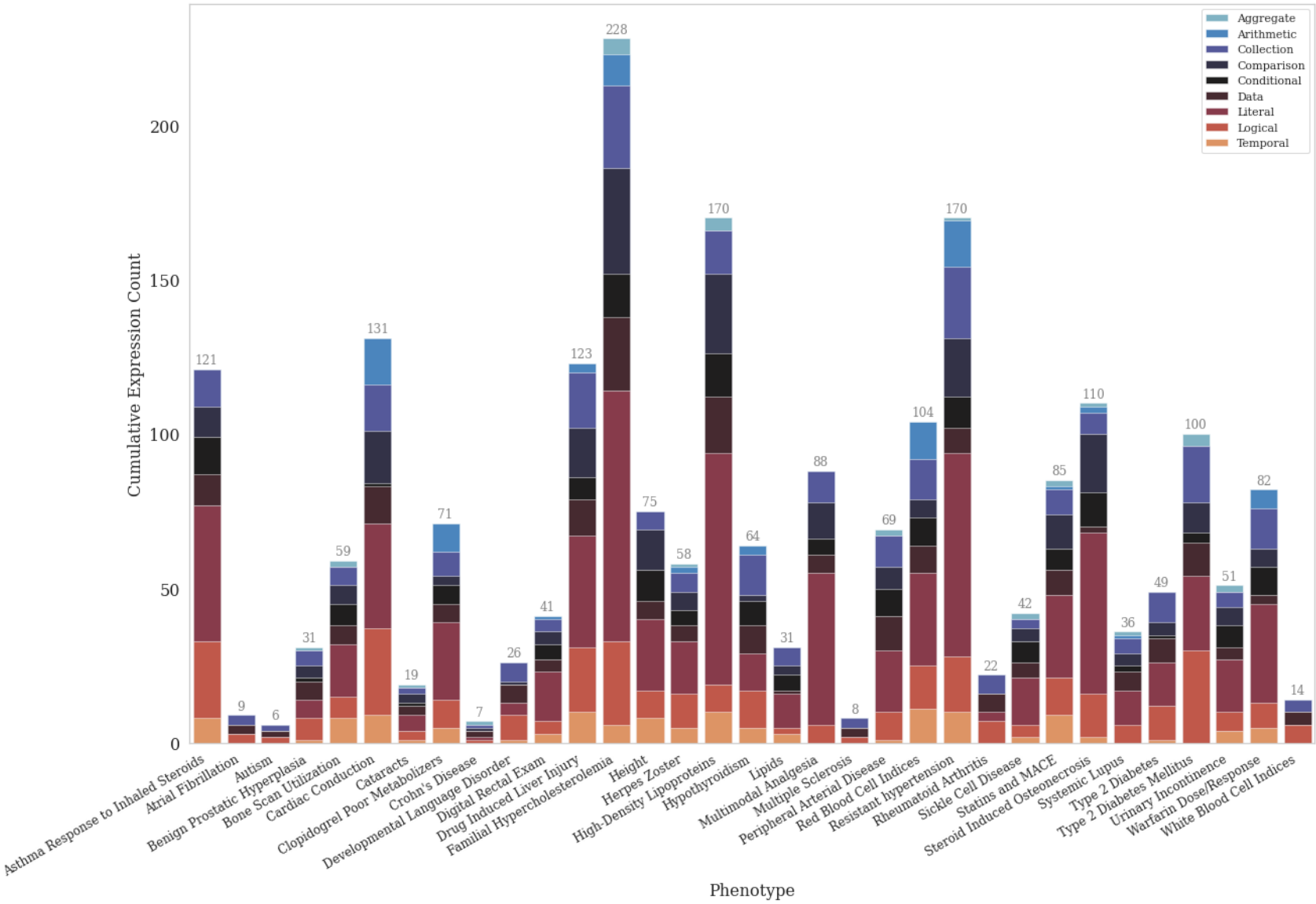


Figure 4. Cumulative expression counts per category.

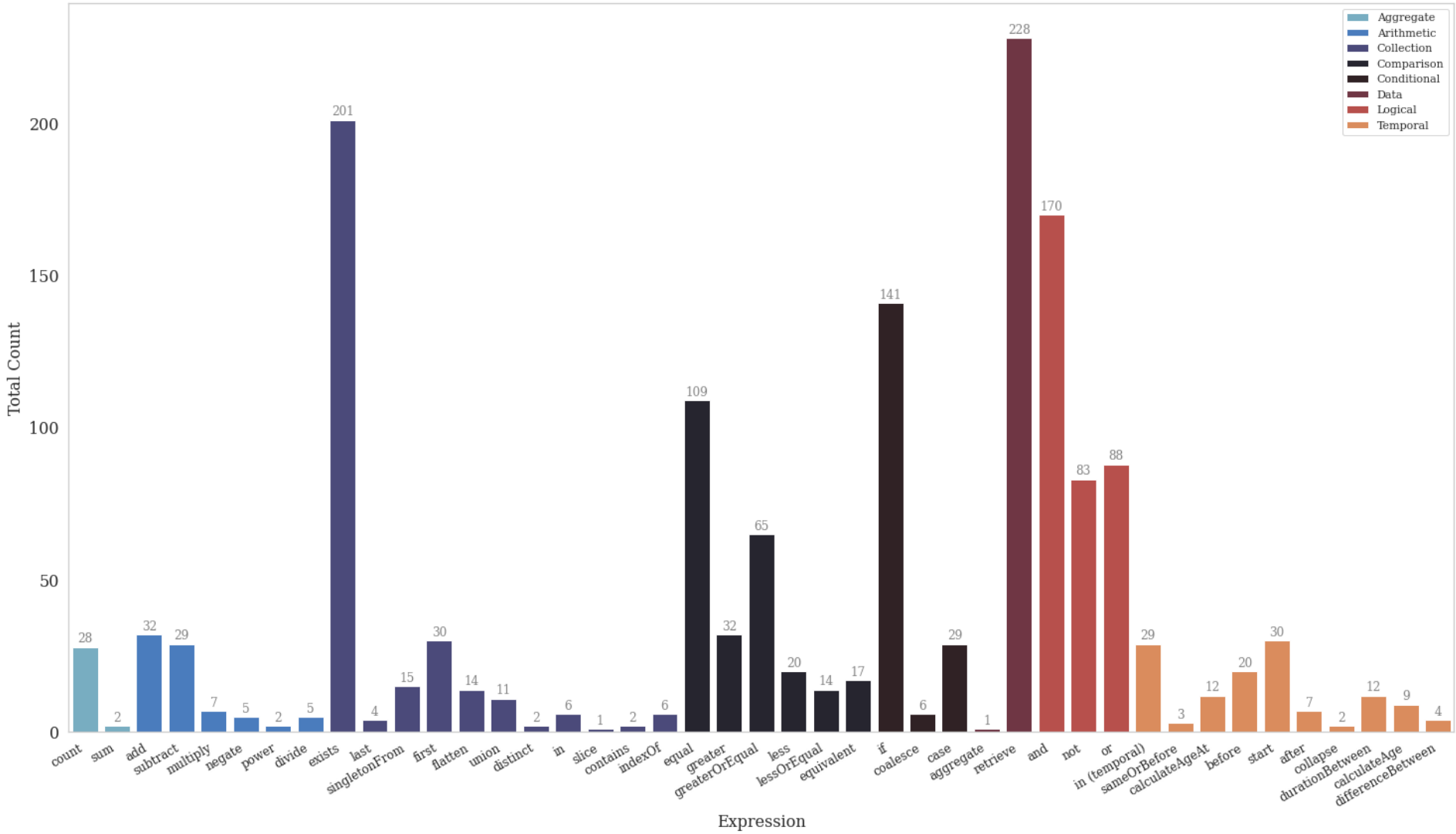


Figure 5. Total numbers of individual expression types by category (excluding literal expressions).

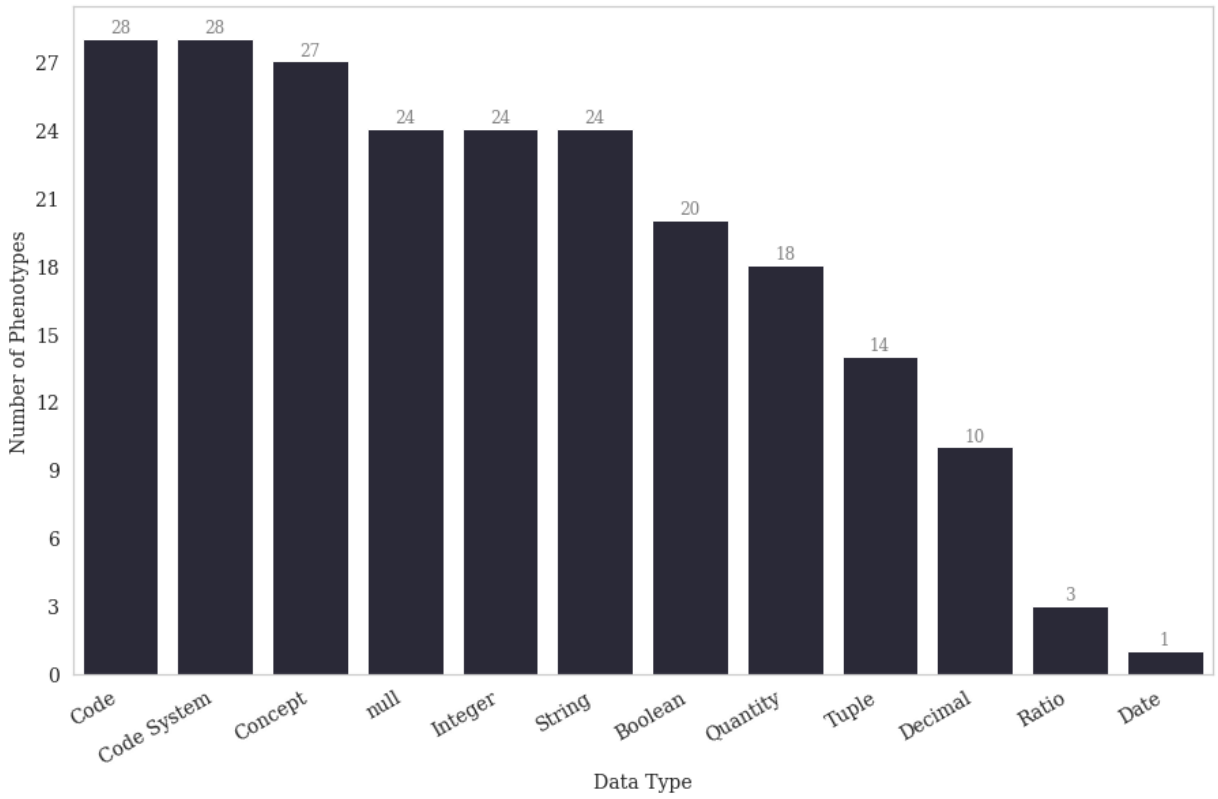
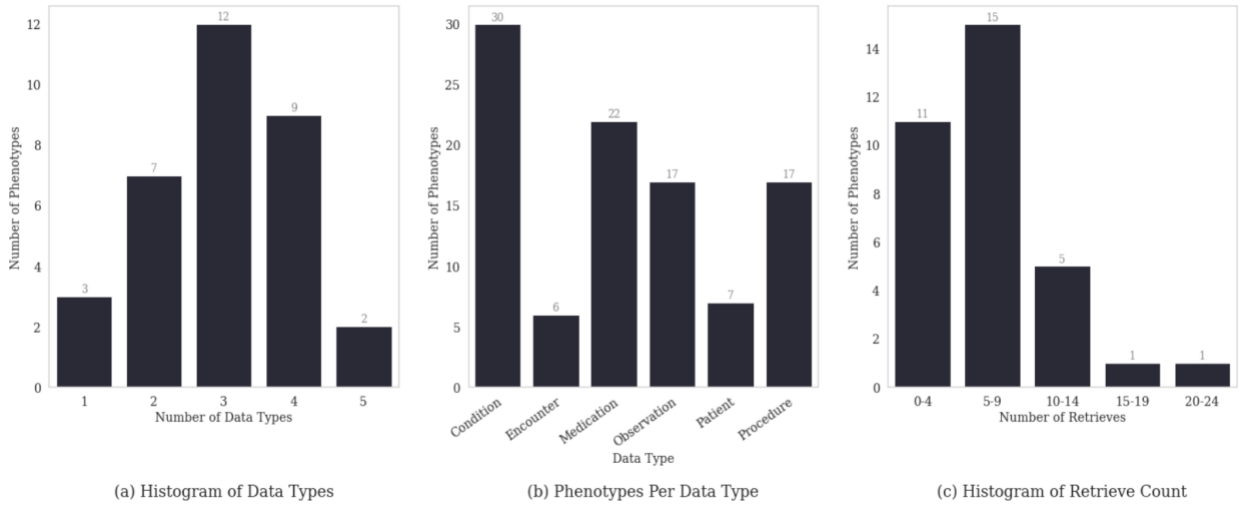


Figure 6. Utilization of various data types by literal expressions.

Data Sources

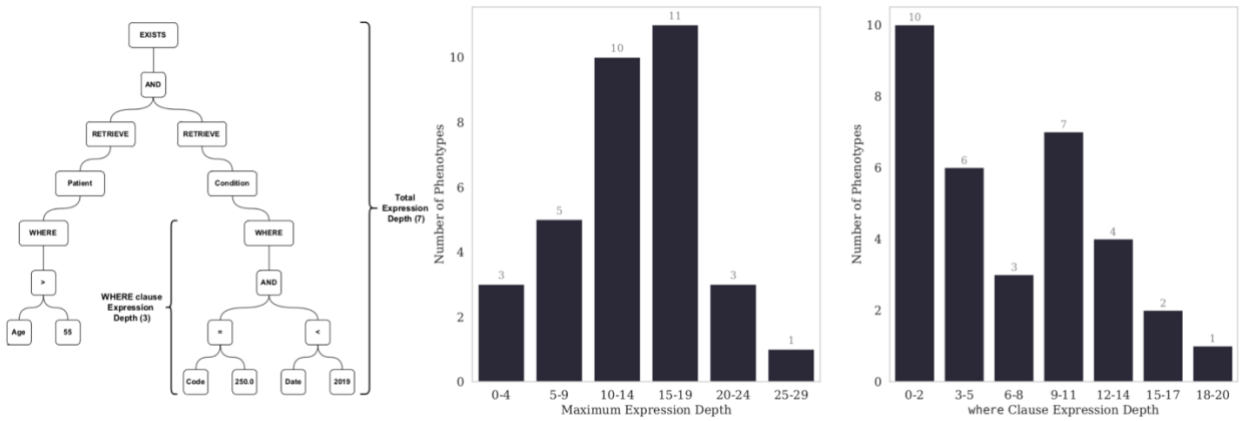
Figure 7 illustrates how many data types (distinct FHIR resources, such as Condition or Encounter, given the use of FHIR as the data model) are used per phenotype definition and how many phenotypes used each data source. The majority of phenotypes (n=22) used 3 data types or fewer. Condition was the most common data type, used by almost all (n=30) phenotypes, followed by medications (n=22), procedures (n=17), and observations (n=17). Demographic and encounter data were the least frequently used.

Also provided are histograms of the numbers of **`retrieve`** operations (used to fetch records for data sources). The majority of phenotypes use fewer than ten retrieves. The outliers are *Familial Hypercholesterolemia* (n=24) and *High-Density Lipoproteins* (n=18). The number of retrieves vary depending on the phenotype logic. For example, *Resistant Hypertension* has few retrieves (n=7). The retrieve count indicates how many distinct data requests are made, such as querying for all medications using codes from a specific value set, which can then be filtered or shaped in multiple ways.



(a) Histogram of Data Types (b) Phenotypes Per Data Type (c) Histogram of Retrieve Count

Figure 7. Data retrieval expressions and data types.



(a) Simplified Expression Depth Example (b) Phenotypes Per Maximum Expression Depth (c) `where` Clause Expression Depth

Figure 8. Example of expression depth (a), along with histograms of total (b) and **`where`** clause (c) expression depths.

Expression Depths

Expression depth is an indicator of how many logical expressions are applicable concurrently, which is roughly correlated with how many phenotype definition criteria are concurrently applicable. Figure 8 provides a histogram of total expression depth as well as **`where`** clause expression depth. **`where`** clause expression depth is an indicator of how complicated data filtering expressions are. Finally, Figure 9 illustrates expression depth per expression category, which shows how many expressions of each different category are concurrently applicable.

The lowest total expression depth is 4 (*Multiple Sclerosis*, *Crohn's Disease*, and *Autism*) and the highest is 27 (*Familial Hypercholesterolemia*), with a median of 14 (mean: 13.5, std: 5.7). Seven phenotypes have a **`where`** clause expression depth of zero (*White Blood Cell Indices*, *Rheumatoid Arthritis*, *Multiple Sclerosis*, *Crohn's Disease*, *Atrial Fibrillation*, *Autism*, and *Developmental Language Disorder*), meaning that data is only filtered by value set and no other criteria. *Red Blood Cell Indices* has the highest **`where`** clause expression depth of 18, and the median **`where`** clause expression depth is 7 (mean: 6.6, std: 5.4).

Many phenotypes have expression depths of zero or one for aggregate, arithmetic, collection, comparison, and conditional expression. Logical expressions always have a depth of at least one. There are some instances of expression depths in the two to four range for conditional, collection, comparison, and logical expressions, but only logical and arithmetic expressions have a depth of five or greater. Only logical expressions have a depth greater than six, with a maximum of 10 in two cases (*Red Blood Cell Indices* and *Familial Hypercholesterolemia*).

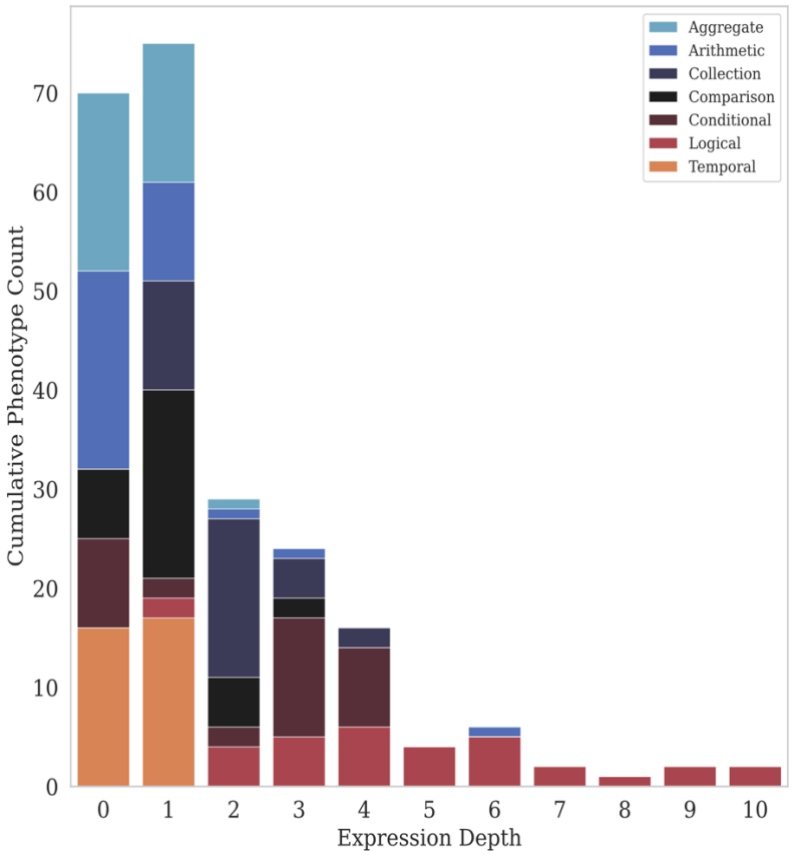


Figure 9. Expression depth utilization per category.

Discussion

In this work we found that all of the 33 phenotypes evaluated use a combination of clinical logic and operational logic as part of their definitions. Similarly, despite the range of conditions, all of the definitions consisted of logical criteria, representing both clinical and operational logic, and tabular data, consisting of codes from standard terminologies and keywords for natural language processing. The original metadata associated with phenotype algorithms we evaluated was found to be irrelevant for this analysis or incorrect, demonstrating the need not only to identify more relevant metadata elements and ensure they are appropriately curated and verified, but also ensure metadata review and curation is ongoing to ensure accuracy. Some metadata elements that were originally curated by hand in PheKB, such as the categories of data or medical vocabularies used, could be extracted and updated programmatically if phenotype algorithms were represented in a standard computable format, such as CQL and FHIR. Overall, ongoing work is needed to improve metadata collection for phenotype algorithms, with some groundwork proposed in the field.[Chapman2021, Alper2022]

The artifacts that comprised the phenotype definitions can be divided into two high-level categories: logic and tabular data. The tabular data consists of value sets of codes from various code systems and lists of keywords or regular expressions used for NLP, although we did not analyze the latter in depth. We identified that diverse vocabularies are used for phenotype algorithms - all of which could be represented in CQL and FHIR. The OMOP common data model standardizes the terminologies used to a smaller subset, however, and comprehensive phenotyping platforms should anticipate accommodating a broad set of vocabularies.

Phenotype logic can be further divided into two categories: clinical logic and operational logic. Clinical logic is the core of the phenotype definition, and describes the clinical definition of the phenotype. Clinical logic includes things such as which diagnoses are relevant, which procedures, medications, and laboratory orders are associated with the phenotype, as well as patient demographic criteria that should be considered. Operational logic is also important, and while it contributes to the clinical definition, it is typically a bridge to how data are recorded in the EHR. Patterns in operational logic have been observed,[Rasmussen2014] but they can be difficult to express accurately using universally applicable logical expressions. For example, the *Herpes Zoster* phenotype requires that a matching patient have at least 5 years of continuous enrollment. The reason for this requirement is to "increase the probability that a subject's status with respect to herpes zoster infection is known by the health care system." This does not necessarily increase the correctness of the phenotype definition, but may nevertheless increase the negative predictive value. Another very common operational criterion is the requirement that a patient have at least 2 diagnoses of a given condition. This criterion is relatively simple to define using universally applicable CQL logic, while the concept of enrollment is determined differently at different institutions. One solution to this problem, which is available when using a modular formal representation, is to have local implementations for common operational criteria that are used during cohort execution. This is the same approach used in computer software, where system libraries provide routines with known names and well-defined parameters, but the implementation varies according to the operating system. However, this highlights the need for more broadly accepted metrics applicable to a wide range of health data that convey information about the completeness of capture of patient data, for a given data source. Completeness has been explored and reported in data quality frameworks,[Kahn2016, Schmidt2021] but work is needed to integrate these considerations more readily into phenotype authoring.

The types of expressions observed were simple, and required more complex calculations and aggregations of data to take place within the phenotype. The complexity of the phenotypes then comes from the topology - many simple expressions linked together in increasingly complex ways. The prevalence of certain expressions such as data retrieval, manipulation, and literal expressions is expected, but we were surprised by the relatively low usage of arithmetic and aggregate expressions, which indicates that in most cases data values are used directly, not used to construct derived values. Both the **`count`** and **`sum`** aggregate expressions are used as cardinality constraints (e.g., at least 2 diagnoses required) and not in an arithmetic context. The highly used existential operator (**`exists`**) is used for the same purpose (e.g., does an observation exist that meets certain requirements). Additionally, even though about two thirds of phenotypes use temporal expressions, the total number used is relatively small. The fact that the **`and`** operator is used about twice as much as the **`or`** and **`not`** operators makes sense, since conceptually, many phenotypes are defined by clusters of concurrent criteria rather than by the disjunction of many different criteria.

Overall expression depths seem to be normally distributed around 15, while **`where`** clause expression depth appears to be bimodal, but generally trends down at high values. The downward trend implies that complicated data filtering expressions are uncommon. Most expression categories are not very deeply nested, and only logical expressions (and in one case arithmetic expressions) have a depth of 5 or higher. This can be interpreted to mean that conceptually simple criteria are combined in complex ways using **`and`** and **`or`** expressions. This can be confirmed by looking at the flowcharts provided with some phenotype definitions, which may have a complicated topology, but the criteria represented by each node are relatively simple.

The total number of expressions per phenotype is generally not very high, with a median of just 59. The total number of expression types is also quite low, at just 43 (CQL has over 200 expression types). There are several possible reasons for the simplicity of the phenotypes in the data set. First, in our experience, implementing even simple phenotype definitions is quite challenging, so authors may choose to keep definitions simple to make implementation practical. Second, the phenotypes generally restrict themselves to data available in the EHR, which may be simplified, as this data is often for billing purposes. Further, since the PheKB phenotypes are designed to be shared, authors may limit themselves only to data available to most implementers, such as the most basic data elements. Finally, since phenotypes are created as narrative text, the lack of a formal expression language may be a factor that limits the level of detail provided.

There is no clear correlation between the severity or complexity of presentation of a disease and the number of expressions used. For example, something ostensibly simple like *Height* has ten times as many expressions as *Multiple Sclerosis*. There are also two Type 2 Diabetes phenotypes, one with 49 expressions and one with 100, so even the same disease can be represented in vastly different ways. This implies that a large part of phenotype definition complexity is determined by the level of detail that the author decided to use. This is independent of any formal representation, and may depend on the intended use of the phenotype definition.

Limitations

We note the following limitations in this work. First, the data set used is relatively small, consisting of only 33 phenotype definitions, and there may exist additional phenotype definitions that would alter our conclusions. Additionally, many of the phenotypes evaluated were developed for conducting genomic studies, and the implementation decisions may have been tuned specifically for that purpose, biasing our findings. We also note that all the definitions in our dataset were designed to detect patients with a single condition. Even though EHR data provides the opportunity to conduct research on patients with multiple simultaneous conditions, the phenotype definitions we used were not designed to identify cohorts of such complex patients. The phenotypes analyzed are reminiscent of those that might be used for recruiting patients for randomized controlled trials, but we hope this work will provide some insights that lead to methods for developing higher fidelity phenotype definitions, and that these definitions can be used for so-called *deep* phenotyping.

The decision not to include NLP directives in our analysis was purposeful, but impacts any conclusions we would wish to draw about individual phenotypes. We know that most (n=28) phenotype definitions make use of some form of NLP, so we are omitting data from a significant number of definitions. However, almost all definitions including NLP directives simply provide a keyword list or regular expressions, and not higher-level entities, negation, and/or temporal constructs. In some cases, no details are given about how the NLP should be implemented. For example, the *Red Blood Cell Indices* phenotype definition says only "NLP was implemented to that regard" (referring to identifying patients taking specific medications). So, we believe that including a more detailed analysis of NLP data elements would not be informative due to their underspecificity in the dataset. A substantial amount of data useful for EHR-driven phenotyping may be stored in clinical notes. The body of research focused on extracting this data is growing,[Zeng2018] but to our knowledge, no widely accepted standard representation for NLP metadata and processes has yet emerged. The PhEMA research team is working on methods for integrating NLP into both FHIR [Hong2019] and CQL [Wen2021] and in future work we hope to integrate this research into our analysis of phenotype definitions using formal representations.

Conclusion

In this work, we characterize a set of 33 validated research phenotype definitions, identifying how phenotype implementations use a combination of clinical logic and operational logic. The phenotypes analyzed are composed of the same high-level components, namely tabular data and logical expressions. The most important type of tabular data analyzed here are value sets, which can readily be represented in a standard format. Despite the limited number of expressions used from those available, individual phenotype algorithms could become complex in the total number of expressions needed and depth of nested operations, often to accommodate the needed operational bridge from how EHR data is collected and recorded. This complexity points to the need to use standard-based representations for expressing, sharing and implementation of phenotype definitions.

Acknowledgements

The authors wish to thank Frank Mentch from the Children's Hospital of Philadelphia for providing feedback on an earlier draft of the manuscript.

Funding Statement

This work was conducted during the third phase of the eMERGE Network, which was initiated and funded by the NHGRI through the following grants: U01HG008657 (Group Health Cooperative/University of Washington); U01HG008685 (Brigham and Women's Hospital); U01HG008672 (Vanderbilt University Medical Center); U01HG008666 (Cincinnati Children's Hospital Medical Center); U01HG006379 (Mayo Clinic); U01HG008679 (Geisinger Clinic); U01HG008680 (Columbia University Health Sciences); U01HG008684 (Children's Hospital of Philadelphia); U01HG008673 (Northwestern University); U01HG008701 (Vanderbilt University Medical Center serving as the Coordinating Center); U01HG008676 (Partners Healthcare/Broad Institute); U01HG008664 (Baylor College of Medicine); and U54MD007593 (Meharry Medical College). LVR, JBS, AK, YL, JAP, and TLW received additional support from NHGRI grant U01HG011169. PSB was funded by the Fulbright Foreign Student Program and the South African National Research Foundation.

Conflict Of Interests Statement

PSB is a consultant for Commure, Inc. TLW has received research funding from Gilead Sciences unrelated to the work presented here. The other co-authors have no competing interests to declare.

References

[Banda2018] Banda JM, Seneviratne M, Hernandez-Boussard T, Shah NH. Advances in Electronic Phenotyping: From Rule-Based Definitions to Machine Learning Models. *Annu Rev Biomed Data Sci*. 2018. doi:10.1146/annurev-biodatasci

[Fleurence2014] Fleurence RL, Curtis LH, Califf RM, Platt R, Selby J V., Brown JS. Launching PCORnet, a national patient-centered clinical research network. *J Am Med Informatics Assoc*. 2014;21(4):578-582. doi:10.1136/amiajnl-2014-002747

[Hripcsak2015] Hripcsak G, Duke JD, Shah NH, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. *Stud Health Technol Inform*. 2015;216:574-578. doi:10.3233/978-1-61499-564-7-574

[McCarty2011] McCarty CA, Chisholm RL, Chute CG, et al. The eMERGE Network: A consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med Genomics*. 2011;4(1):13. doi:10.1186/1755-8794-4-13

[Gottesman2013] Gottesman O, Kuivaniemi H, Tromp G, et al. The Electronic Medical Records and Genomics (eMERGE) Network: Past, present, and future. *Genet Med*. 2013;15(10):761-771. doi:10.1038/gim.2013.72

[Zouk2019] Zouk H, Venner E, Lennon NJ, et al. Harmonizing Clinical Sequencing and Interpretation for the eMERGE III Network. *Am J Hum Genet*. 2019;105(3):588-605. doi:10.1016/j.ajhg.2019.07.018

[Hripcsak2015] Hripcsak G, Duke JD, Shah NH, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. *Stud Health Technol Inform*. 2015;216:574-578. doi:10.3233/978-1-61499-564-7-574

[Ahmad2020] Ahmad FS, Ricket IM, Hammill BG, et al. Computable Phenotype Implementation for a National, Multicenter Pragmatic Clinical Trial: Lessons Learned From ADAPTABLE. Circ Cardiovasc Qual Outcomes. 2020;13(6):e006292. doi:10.1161/CIRCOUTCOMES.119.006292

[Hripcsak2019] Hripcsak G, Shang N, Peissig PL, et al. Facilitating phenotype transfer using a common data model. J Biomed Inform. 2019;96:103253. doi:10.1016/j.jbi.2019.103253

[Burn2020] Burn, E., You, S.C., Sena, A.G. et al. Deep phenotyping of 34,128 adult patients hospitalised with COVID-19 in an international network study. Nat Commun 11, 5009 (2020). https://doi.org/10.1038/s41467-020-18849-z

[Richesson2016] Richesson RL, Smerek MM, Blake Cameron C. A Framework to Support the Sharing and Reuse of Computable Phenotype Definitions Across Health Care Delivery and Clinical Research Applications. EGEMS (Wash DC). 2016 Jul 5;4(3):1232. doi: 10.13063/2327-9214.1232. PMID: 27563686; PMCID: PMC4975566

[Kirby2016] Kirby JC, Speltz P, Rasmussen L V., et al. PheKB: A catalog and workflow for creating electronic phenotype algorithms for transportability. *J Am Med Informatics Assoc*. 2016;23(6):1046-1052. doi:10.1093/jamia/ocv202

[Denaxas2019] Spiros Denaxas, Arturo Gonzalez-Izquierdo, Kenan Direk, Natalie K Fitzpatrick, Ghazaleh Fatemifar, Amitava Banerjee, Richard J B Dobson, Laurence J Howe, Valerie Kuan, R Tom Lumbers, Laura Pasea, Riyaz S Patel, Anoop D Shah, Aroon D Hingorani, Cathie Sudlow, Harry Hemingway, UK phenomics platform for developing and validating electronic health record phenotypes: CALIBER, Journal of the American Medical Informatics Association, Volume 26, Issue 12, December 2019, Pages 1545–1559, https://doi.org/10.1093/jamia/ocz105

[Pathak2013a] Pathak J, Kho AN, Denny JC. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. *J Am Med Informatics Assoc*. 2013;20(e2):e206-e211. doi:10.1136/amiajnl-2013-002428

[Newton2013] Newton KM, Peissig PL, Kho AN, et al. Validation of electronic medical record-based phenotyping algorithms: Results and lessons learned from the eMERGE network. *J Am Med Informatics Assoc*. 2013;20(E1). doi:10.1136/amiajnl-2012-000896

[Shivade2014a] Shivade C, Raghavan P, Fosler-Lussier E, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. *J Am Med Inform Assoc*. 2014;21(2):221-230. doi:10.1136/amiajnl-2013-001935

[Adekkanattu2020] Adekkanattu P, Jiang G, Luo Y, et al. Evaluating the Portability of an NLP System for Processing Echocardiograms: A Retrospective, Multi-site Observational Study. *AMIA Annu Symp Proc*. 2020;2019:190-199. Published 2020 Mar 4

[Yu2021] Yu J, Pacheco JA, Ghosh A, et al. Under-specification as the Source of Ambiguity and Vagueness in Narrative Phenotype Algorithm Definitions. Research Square; 2021. DOI: 10.21203/rs.3.rs-723873/v1

[Peterson2014a] Peterson KJ, Pathak J. Scalable and High-Throughput Execution of Clinical Quality Measures from Electronic Health Records using MapReduce and the JBoss® Drools Engine. *AMIA Annu Symp Proc*. 2014;2014:1864-1873.

[Pathak2013] Pathak J, Bailey KR, Beebe CE, et al. Normalization and standardization of electronic health records for high-throughput phenotyping: The sharpn consortium. *J Am Med Informatics Assoc*. 2013;20(E2):341-348. doi:10.1136/amiajnl-2013-001939

[Mo2015b] Mo H, Pacheco JA, Rasmussen L V, et al. A Prototype for Executable and Portable Electronic Clinical Quality Measures Using the KNIME Analytics Platform. *AMIA Jt Summits Transl Sci proceedings AMIA Jt Summits Transl Sci*. 2015;2015(Icd):127-131. http://www.ncbi.nlm.nih.gov/pubmed/26306254.

[Mo2016] 1. Mo H, Jiang G, Pacheco JA, et al. A Decompositional Approach to Executing Quality Data Model Algorithms on the i2b2 Platform. *AMIA Jt Summits Transl Sci proceedings AMIA Jt Summits Transl Sci*. 2016;2016:167-175. http://www.ncbi.nlm.nih.gov/pubmed/27570665.

[Jiang2017] Jiang G, Prud'Hommeaux E, Xiao G, Solbrig HR. Developing A Semantic Web-based Framework for Executing the Clinical Quality Language Using FHIR. *CEUR Workshop Proc*. 2017;2042:1-5

[Chapman2020a] Chapman M, Rasmussen L V., Pacheco JA, Curcin V. Phenoflow: A Microservice Architecture for Portable Workflow-based Phenotype Definitions. *medRxiv*. 2020. doi:10.1101/2020.07.01.20144196

[Hong2019] Hong N, Wen A, Stone DJ, et al. Developing a FHIR-based EHR phenotyping framework: A case study for identification of patients with obesity and multiple comorbidities from discharge summaries. *J Biomed Inform*. 2019;99(April):103310. doi:10.1016/j.jbi.2019.103310

[Brandt2020] Brandt PS, Kiefer RC, Pacheco JA, et al. Toward cross-platform electronic health record-driven phenotyping using Clinical Quality Language. *Learn Heal Syst*. 2020;4(4):1-9. doi:10.1002/lrh2.10233

[Conway2011] Conway M, Berg RL, Carrell D, et al. Analyzing the heterogeneity and complexity of Electronic Health Record oriented phenotyping algorithms. *AMIA . Annu Symp proceedings AMIA Symp*. 2011;2011:274-283. doi:PMC3243189

[Dorr2011] Dorr DA, Cohen AM, Williams MP-J, et al. From simply inaccurate to complex and inaccurate: complexity in standards-based quality measures. *AMIA . Annu Symp proceedings AMIA Symp*. 2011;2011:331-338. http://www.ncbi.nlm.nih.gov/pubmed/22195085

[VanSpall2007] Van Spall HGC, Toren A, Kiss A, Fowler RA. Eligibility Criteria of Randomized Controlled Trials Published in High-Impact General Medical Journals. *JAMA*. 2007;297(11):1233. doi:10.1001/jama.297.11.1233

[Ross2010a] Ross J, Tu S, Carini S, Sim I. Analysis of eligibility criteria complexity in clinical trials. *Summit on Translat Bioinforma*. 2010;2010:46-50. http://www.ncbi.nlm.nih.gov/pubmed/21347148

[Richesson2013] Richesson RL, Rusincovitch SA, Wixted D, et al. A comparison of phenotype definitions for diabetes mellitus. J Am Med Inform Assoc. 2013;20(e2):e319-e326. doi:10.1136/amiajnl-2013-001952

[Brandt2021] Brandt PS, Pacheco JA, Rasmussen LV. Development of a repository of computable phenotype definitions using the clinical quality language. JAMIA Open. 2021 Dec 3;4(4):ooab094. doi: 10.1093/jamiaopen/ooab094. PMID: 34926996; PMCID: PMC8672934

[Parr2009] Parr T. *Language Implementation Patterns: Create Your Own Domain-Specific and General Programming Languages*. Pragmatic Bookshelf; 2009

[Gamma1995] Gamma E, Helm R, Johnson R, Vlissides J. *Design Patterns: Elements of Reusable Object-Oriented Software*. USA: Addison-Wesley Longman Publishing Co., Inc.; 1995. doi:10.5555/186897

[Chapman2021] Martin Chapman, Shahzad Mumtaz, Luke V Rasmussen, Andreas Karwath, Georgios V Gkoutos, Chuang Gao, Dan Thayer, Jennifer A Pacheco, Helen Parkinson, Rachel L Richesson, Emily Jefferson, Spiros Denaxas, Vasa Curcin, Desiderata for the development of next-generation electronic health record phenotype libraries, GigaScience, Volume 10, Issue 9, September 2021, giab059, https://doi.org/10.1093/gigascience/giab059

[Alper2022] Alper, BS, Flynn, A, Bray, BE, et al. Categorizing metadata to help mobilize computable biomedical knowledge. Learn Health Sys. 2022; 6:e10271. https://doi.org/10.1002/lrh2.10271

[Rasmussen2014] Rasmussen L V., Thompson WK, Pacheco JA, et al. Design patterns for the development of electronic health record-driven phenotype extraction algorithms. *J Biomed Inform*. 2014;51:280-286. doi:10.1016/j.jbi.2014.06.007

[Kahn2016] Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, Estiri H, Goerg C, Holve E, Johnson SG, Liaw ST, Hamilton-Lopez M, Meeker D, Ong TC, Ryan P, Shang N, Weiskopf NG, Weng C, Zozus MN, Schilling L. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. EGEMS (Wash DC). 2016 Sep 11;4(1):1244. doi: 10.13063/2327-9214.1244. PMID: 27713905; PMCID: PMC5051581

[Schmidt2021] Schmidt, C.O., Struckmann, S., Enzenbach, C. et al. Facilitating harmonized data quality assessments. A data quality framework for observational health research data collections with software implementations in R. BMC Med Res Methodol 21, 63 (2021). https://doi.org/10.1186/s12874-021-01252-7

[Zeng2018] Zeng, Z., Deng, Y., Li, X., Naumann, T. and Luo, Y., 2018. Natural language processing for EHR-based computational phenotyping. IEEE/ACM transactions on computational biology and bioinformatics, 16(1), pp.139-153

[Wen2021] Andrew Wen, Luke V. Rasmussen, Daniel Stone, et al. CQL4NLP: Development and Integration of FHIR NLP Extensions in Clinical Quality Language for EHR-driven Phenotyping. AMIA 2021 Informatics Summit; p624-633