Received: 30 July 2021 | Revised: 10 September 2021 | Accepted: 21 September 2021

DOI: 10.1002/lrh2.10293

LEARNING FROM DATA

Learning Health Systems

Developing real-world evidence from real-world data: Transforming raw data into analytical datasets

Lisa Bastarache[1] | Jeffrey S. Brown[2] | James J. Cimino[3] | David A. Dorr[4] | Peter J. Embi[5] | Philip R.O. Payne[6] | Adam B. Wilcox[7] | Mark G. Weiner[8]

[1]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee, USA

[2]Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, Massachusetts, USA

[3]Informatics Institute, University of Alabama at Birmingham, Birmingham, Alabama, USA

[4]Department of Medical Informatics and Clinical Epidemiology, Oregon Health Sciences University, Portland, Oregon, USA

[5]Center for Biomedical Informatics, Regenstrief Institute, Indianapolis, Indiana, USA

[6]Institute for Informatics, Washington University in St. Louis, St. Louis, Missouri, USA

[7]Institute for Informatics, Washington University in St. Louis School of Medicine, St. Louis, Missouri, USA

[8]Department of Population Health Sciences, Weill Cornell Medicine, New York, New York, USA

Correspondence
Mark G. Weiner, Department of Population Health Sciences, Weill Cornell Medicine, 402 E.67th Street, Rm 237, New York, NY 10-65, USA.
Email: mgw4001@med.cornell.edu

Abstract

Development of evidence-based practice requires practice-based evidence, which can be acquired through analysis of real-world data from electronic health records (EHRs). The EHR contains volumes of information about patients—physical measurements, diagnoses, exposures, and markers of health behavior—that can be used to create algorithms for risk stratification or to gain insight into associations between exposures, interventions, and outcomes. But to transform real-world data into reliable real-world evidence, one must not only choose the correct analytical methods but also have an understanding of the quality, detail, provenance, and organization of the underlying source data and address the differences in these characteristics across sites when conducting analyses that span institutions. This manuscript explores the idiosyncrasies inherent in the capture, formatting, and standardization of EHR data and discusses the clinical domain and informatics competencies required to transform the raw clinical, real-world data into high-quality, fit-for-purpose analytical data sets used to generate real-world evidence.

KEYWORDS
data science, real-world data, real-world evidence

1 | REAL-WORLD DATA

EHRs are a valuable resource for researchers that can be analyzed with variety of methods, from multivariate regression to machine learning, and may be used to support both cross-sectional and longitudinal studies.[1-3] But regardless of study design and method, all EHR-based research requires recognition of data quality issues[4] as well as data curation and cleaning.[5] Raw EHR data must be structured into analytical datasets that are formatted to fit the input requirements of an analytical function. This task is difficult given the chaotic nature of EHR data that are collected for nonresearch purposes, namely patient care and hospital administration.[6] Some challenges include erroneous or ambiguous recording of information, ascertainment bias, the temporality, or provisional nature of some findings and shifting context that can change the meaning of a given data point.

Decisions about data included in the EHR are rarely, if ever, neutral. Facts about patients are almost exclusively recorded when they are in contact with the healthcare system, and the nature of what is recorded

is highly dependent on who cares for the patient, the intended use of the information, and why they are seeking care in the first place.[7] If a patient goes to the emergency room for a broken arm, the fact that they have insomnia is unlikely to be recorded. If they go to a neurologist, recording of this diagnosis is much more likely to be noted. Rather than causing these conditions, a primary care visit is simply an opportunity to ascertain their existence. Ascertainment bias can be difficult to detect and may have profound consequences to downstream analyses.

The transformation of raw data into analytical data sets requires multiple decisions, and it is difficult to know a priori which decisions will impact an analysis. Making these decisions should be a collaborative process, one that combines the data scientist's understanding of underlying data as well as the clinician's domain expertise. However, these decisions are often delegated to the data scientist alone.[8] When clinical expertise is sought, the expert may make decisions based on their knowledge that is clinically correct, but nevertheless are inappropriate for the analysis because they are not informed by the actual content of the data. The data scientist may make decisions that are technically reasonable but insensitive to clinical realities.

The following sections describe issues that the data scientist and the domain expert must recognize to produce high-quality research. The sections are organized by data types that are commonly used in analytical datasets—diagnoses, measurements, and medications—and describes common issues that arise in using real-world EHR data for research. The examples provided below will have varying relevance to a specific research application of EHR data, depending on study design and the input requirements of the analytical methods. But we hope that they convey a sense of the complexity of EHR data and emphasize the importance of data literacy and the need for collaboration between clinical domain experts and data scientists in EHR research.

2 | DIAGNOSES

Analytical data sets frequently require information on the presence or absence of diagnoses. To populate such fields, data scientists often look for assertions of a particular diagnosis, using data like International Classification of Diseases (ICD) codes, problem lists, or text mentions in notes. However, information in these data sources can be misleading. Errors may be introduced through simple typos or miscommunications. But even accurately recorded information can be misleading due to the nature of diagnosis itself. Rather than a statement of fact that can be expressed as a yes/no value, a diagnosis is a statement of probability that can fluctuate over time. For some diseases, there is disagreement even among experts about the correct clinical criteria that should be used to establish a diagnosis.[9] Because the EHR captures data throughout the diagnostic process, a patient's record may accumulate evidence for a diagnosis that is ultimately ruled out.

Diagnoses can be extracted from clinical notes using natural language processing.[10-12] Unstructured text contains detailed information about patients that may be essential to accurately ascertain a disease phenotype, but structured data like International Classification of Diseases (ICD) codes are often used as an alternative. Compared with unstructured text, ICD codes are easy to manipulate and a ubiquitous component of EHR systems. However, data scientists should be aware of potential biases. Despite concerted efforts to standardize ICD coding across health systems, financial incentives and clinician styles can also distort coding practices.[13] Moreover, ICDs are subject to semantic drift, whereby the meaning of a code changes over time.[14] The transition from ICD-9 to ICD-10 in 2015 led to a heightened awareness of this problem, and several studies demonstrated changes in frequencies of some diagnoses during that period.[15,16] But semantic drift can happen for more subtle reasons, including minor revisions to the ICD coding structure or changes in local EHR tools used to translate unstructured-text to ICDs.[17] Figure 1 provides a case in point whereby the introduction of a new ICD code disrupted the apparent frequency of sepsis. These data anomalies can complicate the interpretation of codes over time or across institutions.

To deal with the inherent ambiguity of diagnoses in the EHR, supporting information can be used to improve accuracy. Researchers can use information **redundancy** to refine a phenotype. For example, for phenotypes based on diagnostic codes, requiring more than one code on different dates can improve accuracy.[18] **Supporting evidence** from orthogonal data sources may also be useful. For example, a diabetes mellitus phenotype may be improved by incorporating objective laboratory parameters such as an elevated hemoglobin A1c. **Clinical interventions** can be used to refine a phenotype. For example, a case

Diagnosis	Original ICD9-CM Code	Revised ICD9-CM code
Cardiogenic Shock	785.51	785.51
Septicemic Shock	785.59	785.52
Hypovolemic Shock	785.59	785.59

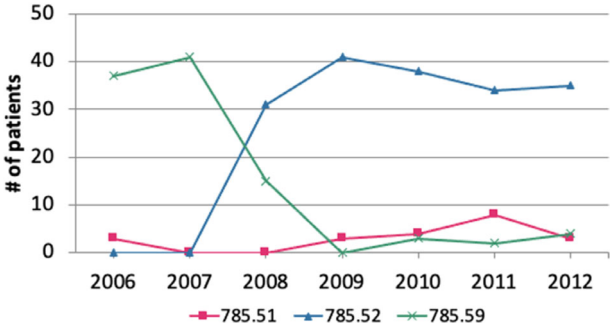

FIGURE 1 Semantic shift in International Classification of Diseases (ICD) code causes shift in prevalence. The National Center for Health Statistics (NCHS) and the Centers for Medicare and Medicaid Services (CMS) periodically adds new code and updates guidance on the use of existing ICD-CM codes, which can radically shift coding practices at an institution. Below is an example in the change of prevalence of codes relating to shock in a 200-bed hospital

for hypothyroidism might be defined as having both diagnostic code evidence and a prescription for levothyroxine.

The above strategies are not foolproof. For example, a patient with type 2 diabetes may have normal laboratory values because their disease is controlled with medication (eg, if blood glucose levels are well-controlled in a patient with diabetes the hemoglobin A1c may be normal).[19,20] Requiring the presence of treatment as a marker for disease can limit the cohort to treated patients. Furthermore, with off-label usage, the data scientist needs to be careful about assuming the appearance of a medication means the patient has the expected condition. Finally, the markers within the EHR that are used to increase certainty may also correlate with disease severity. This conflation of certainty with severity does not invalidate the phenotype but should be recognized in the interpretation of analyses that apply the phenotype.

For these reasons and others, even a carefully designed phenotype algorithm will inevitably generate false positives and negatives. Adjustments to phenotypes designed to improve the positive predictive value such as requiring multiple instances of a diagnosis or a supporting laboratory value often are made at the expense of reducing the negative predictive value, or vice versa. Figure 2 examines the overlap of different sources of evidence for a diagnosis at different sites, illustrating that they do not always agree. The degree of difference in overlap of phenotype components may vary from site to site suggesting that the accuracy of the phenotype may vary across institutions.

3 | LABORATORY RESULTS

Analytic data sets often require laboratory results. EHRs typically store this information in structured fields, so retrieving the data is relatively trivial. In contrast to the uncertainty often attached to diagnoses, laboratory measurements give the impression of relative objectivity and computability. However, laboratory results have their own idiosyncrasies.

3.1 | Repeat measures

Patients often have multiple measurements for a laboratory result. But analytical datasets often require a single value per person, as is the case where a measurement is used as a covariate in a multivariate regression. This leads to a common conundrum of using measurement variables: What is the best way to summarize multiple data points into a single value? Should the median value be used? The maximum? The earliest or most recently measured? How should the timing and cadence of laboratory values be addressed in the decision? The answer to this question has downstream consequences. A study of EHR laboratory results shows that the way the values are summarized affects the ability to replicate known genetic associations. For most laboratory results, the median value had the best performance, but some for laboratory results the maximum or the first performed the best.[21] Related issues exist with longitudinal analyses of multiple measurements where the focus is on change in laboratory parameters over time. While changes of a certain magnitude between two individual values may be easy to define, it is more difficult to assess a sustained change from baseline. As is the case for many EHR-phenotyping conundrums, the question of how to best work with repeated measurements does not have a simple answer, but rather depends on the specific laboratory result and intended use of the values.

3.2 | Context

Laboratory values can be misleading when they are interpreted without an understanding of their context. Some laboratory results, such as a comprehensive metabolic panel or blood count, are ordered routinely, while other tests are ordered in response to patient complaints or to follow up a previously abnormal result. Therefore, the presence of some laboratory results in a patient's record may increase the likelihood that they have a particular disease, even if the test result is normal. For

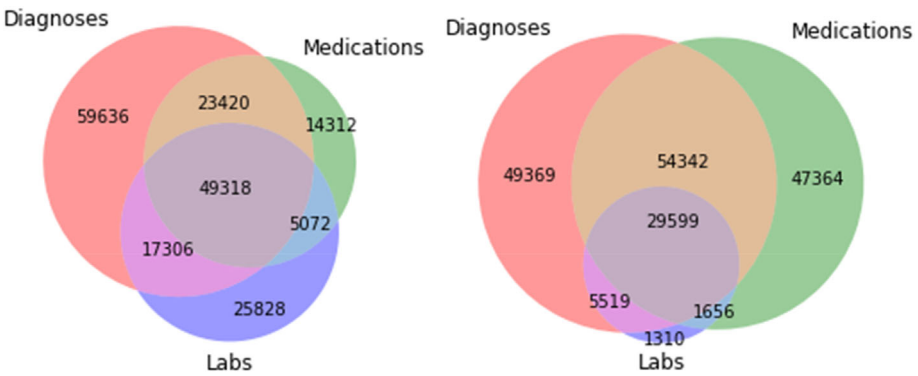

FIGURE 2 Venn diagrams of overlap of suggestive diagnoses, medications, and laboratory results for type 2 diabetes at two different institutions. The different ratios of overlap of data elements for the diabetes computable phenotype suggest algorithm behavior is different between the two sites. These site-level differences in the proportions of patients with different markers of disease may be accompanied by differences in other characteristics that may impact the performance of predictive algorithms developed at one site and applied at another

example, a patient with multiple test results for phenylalanine levels is more likely to have phenylketonuria, even if all the results are in normal range.

Laboratory results are ordered in a variety of clinical settings, including during ambulatory clinic visits, in emergency department visits, and inpatient stays. Laboratory results that are ordered in the inpatient and emergency setting often have different average values compared with those ordered in outpatient clinics. Figure 3 shows systematic differences in laboratory results from the inpatient and outpatient context. Not surprisingly, inpatients are more likely to have elevated laboratory results that indicate acute and serious conditions (eg, troponin) and inflammation (c-reactive protein) compared with those in the outpatient setting. Thus, a liver enzyme measurement in the inpatient setting may not accurately reflect a patient's baseline levels.

3.3 | Secular trends

Changing clinical guidelines and practice styles may alter the use of tests in certain clinical scenarios. Since clinical guidelines change over time, data scientists need to be mindful of the analytical dangers of pooling together laboratory results that span long periods. Trends in the ordering of specific lab results can be observed in longitudinal datasets. For example, Figure 4 shows a precipitous increase of vitamin D testing (Panel A) in the mid-2000s—a trend that has been observed in other datasets[22]—in the context of an increasing interest in the health benefits of vitamin D among researchers and clinicians.[23] In comparison, LDL-C testing has remained fairly stable over the past twenty years (Panel B), though both tests showed a sharp decrease at the onset of the severe acute respiratory syndrome coronavirus 2 pandemic. While the precise reasons for these trends are not always easy to ascertain, data scientists should be aware that such trends are common in longitudinal datasets and thoughtful about the ways they can impact the meaning of the measurements.

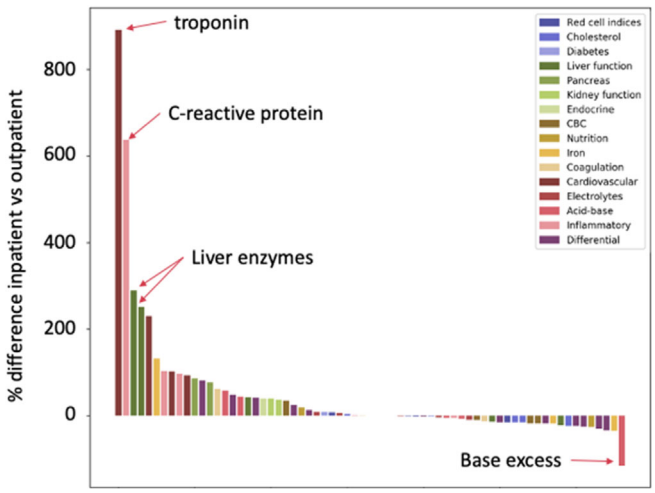

FIGURE 3 Differences in mean values of common laboratory results measured in the inpatient vs outpatient setting

3.4 | Laboratory result names

Finding a particular laboratory result in EHR data can be surprisingly difficult, as laboratory tests have a tremendous degree of variety. Not only are there many different analytes measured in clinical laboratory tests, but the results may be conducted on different types of specimens (eg, whole blood, cerebral spinal fluid) or using different methods. The vast variety of laboratory tests is reflected in the Logical Observation Identifiers Names and Codes (LOINC)[24]—a coding system commonly used for labeling lab measurements—which includes over 96 000 terms to measurements, observations, and document types (2.70).

While the granularity of LOINC terms is necessary to precisely label laboratory results, it can cause confusion for those looking to retrieve values for a particular result. For instance, if a dataset calls for a pH measurement, data scientists must choose from a long list of potentially relevant results, a selection of which are shown in Figure 5. The codes in blue clearly indicate venous blood. The codes in red clearly indicate arterial blood. However, for the codes in purple, it is not clear if the specimen source is arterial or venous. If laboratory results are labeled with these nonspecific codes, proper assessment of their true origin and clinical significance may be difficult.

LOINC terms can be used to label results unambiguously, but the large number of codes for individual tests can cause problems for interoperability. The use of LOINC codes is not uniform or consistently documented across sites or within sites over time, and different institutions may use different codes to label what is, for most purposes, the same laboratory result. If that happens, a query designed at one institution might fail to retrieve all relevant laboratory results a different institution. If the query is part of a larger, complex phenotype, the missing data may go unnoticed.

4 | VITAL SIGNS

Vital signs (temperature, heart rate, blood pressure, respiratory rate, height, weight) are typically collected at ambulatory encounters, which can be regularly scheduled or sporadic—sometimes prompted by specific patient concerns, and other times because the patient presented for a routine checkup. Inpatient vital signs are typically, but not necessarily, ordered on a schedule related to severity of illness—one a day, once a shift, or many more times per day. Vital signs are sometimes obtained by manual methods, and sometimes measured through automated devices with a much greater frequency, including continuous telemetry.

Some of the challenges associated with using vital signs are similar to those of laboratory results, including issues of multiple measurements, context dependent interpretation, incorrect data transcription, and terminology mapping. But vital signs distinguish themselves in terms of the density of the data readout. Some vital signs, like respiratory rate and temperature, can be measured on a near continuous basis during a hospital stay. Vital sign values, particularly with automated measures, can frequently be associated with a great deal of noise, so summary measures that represent the maximum or minimum values may be affected by sporadic, erroneously high or low values. Given the

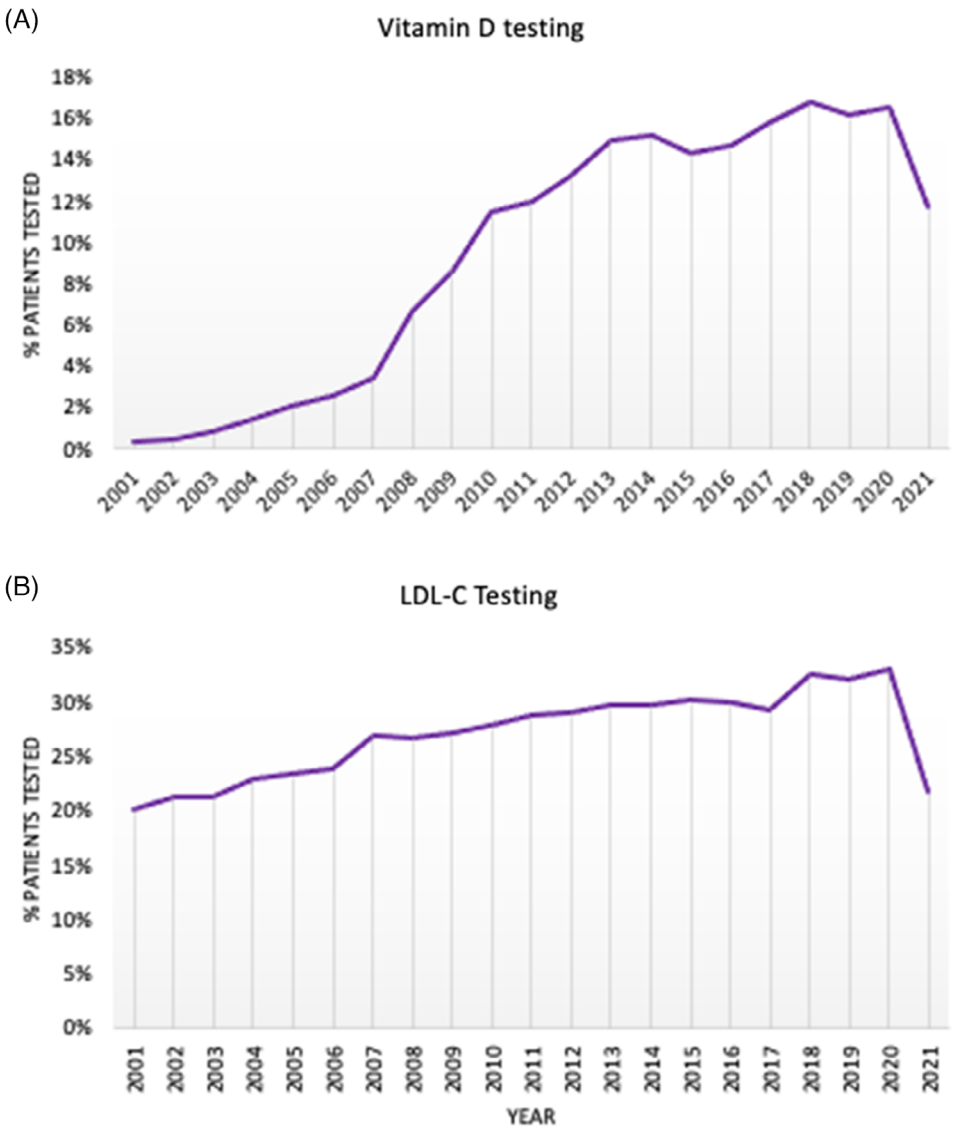

FIGURE 4 Trends in laboratory test ordering over time. Panel A shows the percentage of individuals with a vitamin D measurement among all patients with at least one laboratory measurement in that year. Panel B shows the same for low-density lipoprotein (LDL) cholesterol

50980-2	pH of Mixed venous blood adjusted to patient's actual temperature
39486-6	pH of Venous blood adjusted to patient's actual temperature
2745-8	pH of Capillary blood
49701-6	pH of Blood adjusted to patient's actual temperature
2753-2	pH of Serum or Plasma
33254-4	pH of Arterial blood adjusted to patient's actual temperature
2744-1	pH of Arterial blood

FIGURE 5 A selection of LOINC terms used to specify laboratory results of pH

potential for large numbers of vital signs to be available in the source data, operational decisions are sometimes made to reduce the volume of data stored in data warehouses and recorded in EHRs. The methods by which the source data are filtered can differ to include only the highest, lowest, and median value for a given time interval. Some vital signs like heart rate and blood pressure are typically measured with reasonable precision while others, notably respiratory rates, are often estimated, resulting in unexpected uniformity in many recorded values.[25]

5 | MEDICATIONS

As with measurements, a prescription order for a medication in the clinical record may have the appearance of an unambiguous, objective fact that the provider intended for the patient to take the medication and the patient was administered the treatment or filled the prescription and adhered to the regimen. However, there is still a great deal of uncertainty associated with the medication, especially regarding whether the prescription was filled and the duration over which a medication was taken. It is well known that patients do not always fill prescriptions they are given or take the medications they are dispensed.[26] Methods have been developed to infer adherence based on fill and refill patterns and patterns of prescriptions written.[27] Research on refill adherence measurements have

shown that estimates are highly sensitive to slight changes in definitions.[28] Therefore, not all similar-appearing patterns in medication orders reflect a similar degree of exposure to a drug.

Medications may be discontinued because of side effects, ineffectiveness, cost, or a resolution of the problem they treat.[29] Not all discontinuations represent the intention to stop a medication, as prior iterations of prescriptions need to be discontinued when a renewal is written. The dose of a medication may change when the medication is renewed. Sometimes additional medications are added to the regimen with the intention of maintaining the earlier medication, though other times, the original medication is meant to be stopped. Clinicians should indicate these intended changes though actively continuing or discontinuing older medications, but this may not always happen.

Clinicians are supposed to reconcile the patient's medication list at clinic visits, removing medications that the patient is no longer taking, and adding in new medications. But this process is not always straightforward. Medications may not be removed in a timely manner, making it difficult to infer end dates based on the timing of when they are removed. Some providers are uncomfortable documenting the discontinuation of a medication stopped by the patient since it can imply an endorsement of the medication discontinuation. This scenario is more likely if the stopped medication is related to a clinical domain outside the current provider's area of expertise. The impact of this behavior is that a medication can appear to still be active among medication orders when the patient is no longer actively taking the drug. Analysis of dispensing data may better reflect medications the patient is actually taking, but the availability of dispensing data may be incomplete depending on where the medication was filled and the payment model. With an increasing number of medications available over the counter that previously required a prescription, the range of medications having a dispensed record is changing. For these reasons, generating a list of current medications based on apparently active prescriptions and known dispensing in EHR data is prone to some degree of error.

Once the data scientist has evidence (eg, via refill patterns) that a patient was on a medication, the next issue is how to represent the presence of that medication in the analytical data set. While standards like RxNorm allow encoding of medication information at a very granular level including the ingredient, dose, and form, one can also group medications into those having the same core ingredient or even group related drugs under the same drug class.[30] Combination products pose additional challenges. The data scientist and clinical expert need to collaborate to ensure a balance between creating features that group many drugs too-coarsely into a single category vs applying a very granular coding that distinguishes all medication variations but produces too many features to be well-analyzed.

6 | COMPUTABLE PHENOTYPES AND PHENOTYPING STRATEGIES

Researchers may take different approaches to EHR phenotyping, depending on the requirements of the project. A disease phenotype can often be ascertained rapidly using ICD billing codes along, defining cases as individual who have one or perhaps multiple disease codes. This process can be scaled to generate a phenome-wide snapshot of the patient population, which can be used in a variety of methods that necessitate the capture of hundreds or thousands of different disease labels.[31] Some projects require higher quality phenotypes than can be generate using ICD codes alone. In this case, researchers may develop computable phenotypes that combine various datatypes (eg, diagnosis codes, text mentions, medications) to increase the specificity and/or sensitivity of the phenotype.[32,33]

Ideally, a computable phenotype should be designed with input from both data scientists and clinical experts and within the context of a specific study. A clinical domain expert understands the diagnostic, laboratory, and pharmacological markers for a disease and the specific characteristics that distinguish similar conditions. The data scientist understands how these concepts are represented in the data and whether idiosyncratic data capture or formats need to be considered. The data scientist needs to iterate with the domain expert on the impact of including criteria that may increase the precision of the diagnosis where the presence of characteristics correlates highly with the presence of disease, at the cost of missing some patients who have the disease (high positive predictive value, but low sensitivity), and the sensitivity of characteristics where fewer patients are missed but not all patients with the characteristics really have the disease (high sensitivity, but low positive predictive value).

Having an evaluation process for a computable phenotype is critical. Because of the different perspectives between and among data scientists and clinicians, it is common to have disagreement about the best way to define a computable phenotype. One way to settle these disagreements is to test assumptions against real data. The evaluation process for a computable phenotype usually requires a gold standard of expert reviewed charts. Surrogate markers like genetics may also be used when available to assess phenotype quality.[34,35]

7 | RESEARCH NETWORKS AND COMMON DATA MODELS

Several existing real-world data networks have standardized data across institutions to support clinical research, comparative safety, comparative effectiveness, disease surveillance, and related research, including the Vaccine Safety Datalink, the Health Care Systems Research Network, FDA Sentinel, and PCORnet. Work from these networks demonstrate the feasibility of standardizing data across time and institutions and provide excellent guidance on how to harmonize data, detect problems, and generate quality measurements.[36] Studies conducted in these networks demonstrates that EHR data can enable reliable and clinically meaningful research.

Progress has been made in developing standards that facilitate cross-institutional analyses. The Observational Medical Outcomes Partnership (OMOP) is a common data model (CDM) that allows

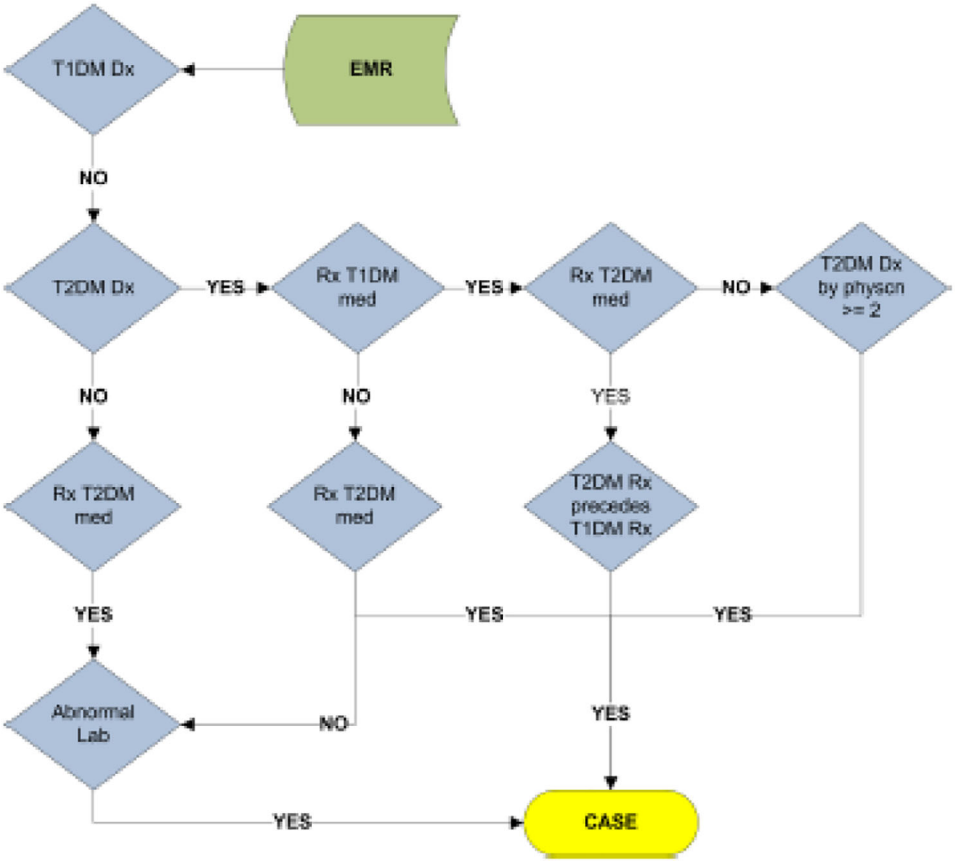

FIGURE 6 Flow chart for identifying type 2 diabetes in PheKB. Developing a computable phenotype for diabetes is illustrative of many of the issues highlighted in this article. Diabetes can be either type 1 or type 2. While type 1 diabetics always require insulin, type 2 diabetics sometimes require insulin. Some patients, especially those who receive insulin, may have accumulated evidence for the diagnosis of both type 1 and type 2 diabetes over time, so identifying the type of diabetes a patient has from diagnosis codes may be challenging. PheKB provides an algorithm for type 2 diabetes that excludes patients who have ever had a diagnosis of type 1 diabetes. That decision likely increases the positive predictive value of the phenotype but lowers the sensitivity. The flow diagram implies that the diagnosis of type 2 diabetes requires the combination of a type 2 diagnosis plus an abnormal laboratory test result or a type 2 diagnosis plus a suggestive medication or two diagnoses of type 2 diabetes. While this is a single definition, it allows for multiple paths for a diagnosis that could be differentially present at different sites. This definition is one of many that could be developed based on the specific data source and use case

institutions to convert their own EHR data, which are often stored in bespoke data structures, into a standardized format, and has been used successfully in multi-institutional EHR studies.[37] National initiatives like the eMerge program have developed standards for computable phenotypes that can be applied across institutions, an example of which is shown in Figure 6.[38]

With these initiatives, the informatics community has grown more adept at harmonizing EHR data. The National COVID Cohort Collaborative (N3C) is a good example of an initiative that used past innovations and knowledge to quickly build a broadly accessible cohort of over 5 million patients from disparate health systems.[39] But while CDMs and phenotypes standards greatly increase the efficiency of cross-institutional research, data scientists using these tools should still be mindful of the caveats described above while analyzing and interpreting results. Phenotypes designed for CDMs may be guaranteed to execute smoothly, but the quality of the phenotypes will vary based on the complexities of EHR data.[40]

8 | CONCLUSION

Developing high-quality clinical evidence from real-world clinical data requires that data scientists and clinicians collaborate, communicate, and iterate on both the development of the analytical data set and the conduct of the analysis. Data scientists in the clinical domain do not need formal clinical education, but to participate deeply in the research process as true collaborators, they require special training into the idiosyncrasies of clinical data and the impact on the interpretability of results.[19] Clinicians should be involved in the process of optimizing clinical phenotypes and need to work with data scientists to support data-driven decisions about appropriate granularity and timing of definitional components. Biostatisticians and epidemiologists also provide important insights into study design and addressing missing or noisy data.[41,42]

To support reproducibility of findings, more work is needed to develop reporting standards for the results of EHR analyses. An analysis of EHR data performed under one set of assumptions, even if well-

informed by expert opinion, may be spuriously correct or incorrect. Worse still, the investigative team may have performed an analysis under multiple assumptions and only reported the version that supports their hypothesis. Therefore, the traditional style of reporting results of analyses of real-world data in a manner similar to randomized controlled trials (RCTs), with a single point estimate and confidence interval for the association between an exposure and outcome, is suboptimal. This approach has led to conflicting literature[43-46] where it is difficult to understand the underlying cause of the different results and the implications for generalizability.[47] Sensitivity analyses should be conducted under different assumptions including handling of missing data, and different definitions of exposure and outcome, with explicit definitions of key variables, and the results of these analyses should be reported. In this way, the robustness of the findings will be more apparent, or if different results arise under different assumptions, these differences can be described sooner and contribute to a more balanced interpretation of the findings.

The practice of repurposing clinical data for research has attracted criticism when conducted without recognition of the idiosyncrasies of clinical practice that impact the recording and interpretation of data. However, there are many examples of well-conducted EHR-based studies that report important and reproducible findings. EHR-based research has earned a place in the research ecosystem. To be successful, researchers must always be mindful of the complexities of EHR data, many of which are described in this article, and remain vigilant for unexpected challenges that could comprise the science.

ACKNOWLEDGMENTS

The authors thank Jeffrey Goldstein (Northwestern Memorial Hospital) for his help creating Figure 3.

CONFLICT OF INTEREST

None of the authors have any conflicts of interest to report nor received any funding in support of the manuscript development.

ORCID

Jeffrey S. Brown https://orcid.org/0000-0002-9340-7189
James J. Cimino https://orcid.org/0000-0003-4101-1622
Philip R.O. Payne https://orcid.org/0000-0002-9532-2998
Mark G. Weiner https://orcid.org/0000-0001-5586-9940

REFERENCES

1. Embi PJ, Richesson R, Tenenbaum J, et al. Reimagining the research-practice relationship: policy recommendations for informatics-enabled evidence-generation across the US health system. *JAMIA Open*. 2019;2(1):2-9. https://doi.org/10.1093/jamiaopen/ooy056
2. Friedman CP, Wong AK, Blumenthal D. Achieving a nationwide learning health system. *Sci Transl Med*. 2010;2(57):57cm29. https://doi.org/10.1126/scitranslmed.3001456
3. Wagner J, Hall JD, Ross RL, et al. Implementing risk stratification in primary care: challenges and strategies. *J Am Board Fam Med*. 2019;32(4):585-595. https://doi.org/10.3122/jabfm.2019.04.180341
4. Kahn MG, Callahan TJ, Barnard J, et al. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. *EGEMS (Wash DC)*. 2016;4(1):1244. https://doi.org/10.13063/2327-9214.1244
5. Hersh WR, Weiner MG, Embi PJ, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. *Med Care*. 2013;51(8 Suppl 3):S30-S37. https://doi.org/10.1097/MLR.0b013e31829b1dbd
6. Cowie MR, Blomster JI, Curtis LH, et al. Electronic health records to facilitate clinical research. *Clin Res Cardiol*. 2017;106(1):1-9. https://doi.org/10.1007/s00392-016-1025-6
7. Boland MR, Alur-Gupta S, Levine L, Gabriel P, Gonzalez-Hernandez G. Disease associations depend on visit type: results from a visit-wide association study. *BioData min*. 2019;12:15. https://doi.org/10.1186/s13040-019-0203-2
8. Moore JH. Empowering the data science scientist. *BioData min*. 2021;14:8. https://doi.org/10.1186/s13040-021-00246-x
9. Lockshin MD, Barbhaiya M, Izmirly P, Buyon JP, Crow MK. SLE: reconciling heterogeneity. *Lupus Sci Med*. 2019;6(1):e000280. https://doi.org/10.1136/lupus-2018-000280
10. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. *J Am Med Inform Assoc*. 2011;18(5):544-551. https://doi.org/10.1136/amiajnl-2011-000464
11. Wang Y, Wang L, Rastegar-Mojarad M, et al. Clinical information extraction applications: a literature review. *J Biomed Inform*. 2018;77:34-49. https://doi.org/10.1016/j.jbi.2017.11.011
12. Koleck TA, Dreisbach C, Bourne PE, Bakken S. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. *J Am Med Inform Assoc*. 2019;26(4):364-379. https://doi.org/10.1093/jamia/ocy173
13. Silverman E, Skinner J. Medicare upcoding and hospital ownership. *J Health Econ*. 2004;23(2):369-389. https://doi.org/10.1016/j.jhealeco.2003.09.007
14. Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. *Methods Inf Med*. 1998;37(4–5):394-403.
15. Khera R, Dorsey KB, Krumholz HM. Transition to the ICD-10 in the United States: an emerging data chasm. *JAMA*. 2018;320(2):133-134. https://doi.org/10.1001/jama.2018.6823
16. Panozzo CA, Woodworth TS, Welch EC, et al. Early impact of the ICD-10-CM transition on selected health outcomes in 13 electronic health care databases in the United States. *Pharmacoepidemiol Drug Saf*. 2018;27(8):839-847. https://doi.org/10.1002/pds.4563
17. Stewart CC, Lu CY, Yoon TK, et al. Impact of ICD-10-CM transition on mental health diagnoses recording. *EGEMS (Wash DC)*. 2019;7(1):14. https://doi.org/10.5334/egems.281
18. Verma A, Ritchie MD. Current scope and challenges in Phenome-wide association studies. *Curr Epidemiol Rep*. 2017;4(4):321-329. https://doi.org/10.1007/s40471-017-0127-7
19. Wiese AD, Roumie CL, Buse JB, et al. Performance of a computable phenotype for identification of patients with diabetes within PCORnet: the patient-centered clinical research network. *Pharmacoepidemiol Drug Saf*. 2019;28(5):632-639. https://doi.org/10.1002/pds.4718
20. 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. https://doi.org/10.1136/amiajnl-2013-001952
21. Goldstein JA, Weinstock JS, Bastarache LA, et al. LabWAS: novel findings and study design recommendations from a meta-analysis of clinical labs in two independent biobanks. *PLoS Genet*. 2020;16(11):e1009077. https://doi.org/10.1371/journal.pgen.1009077
22. Basatemur E, Horsfall L, Marston L, Rait G, Sutcliffe A. Trends in the diagnosis of vitamin D deficiency. *Pediatrics*. 2017;139(3):e20162748. https://doi.org/10.1542/peds.2016-2748
23. Rooney MR, Harnack L, Michos ED, Ogilvie RP, Sempos CT, Lutsey PL. Trends in use of high dose vitamin D supplements

exceeding 1,000 or 4,000 international units daily, 1999-2014. *JAMA*. 2017;317(23):2448-2450. https://doi.org/10.1001/jama.2017.4392
24. McDonald CJ, Huff SM, Suico JG, et al. LOINC, a universal standard for identifying laboratory observations: a 5-year update. *Clin Chem*. 2003;49(4):624-633. https://doi.org/10.1373/49.4.624
25. Badawy J, Nguyen OK, Clark C, Halm EA, Makam AN. Is everyone really breathing 20 times a minute? Assessing epidemiology and variation in recorded respiratory rate in hospitalised adults. *BMJ Qual Saf*. 2017;26(10):832-836. https://doi.org/10.1136/bmjqs-2017-006671
26. Lam WY, Fresco P. Medication adherence measures: an overview. *Biomed Res Int*. 2015;2015:217047. https://doi.org/10.1155/2015/217047
27. Sattler ELP, Lee JS, Perri M. Medication (re)fill adherence measures derived from pharmacy claims data in older Americans: a review of the literature. *Drugs Aging*. 2013;30(6):383-399. https://doi.org/10.1007/s40266-013-0074-z
28. Galozy A, Nowaczyk S, Sant'Anna A, Ohlsson M, Lingman M. Pitfalls of medication adherence approximation through EHR and pharmacy records: definitions, data and computation. *Int J Med Inform*. 2020;136:104092. https://doi.org/10.1016/j.ijmedinf.2020.104092
29. Roborel de Climens A, Pain E, Boss A, Shaunik A. Understanding reasons for treatment discontinuation, attitudes and education needs among people who discontinue type 2 diabetes treatment: results from an online patient survey in the USA and UK. *Diabetes Ther*. 2020;11(8):1873-1881. https://doi.org/10.1007/s13300-020-00843-9
30. Bodenreider O, Rodriguez LM. Analyzing U.S. prescription lists with RxNorm and the ATC/DDD index. *AMIA Annu Symp Proc*. 2014;2014:297-306.
31. Bastarache L. Using Phecodes for research with the electronic health record: from PheWAS to PheRS. *Annu Rev Biomed Data Sci*. 2021;4:1-19. https://doi.org/10.1146/annurev-biodatasci-122320-112352
32. Richesson RL, Smerek MM, Blake CC. A framework to support the sharing and reuse of computable phenotype definitions across health care delivery and clinical research applications. *EGEMS (Wash DC)*. 2016;4(3):1232. https://doi.org/10.13063/2327-9214.1232
33. McDonough CW, Babcock K, Chucri K, et al. Optimizing identification of resistant hypertension: computable phenotype development and validation. *Pharmacoepidemiol Drug Saf*. 2020;29(11):1393-1401. https://doi.org/10.1002/pds.5095
34. Denny JC, Bastarache L, Ritchie MD, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol*. 2013;31(12):1102-1111. https://doi.org/10.1038/nbt.2749
35. Fabbri C, Hagenaars SP, John C, et al. Genetic and clinical characteristics of treatment-resistant depression using primary care records in two UKcohorts. *Mol Psychiatry*. 2021;22. https://doi.org/10.1038/s41380-021-01062-9
36. Raebel MA, Haynes K, Woodworth TS, et al. Electronic clinical laboratory test results data tables: lessons from mini-sentinel. *Pharmacoepidemiol Drug Saf*. 2014;23(6):609-618. https://doi.org/10.1002/pds.3580
37. Papez V, Moinat M, Payralbe S, et al. Transforming and evaluating electronic health record disease phenotyping algorithms using the OMOP common data model: a case study in heart failure. *JAMIA Open*. 2021;4(3):ooab001. https://doi.org/10.1093/jamiaopen/ooab001
38. Shang N, Liu C, Rasmussen LV, et al. Making work visible for electronic phenotype implementation: lessons learned from the eMERGE network. *J Biomed Inform*. 2019;99:103293. https://doi.org/10.1016/j.jbi.2019.103293
39. Haendel MA, Chute CG, Bennett TD, et al. The national COVID cohort collaborative (N3C): rationale, design, infrastructure, and deployment. *J Am Med Inform Assoc*. 2021;28(3):427-443. https://doi.org/10.1093/jamia/ocaa196
40. Brown JS, Bastarache L, Weiner MG. Aggregating electronic health record data for COVID-19 research-caveat emptor. *JAMA Netw Open*. 2021;4(7):e2117175. https://doi.org/10.1001/jamanetworkopen.2021.17175
41. Donders ART, van der Heijden GJMG, Stijnen T, Moons KGM. Review: a gentle introduction to imputation of missing values. *J Clin Epidemiol*. 2006;59(10):1087-1091. https://doi.org/10.1016/j.jclinepi.2006.01.014
42. Lesko CR, Jacobson LP, Althoff KN, et al. Collaborative, pooled and harmonized study designs for epidemiologic research: challenges and opportunities. *Int J Epidemiol*. 2018;47(2):654-668. https://doi.org/10.1093/ije/dyx283
43. Etminan M, Forooghian F, Brophy JM, Bird ST, Maberley D. Oral fluoroquinolones and the risk of retinal detachment. *JAMA*. 2012;307(13):1414-1419. https://doi.org/10.1001/jama.2012.383
44. Pasternak B, Svanström H, Melbye M, Hviid A. Association between oral fluoroquinolone use and retinal detachment. *JAMA*. 2013;310(20):2184-2190. https://doi.org/10.1001/jama.2013.280500
45. Azoulay L, Yin H, Filion KB, et al. The use of pioglitazone and the risk of bladder cancer in people with type 2 diabetes: nested case-control study. *BMJ*. 2012;344:e3645. https://doi.org/10.1136/bmj.e3645
46. Wei L, MacDonald TM, Mackenzie IS. Pioglitazone and bladder cancer: a propensity score matched cohort study. *Br J Clin Pharmacol*. 2013;75(1):254-259. https://doi.org/10.1111/j.1365-2125.2012.04325.x
47. Teng AK, Wilcox AB. A review of predictive analytics solutions for sepsis patients. *Appl Clin Inform*. 2020;11(3):387-398. https://doi.org/10.1055/s-0040-1710525

How to cite this article: Bastarache L, Brown JS, Cimino JJ, et al. Developing real-world evidence from real-world data: Transforming raw data into analytical datasets. *Learn Health Sys*. 2022;6(1):e10293. doi:10.1002/lrh2.10293