Journal of the American Medical Informatics Association, 30(2), 2023, 213–221
https://doi.org/10.1093/jamia/ocac157
Advance Access Publication Date: 7 September 2022
Research and Applications

Research and Applications

Not all phenotypes are created equal: covariates of success in e-phenotype specification

Bashir Hamidi [1], Patrick A. Flume[2], Kit N. Simpson[3], and Alexander V. Alekseyenko [1,3,4,5]

[1]Biomedical Informatics Center, Medical University of South Carolina, Charleston, South Carolina 29425, USA, [2]Department of Medicine, Medical University of South Carolina, Charleston, South Carolina 29425, USA, [3]Department of Healthcare Leadership and Management, Medical University of South Carolina, Charleston, South Carolina 29425, USA, [4]Department of Public Health Sciences, Medical University of South Carolina, Charleston, South Carolina, 29425, USA and [5]Department of Oral Health Sciences, Medical University of South Carolina, Charleston, South Carolina 29425, USA

Corresponding Author: Alexander V. Alekseyenko, PhD, 22 WestEdge St, Rm WG213, MSC 200, Charleston, SC 29403, USA; alekseye@musc.edu

Received 11 March 2022; Revised 31 July 2022; Editorial Decision 20 August 2022; Accepted 22 August 2022

ABSTRACT

Background: Electronic (e)-phenotype specification by noninformaticist investigators remains a challenge. Although validation of each patient returned by e-phenotype could ensure accuracy of cohort representation, this approach is not practical. Understanding the factors leading to successful e-phenotype specification may reveal generalizable strategies leading to better results.

Materials and Methods: Noninformaticist experts ($n = 21$) were recruited to produce expert-mediated e-phenotypes using i2b2 assisted by a honest data-broker and a project coordinator. Patient- and visit-sets were reidentified and a random sample of 20 charts matching each e-phenotype was returned to experts for chart-validation. Attributes of the queries and expert characteristics were captured and related to chart-validation rates using generalized linear regression models.

Results: E-phenotype validation rates varied according to experts' domains and query characteristics (mean = 61%, range 20–100%). Clinical domains that performed better included infectious, rheumatic, neonatal, and cancers, whereas other domains performed worse (psychiatric, GI, skin, and pulmonary). Match-rate was negatively impacted when specification of temporal constraints was required. In general, the increase in e-phenotype specificity contributed positively to match-rate.

Discussions and Conclusions: Clinical experts and informaticists experience a variety of challenges when building e-phenotypes, including the inability to differentiate clinical events from patient characteristics or appropriately configure temporal constraints; a lack of access to available and quality data; and difficulty in specifying routes of medication administration. Biomedical query mediation by informaticists and honest data-brokers in designing e-phenotypes cannot be overstated. Although tools such as i2b2 may be widely available to noninformaticists, successful utilization depends not on users' confidence, but rather on creating highly specific e-phenotypes.

Key words: electronic phenotyping, phenotyped data, validation, translational research services, electronic health record

BACKGROUND AND SIGNIFICANCE

Computable phenotypes using algorithms to specify patient cohorts with desired disease and condition characteristics are advancing efforts in quality improvement, comparative effectiveness research, and clinical decision support.[1,2] Although computable phenotypes are portable across sites and require little human intervention to be implemented, many investigators require the use of custom and specialized noncomputable phenotypes to meet their needs.[3,4] Creating e-phenotypes that represent the desired patient cohorts facilitates extraction of relevant data from the electronic health record (EHR), a task that requires precision and accuracy. Misspecification of the e-phenotype, such as case contamination and misclassification, can have adverse consequences, such as impaired decision-making based on erroneous estimates of the prevalence and burden of a disease on the health system.[5–7]

Promoting the use of phenotyping among noninformaticists requires examination of the factors that enhance the ability of translational researchers to specify phenotypes correctly. Investigators may not be aware of the challenges in mapping clinical data from EHRs to represent the desired phenotypes. This leads to a potential mismatch between the expectations of the investigator and the capabilities and the limitations of the e-phenotypes. Since thorough phenotype development training may not be feasible in all scenarios, the development of effective queries may require involvement of a data scientist or a database analyst. The collaborative communication process among clinical domain experts and database analysts has been previously described as "biomedical query mediation."[8,9]

Informatics for Integrating Biology and Bedside (i2b2) software platform allows investigators to access medical records to query cohorts of patients with specific e-phenotypes for research purposes.[10] Self-service phenotyping tools, such as i2b2, allow noninformaticists to accelerate their access to clinical data. For example, translational researchers may wish to capture phenotyped clinical specimens and specimen assay data using systems such as the Living BioBank[11] and Crimson.[12] Such systems rely on creation of e-phenotypes that can accurately represent patient cohorts. The utilization of i2b2 is prolific in some institutions.[13,14] One study investigated nearly 7000 i2b2 queries across 3 years and presented findings on the overall complexity and use of data domains.[13] The researchers discovered that over 70% of those queries were 'basic' (ie, utilizing 3 or fewer conditions and no temporal constraints). Additionally, the investigators discovered that the data domains that required the most effort to implement were used the least. The most commonly used domains were diagnoses (76.5% of all queries) followed by medications and demographics (24.3% and 23.9%, respectively). However, the study did not include a method for estimating the phenotype to patient match-rate, which could provide an objective measure of successful phenotype specification.

In this paper, we collaborated with noninformaticists and honest data-brokers to understand the factors contributing to success of phenotyping. We adopted an iterative process to phenotyping. The intent and a draft of the phenotype were initially provided by the noninformaticist. The phenotype was then finalized with the help of an honest data-broker familiar with i2b2 and the available data. The i2b2 query patient- and visit-sets were extracted and reidentified. A random sample of charts was used by the noninformaticist for validation of e-phenotype. This design allowed us to track the expert and the query characteristics and estimate the phenotype match-rate. The resulting data allowed us to analyze the factors that contribute to successful specification.

METHODS

Setting

Medical University of South Carolina (MUSC) Health Charleston has over 700 physicians serving more than 1 million patients annually at more than 100 clinics and outreach locations. MUSC Health Charleston is home to South Carolina's only solid organ transplant center and designated cancer center, a Level I Trauma Center, as well as one of only 2 National Telehealth Centers of Excellence. In addition to clinical care, MUSC serves as the medical education home for 6 colleges, trains more than 3000 students and more than 850 residents and fellows annually, and houses a Clinical and Translational Science Award hub and related core facilities.

Project design

MUSC's Institutional Review Board (IRB) designated this project as program evaluation because it did not meet the federal definition of human subjects research. The steps of the simulation project are summarized in Figure 1. Recruitment consisted of snowball sampling of clinical domain experts across MUSC. The experts were invited to an interview, which was also attended by a project coordinator and an honest data-broker with expertise in i2b2. To gauge the experts' level of expertise and their expectations for using i2b2, an assessment was administered at the interview. Likewise, the rationale and the intent behind the experts' phenotype was elicited in an unstructured way. Following the interview, each query was refined using input from the honest data-broker and from the experts. The final query was used to generate a patient- and visit-set (further details are shared in Methods, Cohort generation).

A random sample of 20 charts per e-phenotype of interest was designated for the expert to review and validate using a REDCap survey.[15,16] Medical record numbers (MRNs), Visit ID, and Admit Date were shared with the experts. The experts had unrestricted access to each patient's EHR in order to locate the exact chart, Visit ID, and date and to assess the accuracy of the generated e-phenotype and its relationship to the patient cohort. Each expert was asked to rate the respective sample of charts as a 'Match' to the intended e-phenotype, or a 'Mismatch', or 'Unsure'. The reason for the latter 2 choices was recorded as optional free text. Lastly, an assessment evaluating various aspects of the project including clinical experts' difficulty in using i2b2 and trust in the system was administered.

Data source

This was a retrospective analysis of patient records with at least 1 EHR documented clinical encounter over a span of 4 years (July 1, 2014 to July 1, 2018) at MUSC Health Charleston, SC.

Research data warehouse and the i2b2 system

The MUSC research data warehouse (RDW) is a repository of clinical data obtained from the MUSC EHR Epic (Epic Systems Corporation, Verona, WI, USA); laboratory results; diagnostic codes; clinic, discharge, and radiology text notes; research permissions; and the Hollings Cancer Center registry through linkage with the registry's clinical data source in Epic. Appropriate IRB approval and data governance oversight are typically required for data requests.[17] The RDW is updated daily to support longitudinal research and contains data from over 1.5 million patients comprising more than 17 million visits. The MUSC i2b2 (i2b2 Star schema) exposes a deidentified subset of the RDW data through and makes available to

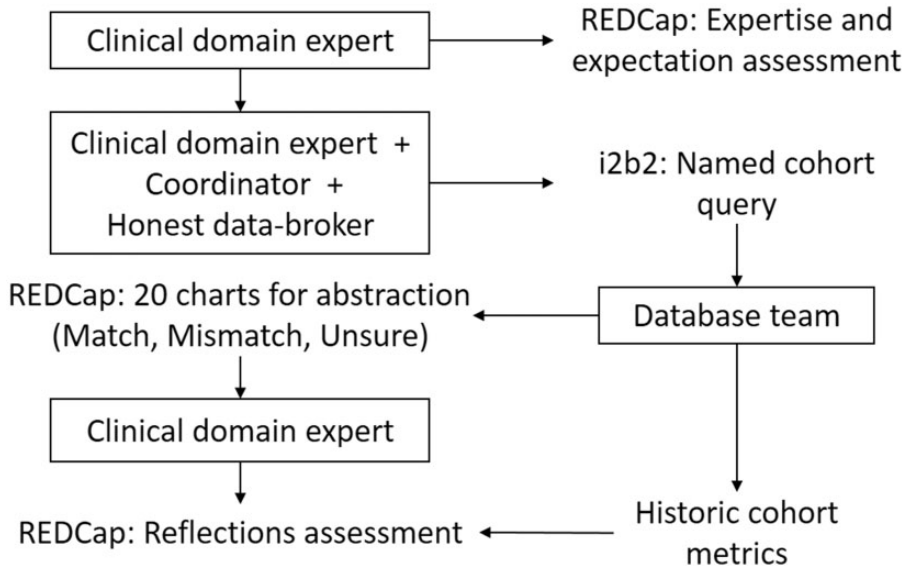

Figure 1. Simulation study outline shows the recruitment of noninformaticist clinical domain experts, drafting of e-phenotypes within i2b2, reidentification of charts matching e-phenotypes, chart-validation performed by and reflections assessment administered to the clinical domain experts.

investigators to query for cohort identification, patient counts, and feasibility studies.

Cohort generation

The honest data-broker provided varying degrees of query mediation depending on the familiarity of the clinical expert with i2b2. The honest data-broker worked with the clinical experts to define the intended patient cohort within i2b2. The honest data-broker has extensive knowledge of clinical data available at the institution and RDW, as well as additional expertise with the functions and capabilities of i2b2. This crucial expertise during the process of cohort generation ensured that the intended patient cohort of the expert was clearly understood and identified a method to query the cohort of interest within i2b2. In a few cases, the process of cohort generation consisted of an ongoing conversation among the study team, the honest data-broker, and the clinical expert. In some cases, a preliminary i2b2 query was generated and assessed for accuracy by generating a patient count within a given window of time or by sharing with the clinical expert a random sample of a patient who matched the query. Because they were familiar with the cohort, the experts could provide feedback on the expected accuracy of the patient count and/or the sampled subject. Through this collaborative process, the query was modified and finalized.

Once the i2b2 queries were generated, the deidentified patient- and visit-set were reidentified. A full list of patient visits matching the e-phenotype of interest was also generated to draw a random sample of 20 patient MRNs with visit identifiers for the expert chart review and validation.

Evaluation of i2b2 phenotype query attributes

To quantify the attributes of the generated i2b2 queries, we utilized several metrics.[13] First, we quantified queries according to the number of groups, logical clauses joined by AND operator in the i2b2 query. Second, we quantified queries on the number of clinical data domains used to generate the phenotype. Third, we assessed the queries on the temporal attributes within the queries.

Queries were assessed for their most granular structural attributes by enumerating the number of groups utilized within the query. These structural attributes are derived from various information sources including demographics, diagnoses, problem lists, computable phenotypes, procedures, medications, labs, vitals, imaging, admission date, clinical department, and length of stay. Each of these information sources are referred to as top-level data domains; the number of top-level domains is recorded for each query. As an example, consider the phenotype "patients with an HIV diagnosis and CD4 count of greater than 200." This phenotype is captured using 2 groups (HIV and CD4 counts) from 2 top-level domains (diagnosis and labs). Sometimes a phenotype may need additional groups. For example, "patients with an HIV diagnosis and stroke, and with a CD4 count of greater than 200." In this case, 3 groups (HIV, stroke, and CD4 counts) were utilized from 2 top-level domains (diagnosis and labs).

Simpler phenotypes will often have only baseline patient characteristics that define the overall population. However, in more complex phenotypes, temporal constraints relate characteristics that change across visits or occur in sequence. For example, we may wish to capture "patients who experienced a traumatic event and subsequently developed depression or PTSD." In this example, our patient population is the general patient population, and our Event 1 is defined as "those without any PTSD ICD codes and with an Emergency Room visit" in the same encounter followed by Event 2 that is defined as "those with any PTSD ICD codes." Another important point is the use of proxy variables. In this example, we have no direct and easily accessible information on whether a patient experienced a traumatic event. However, given that traumatic events are often followed by a visit to the Emergency Room, we utilized that information as a proxy for the patient's experience.

Statistical analysis

We used binomial-family generalized linear models to analyze validation match-rate with respect to confidence of the experts in using i2b2 and query attributes. The Likert-scale factors we expressed as trends in contrasts. All statistical analysis were performed in programming language R (version 3.6.1).[18]

RESULTS

Expert cohort and phenotype characteristics

We identified 21 individuals from the College of Medicine, College of Dental Medicine, and College of Nursing, to serve as clinical domain experts for the simulation project of 21 respective e-phenotypes and cohorts (Table 1). Two e-phenotypes (cancer 20, neurologic 21) were excluded from subsequent data analysis because the chart review for those designated phenotypes was not completed by the clinical experts. The i2b2 e-phenotypes were characterized using temporal constraints (independent, same encounter, and event specified), number of groups, and number of top-level domains in i2b2 ontology. A temporal constraint of 'independent' indicates that all groups are treated independently ($n = 12$), 'same encounter' indicates that select groups must occur at the same financial encounter ($n = 6$), and 'events specified' indicates a temporal query with 2 or more ordered events specified ($n = 3$). Most e-phenotypes were created using 3 groups (median & mode = 3, range = 2–8) and consisted of 3 top-level domains (median & mode = 3, range = 1–7). The 3 domains used most often were diagnostic criteria (eg, ICD10 codes) ($n = 18$), demographics ($n = 12$), and problem lists ($n = 9$). A detailed summary of phenotype attributes is provided in Supplementary Table S1.

Patient counts from i2b2 and RDW show high concordance

The initial patient counts for e-phenotypes were obtained using i2b2. These counts may be discordant with queries against the RDW because of challenges in data mapping as well as the preconfigured noise introduced by i2b2. The RDW contains authoritative data that are used in matching of patients and encounters to respective phenotypes. We observed that the simulated i2b2 queries resulted in concordant patient counts to those obtained from RDW intraclass correlation coefficient ($\rho = 0.998,\ p = 2.46e-24$) (Supplementary Figure S1 and Table S2). The largest relative discrepancies (>20% relative difference) were observed in phenotypes with fewer cases, many groups, and top-level domains (eg, e-phenotype infectious 2, rheumatic 7, neonatal 18).

Phenotype match-rates are related to expert confidence in using i2b2

The chart match-rate to the phenotype varied dramatically (Table 1) with certain phenotypes capturing the desired cohort better than others. On average, infectious, rheumatic, neonatal, and cancer e-phenotypes performed better whereas psychiatric, gastrointestinal, skin, and pulmonary e-phenotypes performed worse at capturing the intended cohort of the clinical experts. The clinical domains with a higher match-rate tended to be inpatient focused, which collect more data in the EHR, while the outpatient phenotypes had inferior match-rates.

The expertise and expectations assessment gauged the experts' level of confidence in specifying an e-phenotype with i2b2. Using a 4-point Likert scale, only 19% of experts were "not at all confident," and 14% were "highly confident." Based on the unstructured interviews, most respondents conflated their ability to use the i2b2 system with their expectations about the quality of the data and their confidence in the project coordinator and honest broker. Interestingly, the mean match-rate was proportional to the expert confidence level (Figure 2 and Table 2).

Changes in clinical characteristics are challenging to capture

The i2b2 system offers 3 options to specify the temporal relationships between patient attributes: (1) to "treat all groups independently"; (2) to specify "selected groups occur in the same financial encounter"; and (3) to "define sequence of events." We assessed the utilization of each of these features and their relationship to chart match-rate (Table 2). We observed that phenotypes that were specified using a sequence of events ($n = 3$) have a lower match-rate as compared to those specified as independent groups ($n = 8$) (linear coefficient=−0.68, $P = .049$). The complexity introduced by using events to specify temporal changes of patient characteristics was challenging to specify and resulted in a high chart mismatch-rate, as indicated by the data.

Expert confidence is related to attributes of the phenotypes they specify

We evaluated statistical associations between expert confidence and query attributes. The use of temporal phenotypes showed a trend toward significance (Fisher exact test, $P = .077$). Further, the number of i2b2 groups and the number of top-level domains used for phenotype specification positively correlated with expert confidence, with a Spearman correlation of 0.52 ($P = .02$) and 0.59 ($P = .007$), respectively. These results indicate that the attributes of the phenotype are at least partially related to the level of confidence the experts expressed toward i2b2 as a phenotyping platform.

Phenotype attributes and expert confidence contribute to match-rate

The attributes of i2b2 e-phenotypes (eg, temporal constraint, number of groups, and number of domains; see Supplementary Figure S2 for overview of i2b2 query attributes) were evaluated for their association with the match-rate (Table 2). As described above, the complexity introduced by the specification of temporal clinical events in an i2b2 query resulted in a decrease in the validated chart march-rate. The increase in specification of the phenotypes assessed by the number of groups and the number of i2b2 top-level domains individually correspond to an increased match-rate, with linear coefficients of 0.349 and 0.254, respectively; clinical expert confidence in i2b2 system negatively affects the match-rate (linear coefficient=−0.77, $P = .005$).

Match-rate relates to expert trust in the data and recommendation of the system

We evaluated the statistical associations between match-rate and experts' reflections including trust in the data and endorsement of the i2b2 system (Table 3). Following chart review, the experts' outlook on future use of i2b2 was significantly associated with the match-rate (linear coefficient 0.95, $P = .006$). Similarly, experts' trust in data derived from the system and recommendation of the system to their colleagues was significantly associated with match-rate (coefficients = 0.55 [$P = .016$] and 0.73 [$P = .005$] respectively). On the other hand, difficulty in using the system was not associated with the match-rate (coefficient=−0.22, $P = .528$).

DISCUSSION

Viewing phenotyping through the prism of the fundamental theorem of biomedical informatics,[19] we propose that the tools are not sufficiently refined and still require involvement of another human in

Table 1. Characteristics of the clinical experts are displayed with the e-phenotype attributes and chart review validation outcomes

				Phenotype attribute complexity			Phenotype validation outcomes	
Clinical domain	E-phenotype #	Expert Seniority (postdoc, early, mid, late)	E-phenotype	Temporal constraints[a]	Number of groups	Number of top-level domains	Match (%)	Unsure (%)
Infectious	1	Mid	HIV infected stable on antiretrovirals	Independent	4	6	100	0
	2	Early	Chronic mycobacterial infections in older adults	Independent	8	7	95	0
	3	Mid	Suspected community-acquired pneumonia	Same encounter	7	6	70	10
	4	Mid	HIV infected with heart diseases or stroke	Independent	2	3	45	25
Rheumatic	5	Mid	Systemic sclerosis with lung involvement without hypertension	Same encounter	3	1	100	0
	6	Late	Nondrug-induced systemic lupus erythematosus	Independent	3	3	80	10
	7	Mid	Nondrug-induced systemic lupus erythematosus with hospitalizations	Independent	4	4	90	0
Psychiatric	8	Late	Uncontrolled diabetes with depression	Events specified	3	3	60	5
	9	Postdoc	Trauma-induced depression or PTSD	Events specified	4	3	30	5
	10	Mid	Working age opioid use disorder with no chronic conditions	Independent	3	3	25	35
GI	11	Mid	Eosinophilic esophagitis with dysphasia in children	Same encounter	3	3	65	10
	12	Postdoc	Alcoholic cirrhosis with ascites	Same encounter	2	2	55	20
	13	Postdoc	Nonalcoholic steatohepatitis	Same encounter	3	2	25	0
Skin	14	Late	Chronic wounds of lower leg	Same encounter	4	3	75	5
	15	Early	Adolescents with acne on antimicrobials	Independent	3	3	20	0
Pulmonary	16	Late	Adults with cystic fibrosis	Independent	4	5	35	0
	17	Late	Adults with indications for bronchoalveolar lavage	Independent	3	3	20	0
Neonatal	18	Late	Vitamin D deficiency in full term pregnancy	Independent	5	4	95	5
Cancer	19	Early	Cancers treated at MUSC Hollings Cancer Center	Independent	2	2	75	15
	20	Early	Graft-vs-host disease following bone marrow transplant	Events specified	6	5	–	–
Neurologic	21	Mid	Aphasia without seizures	Independent	4	4	–	–

GI: gastrointestinal; MUSC: Medical University of South Carolina: PTSD: post-traumatic stress disorder.

[a]'Independent' indicates that all groups were treated independently and the items can occur at any time in patient's history; 'same encounter' indicates that select groups had to occur during the same visit (financial encounter); 'events specified' indicates a temporal query with 2 or more ordered events specified.

order to mediate the queries. Noninformaticists must overcome 2 obstacles to create e-phenotypes: First, (a) they need to communicate their specification for the patient cohort to an individual with the expertise to define the phenotype (eg, database analysts, honest data-brokers); and second, (b) given these specifications with inherent uncertainties, the honest data-brokers need to construct the phenotype using the available tool(s). The collaborative communication process among clinical domain experts and honest data-brokers described under obstacle (a) has been previously described by Hruby et al and labeled as "biomedical query mediation."[8,9] This process focuses on the exchange of information to identify the interests of domain experts, and to refine the interview process. Our study demonstrated the use of query mediation to identify attributes of the expert and the phenotype that predicted successful specification as estimated by direct *post hoc* chart review process.

Data from EHRs may be structured and unstructured, and vary widely in quality and completeness. EHR documentation varies across patient charts, as well as among providers and medical institutions.[20–23] In addition to variations in data heterogeneity and quality, there is a wide range of symptoms and complexities in any given disease among patients who present with the same or different clinical manifestations of the disease.[21,24] When the phenotype is more precise (specific), fewer encounters may be available. However, too broad a phenotype results in cohorts that are hard to capture or irrelevant to the investigator's research question. Although emergence of computable phenotypes helps alleviate some of those issues, rapid development of ad hoc phenotypes for cohort discovery and pilot work is still challenging. Further, phenotypes may not be static; they may represent patients at a particular time in the clinical course of disease. Capturing useful clinical representations of patients' clinical course that are consistent across providers adds additional complexity to e-phenotyping requirements.

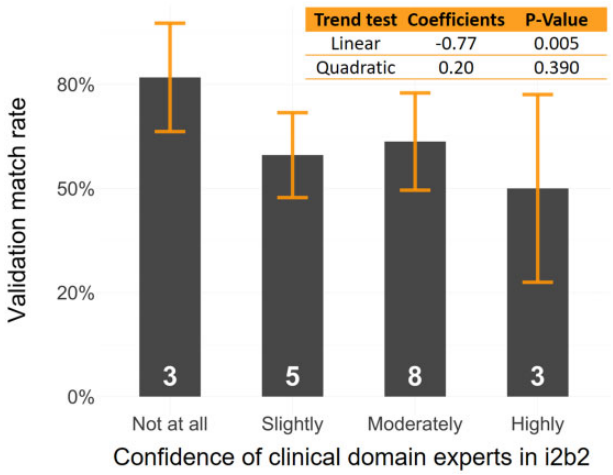

Figure 2. Chart review match-rate is shown as a function of noninformaticist clinical domain experts' confidence in i2b2 e-phenotyping system. Bars display the mean match-rate with standard errors for experts in a given confidence category with number of experts displayed on bars. *P* values are calculated using generalized linear model.

Examining the free-text comments by the experts and unstructured interviews revealed several challenges with processes for defining e-phenotypes:

i. *e-phenotype description needs to differentiate patient characteristics versus clinical events necessary for patient inclusion/exclusion:* We observed confusion by the experts on the difference between identifying in i2b2 the distinct list of patients who met the phenotype criteria over a series of visits and identifying visits that contained the inclusion criteria. Data from i2b2 can be linked to the RDW via the patient or visit identifiers to generate the reports and additional clinical data. The i2b2 system uses an entity-attribute-value (EAV) schema, which is a fundamentally different method of modeling data in comparison to a relational schema. Use of the EAV schema presents a challenge in differentiating between inherent patient characteristics that remain unchanged and clinical events that may only be part of a given visit. As a result, we observed that phenotypes relying on changes of clinical characteristics across visits were harder to capture using i2b2 and resulted in lower chart match-rates.
ii. *Temporal constraints are hard to specify and are often not intuitive even for experienced i2b2 users:* The first step in capturing changes in the temporal characteristics of patient cohorts using i2b2 is to differentiate patient-level and clinical-event characteristics. These patient-level characteristics would identify the

Table 2. Explanatory models of chart-validation match-rate in terms of expert confidence and e-phenotype complexity attributes

		Univariable models[a]				All query complexity covariates[b]				Query complexity and confidence model[c]			
			95% CI				95% CI				95% CI		
Covariates	Contrasts	Effect size	2.5%-tile	97.5%-tile	*P* value	Effect size	2.5%-tile	97.5%-tile	*P* value	Effect size	2.5%-tile	97.5%-tile	*P* value
Temporal constraints	Independent vs same	0.137	−0.323	0.605	.562	0.14	−0.44	0.72	.633	−1.80	−2.72	−0.93	<.001
	Independent vs events	−0.683	−1.372	−0.005	.049	−0.67	−1.40	0.05	.069	0.09	−0.69	0.87	.812
Number of groups		0.349	0.187	0.528	<.001	0.37	0.08	0.68	.015	0.80	0.39	1.22	<.001
Number of domains		0.254	0.107	0.408	.001	−0.02	−0.33	0.28	.876	0.17	−0.19	0.54	.350
Confidence	Linear	−0.77	−1.32	−0.24	.005					−3.26	−4.40	−2.19	<.001
	Quadratic	0.20	−0.26	0.68	.391					0.72	0.14	1.32	.016
	Cubic	−0.36	−0.74	0.03	.068					0.39	−0.07	0.86	.098

[a]Each line corresponds to a univariable model with the named covariate.
[b]Multivariable model including all phenotype complexity covariates.
[c]Multivariable model including all phenotype complexity covariates and expert confidence.

Table 3. Reflections of the clinical domain experts on e-phenotyping using i2b2

		Validation match-rate (%)			
Question and response	*N*	Mean (95% normal CI)	Contrast[a]	Coefficient	Trend test *P* value
In the future, would you use the i2b2 system again to define a cohort?					
Definitely won't use again	0				
Probably won't use again	1	35			
Probably will use again	9	57.8 (42.1–73.5)	Linear	0.95	.006
Definitely will use again	9	67.2 (45.1–89.4)	Quadratic	−0.22	.362
How difficult was the use of i2b2 to define a cohort?					
Not at all difficult	10	64.5 (44.4–84.6)			
Somewhat difficult	5	71 (52–90)			
Moderately difficult	3	30 (24.3–35.7)	Linear	−0.22	.528
Very difficult	1	70	Quadratic	0.7	.025
Would you trust scientific results that derive from a system like this?					
Definitely won't trust	0				
Probably won't trust	6	44.2 (30–58.3)			
Probably will trust	10	70.5 (19.4–107)	Linear	0.55	.016
Definitely will trust	3	63.3 (30–58.3)	Quadratic	−0.59	.001
Would you recommend this system to your colleagues?					
Definitely won't recommend	0				
Probably won't recommend	2	52.5 (18.2–86.8)			
Probably will recommend	10	52.5 (35.4–69.6)	Linear	0.73	.005
Definitely will recommend	7	75.7 (54.2–97.3)	Quadratic	0.42	.027

[a]Only linear and quadratic trends are presented.

overall "population in which events occur," as noted in the i2b2 interface, and would define the subset of observations from which all subsequent *event* data would be drawn. Users are able to create as many event groups as desirable, each group representing a period in the life of the defined population. Lastly, the user would need to specify the order of events between each of these events using drop-down menus defining start/end, first/last-ever, before/on-or-before/simultaneously/after/on-or-after. The user also would need to specify the length of time each event occur in relation to other events.

This process of defining a sequence of events in i2b2 can easily become overly complex even with only 2 events. The highly technical aspect of creating the logic of temporal queries can itself be intimidating for the users. Moreover, there is a challenge in selecting the clinical interests of the experts and utilizing the i2b2 graphical interface to translate the desired phenotype into i2b2 logic. In most cases, this process required abstract thinking about the structure of data on the back-end as well as trial and error.

iii. *Variations in data quality and availability may affect the resulting phenotype and cause the expert to compromise the desired definition based on the mapping status of diagnoses, problem lists, labs, and other variables, as well as the heterogeneity of coding procedures by different clinical teams:* During the phenotype creation phase, we noted the challenge of translating desired phenotypes of clinical experts into an i2b2 query. Although much of this pertained to ideas and challenges discussed in (i) and (ii) above, certain phenotypes were especially hard to capture within i2b2 due to data mapping, incomplete data across observations, and lack of structured data. For example, using phenotype Infectious #3, a clinical investigator desired to capture populations that are "suspected community-acquired pneumonia." Within this phenotype, one criterion was the results of chest imaging (ie, computed tomography, x-ray), which was not available through i2b2, although we were able to include the presence/absence of imaging as a criterion. In many cases where a mismatch was observed, the clinical investigator incorrectly coded or linked concepts to specific encounters (eg, expect imaging on the same encounter yet it occurred several days later). In some cases, imaging data for other parts of the body were included incorrectly (see Supplementary Table S3 for details). We further observed that as the number of inclusion/exclusion groups increased, there was a sharp decrease in the counts of observations matching the criteria. Although this is generally expected, for certain queries the decrease was considerable and unexpected based on clinicians' knowledge of the patient populations, hinting at the possibility of incomplete coding and heterogeneity of data across patients. Where possible, we attempted to capture concepts using multiple equivalent or similar avenues (eg, using ICD codes and problem lists simultaneously). Additional challenges in data quality occurred. For example, we observed that a diagnosis on one encounter incorrectly identified a patient with a condition. In another case patient visits were captured because they matched a historical diagnostic criteria (ICD), but the patient did not truly have the condition (eg, phenotype Pulmonary #16). To ameliorate such challenges, we recommend ensuring that criteria are met on multiple visits across time.

iv. *Medications are perhaps the most challenging phenotype characteristic, especially because the route of administration is currently not mapped in i2b2 although available in RDW:* In concordance with previous reports, we observed challenges in capturing medication lists, prescribed medications, and administered medications.[21] This was especially challenging in outpatient cases where investigators were interested in patients undergoing a particular therapy. For example, we were not able to capture medication administration routes through i2b2 when an investigator was interested in a cohort of "adolescents with

acne on antimicrobials" (phenotype Skin #15). In this case, i2b2 identified topical and oral antibiotics whereas only the former was desired.

CONCLUSIONS, LIMITATIONS, AND FUTURE DIRECTIONS

Using MUSC's existing i2b2 system, we assessed feasibility of accurately defining patient phenotypes using historical medical records. We observed notable challenges in intuitive usability and accuracy, especially for more complex e-phenotypes that included changes in clinical patient characteristics at temporally linked encounters. We observed challenges in e-phenotypes relying on encounter-based characteristics such as medications, labs, and imaging. Chart validation revealed trends between match-rate and e-phenotype attributes, use of temporal constraints, and clinical experts' self-reported confidence in and trust of i2b2 data and peer recommendations of the i2b2 system.

A limited sample size is perhaps the greatest limitation of our study. As a consequence, we could not incorporate the clinical domain and clinical experts' seniority level into the modeling. Other publications have assessed the seniority roles with a large sample size, but they lack the match-rate estimates.[13] A better-powered study may be able to assess the relationship between the investigator seniority roles and match-rate.

Further work is needed to explore the challenges of designing e-phenotypes with encounter-based characteristics, particularly those that rely on changes over time. We recommend that personnel who are proficient with the use of informaticist tool(s), such as honest data-brokers, assist investigators in designing and fine-tuning phenotypes. In addition, we suggest separately exploring patient characteristics that are specific to the encounters versus those that are demographic and historic in nature. Lastly, we suggest making the usability of phenotyping systems by a noninformaticist a design priority.

FUNDING

AVA and BH are supported by NIH/NCATS R21 TR002513, NIH/NLM R01 LM012517, and NIH/NLM T15 LM013977. The project described was supported by the NIH/NCATS UL1 TR001450.

AUTHOR CONTRIBUTIONS

AVA and PAF conceived the essential concepts of the manuscript and directed the research. BH contributed creatively to the implementation of the concept. AVA and BH drafted the manuscript. KNS critically evaluated the design and outcomes of this study. All co-authors have revised and approved the manuscript.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

ACKNOWLEDGMENTS

We would like to thank the Biomedical Informatics Center personnel including Katie Kirchoff and Dr Tami Crawford who contributed to various aspects of this project.

CONFLICT OF INTEREST STATEMENT

None declared.

DATA AVAILABILITY

The data underlying this article are available in the article and in its online supplementary material.

REFERENCES

1. Mo H, Thompson WK, Rasmussen LV, *et al.* Desiderata for computable representations of electronic health records-driven phenotype algorithms. *J Am Med Inform Assoc* 2015; 22 (6): 1220–30.
2. Pacheco JA, Rasmussen LV, Kiefer RC, *et al.* A case study evaluating the portability of an executable computable phenotype algorithm across multiple institutions and electronic health record environments. *J Am Med Inform Assoc* 2018; 25 (11): 1540–6.
3. 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.
4. 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 Inform Assoc* 2013; 20 (e1): e147–e154.
5. Manuel DG, Rosella LC, Stukel TA. Importance of accurately identifying disease in studies using electronic health records. *BMJ* 2010; 341: c4226.
6. Benchimol EI, Smeeth L, Guttmann A, *et al.*; RECORD Working Committee. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. *PLoS Med* 2015; 12 (10): e1001885.
7. Pathak J, Kho AN, Denny JC. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. *J Am Med Inform Assoc* 2013; 20 (e2): e206–11.
8. Hruby GW, Boland MR, Cimino JJ, *et al.* Characterization of the biomedical query mediation process. *AMIA Jt Summits Transl Sci Proc* 2013; 2013: 89–93. eCollection 2013.
9. Weng C, Mir AK, Hanauer D, Cimino J. Dialogue analysis for clinical data query mediation. *Stud Health Technol Inform* 2019; 264: 1398–402.
10. Murphy SN, Weber G, Mendis M, *et al.* Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J Am Med Inform Assoc* 2010; 17 (2): 124–30.
11. Alekseyenko AV, Hamidi B, Faith TD, *et al.* Each patient is a research biorepository: informatics-enabled research on surplus clinical specimens via the living BioBank. *J Am Med Inform Assoc* 2021; 28 (1): 138–43.
12. Murphy S, Churchill S, Bry L, *et al.* Instrumenting the health care enterprise for discovery research in the genomic era. *Genome Res* 2009; 19 (9): 1675–81.
13. Sholle ET, Cusick M, Davila MA, Kabariti J, Flores S, Campion TR. Characterizing basic and complex usage of i2b2 at an Academic Medical Center. *AMIA Jt Summits Transl Sci Proc* 2020; 2020: 589–96.
14. Deshmukh VG, Meystre SM, Mitchell JA. Evaluating the informatics for integrating biology and the bedside system for clinical research. *BMC Med Res Methodol* 2009; 9: 70.
15. Harris PA, Taylor R, Minor BL, *et al.*; REDCap Consortium. The REDCap consortium: building an international community of software platform partners. *J Biomed Inform* 2019; 95: 103208.
16. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform* 2009; 42 (2): 377–81.
17. Obeid JS, Beskow LM, Rape M, *et al.* A survey of practices for the use of electronic health records to support research recruitment. *J Clin Transl Sci* 2017; 1 (4): 246–52.

18. *R: A Language and Environment for Statistical Computing [Program]. R Package Version 3.6.1 Version*. Vienna: R Foundation for Statistical Computing; 2019.
19. Friedman CP. A "Fundamental Theorem" of biomedical informatics. *J Am Med Inform Assoc* 2009; 16 (2): 169–70.
20. Parsons A, McCullough C, Wang J, Shih S. Validity of electronic health record-derived quality measurement for performance monitoring. *J Am Med Inform Assoc* 2012; 19 (4): 604–9.
21. Chan KS, Fowles JB, Weiner JP. Electronic health records and the reliability and validity of quality measures: a review of the literature. *Med Care Res Rev* 2010; 67 (5): 503–27.
22. Gavrielov-Yusim N, Friger M. Use of administrative medical databases in population-based research. *J Epidemiol Community Health* 2014; 68 (3): 283–7.
23. Benin AL, Fenick A, Herrin J, Vitkauskas G, Chen J, Brandt C. How good are the data? Feasible approach to validation of metrics of quality derived from an outpatient electronic health record. *Am J Med Qual* 2011; 26 (6): 441–51.
24. Bennett TD, Moffitt RA, Hajagos JG, *et al.*; National COVID Cohort Collaborative (N3C) Consortium. Clinical characterization and prediction of clinical severity of SARS-CoV-2 infection among US adults using data from the US National COVID Cohort Collaborative. *JAMA Network Open* 2021; 4 (7): e2116901.