Tailoring Rule-Based Data Quality Assessment to the Patient-Centered Outcomes Research Network (PCORnet) Common Data Model (CDM)

Yahia Mohamed, MBChB, M.S[1], Xing Song, PhD[2], Tamara M McMahon, MA[1], Suman Sahil, MBA[1], Meredith Zozus, PhD[3], Zhan Wang, PhD[3], Lemuel R Waitman, PhD[1,2]

[1]University of Missouri-Kansas City School of Medicine, Kansas City, MO, USA; [2]University of Missouri School of Medicine, Columbia, MO, USA; [3]University of Texas Health Science Center at San Antonio

Abstract

Individual researchers and research networks have developed and applied different approaches to assess the data quality of electronic health record (EHR) data. A previously published rules-based method to evaluate the data quality of EHR data provides deeper levels of data quality analysis. To examine the effectiveness and generalizability of the rule-based framework, we reprogrammed and translated published rule templates to operate against the PCORnet Common Data Model and executed them against a database for a single center of the Greater Plains Collaborative (GPC) PCORnet Clinical Research Network. The framework detected additional data errors and logical inconsistencies not revealed by current data quality procedures. Laboratory and medication data were more vulnerable to errors. Hemolyzed samples in the emergency department and metformin prescribing in ambulatory clinics are further described to illustrate application of specific rule-based findings by researchers to engage their health systems in evaluating healthcare delivery and clinical quality concerns.

Introduction

Over the past decade electronic health records (EHR) have become the norm with adoption in the United States reaching 96% in hospitals and 87% in clinics.[1] The National Institute of Health (NIH), Food and Drug Administration (FDA)[2], and other agencies and researchers have increasingly focused on the secondary use of EHR data to advance clinical research. Clinical researchers developed methods to reuse these data in retrospective research to improve the health outcome and safety of care and to conduct pragmatic clinical trials, comparative effectiveness research, and epidemiological studies.[3,4] However, EHR data are not collected according to strict protocols used in research settings limiting secondary use of EHR in clinical research. These limitations are often attributed to various data quality issues such as missingness, differences in format or meaning, error and inconsistency. These and other data quality problems stem from differences in definition, workflow from which clinical data originate, observation or measurement methods and documentation practices among individuals, groups, and organizations. The harmonization of EHR data between departments within the same facility or between different facilities is needed to improve the generalizability.[5,6,7] National research networks such as Patient-Centered Outcomes Research Network (PCORnet) funded by the Patient-Centered Outcomes Research Institute (PCORI) use common data models to gain some harmonization to support multisite observational research.[8] PCORnet does so to integrates seven large Clinical Research Data Networks (CRNs) that represent diverse data for more than 66 million patients across the United States.[9,10]

Harmonization of existing data to a common data model and associated controlled terminology, however, does not address underlying data quality problems. Problems with clinical data quality may impact the validity of results derived from these data.[11,12] Many studies have been conducted to define the data quality and investigate reasons for EHR data quality problems. However, a systematic approach to assess the data quality of EHR is controversial. Weiskopf and Weng[13] conducted a literature review and focused on methods used to evaluate the quality of clinical data. They described the five domains of data quality (completeness, correctness, concordance, plausibility, and currency) and common methods (comparison to a gold standard, data element agreement, element presence, data source agreement, distribution comparison, validity checks, and log review) to assess data quality domains. They found that researchers used inconsistent methods to check for EHR data quality and recommended adopting validated and systemic approaches to evaluate the data quality when using EHR data for research. The recent draft guidance from the Food and Drug Administration similarly recommends systematic assessment, and describes use of accuracy assessment measures including sensitivity, specificity, positive and negative predictive value but not rule-based

approaches, also called validity, discrepancy, or error checks, to identify inconsistencies between data values in a dataset.[2]

Using specific rules that leverage basic logic statements commonly written in Structured Query Language (SQL) combined with encoding clinical data elements using national terminologies are one of the methods used to assess EHR data quality though reports from studies leveraging these methods are limited.[14,15,16] Kahn et al.[5] proposed a conceptual framework to evaluate data quality and discussed the model used in data quality assessment for single-site and multisite. They built this model by modifying and simplifying the Wang and Strong[17] "fit-for-use" model to become compatible for clinical and health services research. This conceptual framework addressed the variability of clinical data and can be translated to become a strategy for data quality checks. Z. Wang et al.[16,18] developed a scalable framework that can organize and manage data quality rules and facilitate their sharing and use for data quality assessment and monitoring in health care facilities. They identified 63,397 rules categorized according to topic and logic using 28 logic rule templates and associated knowledge tables. They evaluated the rules in one center and found that rules identified critical data errors and prompted stakeholders to identify actions to be taken based on the data discrepancies identified by the rules. However, this study was limited to a single center, did not encode the clinical data elements with national terminologies other than RxNORM and the International Classification of Diseases and was not programmed to operate against a national research network's Common Data Model which limited the reuse of the logic templates and SQL scripts in other healthcare facilities.

The Coordinating Center of PCORnet runs quarterly checks for data quality and produces an Empirical Data Characterization (EDC) report summarizing the findings and shares them with the network partners. The EDC uses rules to assess the distribution of information, data model conformance, data plausibility, data completeness, and data persistence.[10] However, the rules used are not specific enough to capture the data errors related to lab values range or units. Also, they did not report the discrepancies related to drug prescriptions versus lab test orders or drug interaction which limits the ability to take data quality findings back to the clinical organizations and align advancing clinical research quality with the goals of a learning health system. Given the promising results of the rules-based approach used by Wang et al. a holistic, effective, reusable, and flexible approach to assess the EHR data quality should be sought. This expansion of that work will (1.) translate and reuse Wang's rules templates and run the rules against the PCORnet CDM, (2.) evaluate the effectiveness of this framework in a second center, and (3.) provide use cases to check for the care delivery and clinical quality problems that resonate with learning health systems.

Methods

The study used a rule-based method to assess the data quality of electronic health records. We leveraged a set of rule templates developed and validated by our collaborator DQ-centric researchers at the University of Texas Health Science Center at San Antonio, which was built based on a variety of knowledge source including but not limited to existing rule sets, EHR screen analysis, crowd sourcing and surveying domain experts.[16,18]

Data source

This study used EHR data from University of Missouri Health Care system, a member of the Greater Plains Collaborative[19] (GPC) network. The GPC is a contributing network of the PCORnet and integrates 13 leading medical center and represent a diverse population of more than 21 million patients across nine states.[19,20] The University of Missouri Health Care system data stored in the PCORnet Common Data Model (CDM) contained 2.28 million patients.

Prior design

Wang et al.[16,18] used many publicly available sources to extract, sort, and build the rule templates that shares the topic and logic structure. Clinical data elements used rules were compiled in knowledge tables. Wang et al. identified and extended the rule templates by presenting their framework at a series of large organizational meeting and with their health system and medical experts to add more rules prompted by their clinical quality interest. Following Khan's conceptual data quality model, they classified their rules into three major categories: Conformance, Completeness and Plausibility.[4,5,18] Rules were composed of logic templates which are categorized into five main categories that check different data quality domains (incompatibility, value out of range, temporal sequence error, incompleteness, duplication), and knowledge tables used to support the implementation of these templates against databases.[18] The resulting knowledge tables included 63,397 rules executed by 28 rule templates.[16]

Implementing of rules against the PCORnet CDM

After receiving the rule templates and knowledge tables from the researcher team, we loaded the knowledge tables to our database using the Snowflake cloud data platform.[21] Using the Structured Query Language (SQL), we rewrote codes according to the Common Data Model format depending on the information provided in the knowledge tables, translated these rules to be compatible to run on PCORnet CDM, and version-controlled rules and scripts in github repository[22] which would also benefit future dissemination across PCORnet sites. We executed the CDM compatible codes and reported the discrepancies. The record with discrepancy is the encounter with inconsistency between data and rules due to data error. In this framework, a discrepancy is defined as an instance of one or more data values matching a knowledge table record such as an age-incompatible diagnosis. Ideally, such an approach would identify all instances of data error and only instances of data error.

The rule templates check for five categories of data quality domains: Some rules used to identify the out-of-range values. The rules in this category determine the range of values of a specific data element and will flag any record that has values that are physiologically or practically impossible. For example, any record with systolic blood pressure higher than 350 mmHg will be flagged as a record with discrepancy. Other rules identify incompatibility. These rules in this category look for any relational inconsistency between related variables and tables. For example, any record for the male patient from the demographic table and shows hysterectomy procedure in the procedure table will be flagged as a record with discrepancy. Some rules identify incompleteness. These rules examine the co-occurrence patterns among related data elements. For instance, certain drugs should be monitored when prescribed to the patient by checking for drug levels in the blood, and we flag any record that has a prescription of these drugs and no record of the recommended drug levels monitoring test in the lab results table. Other rules identify chronological errors. These rules check temporal relationships and identify impossible or incorrect sequences of events. For instance, some lab tests need to be ordered at a specific time to get accurate results. We examine the order time of these tests and flag any record that fails to show the expected temporal relationship. Lastly, some rules will identify duplication. For example, these rules will identify multiple occurrences of procedures that can be performed only once during an individual's life such as total resection of prostate or hysterectomy. The mapping of the conceptual data quality model and the rule templates summarized in (Table 1)

Table 1. Overview of mapping the conceptual data quality model[5] and the rule templates[16]

Conceptual data quality model category	Template category	Template name	Rule template logic
Rules to assess domain constraints	Out of range	Demographic data elements	Flag if demographics data elements are out of valid range
		Observation data elements	Flag if observation data elements are out of valid range
		Valid laboratory values	Flag if laboratory results is out of valid range
Rules to assess relational and attribute dependency	Incompatibility	Age and diagnosis	Flag if age does not meet criteria, diagnosis present
		Age and procedure	Flag if age does not meet criteria, procedure present
		Gender and diagnosis	Flag if gender is equal to invalid gender, diagnosis present
		Gender and procedure	Flag if gender is equal to invalid gender, procedure present
		Gender and clinical specialty	Flag if gender is equal to invalid gender for clinical specialty
		Drug and diagnosis	Flag if drug present, diagnosis present
		Drug and drug interaction	Flag if drug is present, Interaction drug present

Conceptual data quality model category	Template category	Template name	Rule template logic
		Inpatient only procedure	Flag if procedure is inpatient only, location is not inpatient
		Diagnosis and laboratory	Flag if diagnosis present, laboratory absent
	Incompleteness	Drug and laboratory	Flag if drug present, laboratory absent
		Drug and continuous procedure	Flag if drug present, continuous procedure absent
		Drug monitoring	Flag if drug present, drug monitoring absent
Rules to assess historical data and state-dependent objects	Date and time error	Laboratory time	Flag if laboratory time presents at an invalid time of a day
		Date in future	Flag if date is in future
	Duplication	Procedure duplication	Flag if procedure appears more than once

Use cases for rules that discern care delivery issues versus clinical quality

In the course of tailoring and testing the rules against the PCORnet CDM, we naturally uncovered data quality concerns or errors in our CDM. As we looked closer we found our activities went from concern with structural errors in data transformation into the CDM and crossed into concerns that may reflect suboptimal healthcare delivery processes as well as provided insight into the clinical quality and provide two illustrative examples:

a. Laboratory sample hemolysis

One rule template looked at the presence of null laboratory values where a numeric result is expected and conducted analyses to find out the reasons behind the discrepancies whether due to the data transformation issue or a performance of care problem. We selected blood potassium level to conduct this analysis because it is one of the most commonly analyzed electrolytes in the emergency department (ED)and is more prone to error due to sample hemolysis [23]. Such errors in healthcare delivery often require repeat testing and result in delays in clinical monitoring and decision making.

b. Metformin prescription and ordering of hemoglobin A1c (HbA1c)

We used the rule templates to check for drug prescriptions and lab values and ran another detailed analysis looking at encounters that have metformin (oral hypoglycemic drug prescribed for type II diabetes mellitus) prescribed whether they have HbA1c lab test ordered before the medication been prescribed during the ambulatory visit. Drug prescribing without accompanying lab monitoring reflects more of a clinical quality concern as opposed to errors in delivering services and testing.

Results

We translated and programmed the rules to be compatible against the PCORnet CDM and executed 8,208 rules against the University of Missouri Health Care system database that contains 2.28 million patients. We identified the number of patients who have data errors by counting the encounters number that have discrepancies with rules and then counting the number of patients who have data discrepancies. The count of the cohort with observations encompassed by the rule template, the counts of patients with discrepancies for every rule template, and the total percentage for patients with discrepancies are summarized in (Table 2). Number of records with discrepancies summarized in (Table 3)

Table 2. Summary for the result of rules template implementation against the PCORnet CDM (Number of patients)

Template name		Count of patients in the whole cohort (n)	Count of patients with discrepancies (n)	The percent of discrepancies (%)
Out of range values				
Demographic data elements		2,278,706	59	0.003
Observation data elements		781,911	12,576	1.6
Valid laboratory values		379,323	135,981	35.8
Incompatibility				
Age and diagnosis		1,036,512	5,411	0.52
Age and procedure		904,814	27,114	2.9
Gender and diagnosis		597,437	48,050	8.0
Gender and procedure		904,814	358	0.04
Gender and clinical specialty		1,129,798	26,106	2.3
Drug and diagnosis		1,057	31	2.9
Drug and drug interaction		714,897	4,526	0.6
Inpatient only procedure		913,712	12,406	1.4
Diagnosis and laboratory		42,459	1,084	2.6
Incompleteness				
Drug and laboratory		59,856	11,517	19.2
Drug and continuous procedure		5079	226	4.5
Drug monitoring		62,947	32,465	51.6
Date and time error				
Laboratory time		14,727	11,043	74.9
Date in future	Death date	278,048	0	0.0
	Birth date	2,278,706	0	0.0
	Medicine administration date	438,455	0	0.0
	Procedure date	955,964	0	0.0
Duplication				
Duplication		955,964	0	0.0

Table 3. Summary for the result of rules template implementation against the PCORnet CDM (Number of observations)

Template name	Count of observations in the whole cohort (n)	Count of observations with discrepancies (n)	The percent of discrepancies (%)
Out of range values			
Demographic data elements	2,278,706	59	0.003
Observation data elements	47,177,173	19,906	0.04
Valid laboratory values	22,629,324	222,804	0.98
Incompatibility			
Age and diagnosis	55,796,869	11,230	0.02
Age and procedure	39,453,903	49,366	0.13
Gender and diagnosis	30,006,428	98,490	0.33
Gender and procedure	39,453,903	465	0.001
Gender and clinical specialty	1,129,798	32,211	2.85
Drug and diagnosis	7,933	37	0.46
Drug and drug interaction	67,181,927	9,842	0.01
Inpatient only procedure	31,246,770	14,144	0.05

Template name		Count of observations in the whole cohort (n)	Count of observations with discrepancies (n)	The percent of discrepancies (%)
Diagnosis and laboratory		95,878,619	87,736	0.09
Incompleteness				
Drug and laboratory		660,152,713	17,598	0.003
Drug and continuous procedure		3,133,687	63,936	0.02
Drug monitoring		853,753,275	13,413,880	1.6
Date and time error				
Laboratory time		45,783	28,752	62.8
Date in future	Death date	328,017	0	0.0
	Birth date	2,278,706	0	0.0
	Medicine administration date	33,984,471	0	0.0
	Procedure date	41,633,278	0	0.0
Duplication				
Duplication		41,633,278	0	0.0

Rules to assess out of range values

In this category, 43 rules represent 3 templates. The demographic data elements template rules revealed that 0.003% of patients had a birth date before 01/01/1850. Rules in the Observation data elements template that assessed height, weight, and blood pressure found that 1.6% of patients have value out of range for these measurements. The valid lab values template includes rules that assess the low or high invalid lab test values and lab test unit. Implementation of these rules revealed that 35.8% of patients have lab data errors due to invalid lab values or unit mismatch.

Rules to assess the incompatibility

In this category, 8,121 rules represent 9 templates. The age and diagnosis rules template showed that 0.52 % of patients have records that include diagnoses incompatible with their age. Similarly, rules in the age and procedure template revealed that 3% of patients have procedures inconsistent with their age. Rules in gender and diagnosis templates found that 8 % of patients have diagnoses incompatible with their gender. Note however to implement the rules in this template, we restricted the date of diagnosis after January 1, 2015 to reduce the execution time. Rules in gender and procedure templates revealed that 0.4% of patients have procedures not performed in their gender. Rules in gender and clinical specialty template found 2 % of male patients over one year old received services at outpatient obstetric or gynecology clinics. Rules in drug and diagnosis template implemented to check for patients who prescribed Non-Steroidal Anti-Inflammatory Drugs (NSAIDS) on the same date of diagnosis with peptic or duodenal ulcer disease. The rules in this template showed that 2.9% of patients diagnosed with peptic or duodenal ulcer disease were prescribed NSAIDS on the same date of diagnosis. Rules in drug and drug interaction templates examined if any drug was prescribed at the same date with other medications known to interact together. The rules found that 0.63% of patients in the cohort were prescribed two interacting medications simultaneously. The inpatient-only procedure template rules revealed that 1.4% of patients have records that show inpatient procedures were done in the emergency room or during the ambulatory visit. The diagnosis and lab template rules assessed data of patients who have diabetes mellitus diagnosis (DM)and their lab test for blood glucose and HbA1c. The rules found that 2.6 % of these patients have no blood glucose or HbA1c lab test records.

Rules to assess the incompleteness

In this category, 3 rule templates contained 35 rules. The drug and lab template rules examined data of patients who prescribed a specific drug to be followed up with blood lab tests to monitor the body's response while on a continuous regimen of a medication. The rules showed that 19.2% of patients in the cohort who had a recorded prescription for

the three types of drugs assessed by the rules did not have a co-occurring record of the required monitoring lab test. The drug continuous procedure rules examined the data for patients who were prescribed specific drugs and were expected to have a follow-up procedure to assess for potential adverse drug effects. The rules found that 4.5 % of patients missed the follow-up procedure in their records. The rules in the drug monitoring template checked for data of patients who were prescribed a specific drug that needed blood test level monitoring to prevent drug toxicity. The results show that 51.6% of patients in the cohort did not have a record of the blood level test monitoring described the drug labeling.

Rules to assess the date and time error

In this category, 2 rule templates including 6 rules were applied. Rules in the lab time template assessed lab test data for which the sample should be taken at a specified time of day. The rules found that 74.9% of patients in the cohort had specimens that were taken at the time outside the window used in the rules. The rules in the "date in future" template checked for inconsistent future death date, birth date, medicine administration start and stop date, and procedure date. The rules found 0% inconsistent future date for tested tables.

Rules to assess the duplication

In this category, 3 rules for 1 rule template were implemented to assess procedure duplication and examine patients' records that have procedure duplication. The template specified hysterectomy, right leg amputation, and prostate removal in the check. The rules found 0 % of procedure duplication.

Use cases analyses results

Laboratory sample hemolysis

Implementation of valid lab values template revealed 35.8% of patients have lab values out of specified range or have the incorrect unit. We found that most of the discrepancies were due to the presence of NULL value in the lab result number column. The count for distinct encounters of all lab test results with NULL value was 2,361,301. At the same time, the count for distinct encounters of lab test results with a NULL value and hemolyzed sample was 80,342. Also, the count for distinct encounters in the emergency department with lab test results as a NULL value and hemolyzed sample was 15,975. Of these hemolyzed sample distinct encounters, 36,372 were for serum or plasma potassium, and 6,511 were for serum or plasma potassium ordered in the emergency department. The annual percentage of hemolyzed samples decreased from 23.1% in 2010 to reach 16.9% in 2020. The same trend of decrease was associated with hemolyzed potassium samples, which were decreased from 31.6% in 2010 to 13.6% in 2020.

Metformin prescriptions and order of HbA1c

Implementation of diagnosis and lab template showed that 2.5 % of patients in the cohort specified by the rules missed blood glucose or HbA1c. We conducted an analysis using the metformin prescriptions, and order of HbA1c for the ambulatory visit encounters to assess the effectiveness of this rule template. We found that 45,743 patients received metformin prescriptions during an ambulatory visit and 771 of these patients received metformin prescriptions without an order for HbA1c test in their lab test records before the prescription.

Discussion

This study used the rule-based approach to implement 8,208 rules designed and written by Wang et al. against the database for a single center of the PCORnet. The complex nature of EHR data and the temporal relationship between data elements lead to many challenges during the assessment of EHR data quality. However, the rules used in this study were effective in capturing data error that generalize across national networks. Very detailed clinical events use cases make it relatively straightforward for researchers to conduct sub-analyses to uncover the possible causes of these data quality issues.

Domains with high rates of data errors were associated with the execution of valid lab value template that revealed 35.8% of patients in the cohort have lab values inconsistent with rules in the template, drug and lab template showed 19.2% of patients have data error, drug monitoring template showed 51.6% patients have data discrepancies, and lab

time template revealed 74.9% of patients in the cohort have data error. The percentages of observations with discrepancies as shown in Table 3 were less compared to the count of patients because patient may have multiple records for every encounter. The possible cause for data error captured by drug and lab template rules include incomplete lab data due to poor data collection, or missed visits for lab sample collection. Alternatively, the range used for the rule could be too narrow. The drug monitoring template assesses the data for lab orders for specific drugs (e.g., digoxin) that need therapeutic dose monitoring to optimize the dose and prevent drug toxicity. Therefore, the drug should be monitored by measuring drug levels in the blood when prescribed for patients. The rules in this template capture many patients who received prescriptions of these drugs without drug levels test in their lab records. These discrepancies may be due to the resolution of timing of orders and results reflected in the EHR system, drug level monitoring occurring outside the health system, results returned as non-structured data such as a scanned pdf narrative neither of which is accessible to the rules, or quality of care problems. Other possible causes of discrepancies found by the lab time template were because some valid value exceed the specified time in the rule by just minutes, data entry error, or the specimen collection occurred before or after the specified time. This suggests the need to evaluate refining the windows of time where results and activities may be considered clinically appropriate.

When we look at the discrepancies captured by rules in the valid lab value template, several trends emerge. Many encounters that have a NULL value instead of lab result number were due to sample hemolysis (3.4%). Further, we looked at the number of encounters with hemolysis results to check for the pattern and prevalence of hemolysis in the data, which showed that 20% of hemolyzed samples came from ED which was a known area of concern to the health system. The primary hemolyzed test in ED was potassium (41%). This use case represents the issue with the care delivery rather than data error because these discrepancies were due to improper collection, handling, or analysis of blood samples. Others have noted that blood sample hemolysis is a common issue in the high workload settings such as emergency departments.[24] In the second use case, we used the drug and lab rules template to evaluate data quality discrepancies by adding another rule that uses metformin and HbA1c for a patient with ambulatory visit encounters. The percentage of patients receiving metformin during the ambulatory visit without testing their HbA1c before the prescription was 1.7%. This result highlights the importance of thinking about the possibility of clinical quality captured as data errors.

Wang's framework efficiently captured data errors and identified data quality inconsistencies that were very specific. This specificity allows researchers to more easily engage clinical collaborators and health systems interested in improving quality that impacts care. The framework was flexible and adjustable when there was a need to add more rules or delete rules not reflected in the current CDM. The ease of implementation on the different databases is considered another advantage as standard SQL was used to encode the rule templates as the knowledge tables were represented in openly accessible comma separated value format test files. However, knowledge tables for some rules templates were far from complete due to a lack of relevant, good quality, publicly available knowledge sources. Future work in this area might explore integrating rules encapsulated in national resources such as the Unified Medical Language System (UMLS) and open source or commercial drug and clinical safety knowledge databases. A strength of our study was the translation and implementation of rule templates on one site that uses PCORnet CDM which will broaden dissemination, evaluation, and improvement of the rule templates and knowledge tables across multiple institutions. Our study also has several limitations. Some rules were not applied to the PCORnet CDM because of a lack of information in the CDM or limitations in knowledge representation. For example, rules evaluating drugs in same class at same time rules template include over 55,000 rules requiring the Anatomical Therapeutic Chemical Classification (ATC), which we will address in the future. This account for the reduced number of rules in our findings relative to the prior work by Wang et al. Also, although the rules effectively captured the data quality issues, identifying the exact cause of every data rule fired is challenging and may require significant analyses. Notably, laboratory result ranges checks encoded in the rules were adapted from a set used in checking clinical trial data and were not customized to institutional distributions or to reference ranges which may vary between institutions and within institutions over time as new laboratory equipment are implemented and retired. This likely contributed to the high rate of discrepancies based upon in lab values seen in the results. Lastly, even though the used rules were able to detect the physical impossible values as data errors, the risk of false positive still present which need validation of these rules before use.

Future Directions

Our future work will focus on assessing the scalability, variability and generalizability of rule templates by implementing them on more sites in the GPC network, sharing them nationally, and developing rules that can assist in identifying extract, transform, and load (ETL) errors. We will also seek to determine where rule templates and knowledge tables may be populated by curated and maintained knowledge bases such as the UMLS. Another dimension of quality we seek to exploit will be detecting errors of omission by leveraging the GPC environment[25] that combines EHR data from each site with state wide Medicaid and Medicare insurance claims. This will allow us to look at cases where the EHR-based CDM is missing observations that were billed for by the contributing health system. We will also develop a database to store when rules fire for each site's CDM so we can track discrepancies over time and produce statistical summaries to alert local sites' of potential data integration errors or assist in evaluating discrepancies related to care delivery processes (e.g., hemolysis).

Conclusion

Electronic Health Records are a promising source of data for clinical research including providing real world evidence in comparison to study specific research databases. However, assessing the quality of these data is a considerable undertaking, especially as research becomes increasingly conducted across multiple sites participating in research networks. This study advanced a previously published set of rule templates by translating and implementing them against the PCORnet Common Data Model. The rules were effective in identifying data error and are now scalable to other institutions across PCORnet. Specific rule-based data quality assessment promotes follow-up analyses that look at the possible causes related to care delivery and clinical quality that also resonate with health system clinicians and leaders. Future studies to assess the variability and generalizability of the framework are warranted.

Acknowledgement

This work was supported through Patient-Centered Outcomes Research Institute awards for the Greater Plains Collaborative (GPC CRN) (#RI-CRN-2020-003) and the Greater Plains Collaborative Advancement of PCORnet Infrastructure: Clinical Research Network Phase 3 (#RI-MISSOURI-MC)

References

1. Nordo A, Eisenstein Eric L, Garza M, Hammond WE, Zozus MN. Evaluative Outcomes in Direct Extraction and Use of EHR Data in Clinical Trials. *Stud Health Technol Inform*. 2019;257:333-340.
2. Center for Drug Evaluation and. Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision-Making for Drug and Biological Products. U.S. Food and Drug Administration. Published December 10, 2021. Accessed March 9, 2022. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/real-world-data-assessing-electronic-health-records-and-medical-claims-data-support-regulatory
3. Weiskopf NG, Hripcsak G, Swaminathan S, Weng C. Defining and measuring completeness of electronic health records for secondary use. *J Biomed Inform*. 2013;46(5):830-836.
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*. 2016;4(1):1244.
5. Kahn MG, Raebel MA, Glanz JM, Riedlinger K, Steiner JF. A Pragmatic Framework for Single-site and Multisite Data Quality Assessment in Electronic Health Record-based Clinical Research. *Med Care*. 2012;50(0):10.1097.
6. Brown J, Kahn M, Toh S. Data quality assessment for comparative effectiveness research in distributed data networks. *Med Care*. 2013;51(8 0 3):S22-S29.
7. Lee K, Weiskopf N, Pathak J. A Framework for Data Quality Assessment in Clinical Research Datasets. *AMIA Annu Symp Proc*. 2018;2017:1080-1089.
8. Kahn MG, Brown JS, Chun AT, et al. Transparent Reporting of Data Quality in Distributed Data Networks. *eGEMs*. 2015;3(1):1052.
9. Bian J, Lyu T, Loiacono A, et al. Assessing the practice of data quality evaluation in a national clinical data research network through a systematic scoping review in the era of real-world data. *J Am Med Inform Assoc*. 2020;27(12):1999-2010.
10. Qualls LG, Phillips TA, Hammill BG, et al. Evaluating Foundational Data Quality in the National Patient-Centered Clinical Research Network (PCORnet®). *eGEMs*. 2018;6(1):3.

11. Weiskopf NG, Bakken S, Hripcsak G, Weng C. A Data Quality Assessment Guideline for Electronic Health Record Data Reuse. *eGEMs*. 2017;5(1):14.
12. Johnson SG, Speedie S, Simon G, Kumar V, Westra BL. Quantifying the Effect of Data Quality on the Validity of an eMeasure. *Appl Clin Inform*. 2017;8(4):1012-1021.
13. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J Am Med Inform Assoc*. 2013;20(1):144-151.
14. Codd EF. Extending the database relational model to capture more meaning. *ACM Trans Database Syst*. 1979;4(4):397-434.
15. Feder SL. Data Quality in Electronic Health Records Research: Quality Domains and Assessment Methods. *West J Nurs Res*. 2018;40(5):753-766.
16. Wang Z, Talburt JR, Wu N, Dagtas S, Zozus MN. A Rule-Based Data Quality Assessment System for Electronic Health Record Data. *Appl Clin Inform*. 2020;11(4):622-634.
17. Wang RY, Strong DM. Beyond Accuracy: What Data Quality Means to Data Consumers. *J Manag Inf Syst*. 1996;12(4):5-33.
18. Wang Z, Dagtas S, Talburt J, Baghal A, Zozus M. Rule-Based Data Quality Assessment and Monitoring System in Healthcare Facilities. *Stud Health Technol Inform*. 2019;257:460-467.
19. Waitman LR, Aaronson LS, Nadkarni PM, Connolly DW, Campbell JR. The Greater Plains Collaborative: a PCORnet Clinical Research Data Network. *J Am Med Inform Assoc JAMIA*. 2014;21(4):637-641.
20. The Greater Plains Collaborative. Greater Plains Collaborative (GPC). Accessed February 23, 2022. https://gpcnetwork.org/
21. Snowflake's Cloud Data Platform. One Platform for All Your Data. Snowflake. Accessed March 4, 2022. https://www.snowflake.com/cloud-data-platform/
22. Greater Plains Collaborative Data Quality Repository. GitHub. Accessed March 8, 2022. https://github.com/gpcnetwork/DataQuality
23. Khodorkovsky B, Cambria B, Lesser M, Hahn B. Do Hemolyzed Potassium Specimens Need to be Repeated? *J Emerg Med*. 2014;47(3):313-317.
24. Grant MS. The effect of blood drawing techniques and equipment on the hemolysis of ED laboratory blood samples. *J Emerg Nurs*. 2003;29(2):116-121.
25. Waitman LR, Song X, Walpitage DL, et al. Enhancing PCORnet Clinical Research Network data completeness by integrating multistate insurance claims with electronic health records in a cloud environment aligned with CMS security and privacy requirements. *J Am Med Inform Assoc JAMIA*. Published online December 13, 2021:ocab269.