Article published online: 2020-09-23

A Rule-Based Data Quality Assessment System for Electronic Health Record Data

Zhan Wang[1] John R. Talburt[2] Ningning Wu[2] Serhan Dagtas[2] Meredith Nahm Zozus[1]

[1] Department of Population Health Science, University of Texas Health Science Center at San Antonio, San Antonio, Texas, United States
[2] Department of Information Science, University of Arkansas at Little Rock, Little Rock, Arkansas, United States

Address for correspondence Meredith Nahm Zozus, PhD, Department of Population Health Sciences, University of Texas Health Science Center at San Antonio, 7703 Floyd Curl Drive, MC 7933, San Antonio, TX 78229, United States (e-mail: zozus@uthscsa.edu).

Appl Clin Inform 2020;11:622–634.

Abstract

Objective Rule-based data quality assessment in health care facilities was explored through compilation, implementation, and evaluation of 63,397 data quality rules in a single-center case study to assess the ability of rules-based data quality assessment to identify data errors of importance to physicians and system owners.

Methods We applied a design science framework to design, demonstrate, test, and evaluate a scalable framework with which data quality rules can be managed and used in health care facilities for data quality assessment and monitoring.

Results We identified 63,397 rules partitioned into 28 logic templates. A total of 819,683 discrepancies were identified by 4.5% of the rules. Nine out of 11 participating clinical and operational leaders indicated that the rules identified data quality problems and articulated next steps that they wanted to take based on the reported information.

Discussion The combined rule template and knowledge table approach makes governance and maintenance of otherwise large rule sets manageable. Identified challenges to rule-based data quality monitoring included the lack of curated and maintained knowledge sources relevant to data error detection and lack of organizational resources to support clinical and operational leaders with investigation and characterization of data errors and pursuit of corrective and preventative actions. Limitations of our study included implementation within a single center and dependence of the results on the implemented rule set.

Conclusion This study demonstrates a scalable framework (up to 63,397 rules) with which data quality rules can be implemented and managed in health care facilities to identify data errors. The data quality problems identified at the implementation site were important enough to prompt action requests from clinical and operational leaders.

Keywords
- data quality
- electronic health records
- health care

Background and Significance

Electronic health record (EHR) data quality issues have been shown to impact decision support and secondary uses of health care data such as quality improvement and research alike.[1–16] There are many reports of data quality assessment in research data, either data collected prospectively or existing health care (or other) data to be used for research. Feder categorized current data quality assessment methods into rule-based methods, data abstraction-based methods with statistical measures, data comparisons with manual chart review, management of missing data using statistical methods, and data triangulation between multiple EHR databases.[17] Pezoulas et al presented a framework for metadata extraction

received
March 2, 2020
accepted
July 6, 2020

DOI https://doi.org/10.1055/s-0040-1715567.
ISSN 1869-0327.

and use of statistical methods for detection of anomalous values, missing values, and duplicate data.[18] Scholte et al assessed data quality by comparison between EHR data and survey data, which demonstrated EHR data performing better with respect to completeness and accuracy.[19] Callahan et al compared data checks across six data sharing networks to help expand the scope of data quality assessment across clinical data networks.[20]

However, medical institutions are laggards in the use of rule-based approaches, it is proved that several marketed tools or commercial packages do not cover data quality assessment well.[21] Rule-based approaches have proven effective for clinical decision support and clinical research data management. These methods inform but are not directly relevant to data quality assessment and improvement in health care settings and are not further discussed here. In early work, Carlson et al used a rules-based approach to identify instances of incompleteness, invalid values, inconsistent units of measurement, and inconsistent relationships in data from multiple facilities used for clinical decision support in intensive care settings.[22]

In 2002, Brown and Warmington defined data quality probes (rules or logic statements used to detect data problems) and executed them as a query in a clinical information system to find the inconsistency between two associated data items.[23] Data quality probes detected errors, tracked the prevalence of these cases, monitored the EHR quality, and gave feedback to clinicians.

In 2012, Kahn et al proposed a "fit-for-use" conceptual model for assessment of EHR data quality for secondary use in multisite studies.[24] They modified and simplified the Wang and Strong framework[25] into five domains relevant to the secondary use of EHR data use case. In separate work, members of the group used the framework to categorize sets of data quality rules used by large secondary data use initiatives.[26]

Most recently, Hart and Kuo reported rule-based discrepancy identification and resolution in health care data used for direct patient care and management of health services.[27] They reported a greater than 50% decrease in rejected records across three domains in 6 months (from 14.9 to 6.6 errors per 10,000 fields for patient information, from 8.5 to 2.9 errors per 10,000 fields for service information, and from 12.7 to 4.7 errors per 10,000 fields for financial information).

In 2017, Skyttberg et al[28] provided a comparison data quality assessment among paper-based, electronic, and mixed health record against vital sign data. Data quality was assessed in three categories: currency, completeness, and correctness. To estimate correctness, two further categories—plausibility and concordance—were used.

Johnson et al[29] demonstrated the utility of a health care data quality framework by a fit-for-use application in 2017. This study focused on five domains of time constraints, including patient, hospital admission, procedure, medication, and catheter intervention. A linear model was used to describe the impact of each data quality issues.

Based on the limited work to date, the rule-based data cleaning methods used in health care have successfully identified data errors. However, the studies are limited by the application context and scope of the rules. It is not uncommon for EHRs to have data elements in the thousands. Methods to focus on those of clinical, operational, and administrative importance will be critical. Further, knowing about data errors is necessary but not sufficient. Data quality assessment in health care will only have value when processes and resources are in place to trigger improvements in data and processes that translate to improvements in outcomes meaningful to patients, providers, and the clinical enterprise. Pragmatic methods for doing so are needed to support the burgeoning use of clinical data.

Objective

The objective of this research was to probe the basic feasibility and potential benefit of data quality assessment and management in health care:

1. Identifying the necessary functionality for data quality monitoring in health care.
2. Acquiring computationally accessible knowledge for clinically relevant rules.
3. Identifying true data quality problems.
4. Identifying data quality problems important enough to prompt improvement-oriented action.

We achieved these objectives through a design science research approach. The design cycle focused on the first and second research objectives and included designing and developing a system for rule-based data quality assessment in health care. The design was vetted with representative stakeholders during design iterations. The empirical cycle was conducted as a mixed-methods evaluation and focused on the third and fourth research objectives.

Methods

Our local Institutional Review Board (IRB) reviewed the materials and methods and determined that this project was not human subject research as defined in 45 CFR 46.102, and therefore it does not fall under the jurisdiction of the IRB review process (IRB#228157). The committee noted that data collected was about processes, not individuals.

A high-level conceptual architecture diagram is provided as ►Fig. 1. Briefly, the data quality assessment system runs on an institutional data warehouse. There are three system components, (1) rule logic templates and knowledge tables supporting the rules, (2) rule results tables, and (3) ability to output results relevant to stakeholders.

Rule Templates and Knowledge Tables

To broaden applicability and use across disciplines, basic data quality assessment approaches often avoid domain-specific constrains or rules. On the other hand, domain-specific knowledge can be exploited to detect data errors not identifiable through common column statistics. For example, a temperature in the human body over 49°C is incompatible with life and therefore physically impossible in a living subject and definitely a data error in an EHR. There are hundreds of thousands of such

Fig. 1 The data quality assessment system is designed to run on an institutional data warehouse. There are three system components, (1) rule logic templates and knowledge tables supporting the rules, (2) rule results tables, and (3) ability to report results relevant to stakeholders.

relationships in biomedicine and health care that can be used to potentially identify data errors of greater importance to clinicians and health system administrators. We considered writing and managing rules for each unpractical; acquisition and maintenance of this many initial rules conflicted with our goal of scalable rule management and maintenance. Inspired by the rule abstraction in Brown and Warmington's work,[23] we evaluated and sorted potential rules as we came across them. Those sharing topic and logic structure were abstracted into a single rule template (►**Fig. 2**). The rule templates were further categorized into five higher-level types: incompatibility, value out of range, temporal sequence error, incompleteness, and duplication (►**Table 1**).

Knowledge Acquisition

Reuse of Rules from Existing Sets

To identify candidate rules, we first looked to existing rule sets. These included the publicly available Observational Medical Outcomes Partnership rules, the National Patient-

Rules:

Rule 1: If a patient's diagnosis is Benign neoplasm of right ovary and gender is male, then flag

Rule 2: If a patient's diagnosis is Malignant neoplasm of placenta and gender is male, then flag

Rule 3: If a patient's diagnosis is Malignant neoplasm of prostate and gender is female, then flag

Rule 4: If a patient's weight is over than 300kg, then flag

Rule 5: If a patient's body temperature is over then 44°C, then flag

Abstract into

Rule Templates + Knowledge Tables

Rule Template: Flag if **Gender** = *Invalid Gender* and **Diagnosis Code**= *Code*

Knowledge Table:

Code Type	Code	Invalid Gender	Code Description
ICD10	N83.51	M	Benign neoplasm of right ovary
ICD10	C58	M	Malignant neoplasm of placenta
ICD10	C61	F	Malignant neoplasm of prostate

Rule Template: Flag if **Observation Code** = *Code* and **Unit** = *Unit* and (**Observation value** < *Valid Low* or **Observation value**> *Valid High*)

Knowledge Table:

Code Type	Code	Valid Low	Valid High	Units	Code Description
LOINC	3141-9	1	300	KG	Weight
LOINC	8310-5	6	44	°C	Body Temperature

Fig. 2 We evaluated and sorted potential rules as we came across them. Those sharing topic and logic structure were abstracted into a single rule template and the rule knowledge added to the corresponding knowledge table.

Table 1 Rule templates categorization

Type name	Definition	Example
Incompatibility[a]	One data value is logically inconsistent with another data value	Patient gender is incompatible with diagnoses
Value out of range[b]	The actual data value being outside the limits of an established range	Date of birth is before 1880
Date and time[c]	Any two dates occurring in an invalid order	Date of encounter cannot be earlier than date of birth for an adult
Incompleteness[d]	Occurrence of a data value that is expected but missing	A procedure is present but there is no corresponding encounter record
Duplication[e]	Multiple occurrences of events that can physically happen only once	A patient with two hysterectomies

Note: Categories we used for the rules are provided in the table along with the definition and an example of a rule meeting the definition of the category.
[a]Relational integrity rules, state-dependent objects rules, and attribute dependency rules from Kahn's model.
[b]Attribute domain constraints from Kahn's model.
[c]Historical data rules and state-dependent objects rules from Kahn's model.
[d]Relational integrity rules from Kahn's model.
[e]State-dependent objects rules from Kahn's model.

Centered Clinical Research Network (PCORnet) rules, the Healthcare Systems Research Network rules,[30] and the Mini-Sentinel data checking rules,[31] as well as rules written for an internal project using multisite EHR data.[32] Rules from these sources existed as individual logic statements. We also incorporated tables of and age and gender incompatible diagnosis and procedure lists from payers and drafted the corresponding rule templates.

Application of Rule Development Methods Used in Research

We applied rule identification methods commonly applied in clinical trial data cleaning. Each element in our local data warehouse model and a subset of EHR screens used by anesthesiologists was carefully reviewed, taking into consideration as many of the possible relationships with other data elements as possible to identify relationships that could be exploited to identify data errors. Once identified, rule templates and knowledge were written for them.

Crowd Sourcing

Crowd sourcing was also attempted to identify new rules and rule templates. We presented the rule-based data quality assessment system strategy and rule templates at several large meetings with participants from multiple organizations. We shared the existing rules and solicited rule contributions in exchange for sharing our existing rules.

Talking to Experts

We interviewed experts and asked for any rule templates or knowledge they could add based on their medical experience and expertise. Eleven physicians from different clinical specialties were invited to participate in rule and rule knowledge identification during the design phase of the project.

Design Validation

Design validation was completed in three steps. First, we analyzed the rule results from the system, that is, the data quality problems identified by the rules. Second, we interviewed physicians and information system owners about the rule result summaries presented to them. The interviews collected both structured and qualitative data toward evaluating the system's capability to identify data errors of importance to physicians and system owners. Last, we summarize from the design validation where the system design was limited or otherwise fell short.

The consent process was conducted in person via signed informed consent form. A semistructured interview approach and questionnaire comprising six structured questions and two semistructured questions was used to support a mixed methods evaluation rather than imposing an exhaustive and mutually exclusive list of choices. Because this was early formative work, we specifically wanted to impose as little structure on the participant responses as possible to remain open to stakeholder goals and reactions that had not previously occurred to us. At the same time, we equally valued structured questions to obtain clear indication of whether the approach could identify data quality problems of importance to the participants. Thus, open-ended probes were used for each question and participant analysis and articulation of rationale was encouraged.

Eleven participants ultimately participated in the design validation. Some participants had multiple roles. The 11 participants included 9 physicians (6 practicing and 3 nonpracticing), 6 information system owners, and 5 secondary data users. Participants were shown the overall template-level rule list and results reports.

Results

Rule Templates and Knowledge Sources

In the initial rule templates and knowledge source acquisition work, 6,051 rules and 16 rules templates were identified from existing sets (►Table 2). The anesthesia screen analysis produced 58 new temporal sequence error rules for addition into

Table 2 Rule acquisition

Knowledge source	Number of rules added	Number of rule templates added
Existing sets	6,051	16
Data model and EHR screens analysis	78	0
Crowd sourcing	80	2
Talking to experts	57,188	10
Total	63,397	28

Abbreviation: EHR, electronic health record.
Note: Knowledge sources from which we obtained rules are listed with the number of rule templates and rule records obtained from each. In the initial rule templates and knowledge source acquisition work, 6,051 rules and 16 rules templates were identified from existing sets. Review of anesthesia screens produced 58 new temporal sequence error rules for addition into the knowledge table. From crowd sourcing, 2 rule templates and 80 rules were identified. Ten rule templates and 57,188 rules are added based on experts' knowledge, mainly from addition of drug-related rules that leveraged relationships in RxNorm or information in the Structured Product Label.

the knowledge table. From crowd sourcing, 2 rule templates and 80 rules were identified. Ten rule templates and 57,188 rules are added based on experts' knowledge. Experts' knowledge contributed a significant number because one expert mentioned a template: more than one drug in same class at same time, which includes 55,243 rules. The rule identification ultimately resulted in 63,397 rules which were compressed through the use of knowledge tables to 28 rule templates (►Table 3). Although this work evaluated 63,397 rules, a limitation to use today is that many useful rule templates have likely been overlooked. To mitigate the problem and allow identification and adding new rule templates over time, a generic template can be used for two (or more) value logic inconsistencies, for example, Flag if Table.column 1 value does not match Table.column 2 value. With the generic template and four knowledge acquisition methods, rule templates combined with knowledge tables instead of individual rules provide a framework for scalable rule management and maintenance over time. It allows handling a large number of rules up to 63,397 rules and has ability to extend the rule volume.

Rule Results

The 63,397 rules were programmed, tested, and executed over an institutional data warehouse serving a tertiary care hospital and associated clinics using four different EHR systems. At the time of the assessment, the data warehouse contained data from 1.46 million patients and 9 facilities. Summary results from the 23 implemented rule templates are shown in ►Table 4.

Design Validation

Semistructured Interviews

The rules-based approach was also validated through the semistructured interviews conducted based on the rule results. The first structured question, "Did the rules identify data quality problems?" was the primary research question. Nine out of 11 participants (82%) stated that data quality problems were identified by the rules. The remaining two participants stated that the rule results were not specific enough for them to tell if there were actual data quality problems identified (►Table 5). The second question probed the extent to which data in which errors were detected were used in clinical decision-making. Eight out of 11 participants (73%) voiced that the data quality problems identified were in data used in clinical decision making (►Table 5). The remaining structured questions probed the perceived potential impact of the data quality problems (if the participant felt any were identified) on clinical decisions, patient safety, institutional finances, and regulatory compliance. More than 50% of the participants believed the data quality problems have medium to high impact to clinical decision making (►Table 5). Fifty-four percent participants choose medium to high impact to patient safety, the rest of the participants indicated that they were unable to judge (►Table 5). For financial and regulatory impact, 55% and 36% participants, respectively, reported being unable to judge for the same reason (►Table 5). Among the participants (►Table 5), more participants indicated that data quality problems identified had a higher potential impact on institutional finances (45%) than regulatory compliance (36%).

The impact questions were followed by two semistructured interview questions asking the participant's opinion regarding potential reasons for the data quality problems and actions that they would like to take (if any) based on the rule results. In both cases, participants could select from a list or list others (not listed) that came to mind. Participants indicated multiple possible causes for the data errors. Manual input error, incorrect code system use, and incorrect clinical practice were the most often selected potential reasons for data quality problems (►Fig. 3).

Forty-eight percent participants indicated that they would like to initiate an analysis to better understand the root cause, opportunity for correction, and opportunities for prevention of future similar data errors. Twenty-eight percent of participants listed actions other than those preprinted on the questionnaire (►Fig. 4). When participants were asked to select (or add) options that best described actions that they wanted to take based on the rule results, they significantly added to the list on the form. The coded qualitative data are listed with the quantitative data in ►Fig. 4.

The participant indicating that there were not any actions that they wanted to take as a result of the rule results was asked the reason for not taking any action. Only one participant chose no action and stated three reasons: (1) the participant felt that it was unlikely that they would find anything from large volume of data, (2) they reported having no resources to analyze the problems, and (3) the participant saw no likelihood of financial benefit.

The final question in the interview probed differences in perceived data quality across different domains of EHR data by asking participants to rate the adequacy of different data

Table 3 Provides rule templates and the associated logic for each

	Template name (category)	Rule template
1	Age and diagnosis (incompatibility)	Flag if age does not meet criteria, diagnosis present
2	Age and procedure (incompatibility)	Flag if age does not meet criteria, procedure present
3	Age and drug (incompatibility)	Flag if age does not meet criteria, drug present
4	Gender and diagnosis (incompatibility)	Flag if gender is equal to invalid gender, diagnosis present
5	Gender and procedure (incompatibility)	Flag if gender is equal to invalid gender, procedure present
6	Gender and drug (incompatibility)	Flag if gender is equal to invalid gender, drug present
7	Gender and clinical specialty (incompatibility)	Flag if gender is equal to invalid gender for clinical specialty
8	Drug and diagnosis (incompatibility)	Flag if drug present, diagnosis absent
		Flag if drug absent, diagnosis present
		Flag if drug present, diagnosis present
9	Drug and procedure (incompatibility)	Flag if drug present, procedure absent
		Flag if drug absent, procedure present
		Flag if drug present, procedure present
10	In patient only (IPO) procedure (incompatibility)	Flag if procedure is IPO, location is not inpatient
11	Drug and allergy to drug (incompatibility)	Flag if drug is present, allergy to drug present
12	Drug and interaction drug (incompatibility)	Flag if drug is present, interaction drug present
13	Drug dose (value out of range)	Flag if drug dose is out of range
14	Valid laboratory value (value out of range)	Flag if laboratory result is out of valid range
15	Delta laboratory value (value out of range)	Flag if delta of two laboratory results for a same patient in a time period is out of valid range
16	Observation data elements (out of range)	Flag if observation data elements are out of valid range
17	Demographics data elements (out of range)	Flag if demographics data elements are out of valid range
18	Time sequence (date and time error)	Flag if date 1 is after date 2
19	Laboratory time (date and time error)	Flag if laboratory time presents at an invalid time of a day
20	Date in Future (date and time error)	Flag if date is in future
21	Drug monitoring (incompleteness)	Flag if drug present, drug monitoring absent
22	Drug and laboratory (incompleteness)	Flag if drug present, laboratory absent
23	Diagnosis and laboratory (incompleteness)	Flag if diagnosis present, laboratory absent
24	Drug and continuous procedure (incomplete)	Flag if drug present, continuous procedure absent
25	Laboratory and continuous laboratory (incompleteness)	Flag if laboratory present, continuous laboratory absent
26	Drug and necessary cooccurring drug (incompleteness)	Flag if drug is present, necessary cooccurring drug absent
27	Drug in same class at same time (duplication)	Flag if two drugs in a same class, they are prescribed at same time
28	Procedure duplication (duplication)	Flag if procedure appears more than once

types for their decision-making use. Among 13 enumerated data types, laboratory data was rated adequate (for the participant's use) by more than 90% of the respondents. Nurse notes, intake-output data, and patient weight data were rated to have the least adequacy (20% inadequate and less than 40% adequate). Close to 30% of physician respondents indicated that they do not use nursing notes and data from medical devices for clinical decision making (►Fig. 5).

Qualitative Data Collected during the Result Review Portion of the Participant Interviews

The interviews also surfaced multiple preferences for use of rules in practice. We initially posited that physicians would care only about data errors on their patients or for encounters with providers in their specialty. This was not uniformly the case with one participant directly stating that he cared about all care settings and specialties and another stating that they see patients all over the hospital. Some participants stated that monthly receipt and review of the reports would be "about right" for them. From the interviews, we noted a distinct preference regarding the data elements of interest. Practicing clinicians favored data elements for which results reports could surface documentation issues of potential impact to care or could directly surface issues with care quality. For example, one participant noted that not all laboratories are of equal importance; laboratory values like international normalized ratio used for dosing decisions are probably more important because they have large role in immediate care decisions. Another participant further noted that data elements that are used in clinical decision support algorithms could be more impactful

Table 4 Rule results performance summary

Template name	Number of rules[a]	Number of triggered rules[b]	Discrepancies[c]
Age and diagnosis (incompatibility)	130	47	2,701
Age and procedure (incompatibility)	79	18	2,832
Gender and diagnosis (incompatibility)	5,205	622	5,836
Gender and procedure (incompatibility)	640	15	89
Gender and clinical specialty (incompatibility)	5	2	898
Drug and diagnosis (incompatibility)	18	3	798
In patient only procedure (incompatibility)	1,775	8	240
Drug and interaction drug (incompatibility)	11	7	2,497
Drug dose (value out of range)	13,654	2,451	89,906
Valid laboratory value (value out of range)	37	22	7,256
Delta laboratory value (value out of range)	31	19	536,256
Observation data elements (value out of range)	17	7	57,322
Demographics data elements (value out of range)	2	1	5
Time sequence (date and time error)	151	25	2,563
Laboratory time (date and time error)	4	2	8,303
Date in future (date and time error)	4	3	58
Drug monitoring (incompleteness)	23	16	16,718
Drug and laboratory (incompleteness)	10	6	15,339
Diagnosis and laboratory (incompleteness)	2	2	27,934
Drug and continuous procedure (incompleteness)	3	3	4,863
Laboratory and continuous laboratory (incompleteness)	3	3	33,534
Drug in same class at same time (duplication)	55,244	165	3,732
Procedure duplication (duplication)	3	2	3

Abbreviation: EHR, electronic health record.
Note: Overall results from running the rules over EHR data in the data warehouse at our institution.
[a]The number of rules: the number of records in a knowledge table supporting a rule template.
[b]The number of triggered rules: the number of knowledge table records that identified one or more discrepancies.
[c]Discrepancies: the count of the number of times the data were found to be in exception to the rule.

Table 5 Summary of the structured question portion of the design validation

Question	Yes	No	Unable to judge		
Do the discrepancies indicate a data quality problem?	82%	0%	18%		
Is the data used in clinical decision making?	73%	18%	9%		
	High	**Medium**	**Low**	**Not at all**	**Unable to judge**
Rate the potential to impact the decision	45%	9%	0%	9%	36%
Rate the severity of the data quality problem in terms of potential impact to patient safety	45%	9%	0%	0%	45%
Rate the severity of the data quality problem in terms of potential financial impact to the institution	27%	18%	0%	0%	55%
Rate the severity of the data quality problem in terms of potential impact to regulatory and compliance issues	9%	27%	0%	18%	36%

Note: Results from the structured quantitative portion of the design validation are provided.

because clinicians rely on the algorithms to detect exceptions and make clinical decisions based on them.

We considered representation (visualization) of the rule results separately from the results themselves. The former can inform the design of the results reporting, such as rule result grouping, aggregation, visualization, and other aspects of presentation to users. Overall, 9 out of 11 participants provided result representation preferences or suggestion for results representation improvement. Multiple participants requested a drill-down from the frequency distribution for continuous variables because they preferred visual judgment for outliers over the information reduction associated with

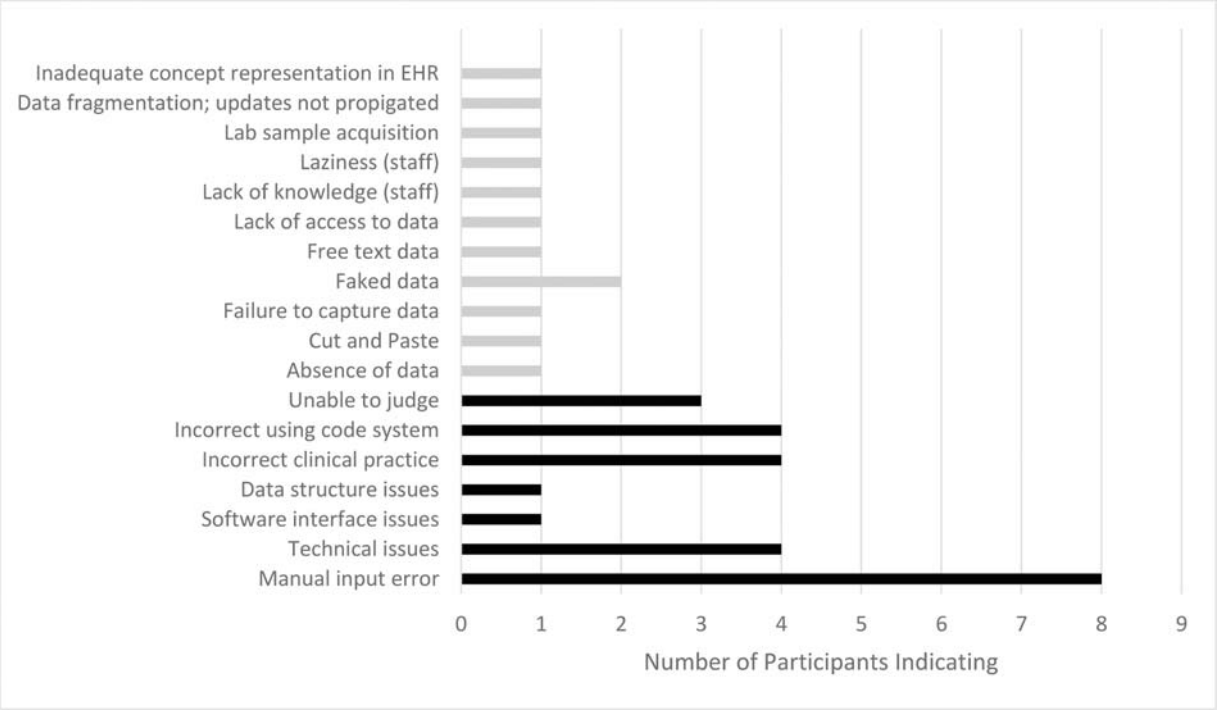

Fig. 3 Dark bars correspond to choices displayed on the interview form from which participants could select all that applied. Light bars correspond to potential reasons for error added by the participants.

deciles. Three participants voiced preferences for different temporal reporting. These included (1) monthly to keep abreast of errors as they occur, (2) by year and 3 years to visually take into account seasonality, and (3) to look for trends year by year.

Overall, 6 of the 11 participants voiced unintended uses of the results reporting. Unintended uses included (1) aiding regulatory compliance, (2) identifying organizational risk, (3) identifying waste and other opportunities for clinical or operational process improvement, (4) to monitor appropriateness of clinical practice, (5) to convert some checks, such as rules based on procedures billable for inpatients only or expected follow-up for laboratory tests, to real-time clinical decision support, and (6) to promote some rules to on-screen error checks for use as guide rails for documentation.

Participant comments helped us discover three aspects where the design fell short. First, several participants commented that the rules and rule results are not specific enough to judge potential impact on clinical decisions. This necessitates planned but not-yet implemented control over alert limits and activating or inactivating rules themselves for individual users. Second, some rules and rule templates were not constrained tightly enough to identifying only physical impossible events. For example, drugs used for different indications, such as, cancer versus immunology are used at significantly different doses; the rules used for this study did not allow sufficient conditionality or specificity of some of the rules for some participants. This again necessitates control over rules themselves for individual users.

Though they did not surface in this study, rule-based data error checking has other limitations. Some data errors are not detectable by rules, that is, wrong but physically plausible errors such as transposition of the last two digits in a measurement that are an inaccuracy, but fall within the physically plausible range. The extent of this blind spot was not characterized by this study. As in other applications of rule-based methods there exists a trade-off between representation depth and return on rule investment. Study participants gave us many suggestions to display rule results more specifically and make rule knowledge and rule templates more specific through deeper or more conditional representation, including (1) separating infants, pediatrics, and adults, (2) separating inpatient versus outpatient, (3) reporting by clinical service line, (4) reporting over recent data only, (5) reporting by data entry source including user role, for example, technician versus nurse or physician, and (6) providing reports for individual providers especially early after having started a new position.

To meet the participant-articulated needs in this study, user-specific conditionality would need to be added to the management of rule templates. User or specialty-specific conditionality suggestions took multiple forms including pruning (turning off) records in the knowledge table to allow exception for accepted practice for a unit or specialty. For example, most, but not all of the time, detecting prescription for two drugs in the same class at the same time would be undesirable; in postoperative care, however, simultaneous opiate orders exist as patients are transitioned from intravenous to oral forms. Likewise, the same compound may be used at very different dosages for different indications. Similarly, the rule templates themselves may need extensibility with respect to conditionality. For example, significantly abnormal values for measured physical quantities are expected to cooccur with certain diagnoses. In these cases,

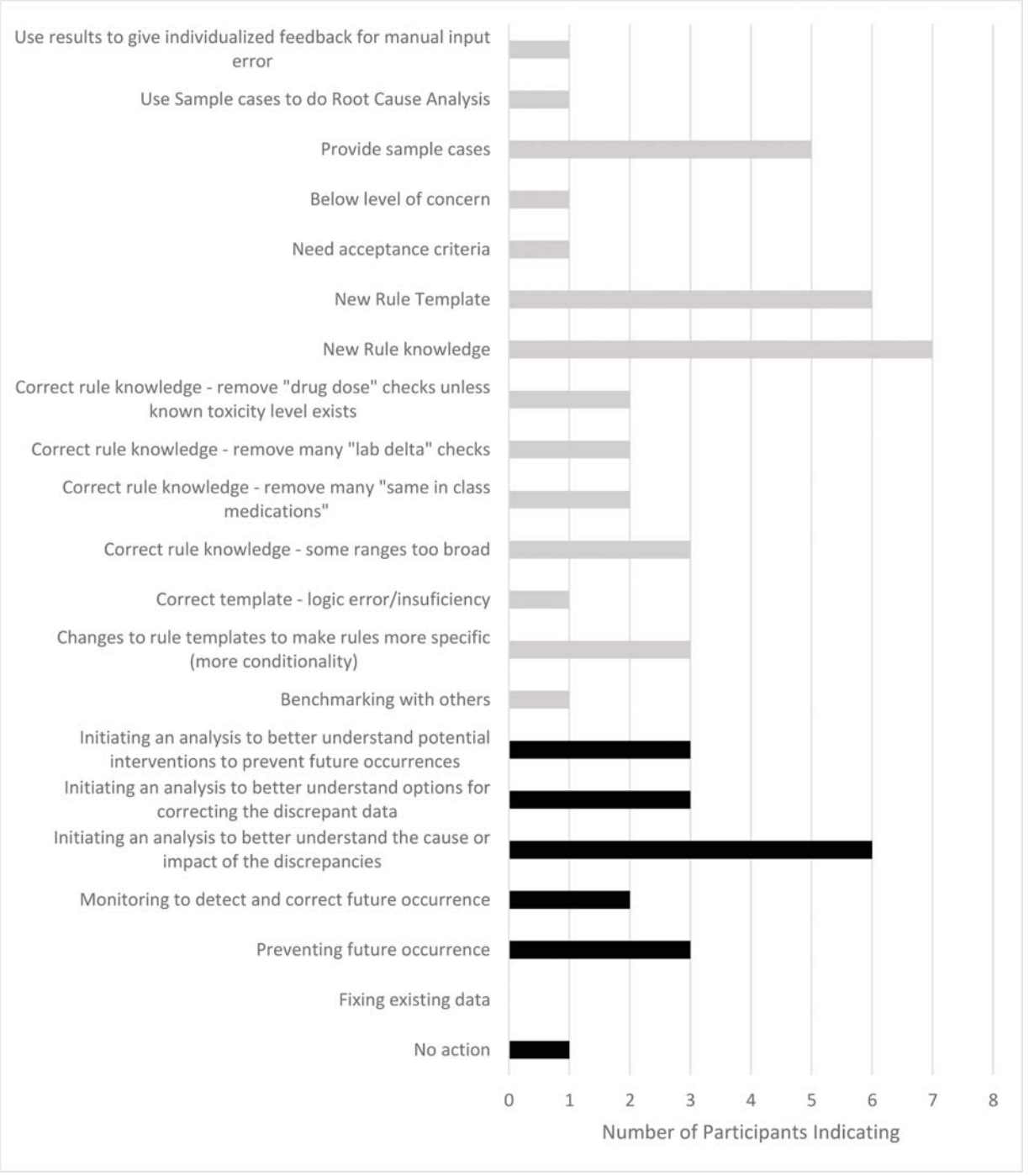

Fig. 4 Dark bars correspond to choices displayed on the interview form from which participants could select all that applied. Light bars correspond to intended actions to be taken added by the participants during the interview.

"except where" conditions would be needed to allow the rule template to conditionally work over different diagnoses. The data structure employed here is extensible toward storing such customization. Making the likely needed use of user-specific conditionality will require additional governance for life-cycle management of user-specific configurations and providing for their graceful existence as rule templates and knowledge tables change over time.

The list of rules and suggestions for customization to rulers and knowledge from study participants is likely not complete over all intended users of such a system. Additional studies are needed to better characterize the extent of user-specific conditionality needed for rule results to be meaningful and useful in practice. At some point the effort associated with comprehensive and deep rule-based medical knowledge representation will become too high or not possible for data error identification return on investment. Additional studies are needed to locate this threshold and the value of rule-based data quality assessment at the threshold. However, the gross results here demonstrated that rule-

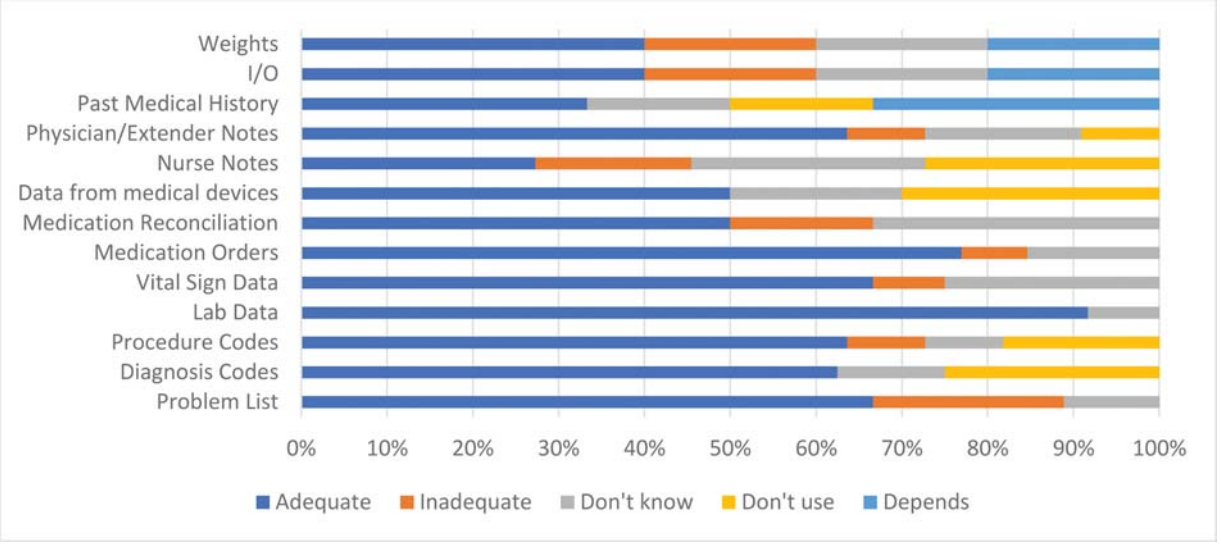

Fig. 5 Differences in perceived data quality across different domains of electronic health record (EHR) data are displayed. Participants were asked to rate the adequacy of different data types for their use in decision-making.

based data quality management is an improvement over snapshot (one and done) approaches to data quality assessment and data profiling.

Three participants voiced four additional barriers to organizational use. One participant (system owner) stated that their group did not have the resources to chase down data errors identified by rules and that they did not see a way to get additional resources for an activity like that. A second organizational barrier noted was the potential for alert fatigue if data quality rules were pushed to the EHR screens at data capture. Another participant noted lack of clear organizational responsibility for data quality in general and for reviewing, prioritizing, and addressing rule results over time. Because this is a new function for health care, these roles and responsibilities are not clear. A fourth organizational barrier well noted in the Clinical Decision Support literature is that rule knowledge changes over time necessitating ongoing rule management activities.[33]

Discussion

Through consolidation of existing rule sets and systematic approaches to identify clinical data elements and applicable multivariate data quality rules, we distilled a set of 63,397 rules across the 28 rule templates described here. Based on our experience in this study, though the number of possible rules is combinatorially large, the actual combinations of interest and utility seem tractable.

The knowledge tables for rule templates must be acquired and brought forward in computationally useful form. For example, narrative text of age and gender incompatible diagnoses is not directly computationally useful whereas data tables of the same that are documented, maintained, versioned, discoverable, and available for automated use are directly consumable by programs and support several thousand data error checks. As previously reported, identification of knowledge sources was challenging. Knowledge sources did not exist and could not easily be identified for 5 of the 28 rule templates initially identified. While this remains a challenge today, collaborative approaches to building and maintaining these knowledge sources are possible as evidenced by multiple publicly identified knowledge sources. The approach of managing rule templates and associated knowledge tables vastly reduces the amount of effort required to maintain rule-based knowledge. Knowledge acquisition, however, is still required, there are definitely more "rules to be discovered" than exist in our set. Based on approaching (not reaching) saturation in our small study, we posit that considering different data sets, new use cases, and exposure of the rules to new specialties and health care functions will help climb the asymptote of useful and possible rules for which knowledge sources can be created or obtained.

Regarding design validation, semistructured interviews were conducted. Semistructured interviews significantly enhanced the information gained from the study; participants identified things that we did not. Semistructured interviewing inspired participants to give qualitative ideas outside of the questionnaire. However, some participants were "unable to judge" whether the discrepancies indicate data quality problem and what effects would be caused by the discrepancies. There are four possible reasons for "unable to judge" responses. First, several participants commented that the rule and rule results are not specific enough in that there were no actual example instances of the actual data to illustrate the potential problem. Second, some rules and rule templates are not identifying physical impossible events or were too broad, leaving some out. Third, there were only 11 participants. The participant panel is too limited and did not include multiple roles present on care teams. Last, rule-based approaches have some limitations for data error checking—(1) though these did not surface in the interviews, we know from first principles that there are some data errors that are not detectable by rules, for example, some laboratory

and observation errors are not out of valid range but recorded incorrectly, and (2) we do not know where the "sweet spot" is with respect to representation depth and return on rule investment. We do not know how significant these are, or how detrimental this gap is to rule-based data quality management as an approach. The first two problems are tractable, and participants gave us many suggestions to display rule results more specifically and make rule knowledge and rule templates more specific. To mitigate the third problem, nurses and other members of the health care team should be included in future studies. Nurses make significant contributions to data documentation and may have more ideas about data discrepancies and how to resolve them. Other team members may offer additional ideas. Regarding the last problem, at some point effort associated with comprehensive and deep medical knowledge representation is too high or not possible for data error identification return on investment. We do know, however, that rule-based data quality management is an improvement over snapshot (one and done) approaches and data profiling.

Another limitation is the fuzzy boundary between data error and odd clinical practice or physiological outliers. For example, a drug may be labeled for use in adults, but a doctor may prescribe it to a 14-year-old. Such off-label uses are common. Thus, if we followed the product label as a strict rule, we would in all likelihood identify many more instances of off-label use than data errors. For this reason, we have segregated the initial rules as identifying a physical impossibility versus those that are possible but implausible. It seems reasonable that rules identifying instances of physical impossibility are much more likely to identify data errors. The rules identifying possible but implausible cases require validation prior to use, that is, some indication that they correlate strongly with known data errors.

A possible application for the rule-based data quality assessment in the future could be as a method to reduce documentation burden. As increasing use of EHR, documentation burden is considered as a significant time-consuming component of EHR use for practitioners. Improving data quality is one of the important strategies to reduce documentation burden. Rule-based data quality method will contribute significantly on clinical decision support and alerting. We still have a long way to go to craft each rule to make sure it is not broad or narrow, so that it overcomes alert fatigue challenge.

Though only a small study, the empirical validation yielded evidence (1) supporting the ability of a rules-based approach to identify potentially important data quality problems, and (2) indicating potential organizational willingness to further evaluate rule-based data quality management. All but one participant was interested in the results. All participants indicated willingness to attend a follow-up meeting to review results of actions taken. Participants give constructive comments and ideas to improve the system. We gained so much useful information that a second iteration with participants to review the results of actions taken would be a logical next step for future research.

Conclusion

Assessing the quality of EHR data is necessary to improve data quality yet doing so is an uncharted territory in health care. To address the three objectives, this study provides a potentially scalable framework with which data quality rules can be organized, shared as rule templates and knowledge tables, and applied in health care facilities to identify data errors. To validate the initial system design, interviews with 11 individual physicians and information system owners demonstrated that the system identifies data quality problems of concern to key stakeholders including physicians, secondary data users, and information system owners. While there is significant additional work to be done in this area, the exploration of the rule template and associated knowledge tables approach here shows the approach to be possible and potentially scalable. This research directly evaluated the potential value of rule-based data quality assessment results and found evidence to support further development and investigation of the approach.

Clinical Relevance Statement

This study demonstrates a scalable framework with which data quality rules can be implemented and managed in health care facilities to identify data errors. The data quality problems identified in implementation site were important enough to prompt action requests from clinical and operational leaders.

Multiple Choice Questions

1. Which of the following methods were not used in this study for acquisition of rule knowledge?

a. Application of rule development methods used in research.
b. Crowd sourcing.
c. Talking to experts.
d. Machine learning.

Correct Answer: The correct answer is option d. Machine learning is able to discover column constraints rules but not very helpful to deep medical knowledge rules that concern the physicians.

2. Which of the following is not one of the rule categorizations in EHR secondary use data quality?

a. Attribute domain constraints.
b. Relational integrity.
c. Data pattern consistency.
d. Historical integrity.

Correct Answer: The correct answer is option c. Data pattern consistency is used in data profiling but not necessarily in rule-based data quality assessment.

3. Which of the following is not one of the rule-based method limitations?

a. Fuzziness.
b. There is no evidence of value.

c. Not scalable.
d. Some data errors that are not detectable by rules.

Correct Answer: The correct answer is option c. Rule-based method is developed as a scalable framework.

4. Which of the following is incorrect according to the results of this study?

a. Rule-based method identified significant data errors.
b. Data discrepancies do not have any effect on clinical decision making.
c. Participants provided multiple ideas for rule improvement.
d. Most of the participants were willing to take actions to identify data discrepancies.

Correct Answer: The correct answer is option b. The study results show that most of the participants agreed that data discrepancies can impact clinical decision making.

Protection of Human and Animal Subjects

The authors declare that human and/or animal subjects were not included in the project.

Conflict of Interest

None declared.

References

1 Forrest WH Jr, Bellville JW. The use of computers in clinical trials. Br J Anaesth 1967;39(04):311–319
2 Kronmal RA, Davis K, Fisher LD, Jones RA, Gillespie MJ. Data management for a large collaborative clinical trial (CASS: Coronary Artery Surgery Study). Comput Biomed Res 1978;11(06): 553–566
3 Knatterud GL. Methods of quality control and of continuous audit procedures for controlled clinical trials. Control Clin Trials 1981;1 (04):327–332
4 Norton SL, Buchanan AV, Rossmann DL, Chakraborty R, Weiss KM. Data entry errors in an on-line operation. Comput Biomed Res 1981;14(02):179–198
5 Cato AE, Cloutier G, Cook L. Data entry design and data quality. 1985
6 Bagniewska A, Black D, Molvig K, et al. Data quality in a distributed data processing system: the SHEP Pilot Study. Control Clin Trials 1986;7(01):27–37
7 DuChene AG, Hultgren DH, Neaton JD, et al. Forms control and error detection procedures used at the Coordinating Center of the Multiple Risk Factor Intervention Trial (MRFIT). Control Clin Trials 1986;7(03):34S–45S
8 Crombie IK, Irving JM. An investigation of data entry methods with a personal computer. Comput Biomed Res 1986;19(06): 543–550
9 Fortmann SP, Haskell WL, Williams PT, Varady AN, Hulley SB, Farquhar JW. Community surveillance of cardiovascular diseases in the Stanford Five-City Project. Methods and initial experience. Am J Epidemiol 1986;123(04):656–669
10 Houston L, Probst Y, Yu P, Martin A. Exploring data quality management within clinical trials. Appl Clin Inform 2018;9 (01):72–81
11 Joukes E, de Keizer NF, de Bruijne MC, Abu-Hanna A, Cornet R. Impact of electronic versus paper-based recording before EHR implementation on health care professionals' perceptions of EHR use, data quality, and data reuse. Appl Clin Inform 2019;10(02):199–209
12 Reimer AP, Milinovich A, Madigan EA. Data quality assessment framework to assess electronic medical record data for use in research. Int J Med Inform 2016;90:40–47
13 Huser V, DeFalco FJ, Schuemie M, et al. Multisite evaluation of a data quality tool for patient-level clinical data sets. EGEMS (Wash DC) 2016;4(01):1239
14 Sengupta S, Bachman D, Laws R, et al. Data quality assessment and multi-organizational reporting: tools to enhance network knowledge. EGEMS (Wash DC) 2019;7(01):8
15 Lee K, Weiskopf N, Pathak J. "A framework for data quality assessment in clinical research datasets." AMIA Annual Symposium Proceedings. Vol. 2017. American Medical Informatics Association; 2017
16 Houston ML. "Defining and Developing a Generic Framework for Monitoring Data Quality in Clinical Research." AMIA Annual Symposium Proceedings. Vol. 2018. American Medical Informatics Association; 2018
17 Feder SL. Data quality in electronic health records research: quality domains and assessment methods. West J Nurs Res 2018;40(05):753–766
18 Pezoulas VC, Kourou KD, Kalatzis F, et al. Medical data quality assessment: on the development of an automated framework for medical data curation. Comput Biol Med 2019;107:270–283
19 Scholte M, van Dulmen SA, Neeleman-Van der Steen CW, van der Wees PJ, Nijhuis-van der Sanden MW, Braspenning J. Data extraction from electronic health records (EHRs) for quality measurement of the physical therapy process: comparison between EHR data and survey data. BMC Med Inform Decis Mak 2016;16(01): 141
20 Callahan TJ, Bauck AE, Bertoch D, et al. A comparison of data quality assessment checks in six data sharing networks. EGEMS (Wash DC) 2017;5(01):8
21 Ehrlinger L, Rusz E, Wöß W. "A Survey of Data Quality Measurement and Monitoring Tools." arXiv preprint arXiv:1907.08138; 2019
22 Carlson D, Wallace CJ, East TD, Morris AH. Verification & validation algorithms for data used in critical care decision support systems. Proc Annu Symp Comput Appl Med Care 1995;•••:188–192
23 Brown PJ, Warmington V. Data quality probes-exploiting and improving the quality of electronic patient record data and patient care. Int J Med Inform 2002;68(1-3):91–98
24 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:S21–S29
25 Wang RY, Strong DM. Beyond accuracy: what data quality means to data consumers. J Manage Inf Syst 1996;12(04):5–33
26 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(01): 1244
27 Hart R, Kuo MH. Better data quality for better healthcare research results - a case study. Stud Health Technol Inform 2017;234(234): 161–166
28 Skyttberg N, Chen R, Blomqvist H, Koch S. Exploring vital sign data quality in electronic health records with focus on emergency care warning scores. Appl Clin Inform 2017;8(03):880–892
29 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(04):1012–1021
30 Bauck A, Bachman D, Riedlinger K, et al. C-A1-02: Developing a Structure for Programmatic Quality Assurance Checks on the Virtual Data Warehouse. Clin Med Res 2011;9(3-4):184
31 Curtis LH, Weiner MG, Beaulieu NU, Rosofsky RA, Woodworth TS, Boudreau DM. 2012. Mini-Sentinel year 1 common data

model—data core activities. 2012. Available at: http://www.mini-sentinel.org/data_activities/details.aspx?ID=128. Accessed July 21, 2020
32 Tenenbaum JD, Christian V, Cornish MA, et al. The MURDOCK Study: a long-term initiative for disease reclassification through advanced biomarker discovery and integration with electronic health records. Am J Transl Res 2012;4(03):291–301
33 Jenders RA, Huang H, Hripcsak G, Clayton PD. Evolution of a knowledge base for a clinical decision support system encoded in the Arden Syntax. Proc AMIA Symp 1998:558–562