Rule Templates and Linked Knowledge Sources for Rule-based Information Quality Assessment in Healthcare

(Research-in-progress, IQ Assessment, Measures and Models)

Zhan Wang and Meredith Nahm Zozus
University of Arkansas at Little Rock, University of Arkansas for Medical Sciences
zxwang2@ualr.edu, mzozus@uams.edu

Abstract: Measuring and managing information quality in healthcare has remained largely uncharted territory with few notable exceptions. A rules-based approach to data error identification was explored through compilation of over 6,000 data quality rules used with healthcare data. The rules were categorized based on topic and logic yielding twenty rule templates and associated knowledge tables used by the rule templates. Knowledge sources for the knowledge tables were sought and identified for eleven of the twenty rule templates and have to be created for the remaining nine. This work provides a framework with which data quality rules can be organized and shared as rule templates and knowledge tables. 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 scalable.

INTRODUCTION

Information Quality Assessment (IQA) in healthcare and health-related research is not new. The earliest reports of data processing in clinical research included accounts of data checking (Forrest et al. 1967, Kronmal et al. 1978, Knatterud 1981, Norton et al. 1981, Cato 1985, Bagniewska et al. 1986, DuChene et al. 1986, Crombie et al. 1986, Fortmann et al. 1986). In the therapeutic development industry, with the 1962 Kefauver Harris Amendment to the Food, Drug and Cosmetic act a New Drug Application (NDA) had to show that a new drug was both safe and effective and companies began to use rules to check data submitted in NDAs for consistency. In fact, fear that notice of an errant data value would substantially delay a regulatory submission prompted a process in the therapeutic development industry of running often hundreds of rules for a clinical study and contacting the data provider in attempts to resolve each discrepancy against the source, i.e., the medical record (Estabrook et al. 1999). The discrepancies often numbered in the tens of thousands for a small study of a few hundred patients. It is not uncommon for 10-30% of the cost of a clinical study to be spent on data cleaning (Eisenstein et al. 2005). This practice, albeit mediated by today's risk-based approaches continues in therapeutic development and is the standard of practice (Society for Clinical Data Management 2013).

In healthcare, however, there is usually no source against which to identify or resolve data discrepancies.

With alert fatigue common for critical decision support algorithms, few would consider flagging data discrepancies as clinicians chart patient information. Further, aside from being considered by physicians in decision-making and used by other members of care teams widespread secondary use of routine electronic clinical data is a fairly recent phenomenon. The current national emphasis on secondary use of healthcare data for research has been prompted by the large upswing in Electronic Health Record (EHR) adoption over the last decade and federal support for institutional clinical data repositories over the same period. Today the value of data cleaning in healthcare has not been well studied or articulated. Though there have been reports of fixing data quality problems identified through attempts at data use, institutions have been hesitant to allocate even limited resources toward systematic IQA and improvement. For these reasons IQA in healthcare has received relatively little attention as an institutional priority or as a research agenda.

The research in healthcare IQA described here is motivated by (1) recent increases in national attention towards secondary use of healthcare data for research through broad programs such as the National Institutes of Health (NIH) funded Healthcare Systems Research Collaboratory the NIH funded Clinical and Translational Science Awards and the Patient-Centered Outcomes Research Institute funded through the Affordable Care Act, (2) national emphasis on use of healthcare data for organizational performance assessment and improvement, i.e., Accountable Care Organizations, (3) almost ubiquitous availability of rich healthcare data in most institutions, and (4) lack of methods for IQA, specifically assessment of data accuracy, demonstrated effective in healthcare. We seek to ultimately demonstrate and evaluate rule-based data cleaning in healthcare.

BACKGROUND

There are limited accounts of rule-based data quality assessment in healthcare. In early work, Carlson et al. (1995) 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. Though the total number of rules was not reported, based on the data elements and the reported rule examples there were likely a few hundred rules. Data values found to be discrepant were censored from the database or replaced with imputed values prior to use. To our knowledge, this is the earliest report of rule-based data discrepancy identification in healthcare.

In 2003, Brown et al. presented data quality probes to find data quality problems and improve data quality in EHRs. Data errors can happen at every step in a clinical encounter including assessment, data entry, data retrieval, information interpretation and action. Data quality probes consisted of a rule implemented as a query in a clinical information system to find the inconsistency between two or more associated data items. For example, if a lipid level result was expected for all patients with ischemic heart disease, the data quality probes would identify any patients without a lipid level. The examples given by Brown et al. were checks of clinical consistency similar to the ischemic heart disease example and included rules for a diabetes diagnosis with no recorded glycosylated hemoglobin (HbA1c), diagnosis of asthma or regular prescription for an inhaler with no record of Peak Expiratory Flow Rate (PEFR), cases of anti-glaucoma treatment with no recorded diagnosis and cases of tamoxifen prescription with no relevant diagnosis. EHR information quality was then tracked using the number of flagged exceptions to the rules and results were reported to clinicians, to encourage improvement (Brown et al. 2001, Brown et al 2002, Brown et al 2003). To our knowledge, this is the earliest report of using rule templates in identification of discrepant data in healthcare.

In 2012, Kahn et al. proposed an initial "fit-for-use" framework for data quality assessment (DQA) in EHR-based clinical research. Five categories of DQA rules based on types of data checks were offered and include: (1) Attribute domain constraints defined as rules that validate individual attribute values based on restrictions for allowed values; (2) Relational integrity rules defined as rules that ensure accurate relationships between entities (tables), instances (records), and attributes (fields) across multiple tables, (3) Historical data rules defined as rules involving time varying data, (4) State-dependent object rules defined as rules that ensure that changes in the lifecycle of an object follow expected transitions, and (5) Attribute dependency rules defined as rules for describing real-world objects. Later in 2016, through a large collaborative endeavor, Kahn et al. expanded the work to a conceptual model for rule-based data quality checks categorizing DQA rules into five categories (value conformance, relational conformance, computational conformance, completeness, and three types of plausibility – uniqueness plausibility, atemporal plausibility and temporal plausibility) operationalized through the aforementioned types of data checks and over two contexts, *verification* - not dependent on an external reference and *validation* - dependent on an external reference (Khan et al. 2016). Also in 2016, Dziadkowiec et al. successfully applied the 2012 framework to discrepancy identification and cleaning emergency department data extracted for secondary use. In a later publication by Callahan et al., the Khan et al. 2016 categories were applied to rule sets from six research networks using EHR data to test the categorization scheme and to assess rule variability across different organizations (Callahan et al. 2017).

Most recently, Haart and Kuo (2017) reported rule-based discrepancy identification and resolution in healthcare data used for direct patient care and management of health services. They describe development and implementation of a data warehouse-based system at Island Health of Canada to capture, measure and report on data quality. (Haart and Kuo 2017) In their region, provision of home and community care services requires federal and provincial reporting and is subject to data quality acceptance criteria. To assure consistently meeting these, Island Health initiated a business process for data quality assessment whereby all data in the data warehouse are evaluated against 200-300 data quality rules. Any records failing validation are reported back to the responsible staff for correction. Once corrections have been made the data are re-assessed and submitted. They reported a greater than 50% decrease in rejected records across three domains in six 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).

While all aforementioned work assess the quality of EHR data, the latter (Kahn et al., Dziadkowiec et al., Callahan et al., Haart and Kuo) were motivated by secondary data use rather than data use in patient care. There are certainly many cases where clinical data are assessed and data quality issues are addressed as part of extract transform and load process from source systems into data warehouses. However, these processes tend to use the results to standardize data, monitor incoming records and reject or flag nonconforming records based on format or outlying values rather than employ rules to identify errant data. Further, DQA as part of ETL-based processes don't always report back to data sources to improve institutional quality of source data. Identifying errant data, tracking the results over time, and using results in data quality improvement efforts are our goal. Our research seeks scalable methods of assessing and monitoring data quality over time to improve data quality in areas where such improvement is of value to primary or secondary data users.

Inspired by the examples of rule-based DQA in healthcare data, their effectiveness in the Island Health case, and their long term effectiveness at achieving similarly low discrepancy rates in the therapeutic development industry we sought to (1) identify as many rules as possible, (2) to separate them according to whether they identified a physical impossibility or an unlikely but physically possible inconsistency, (3) to devise a scalable approach to rules management, and (4) to achieve rule interoperability, i.e., to share and reuse executable rules at multiple institutions.

METHODOLOGY

Identification rules

To identify candidate rules, we first looked to existing rule sets. These included the publically available Observational Medical Outcomes Partnership (OMOP) rules, the Healthcare Systems Research Network (previously HMORN) rules (Bauck et al. 2011), and the Mini-sentinel data checking rules (Curtis et al. 2012), and publicly available age and gender incompatible diagnosis and procedure lists from third party payers. We also utilized rules written for an internal project using multi-site EHR data (Tenenbaum et al. 2013). All combined, these activities produced over 6,000 individual logic statements or rules.

Management of this many initial rules conflicted with our goal of scalable rule management and maintenance over time. Inspired by the rule abstraction in Brown's work, we evaluated each rule and sorted the rules. Rules sharing a topic and logic structure were abstracted into a single rule template. An example of such a rule template is *Flag the record if GENDER is equal to some invalid gender and DIAGNOSIS is equal to a corresponding invalid diagnosis*. The clinical information in the rules (in the example the list of gender – diagnosis incompatibilities) was extracted and compiled into a knowledge table against which the rule template runs. This categorization yielded twenty different rule templates. The twenty rule templates were further categorized into five higher-level types: incompatibility, value out of range, temporal sequence error, incompleteness and duplication. These correspond to the following Kahn et al. 2016 criteria value conformance, relational conformance, completeness and plausibility.

Incompatibility means one data value is logically incompatible with another data value, such as patient gender is incompatible with a diagnosis. Some incompatibility rule templates have more complicated logic relationships than simple incompatibility. In these cases, we used a 2x2 table to document their logic relationships. For example, there are four relationships possible between drug and diagnosis (Figure 1), three of which are useful for identifying potentially errant data depending on the particular drug and diagnosis. The rule template for the top left quadrant (Figure 1) would be, *Flag if drug is present and diagnosis is present*. For example, if thalidomide, a known teratogen, is prescribed for a pregnant patient. The rule template associated with the top right quadrant is, *Flag if drug is present and diagnosis is absent*. For example, long-acting nitrates is present but there is no diagnosis of heart disease. The rule template associated with the bottom left quadrant is, *Flag if drug is absent and diagnosis is present*. For example, Aspirin is absent but a diagnosis of ischemic heart disease is present.

		DIAGNOSIS	
		Present	Absent
DRUG	Present	1	1
	Absent	1	\

1: Flag; 0: Do not Flag; \: scenario is not useful for IQA.

Figure 1: Logic Relationship between Drug and Diagnosis

The out of range value template when used to identify data errors, flags data values compatible with life or grossly incompatible with product labeling, such as a drug dose incompatible with product labeling, a lab result incompatible with life, or an impossible date of birth. Typos or wrong units could cause these issues and they are highly likely to be actual errors.

Temporal sequence templates focus on dates occurring in an invalid order. For example, a clinical encounter date before the date of birth for an adult.

Incompleteness is defined as occurrence of a data value that is expected but missing. While we didn’t include univariate checks for missing values because they are easily quantified through data profiling approaches, we did include multivariate and record-level incompleteness checks, i.e., when one record is present, but a corresponding ad expected record is absent. For example, a procedure is present but there is no corresponding encounter record.

Lastly, the duplication template identifes multiple occurrences of events that can physically happen only once, for example a patient with two hysterectomies.

Identification of knowledge tables

As described above, we compiled or identified a knowledge table to support each rule template. Figure 2 shows the structure of the knowledge table built for gender and diagnoses incompatibility. The actual knowledge table for gender-diagnosis incompatibility contains 5,250 records. The purpose of the knowledge table is to condense what may eventually be thousands of individual rules down to one template and a knowledge table that can be expanded or edited as medical coding systems change or new knowledge becomes available. In this way, we purposely separated the rules from the knowledge. Twenty rule templates are easier to develop and maintain than 6,000 rules.

ICD9 Code	ICD10 Code	Invalid Gender	Preferred Name (ICD-9)
181	C58	M	Malignant neoplasm of placenta

…			

Figure 2: Knowledge Table of Gender and Diagnoses Incompatibility

RESULTS

The rule identification resulted in over 6,000 rules which were compressed through the use of knowledge tables to twenty rule templates. Fifteen templates pertained to incompatibility while the remaining five templates fell into the value out of range, temporal sequence error, incompleteness and duplication categories. (Table 1)

The knowledge acquisition phase identified knowledge sources for eleven of the twenty rule templates. (Table 1) As others have noted (Smith et al. 2014) well-curated knowledge sources in biomedicine are not common today. Knowledge sources were classified as structured with little processing required, semi-structured with text processing required, unstructured with significant processing required such as natural language processing, and no knowledge source identified (Table 1).

Structured knowledge sources not requiring significant processing were identified for five rule templates (Table 1). Of the structured knowledge sources, publically available rule sets used by third party payers covered four rule templates, and the Drug-drug Interaction Evidence Ontology (DIDEO) covered a fifth rule template.

One semi-structured knowledge source, the Structured Product Label (SPL) for medicinal products, a semi-structured xml file (one file per medicinal product) was identified and covered six rule templates. However, acquisition of the knowledge from SPL requires processing narrative text in the xml files. For example, the knowledge of drug dose out range comes from the text description in the Dosage and Administration section of SPL. The following is an example of text from an SPL file: "*The usual dosage for persons 13 years of age and over is 1 mL buprenorphine hydrochloride injection (0.3 mg buprenorphine) given by deep intramuscular or slow (over at least 2 minutes) intravenous injection at up to 6-hour intervals, as needed"*. This description is unstructured and the content in this section is highly variable from SPL file to SPL file. Thus, the drug dose knowledge contained in SPL requires natural language processing for extraction into a knowledge table.

Knowledge sources did not exist for nine of the rule templates. For five of the rule templates, we were able to build the knowledge tables from the individual rules and note that these knowledge tables are based on identified rules, for example out of range value rules. As such they are incomplete and could easily grow to contain an order of magnitude more records. Four of the rule templates remain without knowledge tables. These will also need to be built to support use of the templates. These four templates b rely on more clinically detailed knowledge and are at risk of change over time.

Some knowledge tables such as age – diagnosis incompatibility exist today and are maintained and publically available. Others such as gender - clinical specialty incompatibility (Table 2) consist of only a few records and thus take little effort to create and maintain. Still others such as valid ranges for physical quantities while now fewer than 100 records based only on the identified rules will ultimately contain thousands of records to support multiple combinations of measurements, applicable genders and ages, and corresponding units. Such a publically available knowledge source for valid ranges of measurements does not exist today. While this effort is tractable and we have started, the initial knowledge tables are far from complete.

Template Name (Category)	Rule Template	Knowledge Source	Extent of Processing Required
Age and DIAGNOSIS (incompatibility)	Flag if AGE does not meet criteria, DIAGNOSIS is present.	Payer rule sets; ICD-9/10	Exists in structured format & requires little processing
Age and PROCEDURE (incompatibility)	Flag if AGE does not meet criteria, PROCEDURE is present.	Payer rule sets; CPT	Exists in structured format & requires little processing
Age and DRUG (incompatibility)	Flag if AGE does not meet criteria, DRUG is present.	Indication and Usage of SPL	Exists and requires text or XML processing
Gender and DIAGNOSIS (incompatibility)	Flag if GENDER is equal to invalid gender, DIAGNOSIS is present.	Payer rule sets; ICD-9/10	Exists in structured format and requires little processing
Gender and PROCEDURE (incompatibility)	Flag if GENDER is equal to invalid gender, PROCEDURE is present.	Payer rule sets; CPT	Exists in structured format and requires little processing
Gender and DRUG (incompatibility)	Flag if GENDER is equal to invalid gender, DRUG is present.	Indication and Usage of SPL	Exists and requires text or XML processing
Gender and clinical specialty (incompatibility)	Flag if GENDER is equal to invalid gender for clinical specialty.		Did not exist; we built it and it is expandable
DRUG and DIAGNOSIS (incompatibility)	Flag if DRUG present and DIAGNOSIS absent. Flag if DRUG absent and DIAGNOSIS present. Flag if DRUG present and DIAGNOSIS present.	Indication and Usage of SPL	Exists and requires text or XML processing
DRUG and PROCEDURE (incompatibility)	Flag if DRUG present and PROCEDURE absent. Flag if DRUG absent and PROCEDURE present. Flag if DRUG present and PROCEDURE present.	Indication and Usage of SPL	Exists and requires text or XML processing
DRUG and ALLERGY TO DRUG (incompatibility)	Flag if DRUG is present and ALLERGY TO DRUG is present.	Adverse Reaction section of SPL	Exists and requires text or XML processing
DRUG and INTERACTION DRUG (incompatibility)	Flag if DRUG is present and INTERACTION DRUG is present. Flag if DRUG is present and INTERACTION DRUG is present.	Drug-drug Interaction Evidence Ontology	Exists in structured format and requires little processing
DRUG and NECESSARY CO-OCCURING DRUG (incompatibility)	Flag if DRUG is present and NECESSARY CO-OCCURING DRUG is absent. Flag if DRUG is absent and NECESSARY CO-OCCURING DRUG is present.		No knowledge source exists
DRUG and RELEVANT LAB VALUE (incompatibility)	Flag if DRUG is present and RELEVANT LAB VALUE does not meet criteria.		No knowledge source exists
DRUG and RELEVANT DATA ELEMENT (incompatibility)	Flag if DRUG is present and RELEVANT DATA ELEMENT is absent.		No knowledge source exists
DIAGNOSIS and RELEVANT DATA ELEMENT (incompatibility)	Flag if DIAGNOSIS is present and RELEVANT DATA ELEMENT is absent.		No knowledge source exists
DRUG DOSE (value out of range)	Flag if DRUG DOSE is out of range for DRUG.	SPL Dosage & Administration	Exists and requires text or XML processing
Measured Physical Quantity Result (value out of range)	Flag if Measured Physical Quantity Result is outside valid range.		Does not exist; we built it and it is expandable
Two Dates in invalid order (temporal sequence error)	Flag if Date 1 and Date 2 are in an invalid order.		Does not exist; we built it and it is expandable
Multivariate / conditional completeness violation	Flag if Table 1 is present and Table 2 is absent.		Does not exist; we built it and it is expandable

(incompleteness)			
Illogical duplication (duplication)	Flag if PROCEDURE appears more than once.		Does not exist; we built it and it is expandable

Table 1: Rule Templates and Corresponding Knowledge Table Sources

DISCUSSION

Our rule templates run on individual data values rather than columns. Thus while the rules described here are stated at the data element level like column statistics in data profiling tools, a result for each rule is produced for each data value. Operating at the data value-level vastly increases the number of possible rules. Doing so is necessary to gain the ability to apply any logically possible check to a data value and enable multiple checks of any one data value rather than only column statistics. Many data profiling tools are extendable through user-defined rules. In this way, the rules-based approach proposed here is compatible with and expands data profiling tools.

Medicine presents significant scalability challenges to DQA. Data elements in common electronic health record systems number over one hundred thousand and new medical knowledge is generated every day requiring new data and new ways of understanding existing data. Our rule template approach is an attempt at scalability of rule management through managing rules at a high level of abstraction (the template level) while gaining rule results at the lowest level of abstraction (the data value-level). Compressing the rules into templates with associated knowledge tables provides scalability in that additional checks can be added by adding records to knowledge tables rather than by writing new rules as computer programs. Use of a common data model further extends scalability to multiple organizations through enabling sharing of executable rule templates. Further, due to the aforementioned data diversity and volume scalability challenges in medicine, commercial knowledge sources are common, for example health systems license commercially curated and managed drug formularies so that their list of available drugs including variants of brand names, forms, dosages, and packaging are always current. Publically available knowledge sources are becoming available for less volatile information. If and when data quality management is recognized as a need in healthcare, incentive will exist for sharable knowledge sources to support DQA.

While the lack of knowledge sources for multiple templates is unfortunate, it is not a roadblock. First, we were able to build the knowledge sources for five of the rule templates. We note that acquiring new knowledge, updating and maintaining these knowledge tables is no more time consuming than the current practice in clinical decision support of managing individual rules, and probably less time intensive due to having to only update a table rather than program and test a new rule. Though we have not built significant knowledge tables for the four more clinically intensive rule templates, this philosophy holds here as well. Further, it is possible to work with Standards Development Organizations (SDOs) such as Health Level Seven (HL7) that maintains the SPL standard to add more structure to expand use of the standard as a computable knowledge source; doing so would cover three additional rule templates. Lastly, as of this writing commercial knowledge sources were not explored. Commercial knowledge sources are being developed to better support health information systems and it is possible that they may be leveraged for rule-based data cleaning as well.

Although this work evaluated over 6,000 rules, a limitation to use today is that many possible and useful rule templates have not yet been identified. To mitigate the problem and allow identification and adding new rule templates over time, a generic template leveraging executable syntax can be used for two (or more) value logic inconsistencies, e.g., *Flag if Table_X.column_Y value is inconsistent with Table_Z.column_W value.* Some examples identified from the initial rule set are provided in Table 2.

Table 1	Table 1 Value	Table 2	Table 2 invalid value
death_date	Null	death_indicator	Yes
death_date	Not Null	death_indicator	No
tobacco use indicator	No	cigarettes_indicator	Yes
tobacco use indicator	No	cigars_indicator	Yes
tobacco use indicator	No	chew_indicator	Yes
tobacco use indicator	No	pipes_indicator	Yes
tobacco use indicator	No	snuff_indicator	Yes
tobacco use indicator	No	smokeless_tobaccco_use_indicator	Yes
tobacco use indicator	No	smoking_tobacco_use_indicator	Null
tobacco use indicator	Yes	smoking_cessation_counceling_date	Null
most_recently_reported_alcohol_use	Null/No	ever_reported_alcohol_use	Not Null
most_recently_reported_alcohol_use	Null/No	ounces_of_alcohol_per_week	Not Null
HIV_indicator	Yes	Problem List	Not Null
Procedure_type	diagnostic test	diagnostic text report	Null
Partial Thromboplastin Time (PTT, APTT)	15.0-50.0 seconds	anticoagulation therapy	No
Partial Thromboplastin Time (PTT, APTT)	15.0-400.0 seconds	anticoagulation therapy	Yes
Prothrombin Time (PT, Pro Time)	8.0-25.0 seconds	anticoagulant therapy	No
Prothrombin Time (PT, Pro Time)	8.0-400.0 seconds	anticoagulant therapy	Yes

Table 2: Other Table Value Incompatibility Knowledge Table

Another limitation that plagues DQA in healthcare in general 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 fourteen year old. Such off-label uses are common in practice. 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 partitioned the initial rules into two groups; the first identifies instances of physical impossibility while the second identifies instances that are merely possible but implausible. It seems reasonable that the former are much more likely to identify data errors. The rules identifying possible but implausible cases require validation prior to use, i.e., some indication that they correlate strongly with known data errors.

Similar to clinical decision support, sharing of rules is a challenge. It is tempting to write rules against an institutional data model or information system. However, doing so is wasteful. Instead, the rules here will be demonstrated against the Observational Medical Outcomes Partnership (OMOP) common data model. This will easily extend their use to any organization able to implement the OMOP data model.

Our future plans include demonstrating an infrastructure that includes a rules engine, knowledge table management structure, and rule result logging to identify errors in EHR data monitor their frequency over time.

CONCLUSION

Assessing the quality of EHR data is necessary to improve data quality yet doing so systematically represents uncharted territory in healthcare. This study provides a potentially scalable framework with which data quality rules can be organized and shared as rule templates and associated knowledge tables. While there is significant additional work to be done in this area, the exploration of the rule template and associated knowledge table approach was shown here to be possible and potentially scalable.

REFERENCE

Bagniewska, A., Black, D., Molvig, K., Fox, C., Ireland, C., Smith, J., Hulley, S. and SHEP Research Group, 1986. Data quality in a distributed data processing system: the SHEP pilot study. *Controlled clinical trials* (7:1), pp.27-37.

Bauck, A., Bachman, D. and Riedlinger, K., 2011. Developing a consistent structure for VDW QA checks. Available at: http://www.hmoresearchnetwork.org/archives/2011/concurrent/A1-02-Bauck.pdf.

Brown, P. J. and Warmington, V., 2002. Data quality probes—exploiting and improving the quality of electronic patient record data and patient care. *International journal of medical informatics* (68:1), pp.91-98.

Brown, P. J. and Warmington, V., 2003. Info-tsunami: surviving the storm with data quality probes. *Journal of Innovation in Health Informatics* (11:4), pp.229-237.

Brown, P. J., Harwood, J. and Brantigan, P., 2001. Data quality probes--a synergistic method for quality monitoring of electronic medical record data accuracy and healthcare provision. *Studies in health technology and informatics* (84: 2), pp.1116-1119.

Callahan T. J., Bauck A, Bertoch D, Brown J. S., Khare R, Ryan P. B., Staab J, Zozus M. N., Kahn M. G., 2017. A comparison of data quality checks in six data sharing networks. Submitted to eGEMS, in press.

Carlson D, Wallace CJ, East TD, Morris AH.,Verification &amp; validation algorithms for data used in critical care decision support systems. Proc Annu Symp Comput Appl Med Care. 1995:188-92.

Cato, A. E., Cloutier, G. and Cook, L., 1985. Data entry design and data quality.

Crombie, I. K. and Irving, J. M., 1986. An investigation of data entry methods with a personal computer. Computers and Biomedical Research (19:6), pp.543-550.

Curtis, L.H., Weiner, M.G., Beaulieu, N.U., Rosofsky, R.A., Woodworth, T.S. and Boudreau, D.M., 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.

DuChene, A. G., Hultgren, D. H., Neaton, J. D., Grambsch, P. V., Broste, S. K., Aus, B. M. and Rasmussen, W. L., 1986. Forms control and error detection procedures used at the Coordinating Center of the Multiple Risk Factor Intervention Trial (MRFIT). Controlled clinical trials (7:3), pp.34-45.

Dziadkowiec O, Callahan T, Ozkaynak M, Reeder B, Welton J., Using a Data Quality Framework to Clean Data Extracted from the Electronic Health Record: A Case Study. EGEMS (Wash DC). 2016 Jun 24;4(1):1201.

Eisenstein, E. L., Lemons, P. W., Tardiff, B. E., Schulman, K. A., Jolly, M. K. and Califf, R. M., 2005. Reducing the costs of phase III cardiovascular clinical trials. *American heart journal* (149:3), pp.482-488.

Estabrook, R. W., Woodcock, J., Nolan, V. P. and Davis, J. R. eds., 1999. Assuring data quality and validity in clini*cal trials for regulatory decision making: workshop report*. National Academies Press.

Forrest JR, W. H. and Bellville, J. W., 1967. The Use of computers in clinical trials. *BJA: British Journal of Anaesthesia* (39:4), pp.311-322.

Fortmann, S. P., Haskell, W. L., Williams, P. T., Varady, A. N., Hulley, S. B. and Farquhar, J. W., 1986. Community surveillance of cardiovascular diseases in the Stanford Five-City Project: methods and initial experience. American journal of epidemiology (123:4), pp.656-669.

Hart, R. and Kuo, M. H., 2017. Better Data Quality for Better Healthcare Research Results-A Case Study. *Studies in health technology and informatics* (234), p.161.

Kahn, M.G., Raebel, M.A., Glanz, J.M., Riedlinger, K., Steiner, J.F., A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. Med Care. 2012 Jul;50 Suppl:S21-9.

Kahn, M. G., Callahan, T. J., Barnard, J., Bauck, A.E., Brown, J., Davidson, B. N., Estiri, H., Goerg, C., Holve, E., Johnson, S. G., Liaw, S. T., Hamilton-Lopez, M., Meeker, D., Ong, T.C., Ryan P., Shang, N., Weiskopf, N.G., Weng, C., Zozus, M.N., Schilling, L., 2016. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. *eGEMs* (4:1).

Knatterud, G. L., 1981. Methods of quality control and of continuous audit procedures for controlled clinical trials. *Controlled clinical trials* (1:4), pp.327-332.

Kronmal, R. A., Davis, K., Fisher, L. D., Jones, R. A. and Gillespie, M. J., 1978. Data management for a large collaborative clinical trial (CASS: Coronary Artery Surgery Study). *Computers and Biomedical Research* (11:6), pp.553-566.

Norton, S. L., Buchanan, A. V., Rossmann, D. L., Chakraborty, R. and Weiss, K. M., 1981. Data entry errors in an on-line operation. *Computers and Biomedical Research* (14:2), pp.179-198.

Observational Medical Outcomes Partnership, 2012. Generalized Review of OSCAR Unified Checking. Available at: http://omop.fnih.org/GROUCH.

Observational Medical Outcomes Partnership, 2012. OSCAR - Observational Source Characteristics Analysis Report Design Specification and Feasibility Assessment. Available at:

http://omop.fnih.org/OSCAR.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C. J. and Leontis, N., 2007. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology* (25:11), p.1251.

Society for Clinical Data Management (SCDM), Good Clinical Data Management Practices (GCDMP), 2013. Available from www.scdm.org.

Tenenbaum, J. D., Christian, V., Cornish, M. A., Dolor, R. J., Dunham, A. A., Ginsburg, G. S., Kraus, V. B., McHutchison, J. G., Nahm, M. L., Newby, L. K. and Svetkey, L. P., 2012. The MURDOCK Study: a long-term initiative for disease reclassification through advanced biomarker discovery and integration with electronic health records. *American journal of translational research* (4:3), p.291.