Article published online: 2021-10-18

A Systematic Approach to Reconciling Data Quality Failures: Investigation Using Spinal Cord Injury Data

Nandini Anantharama[1] Wray Buntine[1] Andrew Nunn[2]

[1] Faculty of IT, Monash University, Clayton, Victoria, Australia
[2] Victorian Spinal Cord Service, Austin Health, Heidelberg, Victoria, Australia

Address for correspondence Nandini Anantharama, Monash University, Wellington Road, Clayton, Victoria 3800, Australia (e-mail: nandini.anantharama1@monash.edu).

ACI Open 2021;5:e94–e103.

Abstract

Background Secondary use of electronic health record's (EHR) data requires evaluation of data quality (DQ) for fitness of use. While multiple frameworks exist for quantifying DQ, there are no guidelines for the evaluation of DQ failures identified through such frameworks.

Objectives This study proposes a systematic approach to evaluate DQ failures through the understanding of data provenance to support exploratory modeling in machine learning.

Methods Our study is based on the EHR of spinal cord injury inpatients in a state spinal care center in Australia, admitted between 2011 and 2018 (inclusive), and aged over 17 years. DQ was measured in our prerequisite step of applying a DQ framework on the EHR data through rules that quantified DQ dimensions. DQ was measured as the percentage of values per field that meet the criteria or Krippendorff's α for agreement between variables. These failures were then assessed using semistructured interviews with purposively sampled domain experts.

Results The DQ of the fields in our dataset was measured to be from 0% adherent up to 100%. Understanding the data provenance of fields with DQ failures enabled us to ascertain if each DQ failure was fatal, recoverable, or not relevant to the field's inclusion in our study. We also identify the themes of data provenance from a DQ perspective as systems, processes, and actors.

Conclusion A systematic approach to understanding data provenance through the context of data generation helps in the reconciliation or repair of DQ failures and is a necessary step in the preparation of data for secondary use.

Keywords

- data quality
- electronic health record
- machine learning modelling
- trust
- data quality assessment
- specialized care

Background and Significance

The widespread adoption of electronic health records (EHRs) has been followed by its adoption as a source of research data in multiple domains,[1–4] commonly termed secondary use. The validity and robustness of such secondary use is dependent on the quality of the underlying EHR data, and multiple data quality (DQ) frameworks[5–8] have been articulated for this purpose. These frameworks provide assessment methods for analyzing EHR quality in terms of DQ dimensions.[9,10] However, the assessment of DQ remains specific to and dependent on the secondary use case, and this is termed

received
January 28, 2021
accepted after revision
August 13, 2021

DOI https://doi.org/10.1055/s-0041-1735975.
ISSN 2566-9346.

Georg Thieme Verlag KG, Rüdigerstraße 14, 70469 Stuttgart, Germany

"fitness for use."[11,12] The failures of a DQ assessment manifest as a typology of challenges documented in the literature such as inconsistencies, missing variables, lack of temporality, and lack of standardization among others.[13–20] A few of the studies suggest using supplementary data sources, surrogate fields, the Natural Language Processing (NLP) techniques, and shared metadata documentation as possible solutions.[14,16,17,21,22] Even when a task is well defined and the corresponding DQ dimensions for the assessment of DQ are clear, data are seldom error free.[23] This results in mistrust of the data,[24] and consequently the level of confidence in its viability for research use.[23] This is more relevant for a retrospective study where the data collection and the intended usage do not align as the data usage has possibly been defined some years after the data were stored.

The interpretation of DQ results, when there are DQ failures, is a necessary next step following DQ assessment, but there are no established frameworks that define how such an interpretation can be structured. In this study, we define a systematic approach for the analysis of DQ failures that are surfaced through the application of a DQ framework. DQ is a function of the point in time and context in which it was recorded. An EHR is a collation of data points of differing provenances generated by system users like clinicians and hospital administrators, and recorded for use across multiple systems spanning different care types and facilities.[8] Information generation, storage, and propagation are parts of a dynamic process, and the information transforms as it navigates across departmental boundaries and during the interpersonal communication between stakeholders.[25] Further, when considered retrospectively, the processes that recorded the data could have evolved significantly over the intervening time period, and the generated data reflect this evolution.[14,26–28] Thus, the retrospective assessment of DQ failures requires an awareness of not just the information transformation between generation and use[29] but also the dynamics of data generation itself. While the need for such contextual awareness has been studied for dimensions of DQ,[25] an equivalent emphasis on the importance of data provenance in interpreting DQ results is lacking. Data provenance of EHR data is the process of understanding the origin or source of the data, data transformations, and the metadata that provide the underlying context of data generation and collection.[8,35]

Objectives

The study objective is to develop and describe a systematic approach to assess DQ failures by understanding data provenance and provides guidelines for reconciling or repairing DQ failures by better understanding the context of data generation. The approach is grounded in a study on spinal cord injury (SCI) patients.

Methods

The research is based on longitudinal inpatient EHR data sourced from different care types for SCI patients. The preliminary step in our systematic approach is the quantitative study of DQ of the data using a suitable DQ framework. We then assess the DQ failures by a qualitative study consisting of multiple semistructured interviews with clinicians and data custodians with the objective of understanding the data provenance, and therefore resolve, mitigate, or reconcile the DQ failures revealed during the quantitative study. All analyses were performed using R (3.6.1) and the Multivariate Imputation by Chained Equations (MICE) package.[30]

Guiding Aim

Our secondary use of EHR is exploratory machine learning modeling to discern temporal patterns of complications and infections, so as to identify subgroups of the high-cost and high-need SCI patients. The creation of a comprehensive and sequential record of each SCI patient requires unification from heterogeneous data sources, as they are typically admitted for multiple months at a time, and their medical records necessarily span multiple care types, wards, and laboratories. DQ evaluation would be required to ensure unbiased data. We hypothesized that while the DQ evaluation would identify quality failures, we would need to evaluate these failures through an understanding of data provenance to ascertain if the failure was fatal, recoverable, or not relevant to the field's inclusion in our dataset.

Data

The research analyses the DQ of EHR representing the medical records of SCI patients at Austin's State Spinal Centre. We define EHR to be inpatient data retrieved from data warehouse, care type data silos, external registries and diagnosis recording (International Classification of Disease,10th revision Australian Modification [ICD]-10 AM coding). The cohort comprised patients older than 17 years, admitted from 2011 to 2018, and having the ICD-10 AM code for SCI (►Supplementary Table S1, available in the online version). The dataset is a comprehensive record of a patient's progression from the time of injury (usually ambulance), through the hospital stay, that is, the different care types up until discharge. The data included demographics, injury etiology, pathology, and radiology results, microbiology, medication, diagnosis, and discharge summaries (DS), and these were retrieved from one or more data sources (►Supplementary Figure S1, available in the online version). ►Table 1 illustrates the sources of data using injury etiology as an example. The data were linked using patient identifiers and arranged sequentially by event recording for all the inpatient encounters that the patients had during the study period.

We had a total of 1,382 patients in our dataset (►Supplementary Table S2, available in the online version). The linking of patients across data sources varied, with around 80% of patients attributable across all internal data sources, and only around 40% present in external clinical registries (CR). Aside from the patient count, the number of records varied across sources, ranging from 15,397 (microbiology records) to 1,628,070 (medication records). The number of unique features (different types of tests/medication names/ICD codes, etc.) in the data warehouse alone was 5,593, thus resulting in a

Table 1 Sources for injury etiology

Fields	Possible sources	Temporal
Patient_ID, DOB	Business Intelligence team	No
DOIJ Injury type (traumatic, nontraumatic) Injury cause	CR1, CR2, DS	No
Injury cause	CR1, CR2, DS	No
Injury level (quadriplegia, paraplegia, etc.)	DS, ICD-10 AM coding	No
Neurology (C1–S5)	CR1, DS	Yes
ASIA (A-E)	CR1, DS	Yes
Catheter (permanent, intermittent, etc.)	DS	Yes

Abbreviations: ASIA, The American Spinal Cord Injury Association (ASIA) impairment scale; CR, clinical registry; DOB, date of birth; DOIJ, date of injury; DS, discharge summary.

feature scale that is much larger in comparison to the cohort count. This mismatch further required us to reduce the cardinality of the features by using generic or standardized terminology, and to use higher level constructs where hierarchical representations were possible.

Quantitative: Evaluating Data Quality

For DQ evaluation, the study employed the 3 × 3 DQ assessment (DQA) guidelines defined by Weiskopf et al.[10] The "3 × 3 DQA" provides guidelines for three core constructs, completeness, correctness, and currency, along the granularity of patient, variables, and time. Completeness verifies that the data are sufficient for the specific secondary use. Correctness focuses on features that describe the plausibility of data values. Currency deals with recording of data at desired time intervals. The fitness for use was determined by the fitness of the EHR for discerning temporal patterns of secondary complications of SCI. For each of the data fields, we defined the appropriate DQ rules (DQRs) and metrics based on 3 × 3 DQA. To get DQ metrics as percentages, we defined Pcount and Rcount corresponding to patient count and record count, respectively. Pcount (DQR) is the percentage of patients who meet the DQR to the total number of patients for whom the DQR applies, and Rcount (DQR) is the percentage of records that meet the DQR to the total number of records for which the DQR applies. The correctness between multiple sources (interrater reliability) are verified using Krippendorff's α.[31] ►Table 2 provides these DQR using injury etiology as an example. Understanding the DQ of injury etiology is critical to measure the effectiveness of our secondary use case through comparison of etiology profiles across generated subcohorts.

Qualitative: Understanding Data Provenance to Reconcile Data Quality Failures

The qualitative step was focused on understanding data provenance, so as to resolve the DQ failures. Understanding data provenance requires identifying the context of data recording which includes the processes that led to data being recorded and the treatment workflow. The DQ failures were evaluated using an interpretive approach of semistructured interviews. The interviewees were chosen using purposive case sampling,[32] with DQ failures driving the sampling. ►Table 3 maps the interviewees per dataset, with the roles of the interviewees reflecting their area of expertise. We interviewed the 14 experts at least once, with some interviewees being interviewed multiple times. We conducted 20 face-to-face interviews with each interview spanning 15 minutes to 2 hours over a period of 6 months (based on interviewees' availability). The interviews were informed by the guiding questions of what do the data represent, how and when were it recorded, and by whom, so as to better understand the context.[33] The responses from the interviewees were analyzed using the thematic schema[34] which enables the identification and reporting of patterns in the responses.

Results

The results of completeness are shown in ►Fig. 1 (►Supplementary Figure S2 for other data sources [available in the online version]). Date of injury (DOIJ) and American Spinal Cord Injury Association (ASIA) impairment scores are critical variables, and these have very low completeness in our dataset. DOIJ is required for temporal alignment of all variables and ASIA scores help assess the sensory and motor levels post-SCI injury.

The results of DQ evaluation are presented in ►Table 4 for the injury etiology data subset, with DQ failures annotated (in bold). Fields are identified as DQ failures if the percentage of values not meeting the criteria is in the majority, or Krippendorff's α indicates poor agreement.

Our qualitative step of semistructured interviews to understand data provenance identified 11 subthemes and 3 main themes of data provenance that affect the DQ of our secondary use case (►Supplementary Tables S3 and S4, available in the online version). The main themes were systems, processes, and actors.

Data Quality Constructs

The interview questions, identified provenance (listed below as "Interview takeaways"), and resolutions and for each of the DQ failures are enumerated below.

Table 2 DQR for injury etiology

<table>
<tr><td colspan="4"><b>Fields: patient ID, DOB, DOIJ, injury type (traumatic, nontraumatic), injury level (quadriplegia, paraplegia, etc.), injury cause, SCI ICD code</b>
Temporal fields: ASIA impairment scale (A–E), neurology (C1–S5), bladder management (<b>e.g.,</b> catheter type permanent, intermittent<b>, and others</b>)</td></tr>
<tr><td>3 × 3</td><td>Complete</td><td>Correct</td><td>Current</td></tr>
<tr><td>Patients</td><td>DQR 1: Pcount (all fields required to construct etiology are present)</td><td>DQR 3: Pcount (field values belongs to allowed set)</td><td>DQR 13: Pcount (date time of the field is within the study duration and patient's inpatient stay)
Field ε {ASIA score, neurology, catheter}</td></tr>
<tr><td></td><td></td><td>DQR 4: Pcount (DOIJ, DOB format is correct)</td><td></td></tr>
<tr><td></td><td></td><td>DQR 5: Pcount (unique (field) is true)
Field ε {patient ID, DOB, DOIJ, injury type, injury level, injury cause}</td><td></td></tr>
<tr><td>Variables</td><td>DQR 2: Rcount(Fields have value)</td><td>DQR 6: α (DOIJ from 3 sources)</td><td></td></tr>
<tr><td></td><td></td><td>DQR 7: α (ASIA score from 2 sources)</td><td>DQR 14: verifying DOB and DOIJ order
Rcount(DOIJ $\geq$ DOB)</td></tr>
<tr><td></td><td></td><td>DQR 8: α (injury cause from 3 sources)</td><td></td></tr>
<tr><td></td><td></td><td>DQR 9: α (neurology from 2 sources)</td><td></td></tr>
<tr><td></td><td></td><td>DQR 10: α (injury level from 2 sources)</td><td></td></tr>
<tr><td></td><td></td><td>DQR 11: Rcount (neurology and injury level follows semantics)</td><td></td></tr>
<tr><td></td><td></td><td>DQR 12: Rcount (field changes in value conforms to expert knowledge)
Field ε {ASIA score, neurology, catheter}</td><td>DQR 15: recorded at admission and discharge
Pcount (count(field) $>=2$) Field ε {ASIA score}</td></tr>
<tr><td></td><td></td><td></td><td>DQR 16: recorded over time
Pcount(count(neurology, catheter) $\geq$ 1)</td></tr>
</table>

Abbreviations: ASIA, The American Spinal Cord Injury Association; CR, clinical registry; DOB, date of birth; DOIJ, date of injury; DQR, data quality rule; DS, discharge summary; ICD, International Classification of Disease; Pcount, patient count; Rcount; record count; SCI, spinal cord injury.
Note: Krippendorff's α.

Table 3 Key interviewees in analyzing data provenance

Data	Departments	Data sources	Roles
Injury etiology	Spinal care team	Data registry 1 Data registry 2 Discharge summary	Clinician (1) Clinical research liaison officer (2)
Diagnosis (ICD coding)	Health information services	Data warehouse	Administrator (1)
Episode information	Business intelligence team	Data warehouse	Data custodians (3)
Pathology and radiology	Pathology and radiology team	Data warehouse, Radiology data silo	Pathologist (1) Radiologist (1) Laboratory technician (1)
Microbiology	Microbiology team	Microbiology data silo	Clinician (1) Laboratory technician (1)
Medication (stewardship and antibiotics)	Infectious diseases team	Data warehouse, Antibiotics dispensing management system	Clinician (1) Pharmacist (1)

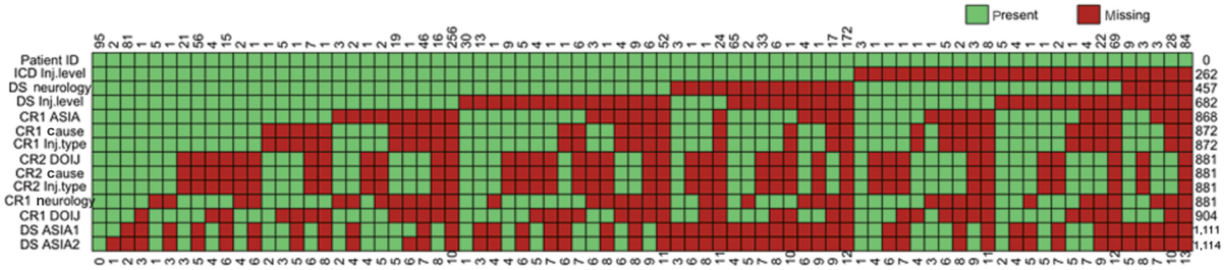

Fig. 1 Injury etiology completeness (top row numbers represents frequency of pattern, bottom row numbers represents missingness count within each pattern, end column numbers represent missingness count in the corresponding column). ASIA, The American Spinal Cord Injury Association; CR, clinical registry; DOIJ, date of injury; DQ, data quality; DS, discharge summaries; ICD, International Classification of Disease; Inj., Injury.

Table 4 DQR results reporting for injury etiology

3 × 3	Complete	Correct	Current
Patients	**DQR 1: 32% (clinical registry 1) Catheter information unrecoverable**	DQR 3: 99%	**DQR 13: 0%**
		DQR 4: 100%	
		DQR 5: 82%	
Variables	**DQR 2: 32% (clinical registry 1) Catheter information unrecoverable**	DQR 6: $\alpha = 0.97$ Field from DS unrecoverable	
		DQR 7: $\alpha = 0.92$	DQR 14: 100%
		DQR 8:$\alpha = -0.12$ Field from DS unrecoverable	
		DQR 9:$\alpha = 0.28$	
		DQR 10:$\alpha = 0.63$	
		DQR 11: 92%	
Time		**DQR 12: 0%**	**DQR 15: 21%**
			DQR 16: 0%

Abbreviations: DQR, data quality rule; DS, discharge summary.
Notes: Krippendorff's α.
Failures in bold.

Completeness

Completeness failures manifest as missingness or incompleteness in the data, and such failures in our injury etiology dataset were identified through DQR1 and DQR2 in our assessment. The evaluation of DQ failures of injury etiology are described below.

Interview questions are as follows:

- How variables are recorded in the system?
- When are the injury etiology variables recorded?
- When is it not recorded?
- How is the ICD coding done for spinal injury?
- What is a typical spinal patient trajectory through the treatment workflow?

Interview takeaways: injury etiology lacked structured recording. The fragmented reporting of care specific information was available as scanned neurology charts, unstructured DS, or form notes when ordering of tests. Further, the hospital reports some of the fields to external registries (trauma and spinal registry databases). The data were retrieved from external sources and DS. Mapping the spinal patients to the external spinal registry database informed us of nonspinal patients in our cohort. This led to the investigation of why nonspinal patients were present in the cohort, requiring understanding of the provenance of ICD codes. Further, the treatment workflow of spinal patients helped us understand that the acute patients always get admitted to the acute ward in the spinal unit.

Resolutions: ►Fig. 2 shows our workflow for addressing completeness DQ failures of DOIJ and admit ASIA fields. Using the knowledge of treatment workflow, and leveraging data from external registries through data triangulation, we filtered out nonspinal patients from the cohort. After reconciling the data sources with high concordance, the DOIJ and admit ASIA still required manual extraction from the records. This achieved 52% completeness (►Fig. 3). Finally, we removed cause as a required attribute as the injury etiology remained well defined without this attribute, and we were able to achieve 83% completeness.

Missingness were also identified in other datasets like medications and episodes. Interviewing the pharmacists revealed medication systems went electronic only in early

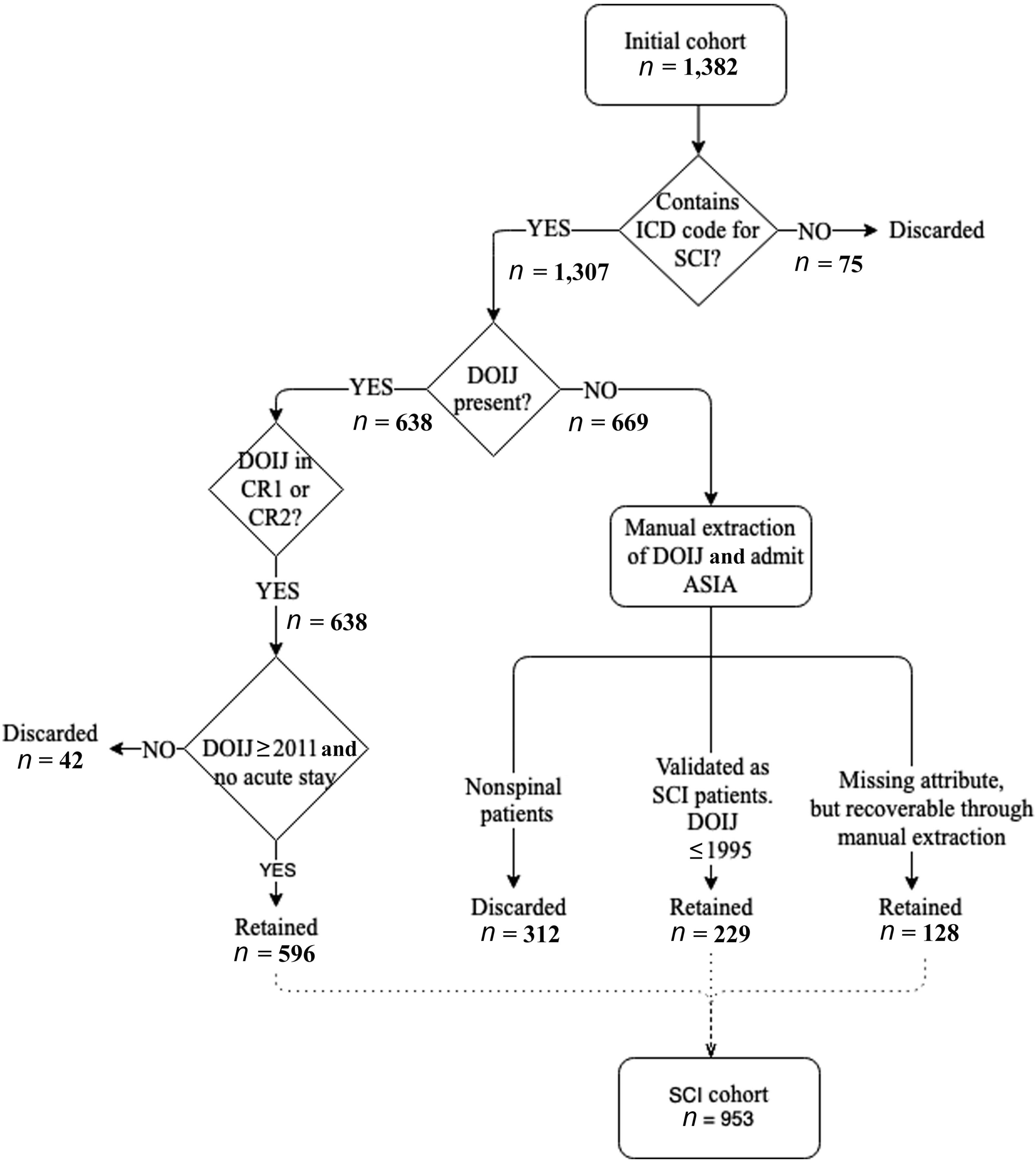

Fig. 2 Injury etiology (DOIJ, admit ASIA) completeness workflow. ASIA, The American Spinal Cord Injury Association; CR 1 and 2, clinical registries 1 and 2; DOIJ, date of injury; ICD, International Classification of Disease; SCI, spinal cord injury.

2013. Thus, we were able to safely mark the entries as missing at random (MAR) and can prevent it from being a spurious signal of "medication not administered." In the case of episodes data, missing variables of the episode dataset could be inferred through the understanding of the processes linked to generation of other variables, similar to our actions in the case of etiology data. For example, we were able to infer ICU admission and discharge times by checking the blood gas tests timestamps, as these tests are typically done early in the morning when the patient is in ICU. Similarly, standard normal ranges retrieved from the pathology department's yearly audit logs complemented the test results by providing the missing metadata for normalizing test results.

Correctness

Correctness failures map to inconsistencies, coding, and concordance errors. We measured correctness across two distinct subdimensions as follows: (1) concordance for fields where we had values from multiple coders (DQR 6–10), and (2) correctness as plausibility validated by experts or common knowledge (DQR 3–5 and DQR 11–12). Of these, DQR 8

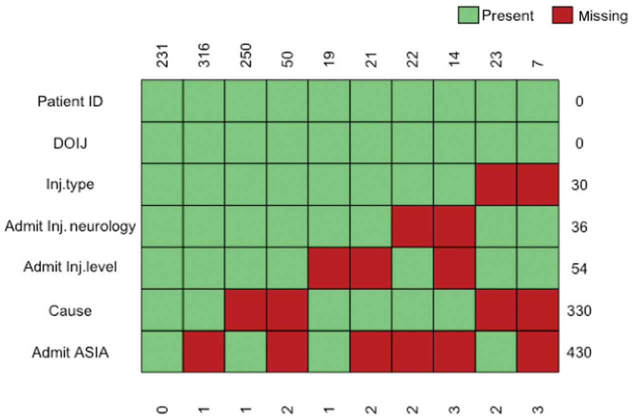

Fig. 3 Injury etiology completeness after data quality (DQ) reconciliation (top row numbers represents frequency of pattern, bottom row numbers represent missingness count within each pattern, end column numbers represent missingness count in the corresponding column). ASIA, The American Spinal Cord Injury Association; DOIJ, date of injury; Inj., Injury.

to 10 identified concordance failures, and DQR 12 identified temporal plausibility failures. The evaluation of these failures and eventual resolution are described hereinafter.

Interview questions: questions for concordance (DQR 8–10)

- How are the injuries categorized?
- How was the coding of neurology done?
- Questions for plausibility (DQR 12)
- How in the system are the variables recorded?
- What is the process by which variables are measured and recorded?

Interview takeaways: clinical reevaluation of inconsistent records revealed complexities including nonstandardization of codes and coder interpretation that result in concordance failures. These may be interpretive in nature, for example, hematoma after a surgery could be interpreted as traumatic or nontraumatic, and push bike over ramp injury could be interpreted as sports or bike. In case of plausibility (DQR 12), the interviews highlighted the lack of first-class support for recording cohort-specific temporal progression for ASIA score, neurology, and catheter type fields.

Resolutions: after terminology standardization, the α score of injury causes improved to 0.85. Though this looked unreliable, both the sources were correct, factoring in the interpretive nature of categorization. During standardization, we cross pollinated the causes. For example, a "fall from bike," categorized as a bike accident in one source and a fall accident in another, was updated to contain both causes in the record. For neurology, the data sources were the CR and DS. Low scores were due to the loss in temporality in the data recording in the DS, and due to coder bias. Further, DS as a source was considered unreliable due to lack of temporality and only ICD codes were used for recording of injury level. The CR was considered gold standard, and thus, the semantic mapping was updated between CR's neurology and ICD injury level. DQ analysis of the progression of fields, such as ASIA scores, neurology, and catheter type, was precluded by the lack of any temporal metadata in the recording of these fields.

Different systems recording data in different forms cause terminology inconsistencies similar to those seen in the injury etiology dataset. These include radiology test names recorded as "X-ray chest" versus "chest," in microbiology, "'tracheostomy swab" versus "ear/nose/throat/eye culture." Multipurpose medication is another such example. Propranolol (Apotex Pty. Ltd., Alphapharm Pty. Ltd., AstraZeneca Pty. Ltd.; New South Wales, Australia) is a dual use drug for suppressing blood pressure and heart rate, with suppressing heart rate being the primary reason, as it is prescribed to SCI patients. While standardized codes, such as Logical Observation Identifiers Names and Codes (LOINC) and Anatomical Therapeutic Chemical (ATC) help in baselining drugs to their generic names, encoding the contextual utility requires domain knowledge, and therefore understanding data provenance is integral to such secondary use.

Currency

Currency measures the quality of temporal information, and in our analysis of the injury etiology, currency measures are crucial for enabling the chronological reconstruction of a patient's record. DQR 13, 15, and 16 reported poor DQ on the associated fields.

Interview Questions

1. What is the frequency of variables recorded?
2. When and where are the injury etiology variables recorded?

Interview takeaways: the system incorporated spinal specific information as scanned documents and in DS. Therefore, SCI -specific variables did not have first-class support of validation and metadata tagging such as timestamps. Temporal information of recording of ASIA scores and catheter information was unavailable.

Resolutions: scores recorded during admission were used as the lack of temporal recording made sequential alignment of the scores impossible. ASIA score and neurology at admission were used, and discharge ASIA score was discarded. The catheter field was discarded from the research dataset due to unavailability of temporal information, even though it is an important signal in analyzing infections.

Discussion

A methodical assessment of data provenance is enabled by our proposed systematic approach of identifying DQ failures through 3 × 3 DQA framework, and addressing these failures through identifying domain experts and conducting semi-structured interviews to understand the context of data generation. The questions were guided by the why, how, and who of data recording, to interpret the data in the context of its generation, and subsequent lifecycle. Data provenance has been highlighted as an important component of EHR secondary use,[35–37] and other studies have documented the consideration of information flow as necessary for robust secondary use and avoiding biases.[25,38] Further, evaluating data provenance enables the reporting of data characteristics in EHR centric studies through presenting data completeness, data collection and handling, and the types of data.[36]

The DQ failures identified in our data were of the typology well represented in the literature.[13–20] Missingness is a common challenge in the secondary use of EHR, and its causes include data being digitized after start of study,[14] and recording in free form clinical notes.[16] Such missingness was identified in our study through understanding the systems through which data were recorded. The semistructured interviews with the actors recording it further clarified the unreliability of these values which led us to eliminate the fields altogether from our study. Data provenance is also essential to categorize observed missingness under the standardMAR, missing not at random, and missing completely at random categories, which Haneuse and Daniels[37] formulated as "what data are observed and why." We identified the categories of missingness through triangulation and understanding recording workflows, such as distinguishing between missing test values and tests not done, a crucial distinction as absence of test values is a valid signal by itself.[39,40] Lack of system flexibility in recording specialized care group information or disease progression[13,18,19,41] causes DQ failures. DQ failures increase as we traverse the data spectrum away from common, well-supported variables, such as pathology and radiology, and toward the variables of specialized care groups such as injury etiology in the SCI cohort. The data specific to specialized care groups are seldom appropriately represented in processes or systems,[13,16,19] leading to workarounds such as unstructured recording and derivative extraction. A systematic approach to understanding data provenance is particularly necessary for such specialized care groups as the distinguishing features of the cohort are also the features with the least built-in quality scaffolding, and therefore need an investment toward DQ and provenance analysis prior to secondary use. The identification of valid sources for such data, context of the recording, and any associated implicit conventions requires domain expertise, and it is made possible through data provenance–focused semistructured interviews.

Secondary use of EHR needs to consider why a particular field was recorded, and if the recording intent is in alignment with, and sufficient for, the intended secondary use.[20] In our initial data extraction, using ICD coding for spinal resulted in a dataset where 30% of the patients did not have spinal injuries and had been coded as spinal. This is specifically relevant in phenotype studies that build cohorts using ICD codes. Biased ICD coding can cause noncompatible patient records to be included in the dataset, and subsequent identification of such records is not feasible without a comprehensive understanding of data provenance as the "signature attributes" of the cohort could be naively interpreted as missing data points in such spurious records, masking the fact that these records do not belong in the cohort in the first place. Similarly, secondary use also needs to consider how a particular field was interpreted in clinical use, and what metadata were necessary to appropriately encode the field for secondary use. Pivovarov et al,[42] for instance, have documented the importance of context in the secondary use of laboratory tests' data and possibilities of bias due to lack of associated documentation. We observed such documentation bias in our test results data through lack of documentation metadata like normal ranges which skewed normalization of results, and were able to mitigate this through associating appropriate normal ranges to test results as identified through our provenance investigation.

Finally, the DQ failure analysis performed in this study spanned 6 months, largely a function of availability of domain experts. While this is not necessarily representative, it does emphasize the cost of such retrospective failure reconciliation, and motivates the need for better recording of metadata with a view toward eventual secondary use. The study also highlights the need for continuous collaboration with clinicians and domain experts during secondary use, such that the context associated with the data is incorporated into secondary use of the dataset.

Limitations

The nature of the research and the single-site focus of the study impose some limitations. Our study is based on the EHR of a specialized care group cohort, therefore the DQ failures and approaches to ascertain data provenance could be broader than what our dataset could illustrate. DQ failures and biases could also be introduced by the specifics of our study, such as the EHR systems, and institute specific workflows. A potential avenue of future research would be the comparison of the performance of models trained on such a DQ-validated data versus raw data.

Conclusion

The paper presents a systematic approach for the analysis of EHR DQ failures through understanding data provenance, and documents the resulting improvements in DQ for secondary use. Data provenance is investigated through semistructured interviews with domain experts, and the understanding of data provenance helps in reconciling DQ failures: evaluate if the issues can be fixed, mitigated, or if the records have to be discarded. Such reconciliation builds trust in the data for secondary use. Further, our semistructured interviews identified the three main themes of data provenance for secondary use of EHR data to be systems, processes, and actors.

Clinical Relevance Statement

Electronic health record (EHR data are reused for clinical research, quality assurance, and modeling among other secondary use cases and data quality assessment (DQA) is a critical step in determining its "fitness for use." DQ failures revealed through the application of these frameworks should be followed-up with a contextual analysis that situates the data in the context of when, how, and who generated the data and therefore help in further evaluating its correctness, readiness for use, and trust in the data. The proposed approach enables such analysis, thereby allowing for the construction of robust datasets for secondary use.

Protection of Human and Animal Subjects

The study was performed in compliance with the Austin Health Human Research Ethics Committee Ethical Approval (HREC). Ethics approval obtained from Austin on August 18, 2017, reference no LNR/17/Austin/408.

Authors' Contributions

N.A.: Study design, data collection and formatting, analyses and evaluation, interviews, manuscript preparation, tables, and figures. W.B.: Study design, intellectual input on statistical analyses and modeling, manuscript preparation, and review. A.N.: Study design, interviews, clinical interpretation and validation of results, manuscript preparation, and review.

Funding

None.

Conflict of Interest

None declared.

References

1 Wei W-Q, Teixeira PL, Mo H, Cronin RM, Warner JL, Denny JC. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. J Am Med Inform Assoc 2016;23(e1):e20–e27

2 Callahan A, Fries JA, Ré C, et al. Medical device surveillance with electronic health records. NPJ Digit Med 2019;2:94

3 Hribar MR, Read-Brown S, Goldstein IH, et al. Secondary use of electronic health record data for clinical workflow analysis. J Am Med Inform Assoc 2018;25(01):40–46

4 Casey JA, Schwartz BS, Stewart WF, Adler NE. Using electronic health records for population health research: a review of methods and applications. Annu Rev Public Health 2016;37:61–81

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(suppl):S21–S29

6 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(01):144–151

7 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):1244–1244

8 Cabitza F, Batini C. Information quality in healthcare. In: Batini C, Scannapieco M, eds. Data and Information Quality: Dimensions, Principles and Techniques. 1st ed. Switzerland: Springer International Publishing; 2016:403–419

9 Kahn MG, Brown JS, Chun AT, et al. Transparent reporting of data quality in distributed data networks. EGEMS (Wash DC) 2015;3(01):1052–1052

10 Weiskopf NG, Bakken S, Hripcsak G, Weng C. A data quality assessment guideline for electronic health record data reuse. EGEMS (Wash DC) 2017;5(01):14–14

11 Juran JM, Gryna FM. Quality Control Handbook. 4th ed. New York, NY: McGraw-Hill; 1988

12 Wang RY, Strong DM. Beyond accuracy: what data quality means to data consumers. Null 1996;12:5–33

13 Botsis T, Hartvigsen G, Chen F, Weng C. Secondary use of EHR: data quality issues and informatics opportunities. Summit On Translat Bioinforma 2010;2010:1–5

14 Bayley KB, Belnap T, Savitz L, Masica AL, Shah N, Fleming NS. Challenges in using electronic health record data for CER: experience of 4 learning organizations and solutions applied. Med Care 2013;51(8, suppl 3)S80–S86

15 Hong CJ, Kaur MN, Farrokhyar F, Thoma A. Accuracy and completeness of electronic medical records obtained from referring physicians in a Hamilton, Ontario, plastic surgery practice: a prospective feasibility study. Plast Surg (Oakv) 2015;23(01):48–50

16 Baier AW, Snyder DJ, Leahy IC, Patak LS, Brustowicz RM. A shared opportunity for improving electronic medical record data. Anesth Analg 2017;125(03):952–957

17 Martin S, Wagner J, Lupulescu-Mann N, et al. Comparison of EHR-based diagnosis documentation locations to a gold standard for risk stratification in patients with multiple chronic conditions. Appl Clin Inform 2017;8(03):794–809

18 Adibuzzaman M, DeLaurentis P, Hill J, Benneyworth BD. Big data in healthcare - the promises, challenges and opportunities from a research perspective: A case study with a model database. AMIA Annu Symp Proc 2018;2017:384–392

19 Cowie MR, Blomster JI, Curtis LH, et al. Electronic health records to facilitate clinical research. Clin Res Cardiol 2017;106(01):1–9

20 Raman SR, Curtis LH, Temple R, et al. Leveraging electronic health records for clinical research. Am Heart J 2018;202:13–19

21 Bae CJ, Griffith S, Fan Y, et al. The challenges of data quality evaluation in a joint data warehouse. EGEMS (Wash DC) 2015;3(01):1125–1125

22 Cohen B, Vawdrey DK, Liu J, et al. Challenges associated with using large data sets for quality assessment and research in clinical settings. Policy Polit Nurs Pract 2015;16(3-4):117–124

23 Zozus MN, Kahn MG, Weiskopf NG. Data quality in clinical research. In: Richesson RL, Andrews JE, eds. Clinical Research Informatics. Switzerland: Springer International Publishing; 2019:213–248

24 Savitz ST, Savitz LA, Fleming NS, Shah ND, Go AS. How much can we trust electronic health record data? Healthc (Amst) 2020;8(03):100444

25 Hausvik GI, Thapa D, Munkvold BE. Information quality life cycle in secondary use of EHR data. Int J Inf Manage 2021;56:102227

26 Panozzo CA, Woodworth TS, Welch EC, et al. Early impact of the ICD-10-CM transition on selected health outcomes in 13 electronic health care databases in the United States. Pharmacoepidemiol Drug Saf 2018;27(08):839–847

27 Raebel MA, Haynes K, Woodworth TS, et al. Electronic clinical laboratory test results data tables: lessons from Mini-Sentinel. Pharmacoepidemiol Drug Saf 2014;23(06):609–618

28 Cholan RA, Weiskopf NG, Rhoton DL, et al. Specifications of clinical quality measures and value set vocabularies shift over time: a study of change through implementation differences. AMIA Annu Symp Proc 2018;2017:575–584

29 Knight S. The combined conceptual life-cycle model of information quality: part 1, an investigative framework. International Journal of Information Quality 2011;2:205–230

30 van Buuren S, Groothuis-Oudshoorn K. Mice: multivariate imputation by chained equations in R. J Stat Softw 2011;45(03):1–67

31 Krippendorff K. Reliability in content analysis. Hum Commun Res 2004;30:411–433

32 Teddlie C, Yu F. Mixed methods sampling: a typology with examples. J Mixed Methods Res 2007;1:77–100

33 Eslami Andargoli A, Scheepers H, Rajendran D, Sohal A. Health information systems evaluation frameworks: a systematic review. Int J Med Inform 2017;97:195–209

34 Braun V, Clarke V. Using thematic analysis in psychology. Null 2006;3:77–101

35 Johnson KE, Kamineni A, Fuller S, Olmstead D, Wernli KJ. How the provenance of electronic health record data matters for research: a case example using system mapping. EGEMS (Wash DC) 2014;2(01):1058

36 Kohane IS, Aronow BJ, Avillach P, et al; Consortium For Clinical Characterization Of COVID-19 By EHR (4CE) What every reader should know about studies using electronic health record data but may be afraid to ask. J Med Internet Res 2021;23(03):e22219

37 Haneuse S, Daniels M. A general framework for considering selection bias in EHR-based studies: what data are observed and why? EGEMS (Wash DC) 2016;4(01):1203–1203

38 Verheij RA, Curcin V, Delaney BC, McGilchrist MM. Possible sources of bias in primary care electronic health record data use and reuse. J Med Internet Res 2018;20(05):e185

39 Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. BMJ 2018;361:k1479

40 Weiskopf NG, Rusanov A, Weng C. Sick patients have more data: the non-random completeness of electronic health records. AMIA Annu Symp Proc 2013;2013:1472–1477

41 Berger ML, Curtis MD, Smith G, Harnett J, Abernethy AP. Opportunities and challenges in leveraging electronic health record data in oncology. Future Oncol 2016;12(10):1261–1274

42 Pivovarov R, Albers DJ, Sepulveda JL, Elhadad N. Identifying and mitigating biases in EHR laboratory tests. J Biomed Inform 2014; 51:24–34