

## Assessing the quality of clinical and administrative data extracted from hospitals: The General Medicine Inpatient Initiative (GEMINI) experience

Short title: Assessing the quality of data extracted from hospitals

### Authors:

\*Sachin V. Pasricha, MD (candidate)<sup>1,2</sup>

\*Hae Young Jung, MSc<sup>1</sup>

Vladyslav Kushnir, MSc, PhD<sup>1</sup>

Denise Mak, PhD<sup>1</sup>

Radha Koppula, MBBS, PG Diploma (Clinical Research)<sup>1</sup>

Yishan Guo, MSc<sup>1</sup>

Janice L. Kwan, MD, MPH, FRCPC<sup>3,4</sup>

Lauren Lapointe-Shaw, PhD, MD, FRCPC<sup>3,5,6</sup>

Shail Rawal, MD, MPH, FRCPC<sup>3,5</sup>

Terence Tang, MD, FRCPC, MSc (candidate)<sup>3,7</sup>

Adina Weirnerman, MD, MSc, FRCPC<sup>3,8</sup>

‡Fahad Razak, MD, MSc, FRCPC<sup>1,3,9</sup>

‡Amol A. Verma, MD, MPhil, FRCPC<sup>1,3,9</sup>

\*S.V.P and H.Y.J are joint first authors.

‡F.R. and A.A.V. are joint senior authors.

<sup>1</sup> Li Ka Shing Knowledge Institute, St. Michael's Hospital, Toronto, Canada

<sup>2</sup> School of Medicine, Faculty of Health Sciences, Queen's University, Kingston, Canada

<sup>3</sup> Department of Medicine, University of Toronto, Toronto, Canada

<sup>4</sup> Department of Medicine, Mount Sinai Hospital, Toronto, Canada

<sup>5</sup> Division of General Internal Medicine, University Health Network, Toronto, Canada

<sup>6</sup> Institute for Clinical and Evaluative Sciences, Toronto, Canada

<sup>7</sup> Institute for Better Health, Trillium Health Partners, Toronto, Canada

<sup>8</sup> Sunnybrook Health Sciences Centre, Toronto, Canada

<sup>9</sup> Institute for Health Policy, Management and Evaluation, University of Toronto, Toronto, Canada

Abstract Word Count: 415  
Manuscript Word Count: 3249

## **Abstract**

**Objective:** Large clinical databases are increasingly being used for research and quality improvement, but there remains uncertainty about how computational and manual approaches can be used together to assess and improve the quality of extracted data. The General Medicine Inpatient Initiative (GEMINI) database extracts and standardizes a broad range of data from clinical and administrative hospital data systems, including information about attending physicians, room transfers, laboratory tests, diagnostic imaging reports, and outcomes such as death in-hospital. We describe computational data quality assessment and manual data validation techniques that were used for GEMINI.

**Methods:** The GEMINI database currently contains 245,559 General Internal Medicine patient admissions at 7 hospital sites in Ontario, Canada from 2010-2017. We performed 7 computational data quality checks followed by manual validation of 23,419 selected data points on a sample of 7,488 patients across participating hospitals. After iteratively re-extracting data as needed based on the computational data quality checks, we manually validated GEMINI data against the data that could be obtained using the hospital's electronic medical record (i.e. the data clinicians would see when providing care), which we considered the gold standard. We calculated accuracy, sensitivity, specificity, and positive and negative predictive values of GEMINI data.

**Results:** Computational checks identified multiple data quality issues – for example, the inclusion of cancelled radiology tests, a time shift of transfusion data, and mistakenly processing the symbol for sodium, “Na”, as a missing value. Manual data validation revealed that GEMINI data were ultimately highly reliable compared to the gold standard across nearly all data tables. One important data quality issue was identified by manual

validation that was not detected by computational checks, which was that the dates and times of blood transfusion data at one site were not reliable. This resulted in low sensitivity (66%) and positive predictive value (75%) for blood transfusion data at that site. Apart from this single issue, GEMINI data were highly reliable across all data tables, with high overall accuracy (ranging from 98-100%), sensitivity (95-100%), specificity (99-100%), positive predictive value (93-100%), and negative predictive value (99-100%) compared to the gold standard.

**Discussion and Conclusion:** Iterative assessment and improvement of data quality based primarily on computational checks permitted highly reliable extraction of multisite clinical and administrative data. Computational checks identified nearly all of the data quality issues in this initiative but one critical quality issue was only identified during manual validation. Combining computational checks and manual validation may be the optimal method for assessing and improving the quality of large multi-site clinical databases.

## BACKGROUND AND SIGNIFICANCE

Routinely collected clinical and administrative health data are increasingly being used in large databases for research and quality improvement.<sup>[1-4]</sup> In many health systems, the electronic data in hospitals are stored in a complex array of repositories with limited central oversight or standardization. Extracting data from these systems may be prone to errors. Ensuring data quality post-extraction is challenging but important,<sup>[5,6]</sup> particularly given the impact that large clinical databases might have on quality improvement, research and policy applications.

There are several widely-cited frameworks used to assess data quality.<sup>[7-9]</sup> A systematic review by Weiskopf and colleagues identified five key data quality dimensions: completeness, correctness, concordance, plausibility, and currency.<sup>[10]</sup> The methodological approach to examining each dimension remains challenging and poorly described.<sup>[11]</sup> Manual data validation is often an important step to ensuring data quality.<sup>[12]</sup> This typically involves manually abstracting a subset of data directly from the source and comparing this to electronically extracted data.<sup>[13]</sup> However, manual chart reviews are resource-intensive<sup>[13-15]</sup> and difficult to scale as databases grow in size.<sup>[6]</sup> Computational data quality assessment has been proposed as an alternative method of ensuring data quality.<sup>[16]</sup> Computational data quality assessment may include application of thresholds to ensure plausibility (e.g. birth date cannot be a future date), visual inspection of linear plots (e.g. to assess temporal trends), and outlier detection formulas.<sup>[7,17,18]</sup>

Our objective was to describe the combination of computational data quality assessment and manual data validation in a large, real-world, clinical dataset - the General Medicine Inpatient Initiative (GEMINI). This case study offers insights for approaches to ensure data quality in large datasets based on routinely collected clinical data.

## **METHODS**

### **Setting**

The General Medicine Inpatient Initiative (GEMINI) database collects administrative and clinical data for all patients admitted to or discharged from the general medicine inpatient service of seven hospital sites affiliated with the University of Toronto in Toronto and Mississauga, Ontario, Canada.<sup>[19]</sup> The seven hospitals include five academic health centres and two community-based teaching sites. At the time of this manuscript, data had been collected for 245,559 patient visits with discharge dates between April 1, 2010 and October 31, 2017.

GEMINI supports both research and quality improvement applications. The Ontario General Medicine Quality Improvement Network (GeMQIN)<sup>[20]</sup> uses data from the GEMINI database to create confidential individualized audit and feedback reports for eligible physicians at participating hospital sites.<sup>[21]</sup> The first version of these reports included six quality indicators (hospital length-of-stay, 30-day readmission, inpatient mortality, use of radiology tests, use of red blood cell transfusions, and use of routine blood-work), and directly informed front-line quality improvement efforts. It was a priority to ensure that high quality data were being used in physician audit-and-feedback reports.

### **Ethics**

Research Ethics Board approval was obtained from each participating institution.

## Data Extraction, Collection, and Processing

Data extraction occurred over two cycles: 1) patients discharged between April 1, 2010 to March 31, 2015 and 2) patients discharged between April 1, 2015 to October 31, 2017.<sup>[4]</sup> A wide range of administrative and clinical data were extracted from a variety of data sources (Table 1): patient demographics, diagnostic codes, intervention procedure codes, admission and discharge times, cost of admission, test results from biochemistry, haematology and microbiology laboratories, radiology reports, in-hospital medication orders, vital signs, and blood transfusions. A template document describing requested data formats and standards was provided to each hospital site to facilitate data extraction. Data were extracted by local hospital staff and stored locally in shared network folders.

After data were extracted, GEMINI staff de-identified the data at each hospital site and then securely transferred the data to a central repository. De-identification was performed by removing personally identifying variables from the dataset (e.g. medical record number, encounter number, date of birth, first name, last name, health card number and emergency room registration number).<sup>[22]</sup> Each patient admission was assigned a unique identifier that allowed for integration of various data tables at the central data repository and re-identification at the local hospital site for the purpose of manual validation.<sup>[23]</sup> A secure hash algorithm was applied to each patient's provincial health insurance number, which allowed us to link encounters for the same patient across multiple institutions.<sup>[24]</sup>

In the central repository, GEMINI data were processed and organized into 21 linkable data tables from each hospital site in order to group data into related categories (Table 1). Data were processed to ensure consistent formatting across hospital sites. Data quality was then iteratively assessed by computational data quality assessment and re-extraction as needed, and finally by manual validation.

### **Computational Data Quality Assessment**

The computational data quality assessments consisted of seven checks. Checks 1-4 were designed to identify errors in data completeness associated with data extraction and transfer procedures and were applied to each data table. Checks 5-7 consisted of detailed inspection of select variables in each table (e.g. sodium tests result values in laboratory data table). Each check assessed different dimensions of data quality (Table 2). The checks are described in detail below:

1. **Admissions over time:** We examined the number of patient admissions meeting the inclusion criteria at each hospital site and that were subsequently used to extract data from source systems. We produced a histogram of the proportion of patient admissions that are contained in each data table over time, based on discharge date. This explored whether data for patient admissions was systematically missing from any data table. (Appendix 2 Figure 1)
2. **Data volume over time:** We examined the amount of data (as measured by number of rows/observations for each data table) by producing a histogram of



total data volume by date and time (if available). In contrast to Check 1, this check assessed missing data, rather than missing patient admissions. Temporal patterns of missingness as well as variations in trends were carefully inspected. (Appendix 2 Figure 2)

3. **Admission-specific data volume over time:** We examined the total volume of data (as measured by number of rows/observations for each data table) per patient admission to eliminate variation in data volume that may be driven by variations in the number of patient admissions over time. We produced a line graph (with a moving average) of data volume per patient admission by date and time (if available). (Figure 1)
  
4. **Distribution of data in relation to admission and discharge times:** We examined the distribution of date and time labels on variables in each data table compared to the patient's admission and discharge date and time (e.g. the difference in time between a radiology test and the patient's admission date and time). This examined whether date and time information collected were plausible and ensured that we were not missing data from a specific portion of a patient's hospitalization (e.g. emergency department stay). (Figure 2)
  
5. **Overall variable presence over time:** We inspected the missingness of every variable from each data table using a heat map. The y-axis was populated with the list of variables in each data table and the x-axis was populated with each row of

data in each table, sequenced by admission date. (Appendix 2 Figure 3). This allowed us to identify temporal patterns and clustering in variable missingness.

6. **Specific variable presence over time:** We assessed the quality of data categorization and standardization for a specific variable. Because naming conventions differ at each hospital site (e.g. for radiology or laboratory tests) and can change within hospital sites over time, data are manually mapped to categories or standard naming conventions to permit multi-site analysis. For each variable, we produced a line graph of the proportion of mapped data volume at each hospital site by date. Visual inspection of data volume over time allowed us to identify problems with mapping and standardization. (Figure 3)
  
7. **Plausibility check:** Each variable was inspected to ensure it contained plausible values. For categorical variables, we examined frequency tables. For numerical interval or continuous variables, we computed distributions and measures of central tendency (minimum, mean, median, maximum, and interquartile range). Specific data tables such as laboratory tests required further ad-hoc analysis. For example for each laboratory test type, we inspected the proportion of non-numeric test results, measurement units, and the numerical distribution of test result values.

A data analyst visually reviewed each computational quality check to identify any potential issues of concern. Potential data quality issues could be identified as unexpected results, outliers, and disruptions in trends. These were subsequently investigated through

an iterative process. We first examined the original de-identified extracted data table to check if there were any errors with data processing (i.e. formatting, mislabelling, etc.). We next examined the extracted files at each hospital to determine whether there were any errors with deidentification or transfer (i.e. not all files transferred correctly). If the issue remained, we then worked with the hospital site to re-extract the problematic data and we then performed the computational data quality check again. If the issue persisted, a data abstractor manually reviewed data of individual problem cases directly from the patient records in the hospital electronic information system to identify differences between extracted data and the electronic information system. They then reported these differences to the local hospital site IT personnel to help fix the data quality issue. These processes typically led to correction of the data quality issue, after which time the computational data quality check was repeated. If these steps did not correct the data quality issue, the data remained in the GEMINI database but that specific data table was flagged such that it would not be used to inform any quality improvement efforts or academic research. Computational data quality checks occurred in the central data repository and did not require direct access to the source data at individual hospital sites.

### **Manual Data Validation**

Manual data validation was performed on the GEMINI database that had been improved based on computational data quality assessment and iterative re-extraction from hospitals. GEMINI data were compared to data that could be accessed through each hospital's electronic medical record (i.e. the information that clinicians see when providing care), which was taken to be the gold standard. We focused primarily on data

that were included in the physician audit-and-feedback reports. Thus, we manually validated data from six data tables (laboratory, radiology, physicians, admission, transfers, and transfusion). Two cycles of manual validation occurred, corresponding to each cycle of data extraction. Because two hospital sites share a single electronic information system, we grouped these two together and therefore refer to six hospital sites (A-F) for manual data verification.

A data abstractor manually reviewed data for a sample of patient admissions at each of the six sites. The data abstractor was not given access to the patient admission data stored in the GEMINI database meaning that they were “blinded” to any expected results. For 2010-2015 data, admissions were sampled randomly and with a more targeted approach to ensure sufficient sampling of rare events (Appendix 1):

- 100 random admissions per site for laboratory data table (specific variables were haemoglobin, sodium, creatinine, calcium, aspartate transaminase, international normalized ratio, and troponin)
- 100 random admissions per site for radiology data table (specific variables were computed tomography, plain radiography, ultrasound, magnetic resonance imaging, and echocardiography)
- 200 random admissions per site for physician data table (specific variables were admitting physician name and discharging physician name)
- 800 selected admissions per site for death, transfers, and transfusions data tables (specific variables being mortality, critical care unit transfer, and red blood cell transfusions, respectively).

Because we performed extensive manual data validation in the first cycle, we adopted a more targeted approach in the second cycle (2015-2017 data). For the second cycle, data were manually abstracted from a random sample of 5 admissions for each physician who would be receiving an audit-and-feedback report (approximately 20-30 physicians at each site corresponding to 100-150 admissions at each site).

The data abstractor recorded the result and date and time for the first (or only) occurrence of each specific variable to be validated in the hospital's electronic medical record for each patient admission. This was considered the gold standard and was compared against the data in the GEMINI database. To avoid potential human recording errors, the data abstractor double-checked the electronic medical record in the event of a discrepancy between the GEMINI database and the manually recorded data. We report both cycles of data validation together. We calculated the sensitivity, specificity, positive predictive value, and negative predictive value of GEMINI database for each data table and for individual variables, overall and stratified by hospital site. We also calculated the overall accuracy (true values / total values) for each data table. Analyses were performed using R version 3.5.2.

## **RESULTS**

### **Computational Data Quality Assessment**

Each of the seven computational data quality assessment checks identified data quality issues, many of which were subsequently corrected (Table 3). For example, check 3 (admission-specific data volume over time) identified a sharp implausible increase in radiology tests at Site B that resulted from the inappropriate inclusion of cancelled radiology tests in extracted data (Figure 1). Check 4 (distribution of data with admission and discharge times) identified an implausible time shift of transfusion data at Site F, which made it appear as if patients received transfusions after discharge (Figure 2). Check 5 (variable presence over time) identified a data processing and mapping issue whereby the raw hospital sodium test code was interpreted as null or missing during the second cycle of data extraction. This occurred because the abbreviation, “NA” (Figure 3), which refers to sodium in chemical notation, was interpreted as missing by our statistical software.

When data quality errors were identified by computational checks, we attempted to determine which step of the data extraction, collection, and standardization process was the source of the error. Computational data quality checks identified errors arising from data extraction/transfer and processing (Table 3) and could hypothetically identify errors that occur during local de-identification, though we did not find an example of this.

### **Manual Data Validation**

Across all sites, data tables, and both cycles of manual validation, 23,419 data points were manually abstracted from 7,488 patient admissions. The specific number of data points for each data table and each specific variable at each site is listed in Appendix 2.

Compared to the gold standard of data manually abstracted from an electronic medical record, the GEMINI database was found to be highly sensitive (Table 4), ranging from 95%-100% across data tables, and highly specific, ranging from 99%-100% across data tables. The database was also found to have high positive and negative predictive values, with overall results ranging from 93-100% and 99-100%, respectively, across data tables. The overall accuracy of the database was found to be 98%-100% across data tables.

Manual data validation identified one important data quality issue that was not flagged by computational approaches. Specifically, blood transfusion data at Site D had poor sensitivity (66%) and positive predictive value (75%), because of problems with the date and time that the transfusions reportedly occurred.

## DISCUSSION

This paper reports the experience of an extensive data quality assessment effort involving a broad range of administrative and clinical data extracted from 7 hospital sites over 8 years. We highlight a feasible and pragmatic approach to computational data quality assessment and illustrate how various data quality issues were identified. After manual validation of over 23,000 data points, the GEMINI database was found to have an overall accuracy of 98%-100% compared with hospital site records. Our experience suggests that although computational data quality checks are effective, they may not identify all important data quality issues. Specifically, we identified crucial data quality issues in blood transfusion data at one hospital site that were not detected through computational data quality checks. The GEMINI experience suggests that computational and manual approaches should be used together to iteratively improve and validate databases that are extracted from clinical information systems.

Data quality assessment is crucial before routinely collected data can be used for secondary purposes, such as research or quality improvement.<sup>[25]</sup> One flexible approach is to ensure that data are “fit-for-purpose” by data consumers, which has informed numerous frameworks and models for data quality assessment.<sup>[7,9,10,26,27]</sup> However, less has been published about how to operationalize these approaches, particularly in multi-site clinical datasets.<sup>[7,28,29]</sup> Kahn and colleagues describe a conceptual model and a number of computational rules to explore data quality in electronic health record-based research.<sup>[7]</sup> Similarly, van Hoeven and colleagues articulate an approach to assessing the



validity of linked data using computational methods and report its application in a specific case using transfusion data.<sup>[28]</sup> Terry and colleagues developed 11 measures of quality for primary health care data extracted from electronic medical records.<sup>[29]</sup> Each of these studies admirably documents the process of operationalizing conceptual data quality frameworks into real-world applications. These studies all focus on computational quality checks. Conversely, Baca and colleagues report a data validation effort of the Axon Registry, a clinical quality data registry, and focus entirely on the validation effort but do not describe computational efforts to assess and improve data quality within the registry.<sup>[30]</sup> We have been unable to find any studies that report the experience of using both computational and manual approaches to data quality assessment. Our study extends the literature by operationalizing broad data quality domains, specifically using computational data checks to inform iterative data quality improvement, and reporting the effectiveness of this approach based on rigorous manual validation across a range of data types and healthcare organizations.

The main implication of our study is that computational quality checks can identify most data quality issues but not all. Blood transfusion data in our study highlight the strengths and weaknesses of computational checks. Computational checks identified that transfusion data were time-shifted at one hospital, because some blood transfusions were apparently administered after patient discharge, which was implausible. However, computational checks missed major inaccuracies in the dates and times of blood transfusions at another hospital because the errors were not systematic and did not create any discernible patterns. The GEMINI experience suggests that although manual data

validation is labour-intensive, it may be necessary to ensure high quality data. Starting with computational checks and targeting manual data validation with a “fit-for-purpose” strategy (as we did in our second data collection cycle when we validated a smaller sample based on physician quality reports) may minimize manual workload without sacrificing data quality.

Our study is limited in several ways. First, we only performed manual validation on certain data tables, stemming from our “fit-for-purpose” approach focused on audit-and-feedback reports.<sup>[26]</sup> Given that this included a breadth of data tables, we feel that our findings are likely generalizable to other data tables within GEMINI. Second, although we describe the need to manually review the visual computational data quality checks, we were unable to quantify the labour hours required for either computational quality assessment or manual validation. Third, our approach to manual data validation cannot address data quality issues at the source system (e.g. missing values or incorrect data entry) as these are already embedded into the data, which form our ‘gold standard’. Finally, we focused on traditional domains of data quality, but future research could further assess non-traditional aspects such as context, representation, and accessibility.<sup>[27]</sup>

## CONCLUSION

The GEMINI experience highlights the importance of an iterative data quality assessment methodology that combines computational and manual techniques. Through this approach, GEMINI has achieved highly reliable data extraction and collection from hospital sites. Our findings demonstrate that computational data quality assessment and manual validation are complementary and combining these should be the ideal method to assess the quality of large clinical databases. Future research should focus on methods to reduce the amount of manual validation that is needed, and to assess non-traditional aspects of data quality.

## Tables & Figures

Table 1. GEMINI database structure.

Source	Data table	Description
Discharge Abstract Database*	Admission	Inpatient administrative information
Discharge Abstract Database*	Diagnosis	Inpatient discharge diagnoses, classified using ICD-10 codes.
Discharge Abstract Database*	Intervention	Inpatient interventions, classified based on CCI code
Discharge Abstract Database*	Case Mix Group	Patient risk groups based on case mix grouping methodology from the Canadian Institute for Health Information
Discharge Abstract Database*	Special Care Unit	Use of critical care and other special care units
Discharge Abstract Database*	HIG Weight	Patient risk groups based on Ontario-specific grouping methodology
National Ambulatory Care Reporting System*	ED Admission	Emergency department administrative information
National Ambulatory Care Reporting System*	ED Consults	Consultations performed in the emergency department
National Ambulatory Care Reporting System*	ED Diagnosis	Diagnoses based on emergency department visit, classified using ICD-10 codes.
National Ambulatory Care Reporting System*	ED Intervention	Interventions in the emergency department, classified based on CCI code
National Ambulatory Care Reporting System*	ED Case Mix Group	Emergency department case mix grouping methodology from the Canadian Institute for Health Information
Hospital ADT system	Admission-Discharge	Admission date, time, and service information
Hospital ADT system	Transfers	In-hospital room transfer information
Hospital ADT system	Physicians	Physician information for each admission
Hospital Electronic Information System	Echocardiography	Echocardiography data
Hospital Electronic Information System	Laboratory	Haematology and biochemistry laboratory data
Hospital Electronic Information System	Microbiology	Microbiology laboratory data
Hospital Electronic	Pharmacy	In-hospital medication orders

Information System		
Hospital Electronic Information System	Radiology	Radiology data
Hospital Electronic Information System	Transfusion	Blood transfusion data
Hospital Electronic Information System	Vitals	Vital signs data

Table 1 Legend: \*Data are extracted from hospital sites, based on what each hospital site reports to the Canadian Institute for Health Information for the Discharge Abstract Database and National Ambulatory Care Reporting System. Abbreviations: ADT = admission, discharge, transfer, ICD-10 = 10th revision of the International Statistical Classification of Diseases and Related Health Problems, CCI = Canadian Classification of Health Interventions, HIG = Health Based Allocation Model (HBAM) Inpatient Group, ED = Emergency Department

Table 2. Summary of computational data quality checks.

Check	Name	Level	Description	Quality Dimensions Assessed
1	Patient admissions over time	Data table	Proportion of all GEMINI patient admissions that are present in the data table, by date	Completeness
2	Data volume over time	Data table	Data volume per data table, by date	Completeness
3	Admission-specific data volume over time	Data table	Data volume per patient admission, by date	Completeness Plausibility
4	Distribution of data in relation to admission and discharge times	Data table	Data volume in relation to admission time and discharge time	Completeness Plausibility
5	Variable presence over time	Variable	Heat map of missing data for each variable, by date	Completeness Plausibility
6	Specific variable presence over time	Variable	Proportion of data within a data table that pertains to a certain variable after data processing, by date	Completeness Plausibility
7	Accuracy check	Variable	Categorical variables: report frequencies Numeric variables: report measures of central tendency and distribution	Correctness Concordance Plausibility

Table 2 Description:

Seven computational data quality assessment checks were conducted on entire data tables of specific variables prior to manual validation to assess different dimensions of data quality (reported in a systematic review by Weiskopf and colleagues).<sup>[10]</sup> We describe each check in detail under the *Methods* section. A data analyst visually reviewed each check to identify issues (unexpected results, outliers, disruptions in trends), which then underwent a step-wise approach including any of data re-extraction, communication with site-specific IT personnel, and manual chart review to correct the issue.

Table 3. Examples of errors identified using computational data quality checks.

<b>Quality Check</b>	<b>Example of Data Quality Issue Identified</b>	<b>Cause of Issue</b>	<b>Manual chart review required to investigate and fix</b>
1. Patient admissions over time	Missing ED data extraction for certain patient admissions in 2015 at one site (Appendix 2 Figure 1)	Data extraction/transfer	No
2. Data volume over time	Missing transfusion data during two time periods at one institution (Appendix 2 Figure 2)	Data extraction/transfer	No
3. Admission-specific data volume over time	Incorrectly included cancelled radiology tests resulting in sharp increase in radiology tests at one institution (Figure 1)	Data extraction/transfer	Yes
4. Distribution of data in relation to admission and discharge times	Transfusion data at one site was time-shifted, leading to the timestamp on certain transfusions being after patient discharge (Figure 2)	Data extraction/transfer	Yes
5. Overall variable presence over time	Glasgow Coma Scale data at one institution was missing (Appendix 2 Figure 3)	Unclear cause at this time	Yes
6. Specific variable presence over time	At site B, “NA” was used to describe sodium tests. At the data processing stage, statistical software recognized this as a null/missing value. (Figure 3)	Data processing at central repository	No
7. Plausibility check	No quality issues identified	Data extraction/transfer	No

Table 3 Description:

Each computational data quality assessment check identified a data quality issue, the causes of which were determined to be data extraction/transfer, data processing, or hypothetically local de-identification (though we did not find an example of this). Some issues that were

identified by the computational checks required a manual chart review to fix, while others did not.

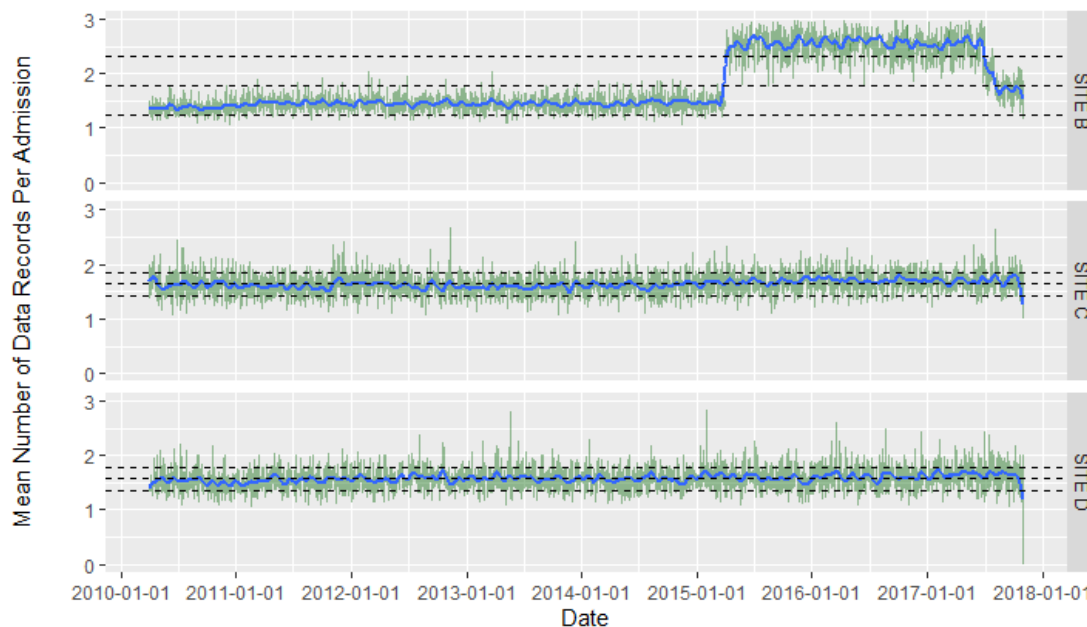


Table 4. Manual data validation results for each data table, overall and stratified by site.

	Variable	Laboratory	Radiology	Physicians	Death	Transfer	Transfusion
Overall	<i>N</i>	5648	5092	2449	3814	3300	3116
	Accuracy	100%	100%	98%	100%	100%	98%
	Sens.	100%	100%	98%	100%	100%	95%
	Spec.	100%	100%	*	100%	99%	99%
	PPV	100%	100%	100%	100%	96%	93%
	NPV	100%	100%	*	100%	100%	99%
Hospital Site A	<i>N</i>	1300	1000	400	800	800	800
	Sens.	100%	99%	94%	100%	100%	100%
	Spec.	100%	100%	*	100%	100%	100%
	PPV	100%	100%	99%	100%	100%	100%
	NPV	100%	100%	*	100%	100%	100%
Hospital Site B	<i>N</i>	950	820	400	90	**	90
	Sens.	100%	100%	98%	100%	**	95%
	Spec.	100%	100%	*	100%	**	97%
	PPV	100%	100%	100%	100%	**	90%
	NPV	100%	100%	*	100%	**	99%
Hospital Site C	<i>N</i>	960	1040	560	980	800	982
	Sens.	100%	100%	100%	100%	100%	100%
	Spec.	100%	100%	*	100%	99%	100%
	PPV	100%	100%	100%	100%	91%	97%
	NPV	100%	100%	*	100%	100%	100%
Hospital Site D	<i>N</i>	860	740	360	280	200	280
	Sens.	100%	100%	98%	100%	100%	66%
	Spec.	100%	100%	*	100%	100%	95%
	PPV	100%	100%	100%	100%	100%	75%
	NPV	100%	100%	*	100%	100%	93%
Hospital Site E	<i>N</i>	588	707	339	769	700	69
	Sens.	100%	100%	99%	100%	100%	100%
	Spec.	100%	100%	*	100%	99%	95%
	PPV	100%	100%	99%	100%	95%	80%
	NPV	100%	100%	*	100%	100%	100%
Hospital Site F	<i>N</i>	890	785	390	895	800	895
	Sens.	100%	99%	99%	100%	100%	96%
	Spec.	100%	100%	*	100%	100%	99%
	PPV	100%	99%	99%	100%	98%	92%
	NPV	100%	100%	*	100%	100%	99%

Table 4 Legend:

*N*=Number of data points, which is distinct from number of patient admissions. For example, at site A, 1000 radiology data points were manually checked, but these came from 200 patient admissions for each type of radiology test. \*For physician variables (accuracy of admitting and discharging physician names), specificity and negative predictive value cannot be calculated given the lack of true negatives (i.e. every individual should theoretically have an admitting and discharging physician, making the number of true negatives equal to 0). \*\*Site B used paper charts for critical care transfer thus manual data validation was not feasible.



*Figure 1. Data volume of radiology tests per patient admission by date at three sites.*

**Figure 1 Description:**

This is an example of Computational Data Quality Check 3: The x-axis represents the date and the y-axis represents the mean number of radiology tests per patient admission. Sites C and D represent the expected situation, where the number of radiology procedures per patient admission (green lines) and the moving average (blue line) are relatively consistent over time, and within the pre-specified thresholds (bottom and top grey dotted lines represent one standard deviation away from the overall mean). The sharp increase in the average number of tests at Site B between 2015-17 identifies a potential issue, as this level of increase is implausible. Re-extraction did not correct the issue initially. Subsequent manual investigation identified that the extraction had inadvertently included cancelled radiology tests. After successful data re-extraction, without the inclusion of cancelled tests, the aberrancy in the plot was no longer present.

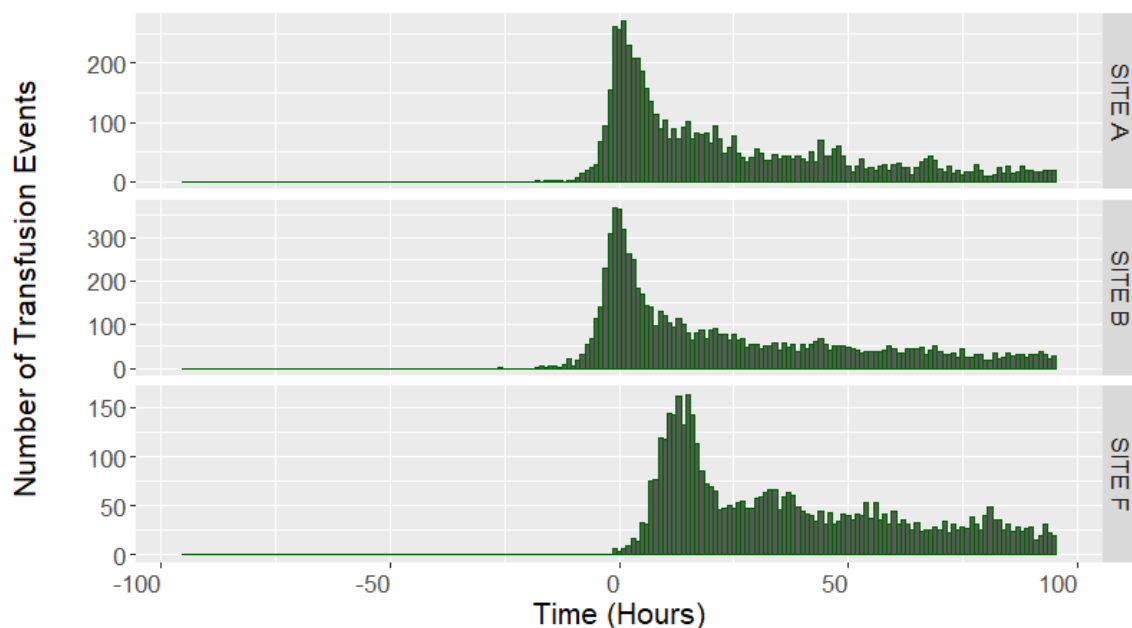


Figure 2A. Distribution of the time between transfusion events and the time of admission across three hospital sites.

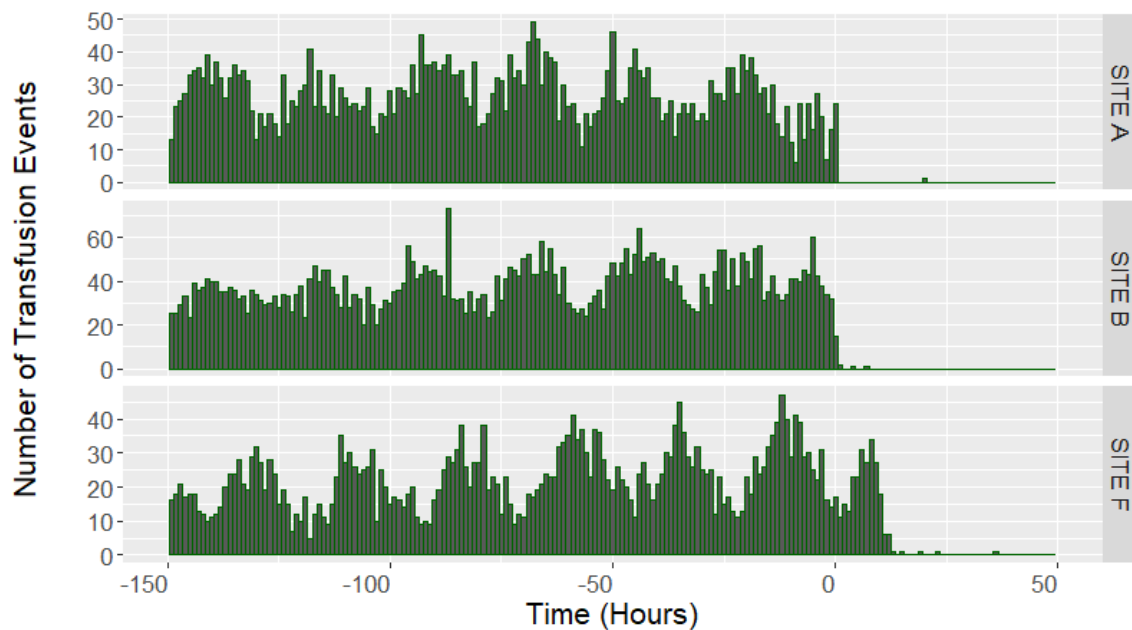


Figure 2B. Distribution of the time between transfusion events and the time of discharge across three hospital sites.

Figure 2 Description:

This is an example of Computational Data Quality Check 4: The x-axis represents the time of the transfusions event in relation to the admission (figure 3A - negative values indicating a time before admission date and time) and discharge (figure 3B – positive values indicate a time after discharge date and time). The y-axis represents the volume of transfusion events that occurred at the time on the x-axis. The plots demonstrate that at Site F, there are no transfusion events prior to admission (at Time<0 Hours). Upon first glance, this suggested missing emergency department transfusion data (i.e. patient in the emergency room and receives a transfusion prior to the admission order) but the issue could not be corrected after re-extraction efforts and discussions with the site data extraction team. The plots also indicate that Site F conducted many transfusions post-discharge (at Time>0 Hours), which would be unlikely. The combination of time from admission, time to discharge, and subsequent manual investigation uncovered a time shift for transfusion data at Site F. This time shift was then corrected in the GEMINI database.

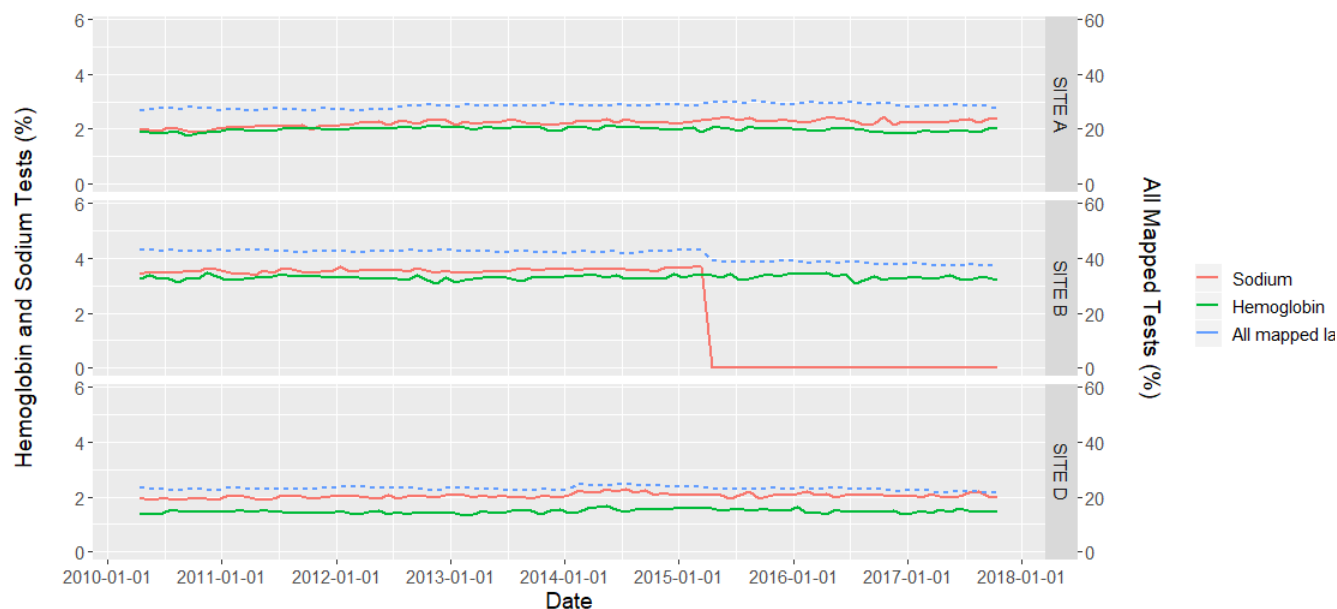


Figure 3: Volume of hemoglobin and sodium laboratory test data over time.

Figure 3 Description:

This is an example of Computational Data Quality Check 6: The x-axis represents the date. The left y-axis represents the percentage of hemoglobin and sodium tests out of all laboratory tests. The right y-axis represents the percentage of all laboratory data that has been standardized across sites (blue line, many less common laboratory tests had not yet been standardized). There is a sudden decrease in the volume of sodium tests after April 1, 2015, at Site B corresponding to the second round of data extraction. Further investigation revealed that statistical software using default settings mistakenly processed the symbol for sodium “Na” as a null/missing value.

## References

- [1] Weiner MG, Embi PJ. Toward reuse of clinical data for research and quality improvement: the end of the beginning? *Ann Intern Med.* 2009;151(5):359-360.
- [2] Herrett E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, van Staa T, Smeeth L. Data Resource Profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol.* 2015;44(3):827-836.
- [3] Birtwhistle R, Queenan JA. Update from CPCSSN. *Can Fam Physician.* 2016;62(10):851.
- [4] Ko CY, Hall BL, Hart AJ, Cohen ME, Hoyt DB. The American College of Surgeons National Surgical Quality Improvement Program: achieving better and safer surgery. *Jt Comm J Qual Patient Saf.* 2015;41(5):199-204.
- [5] Nicholls SG, Langan SM, Benchimol EI. Routinely collected data: the importance of high-quality diagnostic coding to research. *CMAJ.* 2017;189(33):E1054-E1055.
- [6] Cook JA, Collins GS. The rise of big clinical databases. *Br J Surg.* 2015;102(2):e93-e101.
- [7] 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:S21-29.
- [8] Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, Estiri H, Goerg C, Holve E, Johnson SG, Liaw ST, Hamilton-Lopez M, Meeker D, Ong TC, Ryan P, Shang N, Weiskopf NG, Weng C, Zozus MN, Schilling L. A Harmonized Data Quality Assessment Terminology and Framework for the

- Secondary Use of Electronic Health Record Data. EGEMS (Wash DC). 2016;4(1):1244.
- [9] Smith M, Lix LM, Azimaee M, Enns JE, Orr J, Hong S, Roos LL. Assessing the quality of administrative data for research: a framework from the Manitoba Centre for Health Policy. *J Am Med Inform Assoc.* 2018;25(3):224-229.
- [10] Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J Am Med Inform Assoc.* 2013;20(1):144-151.
- [11] Batini C, Cappiello, C., Francalanci, C., Maurino, A. Methodologies for data quality assessment and improvement. *ACM Comput Surv.* 2009;41(3, Article 16):52 pages.
- [12] Williamson T, Green ME, Birtwhistle R, Khan S, Garies S, Wong ST, Natarajan N, Manca D, Drummond N. Validating the 8 CPCSSN case definitions for chronic disease surveillance in a primary care database of electronic health records. *Ann Fam Med.* 2014;12(4):367-372.
- [13] Nissen F, Quint JK, Morales DR, Douglas IJ. How to validate a diagnosis recorded in electronic health records. *Breathe (Sheff).* 2019;15(1):64-68.
- [14] Lee K, Weiskopf N, Pathak J. A Framework for Data Quality Assessment in Clinical Research Datasets. *AMIA Annu Symp Proc.* 2017;2017:1080-1089.
- [15] Cai L, Zhu Y. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Science Journal.* 2015.;14:2.

- [16] Van den Broeck J, Cunningham SA, Eeckels R, Herbst K. Data cleaning: detecting, diagnosing, and editing data abnormalities. *PLoS Med.* 2005;2(10):e267.
- [17] Qualls LG, Phillips TA, Hammill BG, Topping J, Louzao DM, Brown JS, Curtis LH, Marsolo K. Evaluating Foundational Data Quality in the National Patient-Centered Clinical Research Network (PCORnet(R)). *EGEMS (Wash DC).* 2018;6(1):3.
- [18] Sunderland KM, Beaton D, Fraser J, Kwan D, McLaughlin PM, Montero-Odasso M, Peltsch AJ, Pieruccini-Faria F, Sahlas DJ, Swartz RH, Investigators O, Strother SC, Binns MA. The utility of multivariate outlier detection techniques for data quality evaluation in large studies: an application within the ONDRI project. *BMC Med Res Methodol.* 2019;19(1):102.
- [19] Verma AA, Guo Y, Kwan JL, Lapointe-Shaw L, Rawal S, Tang T, Weinerman A, Cram P, Dhalla IA, Hwang SW, Laupacis A, Mamdani MM, Shadowitz S, Upshur R, Reid RJ, Razak F. Patient characteristics, resource use and outcomes associated with general internal medicine hospital care: the General Medicine Inpatient Initiative (GEMINI) retrospective cohort study. *CMAJ Open.* 2017;5(4):E842-E849.
- [20] HQO. General Medical Quality Improvement Network. 2019; <https://www.hqontario.ca/Quality-Improvement/Quality-Improvement-in-Action/The-General-Medicine-Quality-Improvement-Network>. Accessed October 14, 2019.



- [21] HQO. MyPractice: General Medicine. 2019; <https://www.hqontario.ca/Quality-Improvement/Practice-Reports/MyPractice-General-Medicine>. Accessed October 14, 2019.
- [22] Guide to the de-identification of personal health information. [press release]. Boca Raton (FL): CRC Press 2013.
- [23] Neubauer T, Heurix J. A methodology for the pseudonymization of medical data. *Int J Med Inform.* 2011;80(3):190-204.
- [24] Dang Q. Secure hash standard. Federal Information Processing Standards Publication, Gaithersburg (MD): National Institute of Standards and Technology. 2015;Vol. FIPS PUB 1.
- [25] Dentler K, Cornet R, ten Teije A, Tanis P, Klinkenbijn J, Tytgat K, de Keizer N. Influence of data quality on computed Dutch hospital quality indicators: a case study in colorectal cancer surgery. *BMC Med Inform Decis Mak.* 2014;14:32.
- [26] Canadian Institute for Health Information TCDQF, 2009 (Ottawa, Ont.: CIHI, 2009).
- [27] Wang RY, Strong DM. Beyond Accuracy: What Data Quality Means to Data Consumers. *J of Management Information Systems.* 1996;12(4):5-33.
- [28] Hoeven LRV, Bruijne MC, Kemper PF, Koopman MMW, Rondeel JMM, Leyte A, Koffijberg H, Janssen MP, Roes KCB. Validation of multisource electronic health record data: an application to blood transfusion data. *BMC Med Inform Decis Mak.* 2017;17(1):107.
- [29] Terry AL, Stewart M, Cejic S, Marshall JN, de Lusignan S, Chesworth BM, Chevendra V, Maddocks H, Shadd J, Burge F, Thind A. A basic model for

assessing primary health care electronic medical record data quality. *BMC Med Inform Decis Mak.* 2019;19(1):30.

- [30] Baca CM, Benish S, Videnovic A, Lundgren K, Magliocco B, Schierman B, Palmer L, Jones LK, Jr. Axon Registry(R) data validation: Accuracy assessment of data extraction and measure specification. *Neurology.* 2019;92(18):847-858.