

Digital Health Data Imperfection Patterns and Their Manifestations in an Australian Digital Hospital

Kanika Goel
Queensland University of Technology
k.goel@qut.edu.au

Sareh Sadeghianasl
Queensland University of Technology
s.sadeghianasl@qut.edu.au

Robert Andrews
Queensland University of Technology
r.andrews@qut.edu.au

Arthur H. M. ter Hofstede
Queensland University of Technology
a.terhofstede@qut.edu.au

Moe T. Wynn
Queensland University of Technology
m.wynn@qut.edu.au

Dakshi Kapugama Geeganage
Queensland University of Technology
dakshi.kapugamageeganage@qut.edu.au

Sander J. J. Leemans
RWTH University
s.leemans@bpm.rwth-aachen.de

James McGree
Queensland University of Technology
james.mcgree@qut.edu.au

Rebekah Eden
Queensland University of Technology
rg.eden@qut.edu.au

Andrew Staib
Princess Alexandra Hospital
Andrew.Staib@health.qld.gov.au

Rob Eley
Princess Alexandra Hospital
r.eley@uq.edu.au

Raelene Donovan
Princess Alexandra Hospital
rae.donovan@health.qld.gov.au

Abstract

Whilst digital health data provides great benefits for improved and effective patient care and organisational outcomes, the quality of digital health data can sometimes be a significant issue. Healthcare providers are known to spend a significant amount of time on assessing and cleaning data. This paper presents six digital health data imperfection patterns that provide insight into data quality issues of digital health data, their root causes, their impact, and how these can be detected. Using the CRISP-DM methodology, we demonstrate the utility and pervasiveness of the patterns at the emergency department of Australia's major tertiary digital hospital. The pattern collection can be used by health providers to identify and prevent key digital health data quality issues contributing to reliable insights for clinical decision making and patient care delivery. The patterns also provide a solid foundation for future research in digital health through its identification of key data quality issues.

Keywords: Digital Health, Data Quality, Patterns

1. Introduction

Hospitals globally are facing unprecedented demand with an ageing population, an increase in chronic conditions (Duncan et al., 2022), and pandemics causing ripple effects for service delivery,

resulting in major delays to clinical and administrative processes (Sutherland et al., 2020). To help address these pressures, many countries are investing heavily in digital health systems. Digital health systems (e.g., Electronic Medical Records; EMR) enable centralised collection, storage, and management of patient demographic, clinical, and diagnostic data (Weiskopf et al., 2017). Such systems provide timely and efficient access to all relevant information to clinicians and healthcare providers. Advanced solutions (e.g., machine learning techniques and process analytics) may use this information for automated decision support, improving primary care, and making predictions about patient outcomes (Weiskopf et al., 2017). Despite widespread investment, the promise of digital health is yet to be realised (Reisman, 2017), and leveraging digital health data to improve healthcare processes has proven challenging due to the concerns over the reliability of digital health data. The quality of data in a digital health system may be influenced by approaches and practices used to collect, record, extract, collate, and share data (Huang et al., 2016). For example, a clinician may enter a non-sensible value, e.g., a space character, into a mandatory field in order to progress through data fields within the electronic form or even to simply 'save' the form. It is widely acknowledged that digital health data is severely impacted by accuracy and completeness concerns (Afzal et al., 2017).

To redress this situation, a considerable amount

of time is spent on cleaning the data (Miao et al., 2018; Muthalagu et al., 2014). This brings forth the significance of identifying ways to reduce the data pre-processing load on healthcare providers. We respond to this need by addressing the research question *what are the recurring data quality issues in digital health data and what are their root causes?* To address this question we use a patterns-based approach (Lea, 1994) to delineate core recurring data quality issues, which we refer to as ‘digital health data imperfection patterns’. Patterns have the advantage of being specific to a problem at hand, but also general enough to address future problems (Gamma, 1995). We collate and analyse literature to provide pattern descriptions, their examples, their impact, their manifestation, their root causes, and methods of detecting them. Next, we demonstrate the existence of these patterns in the ieMR data of the emergency department of the Princess Alexandra Hospital, a large tertiary digital hospital in Australia.

The digital health data imperfection patterns presented in this paper offer a repository of knowledge about a wide range of issues that digital health data can be exposed to, along with potential root causes. Expressing these issues as patterns makes them more accessible to health providers, data curators, administrators, and users of digital health data. The patterns can also assist in benchmarking the quality of digital health data.

The paper is structured as follows: next, we provide background literature on data quality issues in healthcare; following, we introduce the six digital health data imperfection patterns; followed by introduction to the case site, methodology employed, and findings related to the patterns. The paper concludes with a summary along with avenues for future work.

2. Related Work

Data quality is a notion that is conceptually easy to grasp, but which has proven difficult to properly define (Hoeven et al., 2017). A commonly accepted view is that data quality measures the degree of fitness of a dataset for the intended purpose. Data quality is frequently described as a multi-dimensional concept, with various authors holding different notions of the dimensions by which data quality can be measured (Batini & Scannapieco, 2006; ISO, 2011; Wand & Wang, 1996). Often mentioned dimensions include accuracy, completeness, uniqueness (of identifiers), unambiguity, currency, and timeliness. It is not uncommon to find the same dimension defined differently by different authors.

In healthcare, data quality is critical, as clinical

decision-making processes rely on data stored in digital records, which influences patient safety and care. However, poor data quality is a vital concern in the healthcare domain (Munoz-Gama et al., 2022). Quality dimensions relevant to digital health data are described in Australian Government (2022), Brennan and Stead (2000), Kahn et al. (2016), and Weiskopf and Weng (2013) and include accessibility, accuracy/correctness, completeness, concordance, conformance, consistency, currency, legibility, plausibility, relevance, and timeliness. Of these, the most frequently mentioned are completeness, accuracy/correctness, plausibility, and concordance. All of these works concur that to be high quality, data should be fit for purpose.

Additionally, literature highlights the significance of process-data quality in healthcare. Process-data refers to data about the execution of a process. Mans et al. (2012) proposed a 2-dimensional framework to deal with data generated by different types of hospital information systems. The dimensions of this framework describe log quality according to the level of abstraction of events, and the accuracy of timestamps - the latter being broken up into (i) granularity, (ii) directness of registration, and (iii) correctness. The framework developed by Bose et al. (2013) is based around 4 dimensions (missing, incorrect, imprecise, and irrelevant data) which may be applied to up to 9 event log attributes. Suriadi et al. (2017) provided a patterns-based framework for identifying event log imperfections. In the healthcare setting, the Care Pathways Data Quality Framework (Fox et al., 2018) manually identifies data quality issues of processes discovered from electronic patient records, mitigates the data issues by removing the critical values, and reports the impact of the data quality issues.

Data quality assessment methods commonly assess the dimensions of data quality, root causes of quality issues, data use, or data collection mechanisms (Chen et al., 2014). Kahn et al. (2016) presented a framework for data quality assessment (DQA) with two assessments, i.e., verification and validation. Weiskopf et al. (2017) defined a 3×3 DQA guideline to assess e-health records. They considered completeness, correctness, and currency as the data quality constructs and time, variables, and patients as data dimensions. Both of these data quality assessment frameworks evaluate the suitability of data for the intended purpose. Similarly, Feder (2018) considered accuracy, completeness, consistency, credibility, and timeliness as the relevant data quality dimensions and introduced a DQA method which involved the application of statistical methods to handle missing data and ensure completeness. According to the findings of the

review conducted by Chen et al. (2014) completeness, accuracy, and timeliness were the three most-assessed attributes in healthcare DQA. Recommended data quality improvement techniques include data cleaning, data coherence, and integration of more dimensions such as randomness and data compatibility with statistical software are proposed to improve the data quality in healthcare (Ehsani-Moghaddam et al., 2021). Benevento et al. (2022) suggest integrating domain knowledge with the data for quality improvement.

Prior work highlights the significance of understanding digital health data quality to obtain reliable insights. Digital health systems have resulted in recording a greater amount of bad data than improving the quality of data (Darko-Yawson & Ellingsen, 2016). Addressing data quality challenges and taking them into account has been identified as a characteristic for healthcare data to support management and improvement of real-life healthcare processes (Munoz-Gama et al., 2022). Our observation of data quality literature highlights the need for a consolidated understanding of data quality issues in the digital health context (e.g., EMR). In light of this need, this paper synthesises prior literature and presents six digital health data imperfection patterns, which are presented next. While we describe our patterns using a structure similar to that in Suriadi et al. (2017), our focus is on digital health data imperfection patterns rather than event log data imperfection patterns.

3. Digital Health Data Imperfection Patterns

Digital health data is exposed to recurring data quality issues (Downey et al., 2019; Weiskopf et al., 2017). Knowing these issues and understanding the associated root causes and detection techniques can assist healthcare providers in saving time to clean data contributing to reliable data. This data can then be used for clinical decision making and improved patient care. Patterns describe problems that occur over and over, then describe the core of the solution in such a way that the pattern can be applied many times without ever doing the same thing twice (Alexander, 1977). As pattern-based approaches have proven useful in describing process-data quality issues (Suriadi et al., 2017), we (i) used the keywords ‘data quality’, ‘healthcare’, ‘Electronic Health Records’, and ‘Electronic Medical Records’ to retrieve representative literature from ten databases (i.e., PubMed, Public Health, Cochrane, SpringerLink, EBSCOhost (Medline and PsycINFO), ABI/Inform, AISel, Emerald Insight, IEEE Xplore digital library,

Table 1. Digital Health Data Imperfection Patterns

#	Pattern	Data Quality Dimension
1	Double Trouble	Uniqueness
2	Not The Real Truth	Correctness
3	Not The Whole Picture	Completeness
4	Mixing Sand and Gravel	Granularity
5	Passing The Sniff Test	Plausibility
6	Shifting Shape	Concordance

Scopus), and (ii) conducted a narrative review (Paré et al., 2015) of literature. Thematic analysis (Braun & Clarke, 2006) was used to discern six digital health data imperfection patterns pertinent in the digital health context (see Table 1).

For each pattern, we provide a description of the pattern, followed by a real-life example of the data quality issue represented by the pattern. Then we discuss the way the pattern manifests in data, and hence, how it may be recognised/detected. Next, we provide the impact of the data quality issue, and potential root causes for the occurrence of the data quality issue. This is followed by techniques to detect the data quality issue.

3.1. Pattern 1 - Double Trouble

Description: The *Double Trouble* pattern describes the situation where information about a single physical entity is recorded more than once, and, hence, affects the uniqueness quality dimension. Data uniqueness is defined as the extent to which an entity from the real world is represented once (Pilar Angeles & García-Ugalde, 2009). Duplicated data introduces doubt as to the actual ‘source of truth’.

Example: Multiple health professionals recording details of the same patient in different (or even the same) digital health systems.

Manifestation: This pattern’s signature is the presence of duplicate records in one system, or the presence of the same piece of information in multiple, disparate systems.

Impact: With duplicate records, a clinician can miss important information, present in another record (McCoy et al., 2013). Duplicate data can have negative data storage implications, hindering its effectiveness in supporting collaboration (Chao, 2016) and can result in inconsistent and biased data, culminating in unreliable insights (Ehsani-Moghaddam et al., 2021).

Root Causes: A common reason for the presence of the *Double Trouble* pattern is reception/triage staff simply creating a new patient chart on presentation, rather than correctly identifying and using an already existing chart (McClellan, 2009). *Double Trouble* can also result from combining disjoint datasets that contain overlapping elements or data extraction errors, e.g., incomplete relational joins (Kahn et al., 2016).

Double Trouble can also be a result of the different data documentation style of hospital staff, and the way exchange of information among the staff using tablets occurs (Cifuentes et al., 2015).

Detection: Automated demographics comparison, name similarity comparison, automated charts comparison, and manual checks to detect similarity can be conducted (McCoy et al., 2013). Techniques built to identify duplicates within datasets can also be used.

3.2. Pattern 2 - Not The Real Truth

Description: The *Not The Real Truth* pattern refers to the situation where a recorded value is different from the actual value, and hence, affects the accuracy/correctness quality dimension. Data values are correct when they represent the truth (Reimer et al., 2016). The pattern can affect different types of data, including timestamps and textual data, and domain knowledge may be required to obtain the truth (Weiskopf et al., 2017).

Example: An EMR system having discharge date for the patient recorded as '12/06/2021 9:00PM', which in reality was '12/06/2021 11:00AM'.

Manifestation: This pattern's signature is presence of values that do not represent the real world truth.

Impact: Data exhibiting this pattern prevents a correct understanding of the healthcare process. Incorrect timestamps can result in incorrect order of activities in a process. Choosing the wrong category, e.g., 'diagnosis code', can decrease the hospital's funding (Downey et al., 2019). Incorrect textual data can lead to a wrong interpretation of a patient's conditions (Skyttberg et al., 2017) causing delays in treating patients as the doctors may use triage, presentation complaints, and notes to prioritise them.

Root causes: This pattern can be caused by allowing manual entry or selection in the design of the digital health systems, which can lead to human errors in recording data (Downey et al., 2019). For example, a healthcare professional selects an incorrect option from a list of presented SNOMED codes. The *Not The Real Truth* pattern may arise when incorrect information is provided by a patient either deliberately or inadvertently. Furthermore, recording data after the fact due to the nature of the work in hospitals (Downey et al., 2019; Feder, 2018) can also result in incorrect data.

Detection: The *Not The Real Truth* pattern can be identified by comparing data values within the dataset or with external sources of knowledge (Feder, 2018) and checking the compliance of activity orders with the expected pathways. Incorrect textual and categorical data can be identified through text mining techniques.

3.3. Pattern 3 – Not The Whole Picture

Description: The *Not The Whole Picture* pattern refers to the degree and nature of missing values within a digital health system (Liaw et al., 2013) and, hence, affects the completeness dimension. A record would be considered whole (complete) if all expected data is documented, there exists a sufficient breadth of data elements, there exists a sufficient depth of data over time, and, the data present is sufficient to predict clinical questions of interest (Weiskopf et al., 2017).

Example: At an elementary level, a missing value for a key attribute, e.g., 'Date of Birth' for one or more patients, is an example of this pattern. A patient chart which has not been updated with recent observations, orders, lab results, clinical notes, etc. is a more significant example of *Not The Whole Picture*.

Manifestation: The pattern's signature is the absence of values for mandatory, or related, attributes.

Impact: Data exhibiting this pattern provides a false picture of the treatment of patients and more broadly the activities and processes within the organisation. It can prevent highlighting some immediate actions that need to be taken or identifying appropriate actions for improving healthcare operations (Ehsani-Moghaddam et al., 2021; Weiskopf et al., 2017).

Root causes: This pattern can result from the attitude of personnel towards the type of data (e.g., financial data is more likely to be complete), flexibility in software allowing multiple ways of doing the same task, and variability in the use of standardised vocabulary (Danciu et al., 2014; Fox et al., 2018). It can be due to poor system design (e.g., not enforcing appropriate constraints), which fails to log data at certain times of execution and improper documentation for usage of systems (Bowman, 2013) and because of the time when the data is being examined.

Detection: The *Not The Whole Picture* pattern can be checked by examining the presence/absence of a value, by comparing distributions for values of interest (e.g., analysis of longitudinal data), by conducting a face validity of values, and assessing the changes in values (Weiskopf et al., 2017). The presence of the pattern can also be determined by checking association with existing values in the dataset, using statistical techniques (e.g., maximum likelihood, mean/mode substitution, and pairwise deletion), and triangulating with other sources of evidence (Liaw et al., 2013).

3.4. Pattern 4 - Mixing Sand and Gravel

Description: Data granularity refers to the level of detail of data values in a digital health system and the

consistency of granularity levels within the whole data set (Feder, 2018). The *Mixing Sand and Gravel* pattern refers to the presence of data items at the wrong (or mixed) level of granularity.

Example: An EMR system which records the diagnosis of Type I diabetes without associated complications. An EMR system which records timestamps at all/some of, date, hour, minute, second, millisecond granularity.

Manifestation: The pattern's signature is having too fine, too coarse, or mixed granularity data values for attributes.

Impact: Not recording appropriate detail, e.g., for presenting complaint, may result in a wrong diagnosis. Similarly, assigning less explicit SNOMED codes to patients (Ostropolets et al., 2020) can lead to inappropriate treatment procedures. Lack of detail in a digital health system provides only a general picture of the patient. When analysing hospital processes, mixed granularity or imprecise timestamps will result in incorrect activity ordering. Too-fine grained data prevents capturing a meaningful and holistic view of the healthcare process.

Root causes: *Mixing Sand and Gravel* pattern can be a result of data being captured through multiple systems, each recording data at a different level of detail (Chan et al., 2010). Another possible reason is different data entry habits of healthcare professionals.

Detection: The use of validation rules and comparing data values can help to detect data granularity violations (Kahn et al., 2012) and hence, the presence of the pattern. For instance, the log quality quantification tool (Fischer et al., 2022) can detect timestamps at different levels of granularity across the event log.

3.5. Pattern 5 - Passing the Sniff Test

Description: Data plausibility refers to whether a value makes sense based on external knowledge (Weiskopf et al., 2017). The *Passing the Sniff Test* pattern refers to the existence of data values that may be of the correct datatype, and in an allowable range, but which does not make sense in the current context. Hence, this pattern affects the timeliness and relevancy dimensions. Plausibility can be atemporal or temporal (Kahn et al., 2016). Atemporal plausibility aims at verifying if observed values, distributions, or densities agree with common knowledge or validated from comparisons with external sources which are trusted (Kahn et al., 2016). Temporal plausibility aims at identifying if the time-varying variables alter values as expected based on temporal properties or existing standards (Kahn et al., 2016). We recognise that this

pattern (temporal) is similar to the Inadvertent Time Travel pattern in (Suriadi et al., 2017).

Example: An example of temporal implausibility in an EMR could be a record of a complex surgical procedure with finish time only 5 minutes after the start time. Here, both values represent real date/times, but in the hospital context it is implausible for a complex procedure to be completed so quickly.

Manifestation: This pattern's signature is the presence of data that does not make sense in the current context.

Impact: The *Passing the Sniff Test* pattern can result in biased and questionable insights (Kahn et al., 2016).

Root causes: *Passing the Sniff Test* can be a result of inaccurate measurement techniques, or recording errors such as leaving out, swapping, or adding digits; errors in units (e.g., kg instead of pounds); of documenting one measure as another (e.g., recording weight as height); and of recording measurement for a different patient (Daymont et al., 2017). Lack of an established internal or external standard can also result in implausible values (Kahn et al., 2016).

Detection: Statistical techniques such as calculating weighted moving average and standard deviation score can be used (Daymont et al., 2017). Anomalous trends of data for a variable over time, or atypical data insights when comparing different variables can point towards the existence of the *Passing the Sniff Test* pattern (Kahn et al., 2016). Triangulation of data from multiple sources can also assist in detecting *Passing the Sniff Test*.

3.6. Pattern 6 - Shifting Shape

Description: Data concordance refers to the same language/representation being used for the same real-world element across different systems and applications (Ehsani-Moghaddam et al., 2021). It ensures that there is agreement between data elements (Weiskopf et al., 2017). The *Shifting Shape* pattern refers to the existence of semantically similar, but syntactically different values for the same element. We recognise that this pattern contains elements of the Synonymous Labels and Polluted Label patterns in (Suriadi et al., 2017).

Example: An EMR system where the diagnosis of 'broken arm' is recorded as 'brkn arm', 'fractured arm', and 'broken arm' for three different patient records.

Manifestation: The presence of this pattern is the occurrence of semantically similar, but syntactically different values for the same element.

Impact: This pattern can result in insights which are different or contradictory resulting in misinformation (a record which is not updated may be used) and hence

wrong decisions (Muthee et al., 2018). Concordance can assist in ascertaining the correctness of the digital health system (Weiskopf et al., 2017).

Root causes: *Shifting Shape* can result from miscommunication between digital health system components, the nature of data collected, and hence the way data is recorded, e.g., structured vs unstructured data (Bayley et al., 2013), improper system design, e.g., enforcement of business rules and constraints (Feder, 2018), and lack of awareness of the use of digital health system (Aldosari, 2017).

Detection: Data quality validation rules such as measures of central tendency can be used to detect *Shifting Shape* (Kahn et al., 2012). Data triangulation across multiple sources by using measures of spread such as standard deviation can be used. Frequency distributions of alternate datasets and goodness-of-fit tests for anticipated distributions (Feder, 2018) as well as measures of central tendency such as mean, median, mode, and standard deviation (Weiskopf & Weng, 2013) are other detection techniques.

4. Introduction to the Case

The case site is the emergency department (ED) of a large, tertiary, publicly funded, digital Australian hospital. Annually, the hospital caters to the needs of more than 65,000 patients who present to the ED and to 110,000 admitted patients. An integrated electronic medical record system (ieMR) - comprised of an electronic medical record, computerised provider order entry, ePrescribing, and clinical decision support functionalities - is used throughout the hospital to record the patient care journey. Within ED, administrative and clinical staff rely on the FirstNet module of the ieMR. In addition, administrative staff also use a Hospital-Based Corporate Information Systems. While the ieMR has resulted in some benefits related to accessibility and legibility of documentation, an auditor general report (Queensland Audit Office, 2021) criticized the quality of the ieMR data entered and used by ED staff. A sentiment which was also shared by key stakeholders within the emergency department, who are seeking to improve their data quality.

5. Approach

Following the Cross Industry Standard Process for Data Mining (CRISP-DM) approach (Wirth & Hipp, 2000), we evidenced the utility and pervasiveness of the patterns at the case site. The key stages are discussed next.

1) Business Understanding - Discussions with

stakeholders revealed the need to understand the key data quality issues the system is exposed to and the underlying root causes for improved patient care. Timestamps were communicated to be of added interest because of their use to conduct performance analysis of ED care and report on nationally scrutinised process key performance indicators. It was also communicated that several cleaning tasks related to completeness and accuracy of data attributes are performed by the curator for improving ED's monthly reports. **2) Data Understanding** - The data curator extracted data from FirstNet related to patients who visited ED from 1 October 2019 to 30 September 2021. Patient identifying information was filtered out. The raw data comprised 92 attributes of which 42 were timestamps. **3) Data Preparation** - We created an event log from the raw data using the Filter Tree Java program (Leemans, 2021). An event log consists of information related to execution of processes. Every encounter is considered as a case (with encounter ID being the case ID), and the names of the columns with timestamp values constituted the activities in the process. The output of this step was an event log with 2,329,864 events, 134,846 cases, 42 activities. This was in addition to dataset in csv format. **4) Modeling** - We used process-oriented data mining, R software (survival mode analysis function), association rule mining, correlation, text analysis data quality quantification algorithm, and SQL Server to assess and quantify the data quality issues in the ieMR data. **5) Evaluation** - The findings were analysed to report on data quality issues and associated root causes. Member checking with three stakeholders was performed to validate the findings.

6. Findings

In this section, we outline our findings with regard to the detection and manifestation of digital health data imperfection patterns in emergency department data.

6.1. Double Trouble

We conducted an assessment of 42 timestamps using SQL server and ProM, in particular the plug-in - 'Log Quality Quantification' (Fischer et al., 2022) (see Figure 1) with duplicate timestamps being observed within a case and log. For example, 5502 cases were found to have duplicate 'Arrive At' timestamps which could potentially indicate duplicate records. Additionally, 44,368 cases have the same timestamp for activities 'Edip¹' and 'Departure actual at'.

¹EDIP is the attribute to record the time when the ED patient became an inpatient at the hospital.

Configuration		Event Log Timestamp Quality			
Log Level Quality	Accuracy score 0.9387	Completeness score 0.7202	Granularity score 0.6724	Consistency score 0.5613	
score 0.2415	Medium	Missing Values score 0.8905	Format score 1.0	Duplicates Within Log score 0.5949	
Years Level Quality	Midnight Event Ordering score 0.8476	Missing Activity score 0.9505	Missed Granularity of the Log score 0.3974	score 0.5949	
score 0.5061	Medium	Overlapping Events per Resource score 1.0	Missed Granularity of Years score 0.4434	Duplicates Within Year score 0.808	
Activity Level Quality	score 0.6177	Missing Events score 0.0	Missed Granularity of Activities score 0.803	Duplicates Within Activity score 1.0	
Event Level Quality	Future Entry score 1.0	Missing Timestamp score 1.0			
score 0.882	High	Presence score 0.9981			

Figure 1. Timestamp Quality Assessment.

duplicates within the log are understandable as many patients may have been going through more than one activity at the same time, duplicates within a case require further attention. We investigated this further using SQL queries and found distinct patient records having the same 'Arrive At' timestamp. We limited ourselves to only timestamps and did not use other similarity matching techniques to detect duplicate records, as we did not have patient details such as name and date of birth and the data curator had cleaned the dataset prior.

The root cause of presence of such duplicate data can be attributed to manual entry of timestamps and the way information is recorded and handed over between staff at the ED. Because of manual entry, the difference of seconds may not be recorded for certain activities, resulting in duplicate timestamps. Furthermore, unclear understanding of the fields against which values need to be written, could also have resulted in duplicate records.

6.2. Not The Real Truth

Incorrect ordering of activities stemming from the *Not The Real Truth* pattern (incorrect timestamps) are described. According to the pathway observed at the ED: 1) the patient is ready to leave the ward (depart ready), 2) a bed is booked and administrative admission processes are undertaken (inpatient order), and 3) the patient physically leaves the ED (depart actual). However, 'depart ready' is not observed before 'depart actual' and 'inpatient order' does not happen after 'depart ready' in the journey of 11,068 and 18,037 patients, respectively. For example, for one patient, departure actual happened at 30/1/2021 21:47:00 while depart ready was recorded at 31/1/2021 04:31:35. For another patient, inpatient order happens at 1/10/2019 00:16:47, while depart ready occurs at 1/10/2019 00:32:00.

There are multiple root causes for the aforementioned incorrect timestamps. A common scenario is that the data might be recorded after the fact because the staff member, who was responsible for recording the data, was involved in another (more urgent) activity due to the nature of the emergency

departments. Another possible reason can be manual data entry, which is required for a number of tasks, which can lead to human errors. Not all tasks were evidenced to have a system generated time. Furthermore, there could be constraints in the format in which timestamp needs to be written, which could have further contributed to correctness of timestamps.

6.3. Not The Whole Picture

We found the *Not The Whole Picture* pattern manifesting in the form of missing timestamp values in the dataset. Timestamps could be unavailable due to two reasons— simple failure to record values (e.g., staff had recorded the 'arrival time' but not the 'discharge time') or due to irrelevancy (e.g., if the patient is discharged from the hospital directly from the ED there will be no in-patient 'bed request time', 'bed start time' etc.). Hence, we conducted an analysis to determine the dependency relationships between timestamp columns. We generated association rules between timestamp columns to detect the co-relationships. Association rule mining was used to extract relationships between the items in the dataset and to derive the frequency of occurrence. To limit the number of association rules, we generated association rules with single antecedent and consequent (co-related pairs are extracted). Association rules with *confidence = 1* indicate the existence of *B* always happens with existence of *A*. For example,

$Triaged\ at = 1 \Rightarrow Arrive\ At = 1, \#CONF: 1$ indicates, if *Triaged at* column contains a timestamp value then *Arrive At* also contains a timestamp.

Co-related timestamp column pairs should be consistent (either both co-related columns should be filled or both should be empty). Very high confidence values indicate a strong co-relationship between the columns. We filtered the association rules with confidence value between 0.95 and 1 and extracted incidences of missed or irrelevantly filled timestamps. We retrieved 1,347 association rules, 337 of which indicated incidents of missing values in co-related timestamp column pairs. For example, $Admit\ Bed\ Req\ At = 1 \Rightarrow Inpatient\ Order\ At = 1 \#CONF: 0.95$, denotes 5% records of *Admit Bed Req At* are missed when *Inpatient Order At* values are filled. Similarly, these 337 rules revealed the incidences of missing timestamps based on the co-related timestamp columns.

The root causes for this data quality issue can be related to working style in the ED, and prioritisation of activities by personnel. In ED, the main priority is direct patient care. Consequently, recording data may not be given primary importance, and may be missed at

times. Furthermore, manual entry of data could also be a contributing factor to this issue.

6.4. Mixing Sand and Gravel

We checked the presence of the *Mixing Sand and Gravel* pattern by assessing the granularity of the timestamps in our data set using the Log Quality Quantification ProM plugin (Fischer et al., 2022), as shown in Figure 1. The results show that there are mixed levels of granularity in timestamps with 1,263, 22,226, 898,770, and 1,407,554 instances being at the day, hour, minute, and second levels respectively.

This mixed levels of granularity of timestamps is caused by different habits of data entry across ED staff, with some being more precise than the others. Furthermore, the fact that the system allows for manual entry of timestamps and does not mandate entering precise timestamps plays a role in observing mixed levels of granularity.

6.5. Passing the Sniff Test

We checked the presence of the *Passing the Sniff Test* pattern by analysing unusual activity ordering and unusual activity duration in patient pathways. Figure 2 is an automatically discovered process model representing the different pathways through ED recorded for mental health patients. Nodes represent ED activities. Unusual activity ordering can be recognised by (i) low arc frequency and (ii) arcs between activities (not) existing in conflict with ordering expected from clinical guidelines. Unusual event ordering was observed as (i) ‘Triage’ occurring after ‘Service commencement’ (8 cases), (ii) ‘Treat Nrs Seen’ occurring after ED departure (136 cases), and (iii) ‘Admit bed complete at’ happening before ‘Admit bed req at’ (34 cases).

To discover unusual durations, we used the Kaplan-Meier survival function (Kaplan & Meier, 1958) to estimate the length of stay (LoS, probability the patient has not been discharged after t hrs) of various patient types (see Figure 3). We found Psychiatry admissions tend to have shorter stays in ED (< 10hrs), while Medicine admissions appear to initially stay in hospital longer but then have a sharp discharge rate. Discussing the plausibility of these observations with clinicians revealed these patient types are initially admitted to ED but are then discharged from ED and admitted to purpose-built Short Stay Units which are co-located with ED to allow continuity of care.

It is evident that to deem the plausibility of findings, domain knowledge is required. Nonetheless, the root causes for implausibility include failure to log data or out-of-ordinary executions. Implausibility could also be

Table 2. Irrelevant data identified in columns

Field	Irrelevant Data
RFT Left DT TM, Ramp Left DT TM, Rapid Assement DT TM	Timestamp but numerical values entered, e.g. 164
Time on Ramp	Numerical but string/timestamps values entered, e.g. 10/02/2021 19:06
Service Commencement DT TM	Timestamp column but string values entered, e.g. private
COVID19 Precautions	Yes/No but timestamps entered, e.g. 10/02/2021 01:24
Referred by Code	Code but string values entered, e.g. Cardiologist

because of wrong measurement of value, which can be associated with many reasons such as system design, failure to understand the measure, no defined standard, and not having appropriate data to record.

6.6. Shifting Shape

To check for the existence of *Shifting Shape* pattern in the ieMR data, we analysed the column values, checking timestamps formats, and associations between codes, descriptions, and entered text. We found *Shifting Shape* manifesting as irrelevant records (see Table 2 for some examples). As *15 pairs of co-related code and descriptions from the dataset* (e.g., *Mode of arrival Code, Mode of arrival Desc*), we identified inconsistencies among code and descriptions such as two codes are given for same description (e.g., *Primary diagnosis description, urinary retention is associated with two codes: 210583017 and 397939011*).

The root causes for lack of concordance can be attributed to system design, which doesn’t enforce certain constraints for entering values. Irrelevancy and lack of concordance is also evident because of manual selection. This is because automated data entry would validate the user input, e.g., it won’t allow the user to enter a number for a timestamp. Additionally, automatic selection of diagnosis codes seems missing in the system. E.g., COVID is spelled in multiple ways.

7. Discussion and Future Outlook

Digital health data quality is of critical importance to meaningfully harness insights for improved decision making. We contribute to this need by presenting six digital health data imperfection patterns that were synthesised from literature, which amongst others also illustrate the root causes for the observed data quality issues. The utility and pervasiveness of patterns is evidenced through their applicability in the emergency department of Australia’s tertiary digital hospital. We demonstrate the use of different detection techniques to find evidence for the digital health data imperfection patterns and highlight root causes associated with each

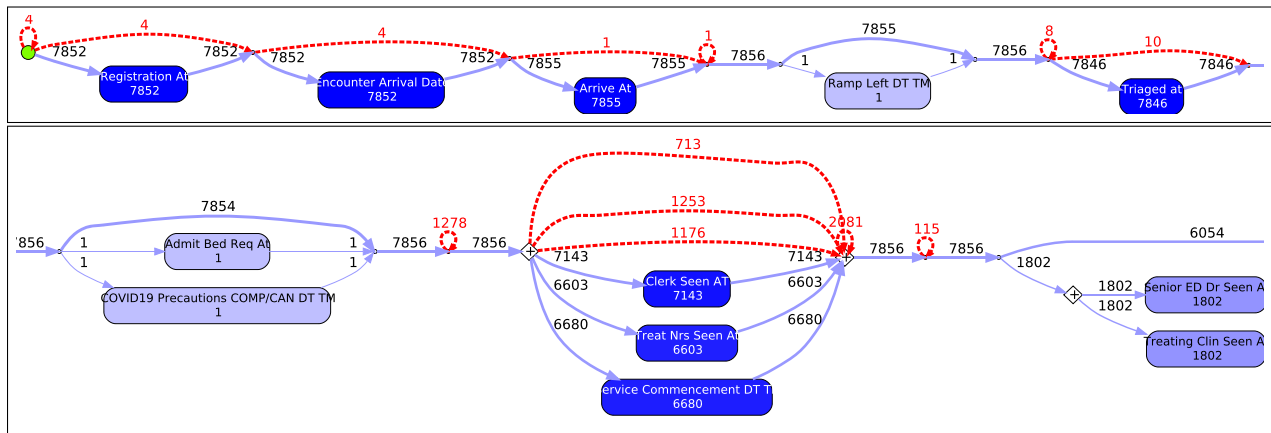


Figure 2. Part of the happy path of the process, capturing 87% of the behaviour.

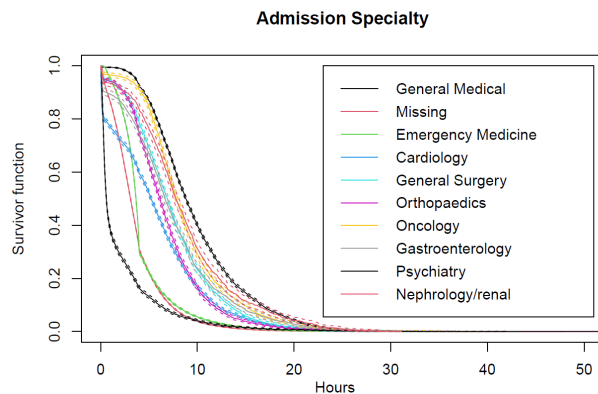


Figure 3. Estimated survivor function with a 95% confidence interval for LoS by admission speciality.

pattern. The patterns outlined in the paper may be a part of the folklore of the data quality community, however, we present them in an accessible way to healthcare providers. We provide a catalogue of digital health data imperfection patterns in a consistent format, providing a new reference point for health care professionals. These patterns assist in smooth communication among healthcare providers, can be used to benchmark the quality of digital health data, and offer a repository of knowledge, contributing to the areas of data quality and healthcare. Use of these patterns can also help healthcare providers in increasing the reliability of insights from their data, contributing to improved patient care delivery.

The work presented in this paper opens avenues for future research. First, we acknowledge that the patterns presented in the paper may not be complete – we present certain examples from literature (e.g., detection techniques and root causes). The patterns can be refined and extended using other forms of validation, e.g., qualitative data. Further, each pattern can be studied in further detail to add prevention and repair techniques. Finally, the paper shows how to use a patterns-based

approach, which can be adopted by future researchers to understand significant areas of interest.

Acknowledgement. We are grateful to QUT Centre of Data Science for funding this project and to PAH for sharing data sources and domain knowledge.

References

- Afzal, M. et al. (2017). Comprehensible knowledge model creation for cancer treatment decision making. *Comput. Biol. Med.*, 82, 119–129.
- Aldosari, B. (2017). Patients' safety in the era of EMR/EHR automation. *Inform. Med. Unlocked*, 9, 230–233.
- Alexander, C. (1977). *A pattern language: Towns, buildings, construction*. Oxford university press.
- Australian Government. (2022). Improve data quality and safety [https://www.myhealthrecord.gov.au/for-healthcare-professionals/howtos/improve-data-quality-and-safety].
- Batini, C., & Scannapieco, M. (2006). *Data quality: Concepts, methodologies and techniques. data-centric systems and applications*. Springer.
- Bayley, K. B. et al. (2013). Challenges in using electronic health record data for cer: Experience of 4 learning organizations and solutions applied. *Med. Care*, 80–86.
- Benevento, E. et al. (2022). How Can Interactive Process Discovery Address Data Quality Issues in Real Business Settings? Evidence from a Case Study in Healthcare. *J. Biomed. Inform.*, 130, 104083.
- Bose, J. et al. (2013). Wanna improve process mining results? It's high time we consider data quality issues seriously. *CIDM Symposium*, 127–134.
- Bowman, S. (2013). Impact of electronic health record systems on information integrity: Quality and safety implications. *Perspect. Health Inf. Manag.*, 10(Fall).
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2), 77–101.
- Brennan, P. F., & Stead, W. W. (2000). Assessing Data Quality: From Concordance, through Correctness and Completeness, to Valid Manipulatable Representations. *J. Am. Med. Inform. Assoc.*, 7(1), 106–107.
- Chan, K. S. et al. (2010). Electronic Health Records and the Reliability and Validity of Quality Measures: A Review of the Literature. *Med. Care Res. Rev.*, 67(5), 503–527.

- Chao, C. A. (2016). A Review of Data Quality Assessment Methods for Public Health Information Systems. *Int. J. Med. Inform.*, 94, 100–111.
- Chen, H. et al. (2014). A review of data quality assessment methods for public health information systems. *Int. J. Environ. Res. Public Health*, 11(5), 5170–5207.
- Cifuentes, M. et al. (2015). Electronic Health Record Challenges, Workarounds, and Solutions Observed in Practices Integrating Behavioral Health and Primary Care. *J. Am. Board Fam. Med.*, 28(Supplement 1), 63–72.
- Danciu, I. et al. (2014). Secondary use of clinical data: The Vanderbilt approach. *J. Biomed. Inform.*, 52, 28–35.
- Darko-Yawson, S., & Ellingsen, G. (2016). Assessing and improving EHRs data quality through a socio-technical approach. *Procedia Comput. Sci.*, 98, 243–250.
- Daymont, C. et al. (2017). Automated identification of implausible values in growth data from pediatric electronic health records. *J. Am. Med. Inform. Assoc.*, 24(6), 1080–1087.
- Downey, S. et al. (2019). Perceptions and challenges of ehr clinical data quality. *Australasian Conference on Information Systems*, 233–243.
- Duncan, R. et al. (2022). Synthesizing dimensions of digital maturity in hospitals: Systematic review. *J. Med. Internet Res.*, 24(3).
- Ehsani-Moghaddam, B. et al. (2021). Data quality in healthcare: A report of practical experience with the canadian primary care sentinel surveillance network data. *Health Inf. Manag. J.*, 50(1-2), 88–92.
- Feder, S. L. (2018). Data quality in electronic health records research: Quality domains and assessment methods. *West. J. Nurs. Res.*, 40(5), 753–766.
- Fischer, D. A. et al. (2022). Towards interactive event log forensics: Detecting and quantifying timestamp imperfections. *Inf. Syst.*, 102039.
- Fox, F. et al. (2018). A data quality framework for process mining of electronic health record data. *International Conference on Healthcare Informatics*, 12–21.
- Gamma, E. (1995). *Design patterns: Elements of reusable object-oriented software*. Pearson Education India.
- Hoeven, L. R. v. et al. (2017). Validation of multisource electronic health record data: An application to blood transfusion data. *BMC Med. Inform. Decis. Mak.*, 17(1), 1–10.
- Huang, Y. et al. (2016). Using primary care electronic health record data for comparative effectiveness research: Experience of data quality assessment and preprocessing in The Netherlands. *J. Comp. Eff. Res.*, 5(4), 345–354.
- ISO. (2011). *ISO/IEC 25010:2011: Systems and software engineering - Systems and software product Quality Requirements and Evaluation (SQuARE) - System and software quality models*.
- Kahn, M. G. et al. (2012). A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. *Med. Care*, 50.
- Kahn, M. G. et al. (2016). A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. *eGEMS*, 4(1).
- Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.*, 53(282), 457–481.
- Lea, D. (1994). Christopher Alexander: An Introduction for Object-Oriented Designers. *ACM SIGSOFT Software Engineering Notes*, 19(1), 39–46.
- Leemans, S. J. J. (2021). Filter tree [http://leemans.ch/leemansCH/filtertree].
- Liaw, S. T. et al. (2013). Towards an ontology for data quality in integrated chronic disease management: A realist review of the literature. *Int. J. Med. Inform.*, 82(1), 10–24.
- Mans, R. S. et al. (2012). Process mining in healthcare: Data challenges when answering frequently posed questions. *Process Support and Knowledge Representation in Health Care* (pp. 140–153). Springer.
- McClellan, M. A. (2009). Duplicate medical records: A survey of twin cities healthcare organizations. *AMIA Annual Symp. Proc.*, 2009, 421.
- McCoy, A. B. et al. (2013). Matching identifiers in electronic health records: Implications for duplicate records and patient safety. *BMJ Qual. Saf.*, 22(3), 219–224.
- Miao, Z. et al. (2018). An assessment and cleaning framework for electronic health records data. *IIE Annual Conference. Proceedings*, 907–912.
- Munoz-Gama, J. et al. (2022). Process mining for healthcare: Characteristics and challenges. *J. Biomed. Inform.*, 127, 103994.
- Muthalagu, A. et al. (2014). A rigorous algorithm to detect and clean inaccurate adult height records within EHR systems. *Appl. Clin. Inform.*, 5(01), 118–126.
- Muthee, V. et al. (2018). The impact of routine data quality assessments on electronic medical record data quality in Kenya. *PLoS One*, 13(4).
- Ostropolets, A. et al. (2020). Characterizing database granularity using SNOMED-CT hierarchy. *AMIA Annual Symp. Proc.*, 2020, 983.
- Paré, G. et al. (2015). Synthesizing information systems knowledge: A typology of literature reviews. *Inf. Manag.*, 52(2), 183–199.
- Pilar Angeles, M. D., & García-Ugalde, F. (2009). A data quality practical approach. *Int. J. Adv. Softw.*, 1(2&3).
- Queensland Audit Office. (2021). *Measuring emergency department patient wait time* (Vol. 2). Queensland Audit Office.
- Reimer, A. P. et al. (2016). Data quality assessment framework to assess electronic medical record data for use in research. *Int. J. Med. Inform.*, 90, 40–47.
- Reisman, M. (2017). EHRs: The challenge of making electronic data usable and interoperable. *Pharm. Ther.*, 42(9), 572.
- Skyttberg, N. et al. (2017). Exploring vital sign data quality in electronic health records with focus on emergency care warning scores. *Appl. Clin. Inform.*, 8(03), 880–892.
- Suriadi, S. et al. (2017). Event log imperfection patterns for process mining: Towards a systematic approach to cleaning event logs. *Inf. Syst.*, 64, 132–150.
- Sutherland, K. et al. (2020). Impact of COVID-19 on healthcare activity in NSW, Australia. *Public Health Res. Pract.*, 30(4).
- Wand, Y., & Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. *Commun. ACM*, 39(11), 86–95.
- Weiskopf, N. G., & Weng, C. (2013). Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research. *J. Am. Med. Inform. Assoc.*, 20(1), 144–151.
- Weiskopf, N. G. et al. (2017). A data quality assessment guideline for electronic health record data reuse. *eGEMS*, 5(1).
- Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. *International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, 1, 29–40.