



UNIVERSITY OF LEEDS

This is a repository copy of *The assessment of data quality issues for process mining in healthcare using Medical Information Mart for Intensive Care III, a freely available e-health record database.*

White Rose Research Online URL for this paper:  
<http://eprints.whiterose.ac.uk/138532/>

Version: Accepted Version

---

**Article:**

Kurniati, AP [orcid.org/0000-0002-4747-1067](https://orcid.org/0000-0002-4747-1067), Rojas, E, Hogg, D [orcid.org/0000-0002-6125-9564](https://orcid.org/0000-0002-6125-9564) et al. (2 more authors) (2019) The assessment of data quality issues for process mining in healthcare using Medical Information Mart for Intensive Care III, a freely available e-health record database. *Health Informatics Journal*, 25 (4). pp. 1878-1893. ISSN 1460-4582

<https://doi.org/10.1177/1460458218810760>

---

© 2018, The Author(s). This is an author produced version of a journal article published in *Health Informatics Journal*. Uploaded in accordance with the publisher's self-archiving policy.

**Reuse**

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

**Takedown**

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing [eprints@whiterose.ac.uk](mailto:eprints@whiterose.ac.uk) including the URL of the record and the reason for the withdrawal request.



[eprints@whiterose.ac.uk](mailto:eprints@whiterose.ac.uk)  
<https://eprints.whiterose.ac.uk/>

# The Assessment of Data Quality Issues for Process Mining in Healthcare Using MIMIC-III, a Freely Available e-Health Record Database

Angelina Prima Kurniati<sup>1,2,\*</sup>, Eric Rojas<sup>3</sup>, David Hogg<sup>1</sup>, Geoff Hall<sup>4</sup>, Owen Johnson<sup>1</sup>

<sup>1</sup> School of Computing, University of Leeds, UK

<sup>2</sup> School of Computing, Telkom University, Bandung, Indonesia

<sup>3</sup> Computer Science Department, Pontificia Universidad Catolica de Chile, Chile

<sup>4</sup> Leeds Institute of Cancer and Pathology, St James's University Hospital, Leeds, UK

\* Corresponding author. Email: [scapk@leeds.ac.uk](mailto:scapk@leeds.ac.uk)

## Abstract

There is a growing body of literature on process mining in healthcare. Process mining of Electronic Health Record (EHR) systems could give benefit into better understanding of the actual processes happened in the patient treatment, from the event log of the hospital information system. Researchers report issues of data access approval, anonymisation constraints, and data quality. One solution to progress methodology development is to use a high quality, freely available research dataset such as MIMIC-III, a critical care database which contains the records of 46,520 Intensive Care Unit (ICU) patients over 12 years. Our paper aims to: (i) explore data quality issues for healthcare process mining using MIMIC-III, (ii) provide a structured assessment of MIMIC-III data quality and challenge for process mining, and (iii) provide a worked example of cancer treatment as a case study of process mining using MIMIC-III to illustrate an approach and solution to data quality challenges. The EHR software was upgraded partway through the period over which data was collected and we use this event to explore the link between EHR system design and resulting process models.

Keywords: data quality, healthcare, MIMIC-III, process mining

## 1. Introduction

Process mining (1) is an emerging approach for discovering and analysing business processes based on event logs extracted from information systems. An event log in an Electronic Health Record (EHR) contains records of timestamped events that have taken place to the patients during their treatment. The approach has been applied to the analysis of healthcare processes (2) with the aims of improving quality of care, patient safety, and optimization of resources (3). Healthcare, with its emphasis on the quality (as well as the cost) of outcomes, is arguably more complex than many other industries. Many healthcare organisations have been slow to adopt large scale information systems to manage and record their clinical data and activities but such EHR systems are becoming more common

(4) and have the potential to provide a rich source of insights for process improvement (5). The data is however sensitive and confidential (6) creating unique challenges.

Reviews of process mining in the general healthcare literature (7) and more specifically in cancer care (8) have reported common issues gaining approval for access to data, difficulty linking data from multiple sources, constraints due to anonymisation, challenges accessing domain experts to help understand business processes and concerns over data quality. Healthcare processes are recognized as complex because the steps are nonlinear, unpredictable and dependent on each individual's medical and personal circumstances - such processes often not follow standard sequences (9). Data quality issues have direct implications for the quality of healthcare provision (10). For researchers, detailed understanding of data quality in health records is essential if we are to use them to draw conclusion about healthcare provision (11). Data quality issues may be explained in terms of underlying bias in the way that users interact with the system (12). Several studies have explored data quality issues related to process mining in healthcare (13–16), and one study has even used process mining as an approach to better understand healthcare data quality (17), but none of them address the Medical Information Mart for Intensive Care III (MIMIC-III) data quality issues for healthcare process mining. The relationship between healthcare processes and data quality or bias were examined by Agniel et al. (12) using only lab test orders. In our paper, we address the data quality issues across the full range of data types from a freely available EHR data so that other researchers can also explore this complex issue.

Weiskopf & Weng (18) developed a framework for assessing data quality in electronic health records using five dimensions linked to seven methods used to address quality issues. This framework has informed our approach. Our own experience of health data quality was shaped by work on process mining chemotherapy care data from a large UK cancer centre which took eight iterations surfacing data quality issues for domain expert review over nine months with each iteration revealing unanticipated further complexity (19). We developed a healthy respect for the data quality challenges in healthcare. Our approach applying the Weiskopf & Weng framework has been to keep careful track of extract criteria and transform data manipulations before loading to standard analytics tools so that strategies to deal with data quality are documented and can be refined through multiple iterations.

One solution to progress methodology development is for process analytics researchers to try out their methods using a freely available research dataset so that they can compare results. In this paper we describe the potential to use the MIMIC-III dataset for process mining in healthcare. The dataset used in this research is MIMIC-III v1.3 released on December 10th, 2015 (available at <https://mimic.physionet.org>). MIMIC-III is a freely available critical care database which contains the de-identified records of 46,520 Intensive Care Unit (ICU) patients at the Beth Israel Deaconess Medical Centre (BIDMC) in Boston, USA covering 58,976 admissions between June 2001 and October 2012. The data was largely sourced from the hospital's Philips CareVue (CV) EHR until that system was replaced in 2008 by a new EHR called iMDSoft MetaVision (MV). Our paper aims to: [i] highlight the opportunities for using MIMIC-III for healthcare process mining, [ii] provide a structured assessment of MIMIC-III using the Weiskopf & Weng EHR data

quality framework (18), and [iii] explore the potential impact of the change in the hospital EHR software in 2008 on data quality and our confidence in resulting process models. The method for de-identification of the data presents several challenges for process mining which we discuss. We provide a worked example of cancer treatment as a case study in process mining to illustrate an approach to data quality challenges.

This paper is structured as follows. Section 2 describes the description of the MIMIC-III used for process mining, Section 3 presents the followed Methodology and the addition of a new stage. Section 4 illustrates the application of the proposed method, the quality assessment and the analysis of the potential impact of the system changes. Finally, Section 5 and Section 6 describe the Discussion and Conclusions.

## 2. Using the MIMIC-III data for process mining

The MIMIC-III data is available as a set of downloadable files which can be used to create a relational database with 26 tables (20). In our previous works (21–23), examples of process mining using MIMIC-III has been published, some details of the data structures relevant to process mining were given, and methods to use MIMIC-III database for process mining were implemented. There were problems with data quality mentioned, but there was no comprehensive discussion on data quality assessment. In this paper we examine the methods for using MIMIC-III and provide a structured data quality assessment. The MIMIC-III dataset contains medical event data such as: vital signs, medications, laboratory results from hospitals and clinics, charted observations during a patient's stay in the ICU and other clinical notes including nursing notes and discharge summaries (20). Patients were included in MIMIC-III if they had at least one period of care in an ICU but the data available covers all of their hospital data for all episodes, it therefore provides a comprehensive example of EHR data from a large and busy hospital.

The data available to researchers has been curated by the MIMIC-III team at the MIT Lab for Computational Physiology to a) address known data quality issues, b) reconcile differences in the data formats of the Philips CareVue (CV) EHR and iMDSoft MetaVision (MV) and c) de-identify the data to make it suitable for external research use (20). The methodology for anonymisation was in accordance with USA Health Insurance Portability and Accountability Act (HIPAA) recommendations to remove all personally identifiable data (names, addresses, phone numbers etc.) with one important exception. HIPAA recommends the removal of event dates and times on the basis that these present unique patterns that could aid re-identification – such an approach would have prevented process mining. The approach to anonymisation used in MIMIC-III was date shifting. All dates were shifted into the future (between 2100 and 2200) by a random offset generated for each patient. Time of the day, day of the week, and approximate seasonality were conserved during date shifting. Time in the MIMIC-III database is stored with one of two suffixes with different resolutions: time (down to the minute) and date (down to the day). Most data are recorded with a time indicating when the observation was made (`charttime`) and when it was validated (`storetime`).

The MIMIC-III dataset does not contain an event log but 16 out of 26 tables do contain timestamped information which can be used to construct the sequence of events

for a patient. Tables are linked by identifiers: `subject_id` refers to a unique patient, `hadm_id` to a unique admission, and `icustay_id` to a unique ICU stay. Process mining focuses on event data but other tables provide supplementary data. For example, when the `chartevents` table is used, we need to refer to `d_item` table to get the corresponding label of `item_id`. Diseases and procedures are encoded using the International Classification of Diseases Version 9 (ICD-9) codes, and the mapping for these can be found in `diagnoses_icd` and `procedures_icd` tables.

One de-identification step was to obstruct identifiable times through random date shifting and this creates a challenge and limitations for process mining. Specifically if dates have been shifted by a random amount, workflow analysis looking at busy days is not meaningful (2). In the MIMIC-III, dates have been shifted so that they are internally consistent for the same patient, but randomly distributed in the future. Similarly the impact of bottlenecks, e.g. of patients waiting for care on a busy day, cannot be deduced. Despite the date shifting, the MIMIC-III dataset still contains detailed information of real healthcare processes for individual patient during their time in the hospital including comprehensive details on administrative activities (admission, discharge, transfer to a ward, etc.) and clinical activities (triage, test and scans, diagnosis, etc.) (24). This data can reveal much about the care processes as they are experienced by each individual patient.

### 3. Methodology

Our goal to understanding suitability of the MIMIC-III dataset for process mining in healthcare is focused on the assessment of data quality. The process mining methodology in this research follows the L\* lifecycle model, suggested by van der Aalst et al (1) but with some adaptation. The standard L\* model consists of five stages: Plan and justify (Stage 0), Extract (Stage 1), Create a control-flow model and connect it to the event log (Stage 2), Create an integrated process model (Stage 3), and Provide operational support (Stage 4). Stage 4 of the L\* life-cycle model is only relevant for researchers who can influence the operation of the organisation and was therefore beyond the scope of our study.

We needed to introduce an additional Stage between Stage 0 and Stage 1, to prepare the MIMIC-III data for process mining. This new Stage involved the preparation steps to reconstruct the MIMIC-III data files into a relational database. This exercise helped us understand the underlying data model before Stage 1 and informed further iterations (see Fig. 1). In earlier work, we develop a question driven methodology for process mining in healthcare (25) and we used this approach to motivate our investigation.

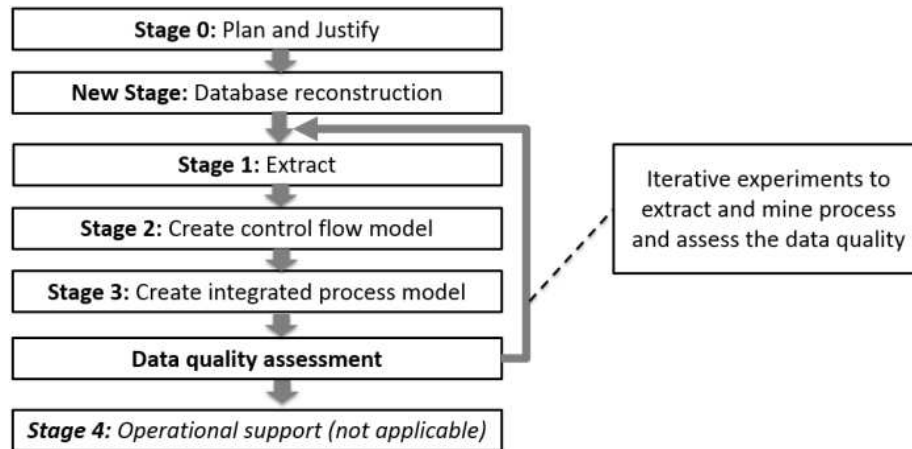


Fig. 1. Adaptation of the L\* life-cycle model to include data quality assessment

Our assessment of data quality follows the Weiskopf & Weng framework (18) with a specific focus on the impact of data quality for process miners. The data quality assessment would mark various issues in the process characteristics and the quality of an event log for process mining research (2,14). In term of process characteristics, EHR databases including MIMIC-III face common challenges due to the voluminous data - large number of cases and events, case heterogeneity - large number of distinct traces, event granularity - large number of distinct activities, process flexibility and concept drifts. Those problems would be identified by showing a few basic metrics such as the number of cases in the event log, the number of unique activities, and the number of distinct traces. While the quality of the event log would be discussed by identifying four broad categories of problems, which are missing data, incorrect data, imprecise data, and irrelevant data. Those four categories would be applied to check problems in case, event, activity name, and timestamp. Methods and dimensions in the Weiskopf and Weng data quality framework could be viewed as the first level quality assessment with the details being the process mining – specific metrics.

The implementation of the L\* lifecycle is presented in section 4.1, with a worked example of process mining using MIMIC-III. Data quality assessment is presented in section 4.2. The effects of the EHR system changes are presented in section 4.3. The current tools used by our group are: (1) Relational database: PostgreSQL version 9.5 using PgAdminIII for SQL-based data extract, (2) Data transformation: Python programming language, (3) Process Mining: the ProM toolkit version 6.4.1 as a standard process mining toolset, (4) Experiment documentation: standard templates in Microsoft Word, (5) Version management: GitHub repository to support rapid and iterations (this project is at <https://github.com/angelinast3/mimic3qualityforpromin>).

## 4. Results

### 4.1 Process mining with the MIMIC-III dataset

#### 4.1.1 Stage 0: Plan and Justify

The justification for this research is based on recognising a need for a high quality, freely available research dataset that can be used internationally to progress process mining methodology development. Planning for this study was based on three research questions:

- Q1. Can the MIMIC-III database be used to better understand data quality issues for process mining in healthcare?
- Q2. What are the data quality issues for process mining with MIMIC-III?
- Q3. How might the change in the EHR system in 2008 affect the data quality?

Q1 was addressed by presenting a cancer specific worked example in this paper to illustrate our approach to data quality in sections 4.1.2 to 4.1.4. Q2 will be addresses by applying Weiskopf and Weng's EHR data quality assessment framework (18) in section 4.2. Q3 will be addressed by investigating the differences in processes and data quality of the CareVue (CV) and MetaVision (MV) systems in the MIMIC-III dataset in section 4.3.

#### 4.1.2 New Stage: Database reconstruction

This stage included the reconstruction of the MIMIC-III dataset to create a relational database in PostgreSQL. This included downloading 26 csv files (more than 6 GB in total) along with scripts to import the data into the PostgreSQL database following the guidelines at <https://mimic.physionet.org/gettingstarted/dbsetup>. The concept-level Entity Relationship (E-R) diagram for the resulting database is presented in **Fig 2**. The interpretation of events and their significance was based on an understanding of the entities and their relations described in the data descriptors paper [15]. It is possible to extract specific table(s) in the MIMIC-III database for process mining, but our approach was to reconstruct and explore the fullest possible set of data and judge relevance in later stages. This was done to maintain the data quality for the next stages, which are the Extract, Create control flow model, Create integrated process model, and Data quality assessment stages. The entities in bold contain timestamped information which can be used to construct event log data for process mining (Fig. 2).

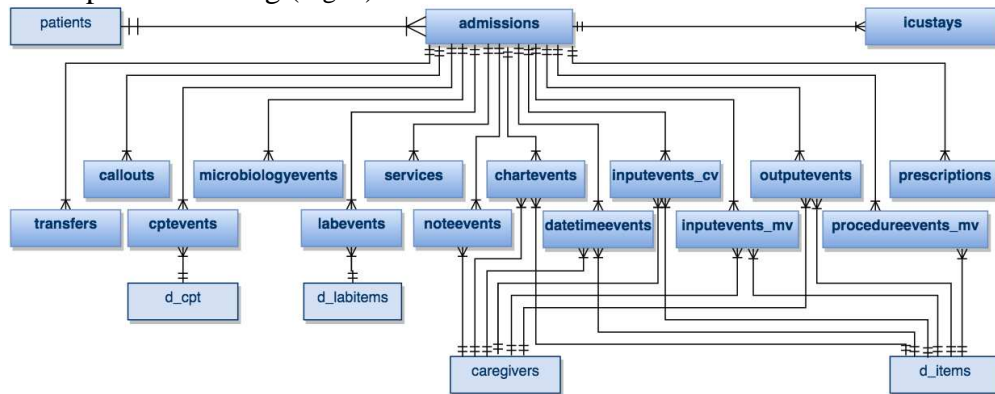


Fig. 2. Concept-level Entity-Relationship (E-R) diagram of the MIMIC-III database

### 4.1.3 Stage 1: Extract

The database reconstruction supports multiple possible extractions so that Stage 1 can be repeated many times. In all experiments we followed a standard documentation template. In the example below (Fig. 3), we select records of cancer patients in the admissions table based on the ICD-9 codes for cancer diagnoses which are 140x-239x (26) creating an event log of [case\_id, activity, timestamp]. The result table was then saved in .csv format and loaded into ProM or other available process mining tools.

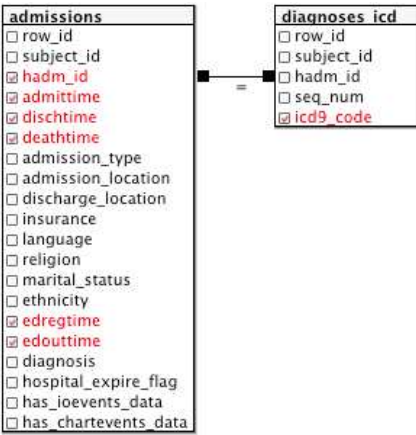
<b>Experiment title:</b> Event log creation of cancer admissions	Date: 17/05/2017 Code: M3C-ADM001
<b>Area of investigation.</b> This experiment is to create an event log of cancer patient records from the admissions table of the MIMIC-III database. The format is [case_id, activity, timestamp].	
<b>Data source.</b> The admissions table in the MIMIC-III database.	
<b>Research question.</b> Is the admissions table in the MIMIC-III database can be used for process mining of cancer patient admissions?	
<b>Hypothesis.</b> The admissions table can be used as it provides at least minimum requirements for process mining, which are case_id, activity, and timestamp.	
<p><b>Method</b></p> <ol style="list-style-type: none"> <li>1) <b>Extract</b> the MIMIC-III database in PostgreSQL by selecting [admittime, disctime, edregtime, edouttime, deathtime] of cancer patient admissions [icd9_codes 140x-239x].</li> <li>2) <b>Transform</b> into event log with [case_id, activity, timestamp] format.              case_id = hadm_id              activity = {admission, discharge, ED reg, ED out, death}              timestamp = {admittime, disctime, edregtime, edouttime, deathtime}</li> <li>2) Save as .csv file</li> <li>3) <b>Load</b> into ProM</li> </ol>	
	
<b>Discussion</b> The admissions table contains minimum required columns for process mining, with additional step of creating transactional table. The resulted event log would be loaded into ProM in the Stage 2.	
<b>Conclusion:</b> Hypothesis proved. The admissions table can be used to create event logs.	

Fig. 3. Example of experiment documentation for a simple extract

### 4.1.4 Stage 2: Create control flow model and connect event log

The next stage in the L\* model is to create control flow models and link these to the event log. For this example, three main plugins were used: (1) “Convert CSV into XES” to convert the event log into the XES format (www.xes-standard.org) required by ProM, (2) “Add artificial events >> START and END events” because there were no explicit start and end events in the event log we have created, (3) “BPMN Analysis using Heuristics Miner” as a commonly used discovery algorithm. The first research question was Q1 (Can the MIMIC-III database be used to better understand data quality issues for process mining in healthcare?). In our example we show how this can be addressed to find the most followed admission paths of cancer patients (Fig. 4).



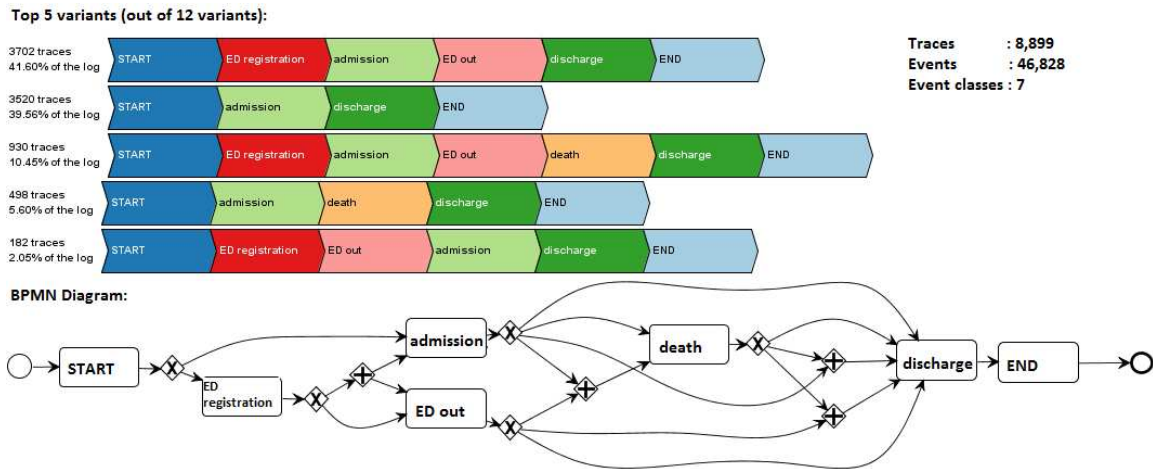


Fig. 4. Five most common variants and BPMN process model of admissions

The top part of **Fig. 4** represents the five most common variants out of 12 variants. The process model in the bottom part of **Fig. 4** was in Business Process Modelling Notation (BPMN). This process model fitness was 100%, precision was 81.713% and generalization 85.275%. We then reviewed the resultant model with a UK-based oncologist for sense-checking. The apparent data quality issue that discharge took place after death was investigated and was found to reflect the hospital’s standard administrative process. There is also variations in administrative steps. For example, when patients were being admitted (admission) after discharged from the Emergency Department (ED out), they might be admitted before ED out (in variant 1) or ED out before admitted (see variant 5). In variant 1, admission – ED out duration are 118 minutes in average with 89 minutes median. In variant 5, ED out – admission duration are 132 minutes in average with 46 minutes median.

### 4.1.5 Stage 3: Create integrated process models

In Stage 3, the models are extended with other perspectives (e.g. time and resources). Our analysis studies the effect of the change of EHR and this is discussed in section 4.3.

## 4.2 Data quality assessment of MIMIC-III for process mining

### 4.2.1 Element presence

The first data quality assessment method is element presence, which was done by checking for the three minimum required attributes for process mining, which are `case_id`, `activity` and `timestamp`. Summary of the event tables is in **Table 1**.

Table 1. Element presence checking: the event tables

Table name	Element presence?				
	case_id			activity	timestamp
	s	h	i		
<b>admissions</b>	✓	✓	✗	{admit, disch, death, edreg, edout}-time	
<b>callout</b>	✓	✓	✗	{create, update, acknowledge, outcome, firstreservation, currentreservation}-time	
<b>chartevents</b>	✓	✓	✓	<b>itemid</b> (links to d_items)   <b>charttime, storetime</b>	

Table name	Element presence?				
	case_id			activity	timestamp
	s	h	i		
<code>cptevents</code>	✓	✓	✗	<code>cpt_cd</code> (links to <code>d_cpt</code> )	<code>chartdate</code>
<code>datetimeevents</code>	✓	✓	✓	<code>itemid</code> (links to <code>d_items</code> )	<code>charttime, storetime</code>
<code>icustays</code>	✓	✓	✓	<b>ICU in, ICU out</b>	<code>intime, outtime</code>
<code>inputevents_cv</code>	✓	✓	✓	<code>itemid</code> (links to <code>d_items</code> )	<code>charttime, storetime</code>
<code>inputevents_mv</code>	✓	✓	✓	<code>itemid</code> (links to <code>d_items</code> )	<code>starttime, endtime, storetime</code>
<code>labevents</code>	✓	✓	✗	<code>itemid</code> (links <code>d_labitems</code> )	<code>charttime</code>
<code>microbiologyevents</code>	✓	✓	✗	<code>spec_itemid</code>	<code>chartdate, charttime</code>
<code>noteevents</code>	✓	✓	✗	<code>description</code>	<code>charttime, chartdate</code>
<code>outputevents</code>	✓	✓	✓	<code>itemid</code> (links to <code>d_items</code> )	<code>charttime, storetime</code>
<code>prescriptions</code>	✓	✓	✓	<code>drug, drugtype</code>	<code>startdate, enddate</code>
<code>procedureevents_mv</code>	✓	✓	✓	<code>itemid</code> (links to <code>d_items</code> )	<code>starttime, endtime, storetime</code>
<code>services</code>	✓	✓	✗	<code>curr_service</code>	<code>transfertime</code>
<code>transfers</code>	✓	✓	✓	<code>eventtype</code>	<code>intime, outtime</code>

Note: s = `subject_id`, h = `hadm_id`, i = `icustay_id`

The MIMIC-III database provides three types of identifiers which can be used as `case_ids`, which are identifier in patient level (`subject_id`), which might have more than one admissions (`hadm_id`), and more than one ICU stays (`icustay_id`). This represents the event granularity problem for process mining, where process miner should aware which level should be used in the analysis. The activity names in the MIMIC-III database are available directly in some tables, e.g. the `admissions` and `callout` tables, or can be made available by referring to the other table, e.g. the `chartevents` and `datetimeevents` table by linking to `d_items` table. Activity names which refer to items in the `d_items` table would have two different level of granularity, which are label (fine-grained level) and category (coarse-grained level). For timestamps, other than date versus time and `charttime` versus `storetime` issues (see section 2), there are four tables recorded the start and end times (`icustays`, `inputevents_mv`, `procedureevents_mv`, and `transfers` tables). When analysis of activity duration is needed, those four tables can be used. Another issue raised by the different granularity of timestamps (**-date** down to the day and **-time** down to the minute) also presents clear issues in discovering proper event sequencing, calculating duration, etc.

#### 4.2.2 Data element agreement

Data element agreement was done to compare two or more elements in the database to see if they report the same or compatible information. For checking data element agreement, we traced back to the MIMIC-III website and data descriptor (20). Our findings were as follows.

(1) When we use `subject_id` or `hadm_id` as the `case_id`, the `admissions` and `transfers` tables are complete. But when we use `icustay_id`, only `icustays` and `transfers` tables are complete. All other tables can be used as needed, but the completeness can be checked by reference to those tables. (2) Time in the MIMIC-III database is stored as datestamps in three tables (`cptevents`, `noteevents`, and `prescriptions`) and as timestamps in the other tables (`admissions`, `icustays`, etc.).

This might encounter issues in process mining when combining tables with different time/date stamps resulting in incorrect temporal order throughout a patient’s encounter. For example, all prescriptions are recorded with time 00:00 in the `prescriptions` table. Combining it with the `admissions` table which does have detailed timestamps would result in prescription appearing to precede the admission. (3) The MIMIC-III data descriptor said that the dates have been shifted into the future to the years 2100 to 2200 but we found some tables had records of events before and after those dates. This happens because some tables also contain historical patient data, for example test results from chemistry and hematology in `labevents` table. This presents no immediate data quality issues but could lead to errors if dates were used in selection.

### 4.2.3 Distribution Comparison

Distribution comparisons were done to check completeness, concordance, and plausibility of the MIMIC-III database to the data descriptor (20). The data descriptor describes charted events such as notes, laboratory tests, and fluid balance as being stored in a series of ‘events’ tables. The number of distinct patients (46,520), hospital admissions (58,976), and ICU stays (61,532) all conform to the data descriptor. For process mining, we identified event tables (see **Table 1**), including `noteevents` which contains all notes, `labevents` which contains all laboratory measurements, and `inputevents` and `outputevents` which contain data about fluid balance. Completeness checking was also done to validate the data distribution by checking on the minimum components needed in process mining (case id, activity and timestamp). The number and percentage of missing components in each table were then recorded and summarised. The summary is presented in **Table 2**.

Table 2. Completeness Checking

Table name	# of row	Missing				
		s	h	i	activity	timestamp
<code>admissions</code>	58,976	0	0	-	-	109,320 (37.07%)
<code>callout</code>	34,499	0	0	-	-	54,366 (26.26%)
<code>chartevents</code>	330,712,483	0	388,836 (0.12%)	3,069,518 (0.93%)	0	0
<code>cptevents</code>	573,146	0	0	-	-	471,601 (82.28%)
<code>datatimeevents</code>	4,486,049	0	1,751 (0.04%)	96,680 (2.16%)	0	0
<code>icustays</code>	61,532	0	0	0	0	10 (0.02%)
<code>inputevents_cv</code>	17,528,894	0	22,658 (0.13%)	30,453 (0.17%)	0	0
<code>inputevents_mv</code>	3,618,991	0	0	57,572 (1.59%)	0	0
<code>labevents</code>	27,872,575	0	5,745,330 (20.61%)	-	0	0
<code>microbiologyevents</code>	328,446	0	430 (0.13%)	-	70 (0.02%)	24,705 (7.52%)
<code>noteevents</code>	2,078,705	0	231,836 (11.15%)	-	0	312,091 (15.01%)
<code>outputevents</code>	4,349,339	0	4,833 (0.11%)	-	0	0
<code>prescriptions</code>	4,156,848	0	0	1,447,706 (34.83%)	0	3,182 (0.08%)
<code>procedureevents_mv</code>	258,066	0	0	4,501 (1.74%)	0	0
<code>services</code>	73,343	0	0	-	0	0
<code>transfers</code>	261,897	0	0	174,176 (66.51%)	0	59,000 (11.26%)

Note: s = `subject_id`, h = `hadm_id`, i = `icustay_id`

**Table 2** shows that `subject_id` is complete in all event tables, but `hadm_id` and `icustay_id` are not. There are 70 missing `spec_itemid` in `microbiologyevents` table, but those can be replaced completely by `spec_type_desc`. In the `admissions` table, there are missing timestamps which represent patients who are not death or not admitted in the emergency department (ED). This was illustrated in **Fig. 4**. The `callout`

table also has some missing timestamps (see **Table 2**). The MIMIC-III documentation mentions that the collection of callout data only began part way through the MIMIC-III database and with date shifting this missing data has been spread at random. This incompleteness should also be considered when using this table. Incomplete timestamps in the `charttime` of `microbiologyevents` and `noteevents` tables can be derived by linking to `chartdate`, but consider that the granularity level would be different. This is also happened in `cpevents`, `prescriptions` and `transfers` tables so process mining would be unreliable.

#### 4.2.4 Validity checking

Validity checking is a data quality assessment method to determine if values ‘make sense’ within the context of the problem domain, following the guidance in (18). It was done in this study by querying data in each table and between related tables. For example, we checked validity of the ICD-9 codes for diagnoses and procedures in the tables and also checked duplicates between different tables.

The first finding was about the ICD-9 codes for diagnoses and procedures. The MIMIC-III dataset provides reference tables, which are `d_icd_diagnoses` and `d_icd_procedures`. But, there are 144 out of 14,711 (0.98%) missing codes in `d_icd_diagnoses` table, and there are 16 out of 258,082 (0.01%) missing codes in `d_icd_procedures` table.

The second finding was that there are duplicates between different tables, such as `datetimeevents` and `admissions` table, `labevents` and `chartevents` tables. For example, there is a `hospital admit date` in the `datetimeevents` table which is duplicated with `admittime` in the `admissions` table. The `admissions` table is specifically providing information of patient admissions in the hospital and has a complete set of admission and discharge times. Meanwhile, the `datetimeevents` table contained all date measurements about a patient in the ICU including hospital admission dates and it was found incomplete. Our investigation compared all records and found that there were 24,549 admissions recorded both in the `admissions` and `datetimeevents` tables with 1,696 records are matched to the duplicate on the other table, but the others are not. In the 24,354 of 24,549 duplicated admissions (99.2%), admission dates in the `admissions` table were recorded earlier than those in the `datetimeevents`. Our suggestion to handle this issue was to use only the valid records and ignore the duplicated records in the other table(s). We suspect there are probably more issues like this with other tables.

### 4.3 How does the 2008 change in the EHR system affect data quality?

Our third research question specified in the Plan and Justify Stage in section 4.1.1, aimed to identify the effects of the change in EHR system in 2008 from CareVue (CV) to MetaVision (MV). Specifically any impact on data quality. Because the MIMIC-III anonymisation used date shifting a simple comparison of dates was not possible. However when MIMIC-III was constructed it was not possible to reconcile `inpuvents` so there

are separate `inputevents_cv` and `inputevents_mv` tables (the only separated tables for the two systems). Our approach was to work backwards from these two tables to identify which hospital admissions had been recorded on which EHR. We used this logic to insert a field in each table that indicated which of the two systems had been used and then used these new fields to extract event logs before and after the new EHR. The checking on `itemid` column in `d_items` table are presented in **Table 3**.

Table 3. Details of `itemid` in `d_item` table

dbsource	linksto	occur	dbsource	linksto	occur
carevue	chartevents	4,982	metavision	chartevents	924
	datetimeevents	52		datetimeevents	141
	inputevents_cv	2,929		inputevents_mv	422
	outputevents_cv	1,087		outputevents_mv	74
hospital	microbiologyevents	436		procedureevents_mv	125

**Table 3** shows that `d_items` table has a column to identify the database source (`dbsource`) and which tables link to those sources (`linksto`). This means that we could identify the differences between the EHR systems through four tables: `chartevents`, `datetimeevents`, `inputevents`, and `outputevents`. We could also identify the sources of each record in the other tables by referring to those four tables. The `dbsource` `hospital` would be ignored at this point because our focus was on the CV and MV systems used. Identifying the original EHR is possible because we marked each admission of cancer patients in the `admissions` table with CV, MV, or both (if the admission is recorded both in CV and MV), and created separate event logs from each data sources. We then compared the models discovered from CV and MV event logs using `DifferenceGraph` plugin in ProM. This plugin supports the identification of differences and commonalities between two process models. The result is presented in **Fig. 5**.

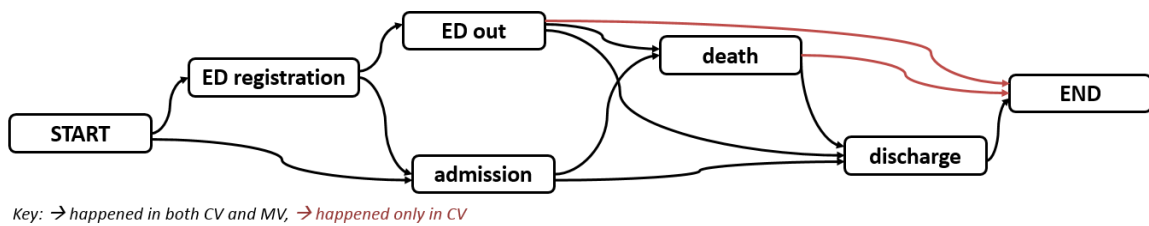


Fig. 5. `DifferenceGraph` of admissions in CV and MV. It is evidence that in the new system, all patients have `discharge` event as the last event in the case.

Fig. 5 shows the differences between admissions in CV and MV, which found changes in the last activities happened in those two system. The admissions in CV (35833 admissions) ended with either `discharge` (35788/ 99.874%), `death` (26/ 0.073%), or `ED out` (19/ 0.053%). But all admissions in MV (19623 admissions) ended with `discharge` (100%), suggesting that MV has better administrative records than CV. The process model of CV fitness was 99.96%, precision was 87.84%, and generalization was 90.06%; while the process model in MV fitness was 99.99%, precision was 93.82%, and generalization was 89.159%. This suggested that both models are able to replay the observed behavior (high fitness), describe the system generally (high generalization), and not allows for too much behavior (high precision). The combination of interpreting event frequencies, the

DifferenceGraph and the conformance measures leads to the conclusion that the EHR system change did affect the process model and quality requiring further investigation.

## 5. Discussion

Process mining has been used in many case studies in healthcare data (15,27,28) and MIMIC-III as a freely available de-identified healthcare database has been made available to researchers and used in many research projects. There are many opportunities for using process mining on MIMIC-III, but only three published examples so far (21–23). Agniel et al. (12) used lab test orders and linked data quality bias to evolving healthcare processes. In this paper we have explored the broader data quality issues that future process mining researchers will need to understand in order to use the full range of EHR data. We have provided a worked example of cancer treatment to illustrate this.

When using MIMIC-III for process mining, the L\* lifecycle can be followed. However, an additional stage between Stage 0 and Stage 1 for database reconstruction is needed. Given that researchers using MIMIC-III are unlikely to have direct access to the Beth Israel Deaconess Medical Centre that provided the data, it should be clear that explicit operational support (Stage 4) is not applicable. In our worked example, we have demonstrated that the MIMIC-III dataset can be used for process mining with many interesting results. For example, detecting the hospital's standard administrative process, investigating variations in the treatment steps, and visualising differences and commonalities between multiple process models. Such analysis can provide insights and learning opportunities for potential clinical improvements and help to progress methodology development in process mining. The ability to compare and publish against a freely available database (such as MIMIC-III) creates opportunities for international collaboration.

We assessed MIMIC-III using the completeness, plausibility, correctness, and concordance dimensions of the Weiskopf & Weng EHR data quality framework (18). We did not assess the data by the **currency** dimension as the MIMIC-III database covers data from 2001 – 2012 and is clearly no longer current. It is important to emphasise that the MIMIC-III dataset does contain **complete** sets of hospital admissions and Emergency Department (ED) registrations. The minimum required attributes for process mining (case ID, activity, time) were available in 16 tables (see **Table 1** which lists these) and therefore there is a rich set of data that can be extracted and converted into event logs for process mining. Potential case IDs are the patient number (`subject_ID`), the admission number (`hadm_ID`), and the ICU stay number (`icustay_ID`). These three options enable process mining to be done at three different levels: patient level, admission level, and ICU stay level. There are some incomplete elements for some tables (see **Table 2** which summarises these) and some historical records outside of the documented time boundary which can be excluded with care. The **provenance** of the MIMIC-III dataset and its careful curation suggests the plausibility is generally good. The clinical diagnoses and procedures are coded using ICD-9 codes and the **correctness** of the coding are depending on the user at the time of entry. Given the human data input, the coding is inevitably subjective, but this reflects the reality of complex healthcare. It is also important to be aware of some concordance

issues because there are some duplicates between tables, as discussed in Section 4.2.4. The description in section 4.2 provides the answer to the Q2 (What are the data quality issues for process mining with MIMIC-III?). Some of the issues addressed in our paper were: missing events, case attributes, activity names or codes, timestamps, and attributes; incorrect events, cases, or timestamps; imprecise resources and timestamps; completeness of the dataset, meaningfulness of the values in the data, correct level of accuracy and format, and repeated events. Despite those issues, the overall data quality of MIMIC-III was found to be good; there is a rich set of detailed event data covering a 10-year period and broadly representative of a real-life hospital. We conclude that the freely availability of MIMIC-III makes it suitable as a basis for reproducible healthcare process mining research, including control-flow, conformance, and performance analysis.

One important data quality issue in using MIMIC-III for process mining is the way that anonymization has been achieved using a date shifting. All real dates have been shifted into the future by a random offset generated for each patient. For example, two patients seen on the same day may have admission dates of 2100-06-07 and 2185-04-08. This anonymization approach makes it impossible to do some analysis in process mining. Public holidays such as Thanksgiving may well impact on processes but its effect will be distributed over 100 year time period. Similarly, it is not possible to identify workflow looking at busy days, comparison of weekend vs weekday workloads, and bottleneck analysis of patients waiting time on a busy day.

A further challenge in the MIMIC-III database is the change of the EHR system that took place in 2008. The CareVue system was used in 2001-2008 and the MetaVision system used in 2008-2012. The two systems have some data in different formats and the item ids differ. Our study found that the EHR system change did affect the process models and therefore quality of process mining. Our partial solution to this problem involves using the two different `inputevents` tables to mark the original EHR on the other tables on our copy of the database. This enabled us to create separate event logs from the two EHR and this made it possible to compare process models from the same process across the two different systems. Our comparison approach identified other differences and commonalities between process models in the two EHR systems in Section 4.3. These differences and commonalities provide insights of the different process execution, specific activities, their order of executions, and those missing. These findings suggest that an awareness of EHR system changes is important to understand the root cause of some data quality issues.

## 6. Conclusion

This paper has answered the three research questions developed in the initial stage of the study. Firstly, we demonstrated the applicability of process mining using the MIMIC-III dataset and performed a data quality assessment. The additional stage in the L\* lifecycle involved database reconstruction and made it possible to do Stage 1 to Stage 3 as many times as needed. We believe this iterative approach to identifying and understanding complex data quality concerns is essential. The use of a standard template for experiment documentation provides control over the iterative approach and helps document each data

quality issue and how it can be mitigated in experiment design. We found that it was possible to mine complex hospital processes with existing techniques to analyse processes for different groups of patients. Secondly, we used the Weiskopf & Weng framework successfully to identify multiple data quality issues in section 4.2 and 4.3. MIMIC-III has been used by EHR researchers to generate a range of insights and the data quality issues we have explored for process mining can therefore be assumed to be generally relevant to EHR research. Finally, we identified the effects of the change of the EHR system on the process and specifically on data quality, and we found evidence that the change of EHR did affect the process model and quality of results. Process mining researchers that have a concern about this may follow our technique to select records from either of the two EHR systems.

The most significant contribution of MIMIC-III to process mining is the opportunity to develop techniques which can be shared with and validated by other researchers and replicated against EHR dataset across the world. Future work will explore several aspects. To improve data quality by using data cleaning strategies, including resolving missing data and aggregation of similar events. To apply different discovery algorithms to find process models with better conformance to the event log. To investigate the effect of EHR system changes on the healthcare processes. Finally, to focus on advanced analysis with more clinically based research questions from medical experts and comparison studies with real-life datasets.

**Acknowledgments.** This research is supported by the ClearPath Connected Health Cities Project and the Indonesia Endowment Fund for Education (LPDP).

## References

1. van der Aalst WMP. Process Mining: Data Science in Action. In: Process Mining. 2nd ed. Springer-Verlag Berlin Heidelberg; 2016. p. 3–23.
2. Mans R, van der Aalst WMP, Vanwersch R. Process Mining in Healthcare: Evaluating and Exploiting Operational Healthcare Processes. 1st ed. Springer International Publishing; 2015. 1-91 p.
3. Mans RS, Schonenberg MH, Song M, van der Aalst WMP, Bakker PJM. Application of Process Mining in Healthcare – A Case Study in a Dutch Hospital. Proc BIOSTEC 2008. 2008;25:425–38.
4. Johnson OA, Fraser HSF, Wyatt JC, Walley JD. Electronic health records in the UK and USA. Lancet (London, England) [Internet]. 2014;384(9947):954. Available from: <http://www.sciencedirect.com/science/article/pii/S0140673614616263>
5. Weber GM, Mandl KD, Kohane IS. Finding the Missing Link for Big Biomedical Data. JAMA. 2014;
6. Homayounfar P. Process mining challenges in hospital information systems. In: Federated Conference on Computer Science and Information Systems (FedCSIS). 2012. p. 1135–40.
7. Rojas E, Munoz-Gama J. Process mining in healthcare: A literature review. J Biomed Inform. 2016;61:224–36.



8. Kurniati AP, Johnson O, Hogg D, Hall G. Process Mining in Oncology: a Literature Review. In: The 6th ICICM, IEEE. Hertfordshire; 2016.
9. Poulymenopoulou M, Malamateniou F, Vassilacopoulos G. Specifying Workflow Process Requirements for an Emergency Medical Service. *J Med Syst*. 2003;27(4):325–35.
10. Bowman S. Impact of electronic health record systems on information integrity: Quality and safety implications. *Perspect Heal Inf Manag*. 2013;10:1c.
11. Weiskopf NG, Bakken S, Hripcsak G, Weng C. A Data Quality Assessment Guideline for Electronic Health Record Data Reuse. *eGems (Generating Evid Methods to Improv Patient Outcomes)*. 2017;5(1):14.
12. Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *BMJ*. 2018;361:k1479.
13. Mans RS, van der Aalst WMP, Vanwersch RJB. *Process Mining in Healthcare: Evaluating and Exploiting Operational Healthcare Processes*. Heidelberg: Springer; 2015. 17-26 p.
14. Bose RPJC. Wanna Improve Process Mining Results ? It ' s High Time We Consider Data Quality Issues Seriously. *Proc 2013 IEEE Symp Comput Intell Data Mining, CIDM 2013 - 2013 IEEE Symp Ser Comput Intell SSCI 2013*. 2013;(1):127–34.
15. Mans R, van der Aalst WMP. Process mining in healthcare: Data challenges when answering frequently posed questions. *Process Support Knowl Represent Heal Care*. 2013;140–53.
16. Lanzola G, Parimbelli E, Micieli G, Cavallini A, Quaglini S. Data Quality and Completeness in a Web Stroke Registry as the Basis for Data and Process Mining. *J Healthc Eng* . 2014;5(2):163–84.
17. Perimal-Lewis L, Teubner D, Hakendorf P, Horwood C. Application of process mining to assess the data quality of routinely collected time-based performance data sourced from electronic health records by validating process conformance. *Health Informatics J*. 2016;22(4):1017–29.
18. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *JAMIA*. 2013;20(1):144–51.
19. Baker K, Dunwoodie E, Jones RG, Newsham A, Johnson O, Price CP, et al. Process mining routinely collected electronic health records to define real-life clinical pathways during chemotherapy. *Int J Med Inform*. 2017;103:32–41.
20. Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. Data Descriptor : MIMIC-III , a freely accessible critical care database. *Sci Data*. 2016;1–9.
21. Alharbi A, Bulpitt A, Johnson O. Improving Pattern Detection in Healthcare Process Mining Using an Interval-Based Event Selection Method. In: *International Conference on Business Process Management*. 2017. p. 88–105.
22. Alharbi A, Bulpitt A, Johnson OA. Towards unsupervised detection of process models in healthcare. *Stud Health Technol Inform*. 2018;247(January):381–5.
23. Kurniati AP, Hall G, Hogg D, Johnson O. Process Mining in Oncology using the MIMIC-III Dataset (accepted version). *IOP J Phys Conf Ser* 971. Bandung,

- Indonesia; 2018;971(012008):1–10.
24. Lenz R, Reichert M. IT support for healthcare processes – premises, challenges, perspectives. *Data Knowl Eng.* 2007;61(1):39–58.
  25. Rojas E, Sepúlveda M, Munoz-Gama J, Capurro D, Traver V, Fernandez-Llatas C. Question-Driven Methodology for Analyzing Emergency Room Processes Using Process Mining. *Appl Sci.* 2017;7(3):302.
  26. Online ICD9/ICD9CM codes - Neoplasms [Internet]. [cited 2017 Mar 16]. Available from:  
<http://icd9cm.chrisendres.com/index.php?action=child&recordid=1059>
  27. Mans R, Schonenberg M. Application of process mining in healthcare—a case study in a dutch hospital. *Int Jt Conf Biomed Eng Syst Technol Springer Berlin Heidelb.* 2008;425–38.
  28. Rebuge A, Ferreira DRD. Business process analysis in healthcare environments: A methodology based on process mining. *Inf Syst.* 2012;37:99–116.