Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00238

Classification and quantification of timestamp data quality issues and its impact on data quality outcome

Rex Ambe

Campus Oefenplein-Etterbeek Ringgold Standard Institution, Solvay Business School, Vrije Universiteit Brussel, Brussel 1050, Belgium

Corresponding author: Rex Ambe (E-mail: rex.che.ambe@vub.be; ORCID: 0000-0002-6168-2354)

Submitted: October 2, 2023; Received: December 1, 2023; Accepted: December 12, 2023

Abstract

Timestamps play a key role in process mining because it determines the chronology of which events occurred and subsequently how they are ordered in process modelling. The timestamp in process mining gives an insight on process performance, conformance, and modelling. This therefore means problems with the timestamp will result in misrepresentations of the mined process. A few articles have been published on the quantification of data quality problems but just one of the articles at the time of this paper is based on the quantification of timestamp quality problems. This article evaluates the quality of timestamps in event log across two axes using eleven quality dimensions and four levels of potential data quality problems. The eleven data quality dimensions were obtained by doing a thorough literature review of

more than fifty process mining articles which focus on quality dimensions. This evaluation resulted in twelve data quality quantification metrics and the metrics were applied to the MIMIC-III dataset as an illustration. The outcome of the timestamp quality quantification using the proposed typology enabled the user to appreciate the quality of the event log and thus makes it possible to evaluate the risk of carrying out specific data cleaning measures to improve the process mining outcome.

Keywords: Timestamp; Process mining; Data quality dimensions; Event log; Quality metrics; Business process.

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00238

1. Introduction

In the information age, the data of an organization is one of its most valuable assets and the growth of the organizations is highly dependent on the management of this data. If the data is not well managed, it may increase the risk of wrong decisions and thus results in potential financial losses. Data from the process executions collected from different information technology (IT) systems of these organizations is stored in event logs. The event log contains the case id, activities, the events and the timestamp of the start and end events. "Process mining is a data-driven analytic method where data from the process executions collected from different IT systems is analyzed to uncover the real behavior and performance of business operations" (Martin et al., 2020). Process mining is therefore aimed at bridging data mining and business process modelling/analysis by putting forward techniques to extract nontrivial information from the data in the event log (Van der Aalst et al., 2012). This nontrivial information when extracted provides organizations with an insight on how their processes performs, thus enabling these organizations to discover, ameliorate and monitor real time processes using the event logs readily available in today's (information) systems (Van der Aalst, 2011; Van Cruchten, 2019). The extent to which we can depend on the outcomes of process mining analyses is directly related to the quality of the input data stored in the event log (Wynn & Sadiq, 2019). This means reliable process mining results are contingent on high quality event log(s) of the business processes mirrored in a running system, but in practice most event logs don't provide the desired data quality. Event logs should therefore be made sure they are of adequate quality before they are used and not naively used for process mining without assuring that they were of adequate quality (Fischer et al., 2020). To ensure that the data in the event log is of adequate quality, we need a metric to quantify the quality and verify

if it meets the required threshold for analysis. Data quality analysis should therefore include an approach for detecting and quantifying timestamp imperfections to help process miners appreciate the state of the information system (Fischer et al., 2020). This article tries to develop timestamp quality metrics from timestamp quality dimensions and the potential data quality problems proposed by van der Aalst (2016) which will serve as an alternative typology to that proposed by Fischer et al. (2020).

Timestamps are crucial for process mining but most research groups the quality issues of timestamps in the event log with that of the activity, case, resource, and others. The quality of event log timestamp has been previously evaluated by Fischer et al., (2020) across two axes with the first axis being the level of abstraction, and it's made up of the event, activity, trace, and log as proposed by Van der Aalst, (2016) and the second axis being the quality dimensions which is made up of the accuracy, completeness, consistency, and uniqueness (table 1). This evaluation resulted in fifteen timestamp quality-related metrics. From Fischer et al.'s (2020) classification, the metrics such as missing trace, missing activity, missing event, and missing timestamp all has to do with missing timestamps at different levels of abstraction; also duplicates within log, duplicates within trace, duplicates timestamp concerns the duplication of timestamps provided it's a timestamp quality issue. In order to resolve this issue, this article evaluates the quality of timestamps in an event log across two axes: the quality dimensions gotten from reviewing more than 50 articles on data quality dimensions and the level of potential data quality problems which are: missing in log, concealed in log and missing in reality as stated by Van der Aalst, (2016) and ambiguous state as cited by Scheepstal, (2016). This article thus seeks to (i) propose a typology for timestamp data quality issues from timestamp related quality dimensions, (ii) explore the various quantifications methods for timestamp-related data quality issues in event logs.

Table 1: Timestamp quality metrics from Fischer's timestamp quality assessment framework

	Accuracy	Completeness	Consistency	Uniqueness
Log level		Missing trace	Mixed granularity of the log	Duplicate within log
			Format	
Trace Level	Infrequent event ordering	Missing activity	Mixed granularity of traces	Duplicate within traces
	Overlapping Events per resource			
Activity level		Missing Event	Mixed Granularity of activities	Duplicate within activity
Event Level	Future entry	Missing timestamp		
	Precision			

Source: Fischer et al., 2020

This article consists of four sections with section one being the introduction, section two is the literature review which gives a general overview of process mining and its advancement especially in the domain of data quality. Section three is the method which investigates the possible quantification methods of timestamp quality. Section four is the results which contains observation from the literature review and from the analysis of the MIMIC dataset from Physionet, and then it ends with section five being the discussion and conclusion which is based on the observations and result.

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00238

2. Literature review

There is an increase in valuable data generated in companies today which can be mined and processed to improve the quality and performance of business processes. Process mining and analysis from good quality data can help companies in taking effective decisions to optimize their business, for this to be possible the quality of the data must meet some standards (Dogan & Gurcan, 2018). This means quality improvement methods need to be applied to solve data quality problems so that the information gotten from data analysis can be reliable. Therefore, event logs, especially its timestamp should be treated with outmost care in the information systems supporting the processes to be analyzed in order to avoid inaccurate data which may give a false representation of a system or state. Unfortunately, most event logs are used as tools for debugging or profiling thus making event logs serve as by-product (Van der Aalst et al., 2011).

Process mining starts with the event logs and the ordering of the events found in this event log is crucial for the process discovery but sometimes the timestamps of the event log may not be reliable or precise, this makes process discovery difficult since the ordering of events in the does not mean the former causes the latter (Van der Aalst & Santos, 2021). Therefore, in order to discover the process models which represents the present state of the system, and which can also be generalized to the population at large, we need high quality event data to provide models with good quality (Pourmasoumi & Bagheri, 2016). Fischer et al., (2020) investigated the scarcity of research that focuses on detecting and quantifying data quality problems in event logs which arises from timestamp. Fischer et al., (2022) identified six papers (Alkhattabi et al., 2011; Kherbouche et al., 2016; Askham et al., 2013; Pipino et al., 2006; Sattler, 2009; Stvilia et al., 2007) that provide information on data quality quantification after doing a search

of 1298 journals in the domain of process mining. These publications investigated the quality of the event log data as a whole without any of them specifically evaluating timestamp data quality which is the key in event log process mining. Fischer et al., (2020) evaluated the quality of timestamps across two axes, with the first axis being the level of abstraction which is made up of the event, activity, trace, and log as proposed by Van der Aalst, (2016) and the second axis being the quality dimensions which is made up of the accuracy, completeness, consistency, and uniqueness, this resulted in fifteen timestamp quality-related metrics. Fischer et al., (2022) framework provided the first step in the quantification of timestamp data quality issues.

2.1. Process mining

A process is a well-organized systemic series of activities designed to produce a specified output with respect to the inputs (Decker, 2019). "Business Process Management (BPM) is the art and science of overseeing how work is performed in an organization to ensure consistent outcomes and to take advantage of improvement opportunities" (Dustdar et al., 2008). Business process according to Van der Aalst, (2011) involves the utilization and unification of knowledge from information technology and that of management science in the domain of operational business process. The execution of these business processes is supported by the Process-Aware Information System (PAIS). PAIS records the task that has been executed for each case in a trace and these traces are stored in an event log.

The main objective of process mining is to monitor, discover, and improve real processes by extracting and analyzing the data from the event logs of these processes (van der Aalst et al., 2011). Event logs are used in process mining and these event logs constitutes of a case id, which is an identification unique to each case; the timestamp which states the time the event

occurred and its activity/event name (Brzychczy et al., 2020). The event logs therefore contain data from observed behavior of processes collected and stored over a period. The chronological ordering of these events for a particular case yields a trace, i.e., a sequence of activities executed for that case, one of such trace represents the process from the initiation/start to its termination/end, thus it represents a simulation run (van der Aalst, 2018). Process mining techniques are used to automatically discover the process models from these event logs, determine the expected processing times for each process or check for the conformance of the process, and also to identify bottlenecks in the process and deviations in order to suggest improvements methods and predict processing times (Bose et al., 2013).

There are different perspectives which process mining can be analyzed from which involves the control flow perspective, the organizational perspective, the case perspective, and the time perspective (van der Aalst, 2016). The point of focus of the control-flow perspective is the ordering of activities to search for a good characterization of all possible paths, therefore it looks at the question of how the process was executed. This perspective is used to compare the designed and observed behavior. The point of focus of the organizational perspective unlike the others is the information hidden in the event log about the resources (Decker, 2019). This perspective looks at which actors (e.g., systems, people, departments, or roles) that are involved in the process and how they related each other, does this perspective search to answer the question; who executed the process? The case perspective focalizes on the properties of cases. This perspective thus seeks to establish what happened in a specific transaction; thus, it permits the scrutiny of a specific part of a business process. The point of focus of the time perspective is on the timing and frequency of events, this perspective uses the timestamps in the event log in process discovery and also to discover bottlenecks in the process (Drakoulogkonas & Apostolou, 2021). Typology is required to identify the data quality issues; thus, we identified

the various data quality dimensions linked to timestamp from literature review and the various data quality issues that affect timestamp.

2.2. Timestamp data quality dimension and quantification

The measurable features of data which can be evaluated against specific standards to determine the quality of data are the data quality dimensions (Askham et al., 2013). Twenty-four articles on data quality dimensions related to timestamp were selected by searching through databases such as IEEE Xplore, JSTOR, ScienceDirect, Scopus, EBSCOhost, and others. This analysis led to twelve timestamp related data quality dimensions as seen below (i.e., nine data quality dimensions with accuracy, time, and clarity each sub divided into two). The timestamp data quality dimensions are discussed below.

Completeness: For a dataset to be complete it must be able to satisfy the breadth, depth, and scope of the task which the data was intended to be used (Firmani et al., 2016; Pipino et al., 2002). Therefore, the completeness shows that all the relevant data in a particular set is present and can be used to represent a real-world scenario (Askham et al., 2013). Completeness can be quantified at the event level using the metric missing timestamp of events thus timestamp is examined if it's recorded for each event (Fischer et al., 2020). The measurement of completeness can also be based on the absence of blank values (null or empty string) or on the presence of non-blank values (Askham et al., 2013). Schema completeness, population completeness and column completeness are the three types of data completeness that exist (Candela et al., 2022). Schema completeness of a data set describes the degree to which the concepts and their attributes are retained or not missing from the data set or event log schema (Jarvis et al., 2018). Column completeness, unlike schema completeness, is the measure of the

missing values in a specific property or column in the event log. Population completeness evaluates the degree to which the event log has missing values in comparison to the total amount of records expected (Firmani et al., 2016; Pipino et al., 2002).

Accuracy: It is the measure of the deviation of the recorded values in the event log to the real value which were expected. The accuracy of the records in the dataset evaluates the degree to which the real-world object correctly describes event being recorded in the event log (Askham et al., 2013; Firmani et al., 2016). Every metric that investigates imprecise timestamps can therefore be allocated to accuracy (Fischer et al., 2020). Suriadi et al., (2017) defines accuracy as a combination of precision and correctness. The correctness part of the accuracy describes the level of semantical accuracy of the information or record while the precision part evaluates the extent to which the information is syntactically accurate (Firmani et al., 2016). Askham et al., (2013) defines data validity as it being conformant to the semantics of its definition, this therefore equates it to the correctness dimension as described by Suriadi et al., (2017) . The granularity of the timestamp determines the precision of the process mining results such as the mean waiting time, thus if the timestamp is not fine grained enough, the precision of the results will be affected (Van der Aalst, 2016).

I. **Generalizability:** Precision and generalizability are two antonyms when it comes to timestamp. Overly general models always have a very poor precision even though they produce a good fit for the process mined. This means the model does not allow the demonstration of all the possible traces even if they are recorded in the event log or occurred in real life (Scheepstal, 2016). Precision and generalizability are tradeoffs thus, the level of precision versus generalization must be in line with the end-user's goal (Van der Aalst et al., 2011).

II. **Uniqueness:** Uniqueness is the measure of the existence of unwanted duplicates in an event log or data set (Fischer et al., 2020). The most frequently detected issues of uniqueness in an event log include identical timestamps for start and end events of the same activity and duplicates within trace at trace level. To evaluate the uniqueness of the timestamps in the event log, the activities that had more than an event with the same timestamp in the event log were evaluated.

III. **Time:** Desplenter, (2018) defines time with respect to data quality dimension as measure of how information is timely and current for the task at hand. The currency aspect of the time in data quality dimension gauges how up to date the information retrieved from the data set is for the analysis which the data is intended. The timeliness aspect of the time in data quality dimension evaluates the age of the data set to determine the extent to which it affects the data quality of the intended analysis with respect to the intended task.

- **Currency:** The currency of data concerns how promptly the data is updated for it to reflect the changes that occurred in the real world (Firmani et al., 2016). While precision is focused on the reproducibility of data, currency focuses on how up to date the data is a representation of the changing reality.
- **Timeliness**: This is the measure of the degree at which a point in time gives a representation of what has occurred in the real world (Askham et al., 2013), thus it expresses how data are current for the task a hand (Pipino et al., 2002). Not all current data is actually useful for the task at hand therefore it's possible to have data that is current but that is not suited for the task at hand. Timeliness therefore checks the discrepancy between the time the real-world event occurred as to when it was recorded. Example a nurse filling forms at the end of a series of task.

IV. **Reliability**: Reliability/trustworthiness measures the probability that the trustee will perform an activity or a set of activity which is beneficial or of interest to the trustee (Firmani et al., 2016). Desplenter, (2018) and Firmani et al., (2016) subdivides reliability/trustworthiness into believability, verifiability, and reputation. Believability gauges the level to which information is considered as truthful and credible subjectively (Firmani et al., 2016; Pipino et al., 2002). The reputation of the information deal with how we regard information with respect to its source or content, with the content part representing the extent to which the information is impartial, unbiased or its unprejudiced (Desplenter, 2018 ;Pipino et al., 2002). Verifiability is the "degree and ease with which the information can be checked for correctness" (Firmani et al., 2016).

V. **Relevancy**: The relevancy of data evaluates the extent to which information in the event log is helpful or applicable for the task (Pipino et al., 2002). Not all information in the event log is helpful for the process mining at that point.

VI. **Clarity:** The clarity of the data represents the extent to which information is concisely and consistently presented. Clarity can be divided into two consistency and conciseness (Desplenter, 2018). Consistency evaluates if the information which the data carries is presented in the same format throughout the dataset and conciseness is the extent to which information is represented in a compact manner (Pipino et al., 2002; Firmani et al., 2016; Desplenter, 2018). The conciseness of the timestamp will not be evaluated since it's impossible to determine if the information was compactly represented as defined.

- **Consistency:** it checks for the absence of difference between two or more representation of things using a base definition for the comparison (Askham et al., 2013). The consistency is thus to captures the violation of semantic rules in an event log (Firmani et al., 2016)**.** Consistency in process mining checks if data

meets these three conditions; the format of the data representation meets the format requirements, two or more sets of data within a database do not conflict with one another, and as last, the level of consistency in data sets correlate to the correctness of data (Fischer et al., 2020).

VII. **Interpretability:** This evaluates the level to which information is presented in both the appropriate language and symbols; and also, the definition clear to decipher by the user (Pipino et al., 2002).

2.3. Timestamp data quality issues in process mining

The quality of the results of the process mining are largely determined by the event log quality (Scheepstal, 2016), thus the reason why the process mining manifesto encourages the development of a good benchmark for quality criteria in order to ensure that the quality of event logs adhere to a sufficient working level (Van der Aalst et al., 2011). This therefore emphasizes the necessity of the quality of the event log to be carefully monitored in order to obtain trustworthy analysis from the data. To achieve this, the systematic identification of the root causes of data quality issues in the data set must be done in order to improve the data quality (Horita et al., 2020). Van der Aalst, (2016) identifies three potential causes of data quality problems which are: missing in reality, missing in log, and concealed in logs, this identification was done using three main entities (case, activity instance, and event) and nine event attributes (case, timestamp, activity, activity instance, position, process, resource, transaction type and any data). Scheepstal, (2016) cites ambiguous state which is different from the above as an addition.

- **Missing in log:** when the reality is not mapped exhaustively into the information system, this results in an information system which is not able to represent the reality as the data is incomplete. There are some entities which exist in the real world but are not recorded in the event log which results in this missing timestamp. For example, a sample collection might occur but the person carrying out the activity failed to record it (Van der Aalst, 2016).
- **Missing in reality:** This occurs when the entity has never existed in reality, but the entity was recorded in the event log. An example of such a case can be a schedule for a particular activity which never took place, but it was in the event log because it was recorded in the information system (Van der Aalst, 2016).
- **Concealed in log/garbling:** For an entity to be concealed in log, it means the entity was recorded or exist/existed, but it's garbled in the event log due to its less structured data set. This situation might have resulted from: data being used for analysis coming from a combination of different data sources; using a broader scope of data for analysis than required; or the same entity being recorded multiple times in the same event log(Van der Aalst, 2016). The above-mentioned deficiency is always as a result of an operational deficiency, which is often from human error (Scheepstal, 2016). This problem leads to incomplete, noisy, and imprecise event logs.
- **Ambiguous representation**: when two different states of reality is mapped into one state of the information system, the information system is not able to assume which state represents the reality, resulting in data whose information system is not reliable anymore (Scheepstal, 2016).

Based on previous literature the data quality dimensions were selected and the timestamp quality issues were obtained by evaluating the cross between the axis of the four data quality problems and the data quality dimensions. This resulted in twelve timestamp quality issues

classified on table 2 and briefly discussed below table 2. Timestamp quality metrics were then suggested for the evaluation of these issues. The time dimension is subdivided into timeliness and currency; the accuracy dimension is subdivided into precision and validity; and the clarity dimension into consistency and conciseness. The conciseness which measures the extent to which information is compactly represented could not be evaluated because it depends on the user of the data, thus eleven quality dimensions were evaluated to decide the timestamp quality metrics.

Table 2: Timestamp quality metrics classification from data quality problems (columns) and quality dimensions (rows)

	Missing in log	**Missing in reality**	**Concealed in log/garbling**	**Ambiguous representation**
Completeness	Missing timestamp			
Consistency		Default timestamp	Mix granularity timestamp	Heterogenous timestamp
Currency	Outdated timestamp		Missing timestamp	
Generalizability			Fine granularity timestamp	
Precision			Coarse granularity timestamp	
Uniqueness				Duplicate/ambiguous timestamps.
Timeliness				Stacked or parallel timestamps
Validity		Incorrect timestamp		
Reliability				Ambiguous timestamps
Relevancy			Irrelevant timestamp	
Interpretability				Ambiguous timestamps

A. **Missing timestamp**: Missing timestamps occur when one or more events do not contain timestamps. Most processes which utilize automatic logging may probably fail to record or introduce deviations of the recorded data from what occurs in the real world, this may be due to machine breakdowns, human errors, resource constraints and bugs in the system (Van Cruchten, 2019). The process mining algorithms responsible for creating process models works on the assumption that the data recorded in the event log is complete and is of high quality, so if the data are missing, the analysis results will be inadequate (Horita et al., 2020). If the position of the event in a case is in order, it results in less serious problems during process mining. This is not the same when the position of the event in the case is not in order which often results in difficulties placing the events in an order which reflects a real-world pattern. This does make missing timestamp a problem to the applicability of the process analysis or a problem on the reliability of the results of the analysis (Van der Aalst, 2011). Missing timestamps also makes the process flow difficult since the position of the event log is not known and thus gives results which are less valid because the order of the events tends to be different from what occurred in reality.

B. **Default/recurrent timestamp**: There may be the occurrence of some values in the event log due to integrity checks. Depending on the system, when an event is created, there are some default values which are generated. If these default values are not adjusted or gotten rid of, it will create a consistent occurrence of these default values in the event log creating default timestamps which are missing reality. As reported by Fischer et al., (2020) most of the systems which exist convert time-related null values into the year 1970 which is called Unix time. In case of default timestamp with 00:00:00 :00 signifying the month, day, hour, minutes; if these default timestamps are

not updated for numerous events, it creates a consistent number of recurrent missing in reality timestamp in the event log.

C. **Heterogenous timestamp**: These are timestamps which has date time represented in different formats. Some of the software used to collect the data in the MIMIC-III dataset is CAREVUE and iMDDOC. Assuming that one registers the time for an activity using the order Day:Month:Year and the other does it by using Month:Day:Year, it will cause heterogenous timestamp issues. This may also occur when some software register time in the twenty-four-hour format while others use PM and AM.

D. **Outdated timestamp**: This occurs when what is recorded in the event log no longer represents the reality or does not fully represent the real state of events at the time. The MIMIC dataset does not have updated info about the state of the patient who ran away from the hospital, this does mean there is a possibility that some patients might not be alive after the data collection.

E. **Fine granularity timestamp**: Occurs when a fine level of granularity is used for the timestamps of events. This may lead to overly precise models and thus the creation of so many webs of unnecessary traces. These overly precise models demonstrate all additional traces which are possible theoretically, thus not giving an abstract view of the process which is needed to be evaluated (Fischer et al., 2020). The precision and generalizability of timestamps impacted the results of process mining (Van der Aalst, 2016).

F. **Coarse granularity timestamp**: This corresponds to a situation where a coarse level of granularity is used for the timestamp's records of events, thus creating a situation where the ordering of events within the log may not be in conformity with the actual ordering in which the events had occurred (Marin-Castro & Tello-Leal, 2021). Process

mining algorithms sometimes have problems with identifying the correct control-flow which can lead to the discovered control-flow model having a substantial number of activities occurring in parallel in the discovered process (Bose et al., 2013). Coarse granularity in the timestamp of event logs may does lead to the problems in process mining such as process performance because the timestamp is not precise for such (Van der Aalst, 2016).

G. **Mix granularity timestamp**: This involves event logs that have timestamps with different levels of precision, e.g., milliseconds, seconds, minutes, days. The effect of event logs with mixed granular timestamps makes the precise ordering not to be clear. This makes process mining algorithms to have problems discovering the correct control-flow for these events (Fischer et al., 2020). The admission time recorded in the admission dataset of the MIMIC-III dataset has a year: Month: day: Hour: minutes time format but the admission time on the ICU dataset gives more fine data including the time in seconds.

H. **Ambiguous timestamps**. These are timestamps/event logs which permit the interpretation of the timestamp/data in more than a single way. An example of this is a scenario where two instances of the same activity are running at the same time, which might have resulted from an instance of an activity being started and before it goes to completion another instance of the same activity is started (Marin-Castro & Tello-Leal, 2021). As there is an ambiguity in associating the complete events with their corresponding start events, process mining algorithms have problems with deciding when a certain instance of an activity is completed based on the available timestamp (Van der Aalst et al., 2011). Ambiguous timestamp can also be due to an undefined activity or an event. For example, in the MIMIC-III dataset, those who escaped from the hospital have no specific discharge time thus at the time of entering the information

that the person is officially discharge, the state of liveliness of the person is undetermined. This can cause issues of reliability in the discovered process if there are a great number of such abnormalities.

I. **Duplicate timestamps**: This corresponds to the scenario where multiple events have the same activity name and thus timestamps. Most process mining algorithms can't identify duplicate task and thus produce inaccurate process models/results (Bose et al., 2013). The presence of (undesired) duplicates in an event log for a particular data set can cause data quality issues which affect the results of the process mining (Fischer et al., 2020). Timestamp duplicates can be classified according to level they occur:

- **Timestamp duplicates within logs**: Some events may not belong to the same trace but use the same timestamp. Event storage which uses electronic form-based often creates such problems. Such problems may include recording multiple trace start event in an electronic form (Fischer et al., 2020).
- **Timestamp duplicates within trace**: Multiple trace may show the same timestamp resulting in data quality issues. This problem can be caused by form-based event capture which involves recording several events of a trace using the same timestamp through e-forms, even though they happened in a sequence (Fischer et al., 2020)
- **Timestamp duplicates within activity**: These are activities that have more than one event with the same timestamp in the event log thus creating quality issues. They often have the same start and end timestamp for the same activity (Fischer et al., 2020).

J. **Stacked or parallel timestamps**: This is due to multiple events recorded at the same time and are thus presented as if they occurred at once. An example of this can be the submission of clinical forms at the end of each examination by a health practitioner in

which he/she declares having recorded clinical signs and symptoms and administering mediation. This form may lead to the all the recorded events in the event log having the same timestamp, i.e., the time the form was submitted shows that all the activities were completed at once even though in reality they never occurred at once (Van der Aalst, 2016). Time based process discovery algorithm will not be able to decipher the appropriate trace because of the events being recorded as they occurred in parallel.

K. **Incorrect timestamp**: This involves events whose dates do not exist or where there exist conflicts between events timestamps and ordering information (Bose et al., 2013). Incorrect timestamps can lead to the control-flow relations becoming unreliable or even incorrect. The data may be from a variety of sources which the information needs to be merged in order to make it useful, this means every event corresponding to a case are scattered among systems in the organization (Van der Aalst, 2016). If different identifiers were used, this tends to cause problems merging the different data sources and may results in incorrect timestamps (Van der Aalst et al., 2011). This is different from recurrent timestamp error in that, for incorrect timestamp the wrong timestamp of an event is recorded.

L. **Irrelevant timestamp**: This is caused by the occurrence of irrelevant activities and their timestamp in the event log on which the analysis is to be done. The cases and events in an event log can be influenced by the irrelevant data.

Assessing the quality dimensions against the event log deficiencies resulted in twelve timestamp quality metrics. The various metrics of quantifying them will be explored in the following section and the metrics proposed will be tested using the MIMIC-III v1.4 dataset (available at https://physionet.org/content/mimiciii/1.4/).

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00238

3. Method

❖ **Analysis**

A comprehensive literature search conducted on databases of journal publishers such as IEEE Xplore, JSTOR, ScienceDirect, Scopus, EBSCOhost, and other data bases was done to identify data quality dimensions and issues, 24 articles on data quality dimensions were selected and analyzed from which the quality dimensions affecting timestamp were selected. The possible data quality metrics were obtained by doing an evaluation across two axes: the possible data quality dimensions and the level of potential data quality problems gotten from reviewing the articles.

There are two types of data quality assessments: the objective measurement and the subjective perceptions, the objective measurement of data quality assessment is based on the dataset in question while the subjective perception of the data quality assessment focus on the individual involved in the data set (Pipino et al., 2002). In carrying out the subjective data quality assessments, the needs and experiences of the stakeholders are mostly taken into consideration while the objective data quality assessments are often task-independent or task-dependent. The difference between the task independent metric and the task dependent metric is that the task independent metric reflects the state of the data without contextual knowledge and can be applied to any data set while the task dependent metrics requires knowledge of the company and government regulations, organization's business rules, and constraints provided by the database administrator (Pipino et al., 2002). This article uses the task independent specific metrics as outlined by Pipino et al., (2002), this therefore implies some quantifications such as conciseness which are subjective have not been quantified. The outdatedness of the data was measured using the currency equation (see equation 4 below) cited by Heinrich et al.,

(2009) adapted from Ballou et al., (1998). Following the updates of Heinrich et al., (2009) of the age of the attribute values, we can define timeliness from Ballou et al., (1998) as the difference between delivery time of the information products and the input time when the data is obtained, and the results added to the age of the data unit when received (see equation 1 below). The mean update frequency of the data which is defined as the number of updates per a time interval (equation 3) and the timeliness (equation 1) factors into the equation of the outdatedness of the data. The duration which an activity was stacked or ran in parallel is measured by the difference of the delivery time and the input time (equation 2). The other issues were quantified using the fraction/percentage of the undesired data in the dataset (see table 3).

Timeliness = (delivery time-input time) +Age Equation 1

Stack or parallel timestamp = delivery time-input time Equation 2

Mean update frequency = number of updates/time interval of updates Equation 3

$$\text{Outdated timestamp} = \frac{1}{(mean\ \ attibute\ update\ frequency)(timeliness)+1} \quad \text{Equation 4}$$

Table 3: Dimension quantification methods and source

Dimension		**Quantification**	**Sources of citations**
Completeness		Percentage of missing timestamp	Fischer et al., 2020 Askham et al., 2013
Consistency		Heterogenous timestamp: percentage with the correct format. Mix granularity timestamp: percentage of case with timestamps without the correct granularity.	Askham et al., 2013 Fischer et al., 2020
Accuracy	Correctness/ Validity	Percentage of data which represent a real activity/event	Askham et al., 2013 Fischer et al., 2020
	Precision	Percentage of data which meets specifications	Suriadi et al., 2017 Fischer et al., 2020
Time	Currency	As per formula proposed by Heinrich et al., 2009	Heinrich et al., 2009
	Timeliness	As per formula proposed by Heinrich et al., 2009	Heinrich et al., 2009
Generalizability		Ratio/fraction of timestamp fine granular timestamp	Van der Aalst et al., 2011
Uniqueness		Percentage of events with duplicate timestamp	Askham et al., 2013 Fischer et al., 2020
Reliability		This is a subjective factor relative to the person metering it.	Desplenter, 2018
Relevancy		Percentage of data relevant for the operation	Desplenter, 2018
Consistency		Percentage of information presented in the same format:	Desplenter, 2018 Fischer et al., 2020
Interpretability		Percentage of information that can decipher	

❖ Data source

The data used in this article is the MIMIC-III clinical database version 1.4 (MIMIC-III v1.4) which was released on the second of September 2016 (available at https://physionet.org/content/mimiciii/1.4/). The MIMIC-III dataset is a database of 46520 anonymized patients records accounting for 58,976 unique visits at the intensive care unit (ICU) of the Beth Israel Deaconess medical Centre (BIDMC) located in Boston, USA. The records contain the information of patients who visited the hospital between June 2001 and October 2012 (Johnson et al., 2016a). The MIMIC-III dataset is comprised of the medical records of patients who had been admitted or have been offered care in the intensive care unit for at least one period. The data collected for such patients cover all their hospital stay for all care (Johnson et al., 2016b). The MIMIC-III team at the MIT Lab for computational physiology have anonymize and curated the data, specifically the timestamps in the data set were anonymized by shifting them into the future between 2100 and 2200 using a random offset with each patient having a different offset date. This shifting was carried out in a manner to preserve the seasonality in the data, the week, and the time of the day which the event took place (Kurniati et al., 2019).

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00238

4. Results

The first objective of this article was to propose a typology for the timestamp quality issues from timestamp related quality dimensions. The classification of timestamp quality issues was done based on four potential data quality problems across eleven different data quality dimensions which resulted in twelve timestamp quality problems metrics: ambiguous timestamps, coarse granularity timestamp, duplicate timestamps, fine granularity timestamp, heterogenous timestamp, incorrect timestamp, irrelevant timestamp, missing timestamp, mix granularity timestamp, outdated timestamp, recurrent timestamp errors, stacked or parallel timestamps.

The second objective was to explore techniques to quantify the various timestamp quality problems gotten from the classification. The techniques outlined above in the methodology were applied to a public dataset (MIMIC-III) to assess the timestamp quality issues. The timestamps related to the admission time, discharge time, emergency (ED) registration time, emergency (ED) discharge/out time, death time and date of death (DOD) which are the main timestamps of the BIDMC hospital process were analyzed. Table 4 shows the quantification of the timestamp quality issues in the MIMIC-III data set using the timestamp of the BIDMC hospital process (represented as rows on the table) and the timestamp quality quantification metric developed above (represented as columns on the table)

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00238

Table 4: Data quality quantification of the MIMICIII dataset timestamp

Activity / time	Ambiguous timestamps	Coarse granularity timestamp	Fine granularity timestamp	Mix granularity timestamp ratio	Duplicate timestamp	Heterogenous timestamp	Incorrect timestamp	Irrelevant timestamp	Missing timestamp	Outdated timestamp	Default timestamp	Stacked or parallel timestamps (seconds)
Admission Time	0.00%	52.57%	47.43%	0.902	1.08%	0.00%	0.31%	0.00%	0.00%	0.4999	0.00%	8188.625
Discharge time	0.00%	57.82%	42.18%	0.73	1.08%	0.00%	0.32%	0.00%	0.00%	0.5002	0.00%	27.357
ED registration time	47.65%	50.81%	49.19%	0.968	0.02%	0.00%	0.10%	0.00%	0.00%	0.5	47.65%	8188.625
ED out time	47.65%	52.34%	47.66%	0.911	0.08%	0.00%	0.14%	0.00%	0.00%	0.5	47.65%	27.357
DOD national register	66.12%	100.00%	0.00%	0	33.63%	0.00%	0.02%	99.98%	0.00%	0.5	61.70%	407.271
Dead time hospital	90.07%	59.88%	40.12%	0.67	0.68%	0.00%	0.07%	99.98%	0.00%	0.5	90.07%	407.271

i. **Ambiguous timestamps**: The ambiguous timestamp was considered as any information which could be interpreted by anyone in numerous ways. The NaN (Not a number) was considered as ambiguous because it can equally represent the absence of the occurrence of the event in reality or the failure to record the value in the event log or any other thing. The death time registered by the hospital only has twenty five percent of the deaths in the national register which has the date of dead, this ambiguity thus cast doubts on its believability which refers to the extent to which information is regarded as true and credible subjectively (Firmani et al., 2016; Pipino et al., 2002) and its verifiability which refers to the "degree and ease with which the information can be checked for correctness" (Firmani et al., 2016). The analyzed results show no ambiguity of the admission timestamp and discharge timestamp. The emergency (ED) registration timestamp and emergency discharge/out time had an ambiguity of about 47.64 %. The date of death (DOD) on the national register and the hospital dead time had a timestamp ambiguity of 66.12% and 90.07% respectively.

ii. **Coarse granularity timestamp**: The coarse timestamp is considered as the timestamp whose time of occurrence in minutes or seconds is undefined. Thus, their precision can only be gotten to the nearest hour. The results above show that more than fifty percent of the timestamp could only be approximated to the nearest hour. The coarse granularity in the DOD national register is 100% and all the other activity time had coarse granularity timestamp quality measurements between 50-60%.

iii. **Fine granularity timestamp**: The fine timestamp is considered as the timestamp with the minute and/or second level of time defined. The fine granularity timestamp

of the date of death in the national register was zero because just the day which the person died was recorded and all of the other activity time had coarse granularity timestamp quality measurements between 40-50%.

iv. **Mix granularity timestamp ratio**: The mix granularity timestamp ratio was calculated as the ratio of the fine granularity timestamp to the coarse granularity timestamp recorded from an activity. The mix granularity timestamp ratio was found to be highest in the Emergency (ED) registration time (96.8%), ED discharge/out time (91.1%) and admission time (0.902) with a mix granularity ratio of more than ninety percent in each of them. A mix granularity timestamp ratio of zero was observed in the date of death (DOD) recorded in the national register.

v. **Heterogenous timestamp**: There was no case of heterogeneous timestamp. This may be because the data from the two organizations involved in taking the medical history of patients has been combined and cleaned.

vi. **Duplicate timestamps**: The DOD had a 33.62% of duplicates which is due to the level of granularity of the timestamp which was used and also due to the fact that more 21.12% of the subject who visited the hospital made the visit more than once. The discharge time and the admission time had a duplicate percentage of 1.078% each. The hospital dead time had a duplicate percentage of 0.68%, while the emergency (ED) registration time and the emergency (ED) discharge/out time had a duplicate percentage of 0.019% and 0.084%.

vii. **Incorrect timestamp**: The incorrect timestamp detected here is based on the premise that: the emergency registration cannot occur before the patient is admitted; the patient cannot be discharged from the hospital before he is discharged from the emergency department; the date of dead in the national register cannot be different from that registered in the hospital. Based on the above conditions all of the

incorrect timestamps were less than a percent. The incorrect timestamp of the discharge time and admission time were highest with values of 0.322% and 0.309%. The ED discharge time and ED registration time had 0.14% and 0.1% incorrect timestamp respectively. The DOD national register and death time hospital had incorrect timestamp percentages of 0.019% and 0.071% respectively.

viii. **Irrelevant timestamp**: The degree of irrelevancy was based on the percentage of timestamp data which could be derived from the others. 99.98% of the timestamp of the death time could be derived from the date of dead from the national register.

ix. **Missing timestamp**: There was no instance of missing timestamp found. This may be as a result of all timestamps being filled with 'NaN' which is ambiguous to interpret.

x. **Outdated timestamp**: The level of outdatedness of the timestamp was quantified using the currency equation cited by Heinrich et al., (2009) adapted from Ballou et al., (1998). All the timestamps had a currency of approximately 0.5.

xi. **Default timestamp**: 'NaN' was filled in all spaces where there was no timestamp, and this was assumed to be the default timestamp. This was also considered as the ambiguous timestamp because of the tendency of it resulting from a mistake of forgetting to fill the value. The death time in the hospital had the highest default timestamp of 90.07% seconded by the DOD national register (61.703%). The ED out time and ED registration time had default timestamp of 47.64%.

xii. **Stacked or parallel timestamps**: The estimated average 'parallelness' or average time for two or more activities to run in parallel was estimated from Ballou et al., (1998) timelines equation with the age part of the equation being suppressed. The average stacking time between the admission time and emergency registration time is about 8188.62 seconds. The average stacking time between the discharge time

and emergency discharge time is about 27.36 seconds. The average stacking time between the date of dead registration and the hospital death time is about 407.27 seconds.

From the quantification of the timestamp quality issues, we observe that the cases which brings about timestamp quality issues represent a very low percentage of the event log data, and thus we can confidently eliminate these cases because its results in a low or no risk of a getting a distorted process model because of their elimination. A 'curated event log' was created by removing some of the cases responsible for the timestamp quality issues and a 'non curated event log' was also created which had the original data downloaded. In the 'curated event log', some of the timestamp quality issues quantified above were treated. This was done by removing the cases which had events with date of discharge earlier than their date of emergency, removing the cases whose event's date of emergency registration was earlier than their date of admission and thus the stacked or parallel timestamp. The quantification of the timestamp quality metric therefore acts like a guide in making decisions on the type of cleaning we can possibly do on the data and the consequences which may result from it. The process mining package PM4Py in python was used to mine the process flow in the hospital (based on the activities) utilizing the 'curated event log' (figure 2) and the 'non curated event log' (figure 1). On figure 1 we can see that there are some cases which after going through their first activity (newborn, urgent or emergency) the process was immediately terminated without going through the end activity while on the figure 2, we can observe such patterns were greatly limited. Some cases had numerous transfers to and from the same intensive care unit (ICU) or different ICU. This can be clearly observed by the spaghetti process which it creates when the process is represented by showing the resources which each case goes through rather than the activity.

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00238

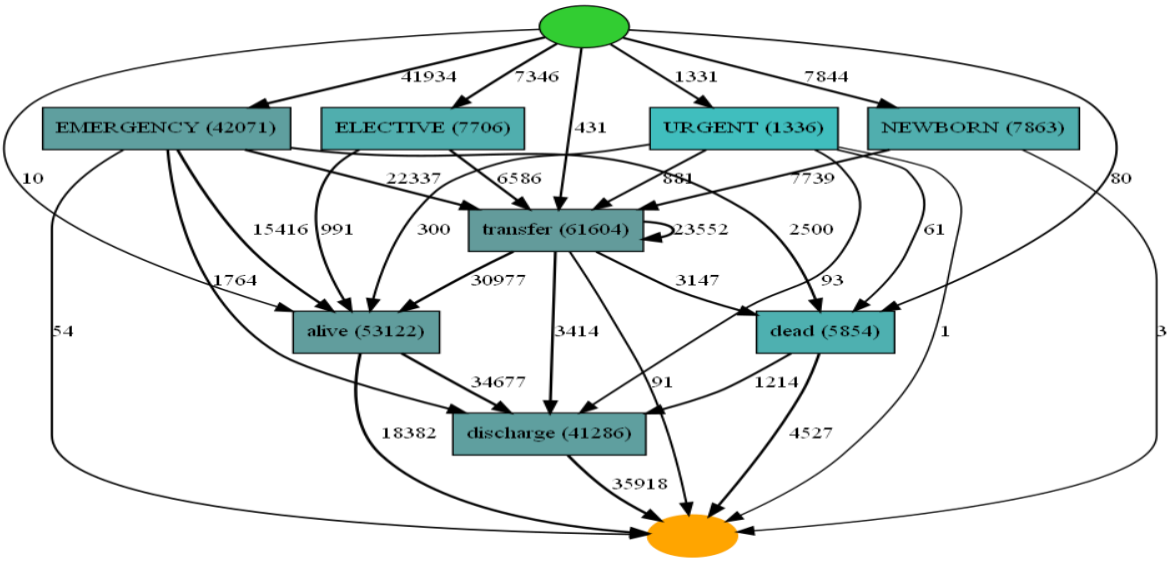

Figure 1: Non curated Activity based process diagram.

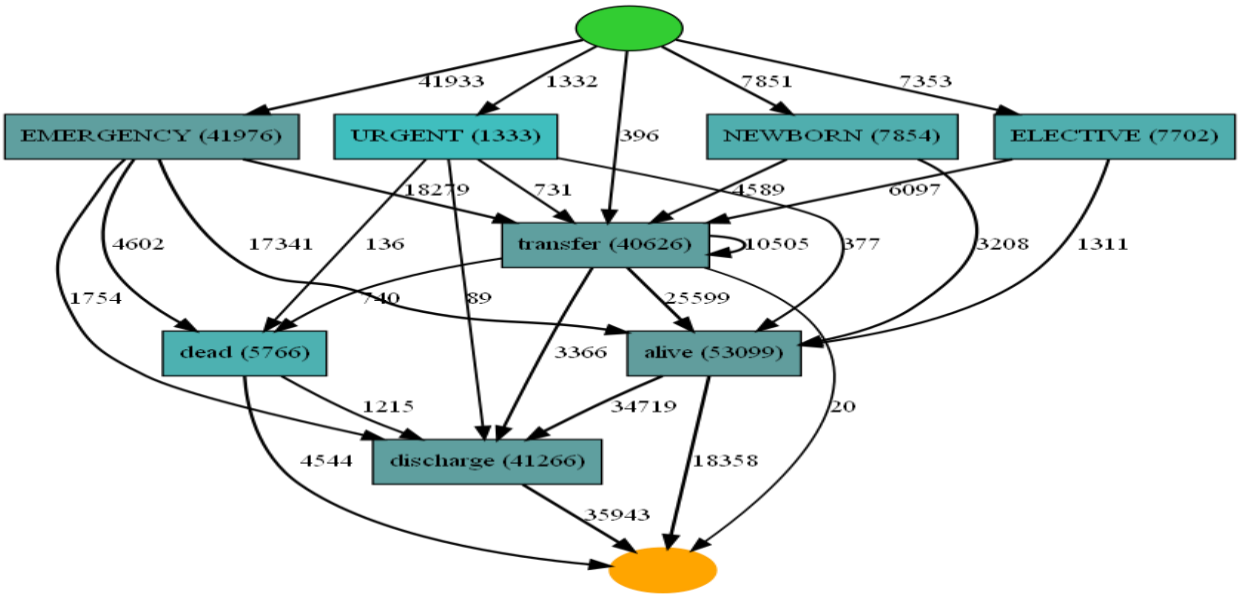

Figure 2: curated Activity based process diagram.

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00238

5. Discussion and conclusion

Timestamp quality quantification is mostly overlooked and sometimes combined with other quality issues even though timestamps play a vital role in process discovery, conformance, and performance. Fischer et al., (2020) evaluated the quality of timestamps across two axes with the first axis being the level of abstraction (composed of the event, activity, trace, and log as proposed by Van der Aalst, (2016)) and the second axis being the quality dimensions (which is comprised of the accuracy, completeness, consistency, and uniqueness), thus resulting in fifteen timestamp quality-related metrics. This framework provided the first step in the quantification of timestamp data quality issues (Fischer et al., 2022). The classification of timestamp into the above four level of abstraction resulted in metrics which identifies timestamp issues and quantify the same issue in numerous metrics. The approach adopted in this article is aimed at reducing the repetition of the same timestamp quality issue in different evaluated metrics. For example, missing trace, missing activity, missing event, and missing timestamp all have to do with missing timestamps (or at least has missing timestamp) at different levels of abstraction. The typology proposed in this article solve the problem by quantifying the quality of timestamps in an event log across two axes: the quality dimensions axis and the level of potential data quality problems (missing in log, concealed in log, missing in reality and ambiguous state). The approach in this article resulted in twelve timestamp quality metrics which we applied the typology to the MIMIC-III public dataset. The timestamp data quality issue which is most prominent in the dataset was the ambiguous timestamp and mix granularity issues. There is an average parallel/stacked timestamp of 8188 seconds when registering the admission time and emergency registration time of the activities. The result of the timestamp quality quantification using the proposed metrics makes it easier to quantitatively appreciate the quality of the timestamp and thus makes

it possible to evaluate the risk of carrying out specific data cleaning techniques in order to improve the process mining outcome.

The data quality dimensions utilized in this article are not assumed to be exhaustive, but rather this article tries to present an alternative method on how timestamp quality issues quantification can be done. This method provides information on how the timestamp quality issue affects the general timestamp and not specific level of abstraction which its sum doesn't reflect the whole data set. Some of the short comings of this method is that analysis can over generalized the results because the levels of log abstraction are not differentiated, for example the occurrence of the duplicate at different levels is unknown. This work can be extended by investigating the acceptable data quality threshold for the timestamp to be use for process mining in various sectors and also by implementing semi-automated methods of curating the data while evaluating the loss of information. This article had as objective of classifying timestamp quality issues from timestamp related quality dimensions and to explore the various quantifications methods for timestamp-related data quality issues in event logs.

Reference list

Alkhattabi, M., Neagu, D., & Cullen, A. (2011). Assessing information quality of e-learning systems: A web mining approach. *Computers in Human Behavior*, *27*(2), 862–873. https://doi.org/10.1016/j.chb.2010.11.011

Askham, N., Cook, D., Doyle, M., Fereday, H., Schwarzenbach, J., Palmer, G., Maynard, C., Lee. Rob, Gibson, M., & Landbeck, U. (2013). *THE SIX PRIMARY DIMENSIONS FOR DATA QUALITY ASSESSMENT Defining Data Quality Dimensions Defining Data Quality Dimensions DEFINING DATA QUALITY DIMENSIONS BACKGROUND*.

Ballou, D., Wang, R., Pazer, H., & KumarTayi, G. (1998). Modeling Information Manufacturing Systems to Determine Information Product Quality. In *Science* (Vol. 44, Issue 4). https://about.jstor.org/terms

Bose, R. P. J. C., Mans, R. S., & van der Aalst, W. M. P. (2013). Wanna improve process mining results? *Proceedings of the 2013 IEEE Symposium on Computational Intelligence and Data Mining, CIDM 2013 - 2013 IEEE Symposium Series on Computational Intelligence, SSCI 2013*, 127–134. https://doi.org/10.1109/CIDM.2013.6597227

Brzychczy, E., Gackowiec, P., & Liebetrau, M. (2020). Data analytic approaches for mining process improvement-machinery utilization use case. *Resources*, *9*(2). https://doi.org/10.3390/resources9020017

Decker, S. (2019). *Data-driven business process improvement*. *May*.

Desplenter, J. (2018). *THE DESIGN OF A DATA QUALITY FRAMEWORK FOR COBIT*.

Dogan, O., & Gurcan, O. F. (2018). Data perspective of lean six sigma in industry 4.0 era: A guide to improve quality. *Proceedings of the International Conference on Industrial Engineering and Operations Management*, *2018*(JUL), 943–953.

Drakoulogkonas, P., & Apostolou, D. (2021). On the selection of process mining tools. *Electronics (Switzerland)*, *10*(4), 1–24. https://doi.org/10.3390/electronics10040451

Dustdar, S., Fiadeiro, J. L., & Sheth, A. (2008). Business process management. In *Data and Knowledge Engineering* (Vol. 64, Issue 1). https://doi.org/10.1016/j.datak.2007.06.004

Firmani, D., Mecella, M., Scannapieco, M., & Batini, C. (2016). On the Meaningfulness of "Big Data Quality" (Invited Paper). *Data Science and Engineering*, *1*(1), 6–20. https://doi.org/10.1007/s41019-015-0004-7

Fischer, D. A., Goel, K., Andrews, R., van Dun, C. G. J., Wynn, M. T., & Röglinger, M. (2020a). Enhancing event log quality: Detecting and quantifying timestamp imperfections. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12168 LNCS*, 309–326. https://doi.org/10.1007/978-3-030-58666-9_18

Fischer, D. A., Goel, K., Andrews, R., van Dun, C. G. J., Wynn, M. T., & Röglinger, M. (2020b). Enhancing event log quality: Detecting and quantifying timestamp imperfections. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12168 LNCS*, 309–326. https://doi.org/10.1007/978-3-030-58666-9_18

Fischer, D. A., Goel, K., Andrews, R., van Dun, C. G. J., Wynn, M. T., & Röglinger, M. (2020c). Enhancing event log quality: Detecting and quantifying timestamp imperfections. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture*

Notes in Bioinformatics), *12168 LNCS*, 309–326. https://doi.org/10.1007/978-3-030-58666-9_18

Fischer, D. A., Goel, K., Andrews, R., van Dun, C. G. J., Wynn, M. T., & Röglinger, M. (2022). Towards interactive event log forensics: Detecting and quantifying timestamp imperfections. *Information Systems*, 102039. https://doi.org/10.1016/j.is.2022.102039

Heinrich, B., Marcus Kaiser, & Klier, M. (2009). *A Procedure to develop Metrics for Currency and its Application in CRM*. https://www.researchgate.net/publication/200047416

Horita, H., Kurihashi, Y., & Miyamori, N. (2020). Extraction of missing tendency using decision tree learning in business process event log. *Data*, *5*(3), 1–12. https://doi.org/10.3390/data5030082

Jarvis, A., Morales, L., Jose, J., Silvestrini, R. T., Burke, S. E., Eng, P. L., Corney, P. J., Westfall, L., Zarghami, A., & Benbow, D. (2018). *Quality Experience Telemetry: How to Effectively Use Telemetry for Improved Customer Success Navigating the Minefield: A Practical KM Companion Introduction to 8D Problem Solving: Including Practical Applications and Examples*. http://www.asq.org/quality-press.

Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. W. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, *3*. https://doi.org/10.1038/sdata.2016.35

Johnson, A., Pollard, T., & Mark, R. (2016). *MIMIC-III Clinical Database (version 1.4). PhysioNet*.

Kherbouche, M. O., Laga, N., & Masse, P.-A. (n.d.). *Towards a better assessment of event logs quality*.

Kurniati, A. P., Rojas, E., Hogg, D., Hall, G., & Johnson, O. A. (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), 1878–1893. https://doi.org/10.1177/1460458218810760

Marin-Castro, H. M., & Tello-Leal, E. (2021). Event log preprocessing for process mining: A review. In *Applied Sciences (Switzerland)* (Vol. 11, Issue 22). MDPI. https://doi.org/10.3390/app112210556

Martin, N., De Weerdt, J., Fernández-Llatas, C., Gal, A., Gatta, R., Ibáñez, G., Johnson, O., Mannhardt, F., Marco-Ruiz, L., Mertens, S., Munoz-Gama, J., Seoane, F., Vanthienen, J., Wynn, M. T., Boilève, D. B., Bergs, J., Joosten-Melis, M., Schretlen, S., & Van Acker, B. (2020). Recommendations for enhancing the usability and understandability of process mining in healthcare. *Artificial Intelligence in Medicine*, *109*(March). https://doi.org/10.1016/j.artmed.2020.101962

Pipino, L. L., Lee, Y. W., Wang, R. Y., & Yang, R. Y. (2002). *Data Quality Assessment* (Vol. 45, Issue 4ve).

Pipino, L. L., Wang, R. Y., Funk, J. D., & Lee, Y. W. (2006). *Journey to Data Quality*. The MIT Press. https://doi.org/10.7551/mitpress/4037.001.0001

Pourmasoumi, A., & Bagheri, E. (2016). *ENCYCLOPEDIA WITH SEMANTIC COMPUTING Business Process Mining*. *1*(14), 1–8.

Sattler, K.-U. (2009). Data Quality Dimensions. In *Encyclopedia of Database Systems* (pp. 612–615). Springer US. https://doi.org/10.1007/978-0-387-39940-9_108

Scheepstal, S. V. A. N. (2016). Data quality within process mining in the auditing context. *Google Scholar*, 85.

Stvilia, B., Gasser, L., Twidale, M. B., & Smith, B. C. (2007). A framework for information quality Assessment. In *JASIST* (Vol. 58, Issue 12).

Suriadi, S., Andrews, R., ter Hofstede, A. H. M., & Wynn, M. T. (2017). Event log imperfection patterns for process mining: Towards a systematic approach to cleaning event logs. *Information Systems*, *64*(September 2016), 132–150. https://doi.org/10.1016/j.is.2016.07.011

van Cruchten, R. M. E. (2019). Data quality in process mining: A rule-based Approach. *CEUR Workshop Proceedings*, *2432*(2013), 1–8.

van der Aalst, W. (2016a). Process mining: Data science in action. In *Process Mining: Data Science in Action*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-49851-4

van der Aalst, W. (2016b). Process mining: Data science in action. In *Process Mining: Data Science in Action*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-49851-4

Van Der Aalst, W., Adriansyah, A., De Medeiros, A. K. A., Arcieri, F., Baier, T., Blickle, T., Bose, J. C., Van Den Brand, P., Brandtjen, R., Buijs, J., Burattin, A., Carmona, J., Castellanos, M., Claes, J., Cook, J., Costantini, N., Curbera, F., Damiani, E., De Leoni, M., … Wynn, M. (2012). Process mining manifesto. *Lecture Notes in Business Information Processing*, *99 LNBIP*(PART 1), 169–194. https://doi.org/10.1007/978-3-642-28108-2_19

van der Aalst, W. M. P. (2018). Process mining and simulation: A match made in heaven! *Simulation Series*, *50*(10), 39–50. https://doi.org/10.22360/summersim.2018.scsc.005

van Der Aalst, W. M. P., Adriansyah, A., Karla, A., Medeiros, A. De, Arcieri, F., Blickle, T., Bose, J. C., Brand, P. Van Den, Brandtjen, R., Burattin, A., Carmona, J., Castellanos, M., Claes, J., Cook, J., Curbera, F., Damiani, E., Delias, P., Dongen, B. Van, Dustdar, S., … Wynn, M. (2011). Process Mining Manifesto. *Business Process Management Workshops*, *99*(1), 169–194.

van der Aalst, W. M. P., & Santos, L. (2021). *May I Take Your Order? On the Interplay Between Time and Order in Process Mining*. http://arxiv.org/abs/2107.03937

Wynn, M. T., & Sadiq, S. (2019). Responsible Process Mining - A Data Quality Perspective. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11675 LNCS*, 10–15. https://doi.org/10.1007/978-3-030-26619-6_2