Made available by Hasselt University Library in https://documentserver.uhasselt.be

Interactive Data Cleaning for Process Mining: A Case Study of an Outpatient Clinic’s Appointment System

Peer-reviewed author version

MARTIN, Niels; Martinez-Millana, Antonio; Valdivieso, Bernardo & Fernández-Llatas, Carlos (2019) Interactive Data Cleaning for Process Mining: A Case Study of an Outpatient Clinic’s Appointment System. In: Di Francescomarino, Chiara; Dijkman, Remco; Zdun, Uwe (ed.). Business Process Management Workshops BPM 2019 , Revised Selected Papers. Springer, Cham, p. 532 -544.

DOI: 10.1007/978-3-030-37453-2_43
Handle: http://hdl.handle.net/1942/30239

Interactive data cleaning for process mining: a case study of an outpatient clinic's appointment system

Niels Martin[1], Antonio Martinez-Millana[2],
Bernardo Valdivieso[3], and Carlos Fernández-Llatas[2]

[1] Hasselt University, Agoralaan Building D, 3590 Diepenbeek, Belgium
niels.martin@uhasselt.be
[2] Universitat Politècnica de València, Camí de Vera, s/n, 46022 València, Spain
{anmarmil,cfllatas}@itaca.upv.es
[3] Hospital Universitario y Politécnico de La Fe, Avinguda de Fernando Abril Martorell, 106, 46026 València, Spain valdivieso_ber@gva.es

Abstract. Hospitals are becoming increasingly aware of the need to improve their processes and data-driven approaches, such as process mining, are gaining attention. When applying process mining techniques in reality, it is widely recognized that real-life data tends to suffer from data quality problems. Consequently, thorough data quality assessment and data cleaning is required. This paper proposes an interactive data cleaning approach for process mining. It encompasses both data-based and discovery-based data quality assessment, showing that both are complementary. To illustrate some key elements of the proposed approach, a case study of an outpatient clinic's appointment system is considered.

Keywords: process mining · data quality · interactive data cleaning · process discovery · outpatient clinic

1 Introduction

Hospitals are confronted with a multitude of challenges including reduced budgets contrasted to augmenting care needs [15]. To cope with these challenges, hospitals are becoming increasingly aware of the need to improve their processes [16]. This awareness, combined with the increased availability of data about these processes, led to an increased attention for process mining in the healthcare domain. Process mining refers to the extraction of knowledge from an event log containing process execution information originating from a process-aware information system such as a hospital information system (HIS).

Process mining research has a predominant focus on the development of new techniques or the innovative application of existing techniques [6]. However, when applying these techniques in reality, it has been widely recognized

This is a pre-print of an article published in Lecture Notes in Business Information Processing. The final authenticated version is available online at: https://doi.org/10.1007/978-3-030-37453-2_43

that real-life data tends to suffer from a multitude of data quality problems [2,6,21]. This especially holds in a flexible and dynamic environment as healthcare, where commonly observed data quality issues include missing events and incorrect timestamps [11,15,16]. Given the biases that data quality issues can introduce in the results, research attention on data quality assessment is increasing in the process mining field in recent years [3,13,21,22,23]. Data quality assessment research tends to focus on the identification of problematic patterns in the data, which are candidate for mitigation before proceeding to process discovery. However, process discovery can also be leveraged to detect data quality issues that might otherwise remain hidden from the analyst. Despite its potential, the use of process discovery within the context of data cleaning has not been given explicit research attention.

This paper proposes an interactive data cleaning approach consisting of three key components: data-based data quality assessment, discovery-based data quality assessment, and data cleaning heuristics. To illustrate these key components, a case study of an outpatient clinic's appointment system is used. The available dataset suffers from several timestamp-related data quality issues, requiring data cleaning to make it usable for process mining purposes. The proposed data cleaning approach and the case study position process discovery as an integral part of data cleaning, which is a new angle in literature. Moreover, the interactive character of the approach supports stepwise data quality improvement in close collaboration with domain experts.

The remainder of this paper is structured as follows. Section 2 highlights some key prior research on data quality in the process mining field. Section 3 presents the interactive data cleaning approach. In Section 4 the case study is outlined. The paper ends with a discussion and conclusion in Section 5.

2 Related work

Data quality has been widely studied in several domains such as statistics and data mining [4]. This section will focus on related research within the process mining field, which can be subdivided in: (i) the identification of data quality issues, and (ii) the mitigation of data quality issues.

2.1 Identification of data quality issues

With respect to the identification of data quality issues, Verhulst [23] reviews existing literature to develop a data quality framework for event logs consisting of 12 high-level dimensions such as completeness and consistency. Bose et al. [6] distinguish 27 more specific event log data quality problems in four categories: missing data, incorrect data, imprecise data, and irrelevant data. Examples of issues are missing events, incorrect timestamps, and imprecise resource information. Using the framework of Bose et al. [6] , Mans et al. [15] evaluate the data quality in the HIS-database of the Maastricht University Medical Centre.

Following an interview-based approach, they conclude that the three most frequently occurring issues are missing events, imprecise timestamps, and imprecise resource information. Similar to Mans et al. [15], Kurniati et al. [14] assess the data quality of the publicly available MIMIC-III database for process mining purposes. This involves, amongst others, determining whether case identifiers, activity labels and timestamps are available, and whether duplicated data is present.

While Bose et al. [6] outline event log quality issues at a generic level, Suriadi et al. [21] define 11 event log imperfection patterns in a more detailed way. This involves, amongst others, a description of how an issue manifests itself in the log and its side effects [21]. Similarly, Vanbrabant et al. [22] present a set of specific data quality assessment techniques which can be used to identify data flaws prior to its use for process analysis purposes.

At a methodological level, Andrews et al. [3] describe a cyclical methodology aiming to support the initial stages of a process mining project while taking data quality explicitly into account. Another approach, presented in Fox et al. [13], stresses, amongst others, the importance of keeping a structured data quality register.

2.2 Mitigation of data quality issues

Several authors propose heuristics for data cleaning in an effort to improve the quality of an event log. These approaches range from conceptual recommendations to formal methods. For an extended reference list on event log cleaning methods, the reader is referred to the recent review by Solti [20].

An example of a conceptual approach are the recommendations of Suriadi et al. [21] to tackle each of the defined event log imperfection patterns. When, for instance, multiple events share the same timestamp because they are recorded by saving a single form, they propose to merge these events in a single [21]. Similarly, Martin [16] conceptually argues the potential of data integration to tackle quality problems in HIS-data.

More formal methods are developed to impute missing data in an event log [5,8] or to correct wrong data [9,18]. While doing so, these methods typically require domain knowledge, often in the form of a process model. This holds, e.g., for Rogge-Solti et al. [18], where a process model is used to repair timestamp errors [18]. Similarly, Di Francescomarino et al. [8] assume that a correct and complete process model is available in an effort to complete event log traces using action languages. Besides a process model, Bayomie et al. [5] also requires activity duration information as an input to detect missing case identifiers by means of decision trees. For the purpose of repairing event ordering issues in an event log, Dixit et al. [9] do not require a full process model. In their approach, users are required to establish ordered relationships between activities, which are matched with the event log using alignment techniques [9].

While the aforementioned approaches require domain knowledge to operate, e.g. in the form of a process model, Nguyen et al. [17] recently developed approaches to detect anomalies in an event log and to add missing values without a need for prior knowledge. To this end, autoencoders, which is a specific type of neural networks, are used. Even though preliminary results on structured artificial data are promising, their approaches still experience difficulties to manage the variability and complexity of real-life data [17].

3 Interactive data cleaning

This paper proposes an interactive data cleaning approach, which is visualized in the boxed area in Figure 1. The approach centers around (i) data quality assessment to identify data quality problems and (ii) data cleaning heuristics to mitigate these problems. Taking a raw log as an input, a user will perform both data-based and discovery-based data quality assessment. Existing assessment approaches in the process mining field have a strong data-based focus [21,22], i.e. they only concentrate on retrieving problematic patterns in the dataset. However, discovery-based assessment, aiming to identify data quality issues by discovering process models, can enable the identification of data inconsistencies which might remain hidden during data-based assessment. The discovered process models can relate to the control-flow, but also to other process mining perspectives such as the organizational perspective. The potential of discovery-based assessment will be illustrated in the case study in Section 4.

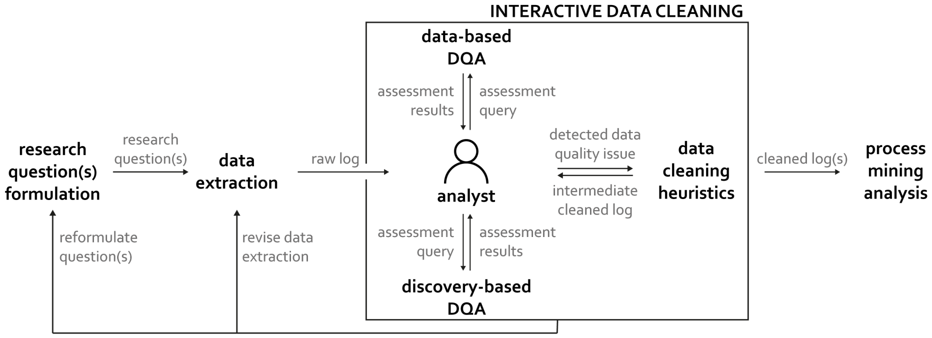

Fig. 1. Interactive data cleaning approach

Based on the assessment results, the analyst can specify appropriate data cleaning heuristics to rectify the detected issues. This generates an intermediate cleaned log, which can, once again, be subject to data quality assessment. After several iterations between assessment and cleaning, a cleaned event log is obtained, enabling the analyst to proceed with the process mining analysis.

Figure 1 positions the interactive data cleaning approach within the broader context of a question-driven process mining project [1] because data cleaning is likely to depend on the question(s) under consideration. When questions relate solely to the process control-flow, the order of activities is essential and the exact timestamp values are of secondary importance. If, in contrast, a process performance analysis is required for a particular question, the analyst should be more reluctant towards changing timestamp values during data cleaning. This shows that, when research questions cover several types of process mining analyses, it might be necessary to generate several cleaned event logs. Data quality assessment can also instigate a reformulation of the research questions as a question might no longer be answerable due to data quality issues. Alternatively, data quality assessment might indicate that the data extraction process should be revised to, e.g., add additional data to the raw log.

The approach in Figure 1 is complementary to existing process mining methodologies. It presents an extension of the log inspection stage of the Process Diagnostics Method, which is the stage that aims to familiarize the analyst with the event log [7]. When considering the L* life-cycle model, interactive data cleaning can be positioned between data extraction (stage 1) and the creation of a control-flow model (stage 2) [1]. Within the PM^2 methodology, it can strengthen the data processing stage as no explicit attention is attributed to data cleaning [10]. Figure 1 is also consistent with the quality-driven process mining preparation methodology of Andrews et al. [3]. In particular, it proposes an interactive approach to operationalize the event quality, pre-study process mining, and evaluation and feedback steps [3].

4 Case study: appointment data of an outpatient clinic

This section illustrates some key elements of the interactive data cleaning approach using a case study with appointment data of an outpatient clinic. First, the case study is described (Section 4.1). Afterwards, data-based data quality assessment (Section 4.2) and discovery-based data quality assessment (Section 4.3) are illustrated. Pointers to data cleaning heuristics are added to these last two sections. Due to space limitations, the case study will focus on the components included in the boxed area in Figure 1 and will center around the control-flow. A more extensive case study, encompassing the entire approach in Figure 1, will be conducted in future work.

4.1 Case study description

The dataset for the case study was provided by University Hospital La Fe (Valencia, Spain), which provides healthcare services to more than 300,000 people and accounts for over 1,100 doctors. The hospital has an area devoted to outpatient services of a wide variety of clinical specialties, which is the specific focus of the case study.

The outpatient clinic is equipped with an automatic registration system using magnetic cards. This magnetic card is used to identify a patient during his/her visit to the hospital. More specifically, four reference times are recorded:

- **Arrival time (At):** time at which the patient arrives at the waiting room and he/she registers using the magnetic card
- **Call time (Ct):** time at which the doctor calls the patient to the consultation

- **Entry time (Et):** time at which the patient enters the room for the consultation
- **Departure time (Dt):** time at which the patient leaves from the consultation

The studied dataset contains anonymized information on 1.6 million appointments that took place at the hospital for 262,061 unique patients. Even though the data shows that a patient can have several appointments during a single visit, the majority of the visits consist of only one or two consultations (87.70% and 10.80% of all visits, respectively).

4.2 Data-based data quality assessment and data cleaning heuristics

The available dataset suffers from several data quality issues. This subsection highlights the key timestamp-related data quality issues that were detected during data-based data quality assessment.

Missing timestamps Missing timestamps imply that one or more of the reference timestamps are absent for an appointment. This quality issue is explicitly recognized in the framework of Bose et al. [6], and belongs to the missing values category defined by Vanbrabant et al. [22].

The call time is absent the most often (for 24.44% of the appointments), followed by the arrival time (15.43%), entry time (4.16%) and departure time (4.15%). Table 1 provides richer insights on this matter as it studies the combinations of missing timestamps. From this table, it follows that all timestamps are recorded for 74.64% of the appointments. For 10.35% of the appointments both the arrival time and the call time are missing, while for 0.92% only the former is missing and for 9.93% only the latter is absent. Note that for 4.4% of the consultations, none of the reference times are recorded. Other missing timestamp combinations occur less frequently.

Table 1. Missing timestamps

Missing timestamps	**n**	**%**
None	1,222,289	74.64
At, Ct	169,534	10.35
Ct	162,589	9.93
At, Ct, Et, Dt	67,766	4.14
At	15,042	0.92
At, Ct, Et	261	0.02
At, Et, Dt	88	<0.01
At, Ct, Dt	13	<0.01
Dt	3	<0.01
Ct, Et, Dt	1	<0.01
Ct, Et	1	<0.01

Overlapping timestamps Another data quality problem are overlapping timestamps, which means that several of the reference times share the same timestamp. This can, at least partly, be attributed to the fact that timestamps are recorded at the granularity level of minutes. When timestamps are expected to be close to each other, e.g. for the call time and the entry time, these times can be correct. In other situations, e.g. between the entry and the departure time, overlapping timestamps are more problematic. In relation to the data quality framework of Bose et al. [6], this problem can belong to either the imprecise or the incorrect timestamp category (depending on whether the issue is problematic). In Vanbrabant et al. [22], it is positioned within the inexactness of timestamps group.

Table 2 provides an overview of the occurrence of overlapping timestamps. When assuming that absent timestamps do not overlap, all reference times differ in 33.46% of the appointments. The most frequently occurring overlap is situated between the call time and the entry time (in 61.89% of the appointments this is the only overlap). While the former could be perceived as normal behavior, this does not hold for an overlap between the entry and departure time (the only overlap in 3.74% of the appointments). As a consultation is unlikely to end in the same minute it started, an overlap indicates that one of these timestamps are incorrect.

Table 2. Overlapping timestamps

Overlapping timstamps	**n**	**%**
Ct, Et	1,013,461	61.89
None	547,943	33.46
Et, Dt	61,259	3.74
Ct, Et, Dt	7,691	0.47
At, Ct, Et	2,660	0.16
At, Et	1,972	0.12
At, Ct	1,347	0.08
{At, Ct} and {Et, Dt}	618	0.04
At, Ct, Et, Dt	315	0.02
Ct, Dt	150	0.01
At, Et, Dt	78	<0.01
At, Dt	73	<0.01
{At, Dt} and {Ct, Et}	17	<0.01
At, Ct, Dt	3	<0.01

Time ordering violation The correct time ordering for an appointment is known, i.e. arrival time → call time → entry time → departure time. Hence, appointments for which this order is not respected are likely to contain incorrect timestamps. This quality issue belongs to the incorrect timestamp category according to Bose et al. [6] and constitutes a violation of logical order in the framework of Vanbrabant et al. [22].

Time order violations occurring at least 250 times are included in Table 3. Note that appointments which violate the correct time ordering due to missing timestamps are not taken into consideration. From Table 3, it can be observed that the most frequent violation constitutes the recording of the arrival time between the start and the completion of a consultation.

Table 3. Time order violations

Time order violation	n	%
Et → At → Dt	12,267	0.75
At → Ct → Dt → Et	1,630	0.10
Et → Dt → At	1,531	0.09
Dt → At → Ct → Et	892	0.05
Ct → At → Et → Ct	463	0.03

Appointment overlap in the same room Besides the aforementioned issues, several other data quality issues were discovered. An example involves the observation of appointment overlaps in the same room, implying that the next appointment in a particular room appears to have started before the current one has ended. For instance: the appointment of patient A in a specific room is recorded from 10:58 until 11:06, while the consultation of patient B in that same room takes place from 11:00 until 11:15. Even though this is not sensible in practice, this pattern has been identified in the dataset for 605,188 appointments.

Data cleaning heuristics Based on the results of the data-based data quality assessment, several data cleaning heuristics are defined taking into account expected behavior and knowledge about the process. Considering the quality problems outlined above, some exemplary cleaning heuristics are:

- **At and Ct missing:** Issue present in 10.35% of the appointments. Under the assumption that these timestamps occur right before Et, both timestamps can be imputed right before Et.
- **Ct missing:** Issue present for 9.93% of the appointments. Under the assumption that Ct happens immediately before the consultation starts, Ct can be imputed right before Et.
- **Ct and Et overlapping:** Issue occurs for 61.21% of the appointments. Under the assumption that both timestamps reflect distinct states, timestamps can be corrected such that entry happens after the call.
- **Appointment overlap in same room:** Issue occurs for 605,188 appointments. When making the assumption that a room can only host one appointment at the same time, timestamps can be corrected such that the departure time of the previous appointment in a particular should precede the entry time of the current appointment in that room.

4.3 Discovery-based data quality assessment and data cleaning heuristics

To illustrate discovery-based data quality assessment, two distinct process models will be considered. The first one, shown in Figure 2, will act as a reference model as it is retrieved from the full dataset without applying data cleaning heuristics. The second model, shown in Figure 3, is mined from the full dataset after applying the data cleaning heuristics mentioned at the end of Section 4.2.

Both models are discovered using PALIA [12]. Arrow colours indicate the number of times that a path is followed (with green referring to less frequent paths and red to more frequent paths). The colors of the circles reflect the average time that elapses between that event and the next event. When mining a process model, we enforced PALIA with two basic conditions to produce a model: (i) reliable time and date formats, and (ii) reject parallel branches as a patient cannot be in several states at exactly the same time. Consequently, Figure 2 is based on 1,090,808 appointments (66.61% of the dataset) and Figure 3 on 1,406,326 appointments (85.88% of the dataset).

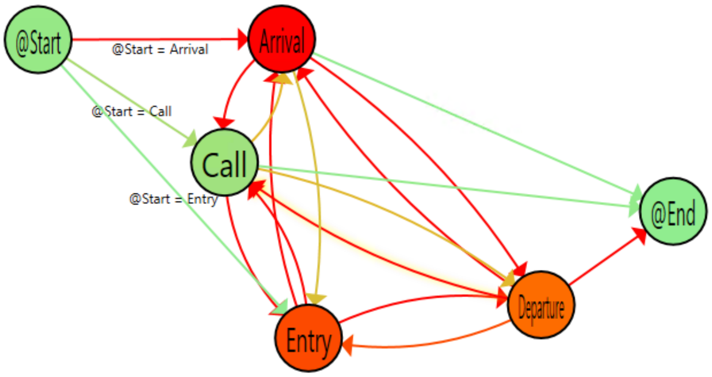

Fig. 2. Discovered process model before data cleaning heuristics

When comparing Figures 2 and 3, it becomes clear that the data cleaning heuristics based on data-based data quality assessment succeed in removing some anomalous behavior caused by data quality problems. Compared to Figure 2, where almost every connection is marked as frequent, the majority of the patients follow the expected trajectory (Arrival → Call → Entry → Departure).

Despite the effectiveness of the data cleaning heuristics, Figure 3 shows patterns that require further investigation to determine whether they originate from further data quality issues or just represent unexpected process characteristics. Examples of such patterns include:

- **At after Ct:** These patients may have been called by the doctor before arriving to the appointment, or they forgot to record their arrival with their magnetic card and only do this once they have been called.

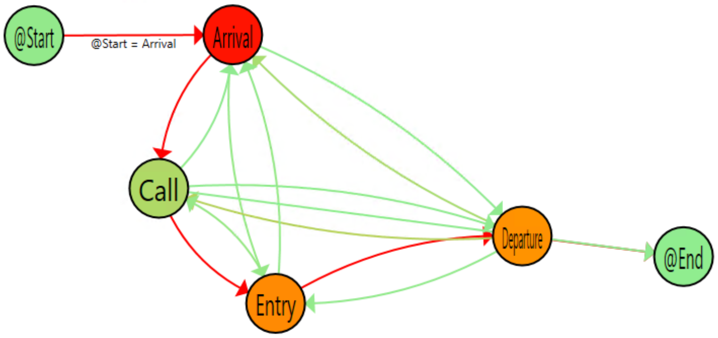

Fig. 3. Discovered process model after data cleaning heuristics

- **At after Et:** This pattern suggests that the arrival takes place after the consultation has started, or that the appointment has ended without the departure being recorded.
- **Ct after Et:** Some appointments show this anomalous behavior as a patient who has entered should no longer be called for that appointment. Further analysis should determine whether the call is related to another appointment of this patient, or whether the doctor performs the calling action once the patient has entered the cabinet.
- **Departure connected to Arrival, Call and Entry:** This exemplifies the process variability. The model shows that some patients with multiple appointments during one visit only seem to arrive after their first appointment has finished. Moreover, some patients are immediately called, without an arrival preceding it. Other patients immediately go to the doctor's cabinet for their next appointment (Departure → Entry).

The potential anomalies that appear from analyzing the mined control-flow model need to be sorted out before a process mining analysis can start. The identified issues exemplify the added-value of discovery-based data quality assessment compared to only using data-based data quality assessment. The need for domain expertise to investigate the aforementioned issues supports the need for an interactive approach, as visualized in Figure 1 by the central position of the analyst. When these issues can be attributed to a data quality issue, suitable data cleaning heuristics need to be specified.

5 Discussion and conclusion

This paper proposed an interactive data cleaning approach consisting of three key components: data-based data quality assessment, discovery-based data quality assessment, and data cleaning heuristics. Existing research on data quality assessment in process mining focuses on the identification of problematic patterns in the data, i.e. data-based data quality assessment. The proposed approach argues that process discovery should also be considered during assessment as this can unveil data quality issues that might otherwise have remained hidden. This was illustrated in a case with a fairly simple process structure. The more complex the process structure becomes, the likelier it becomes that some data problems are overlooked during data-based quality assessment. In such contexts, the added-value of discovery-based insights will become even more significant.

The interactive character of the proposed approach requires an analyst with sufficient domain expertise in control of interactive data cleaning. When applying the approach, an inherent risk for confirmation bias exists. Confirmation bias refers to a person's tendency to search for information that confirms his/her beliefs and to avoid contradicting information [19]. Within the context of data cleaning, this would imply that an analyst might attribute patterns that conflict his/her process beliefs to data quality issues and take according data cleaning measures. Consequently, a prudent approach towards data cleaning is required. Future research could develop guidelines to support analysts and researchers on this matter. A potential guideline could relate to maintaining detailed data cleaning records to ensure that each cleaning step is fully traceable.

Besides guidelines for data cleaning, several other research challenges can be distinguished. Foremost, formal techniques for discovery-based data quality assessment should be developed, similar to existing techniques for data-based assessment. This would enable the development of an integrated toolkit to support interactive data cleaning. Besides enabling the application of assessment techniques and cleaning heuristics, this toolkit should also make the effect of the heuristics explicit. Related to the latter, another promising direction for future work is a benchmark study on the effect of various data cleaning methods on process mining outcomes. Besides Dixit et al. [9] as a notable exception, limited research attention is attributed to this topic.

References

1. van der Aalst, W.M.P.: Process mining: data science in action. Springer, Heidelberg (2016)
2. van der Aalst, W.M.P., Adriansyah, A., ..., Wynn, M.: Process mining manifesto. Lecture Notes in Business Information Processing **99**, 169–194 (2012)
3. Andrews, R., Wynn, M.T., Vallmuur, K., ter Hofstede, A.H., Bosley, E., Elcock, M., Rashford, S.: Leveraging data quality to better prepare for process mining: an approach illustrated through analysing road trauma pre-hospital retrieval and transport processes in queensland. International Journal of Environmental Research and Public Health **16**(7), 1138 (2019)
4. Batini, C., Scannapieco, M.: Data quality: concepts, methodologies and techniques. Springer, Heidelberg (2006)
5. Bayomie, D., Helal, I.M., Awad, A., Ezat, E., ElBastawissi, A.: Deducing case ids for unlabeled event logs. Lecture Notes in Business Information Processing **256**, 242–254 (2016)
6. Bose, R.J.C.P., Mans, R.S., van der Aalst, W.M.P.: Wanna improve process mining results? It's high time we consider data quality issues seriously. Tech. Rep. BPM Center Report BPM-13-02 (2013)

7. Bozkaya, M., Gabriels, J., van der Werf, J.M.: Process diagnostics: a method based on process mining. In: 2009 International Conference on Information, Process, and Knowledge Management. pp. 22–27. IEEE (2009)
8. Di Francescomarino, C., Ghidini, C., Tessaris, S., Sandoval, I.V.: Completing workflow traces using action languages. Lecture Notes in Computer Science **9097**, 314–330 (2015)
9. Dixit, P.M., Suriadi, S., Andrews, R., Wynn, M.T., ter Hofstede, A.H., Buijs, J.C., van der Aalst, W.M.P.: Detection and interactive repair of event ordering imperfection in process logs. Lecture Notes in Computer Science **10816**, 274–290 (2018)
10. van Eck, M.L., Lu, X., Leemans, S.J.J., van der Aalst, W.M.P.: PM^2: a process mining project methodology. Lecture Notes in Computer Science **9097**, 297–313 (2015)
11. Fernández-Llatas, C., Lizondo, A., Monton, E., Benedi, J.M., Traver, V.: Process mining methodology for health process tracking using real-time indoor location systems. Sensors **15**(12), 29821–29840 (2015)
12. Fernández-Llatas, C., Valdivieso, B., Traver, V., Benedi, J.M.: Using process mining for automatic support of clinical pathways design. In: Fernández-Llatas, C., Garcia-Gomez, J. (eds.) Data mining in clinical medicine, pp. 79–88. Springer (2015)
13. Fox, F., Aggarwal, V.R., Whelton, H., Johnson, O.: A data quality framework for process mining of electronic health record data. In: 2018 IEEE International Conference on Healthcare Informatics. pp. 12–21 (2018)
14. Kurniati, A.P., Rojas, E., Hogg, D., Hall, G., Johnson, O.A.: 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 (2018)
15. Mans, R.S., van der Aalst, W.M.P., Vanwersch, R.J.B.: Process mining in healthcare: evaluating and exploiting operational healthcare processes. Springer, Heidelberg (2015)
16. Martin, N.: Using indoor location system data to enhance the quality of healthcare event logs: opportunities and challenges. Lecture Notes in Business Information Processing **342**, 226–238 (2018)
17. Nguyen, H.T.C., Lee, S., Kim, J., Ko, J., Comuzzi, M.: Autoencoders for improving quality of process event logs. Expert Systems with Applications **131**, 132–147 (2019)
18. Rogge-Solti, A., Mans, R.S., van der Aalst, W.M.P., Weske, M.: Repairing event logs using timed process models. Lecture Notes in Computer Science **8186**, 705–708 (2013)
19. Sanderson, C.A.: Social psychology. Wiley, Hoboken (2010)
20. Solti, A.: Event log cleaning for business process analytics. In: Sakr, S., Zomaya, A. (eds.) Encyclopedia of Big Data Technologies. Springer, Heidelberg (2018)
21. Suriadi, S., Andrews, R., ter Hofstede, A.H., Wynn, M.T.: Event log imperfection patterns for process mining: towards a systematic approach to cleaning event logs. Information Systems **64**, 132–150 (2017)
22. Vanbrabant, L., Martin, N., Ramaekers, K., Braekers, K.: Quality of input data in emergency department simulations: Framework and assessment techniques. Simulation Modelling Practice and Theory **91**, 83–101 (2019)
23. Verhulst, R.: Evaluating quality of event data within event logs: an extensible framework. Master's thesis, Eindhoven University of Technology (2016)