# Towards Privacy-Preserving Process Mining in Healthcare

Anastasiia Pika[1⋆], Moe T. Wynn[1], Stephanus Budiono[1], Arthur H.M. ter Hofstede[1], Wil M.P. van der Aalst[2,1], and Hajo A. Reijers[3,1]

[1] Queensland University of Technology, Brisbane, Australia
{a.pika,m.wynn,sn.budiono,a.terhofstede}@qut.edu.au
[2] RWTH Aachen University, Aachen, Germany, wvdaalst@pads.rwth-aachen.de
[3] Utrecht University, Utrecht, Netherlands, h.a.reijers@uu.nl

**Abstract.** Process mining has been successfully applied in the healthcare domain and helped to uncover various insights for improving healthcare processes. While benefits of process mining are widely acknowledged, many people rightfully have concerns about irresponsible use of personal data. Healthcare information systems contain highly sensitive information and healthcare regulations often require protection of privacy of such data. The need to comply with strict privacy requirements may result in a decreased data utility for analysis. Although, until recently, data privacy issues did not get much attention in the process mining community, several privacy-preserving data transformation techniques have been proposed in the data mining community. Many similarities between data mining and process mining exist, but there are key differences that make privacy-preserving data mining techniques unsuitable to anonymise process data. In this article, we analyse data privacy and utility requirements for healthcare process data and assess the suitability of privacy-preserving data transformation methods to anonymise healthcare data. We also propose a framework for privacy-preserving process mining that can support healthcare process mining analyses.

**Keywords:** Process mining · healthcare process data · data privacy

## 1 Introduction

Technological advances in the fields of business intelligence and data science empower organisations to become "data-driven" by applying new techniques to analyse large amounts of data. Process mining is a specialised form of data-driven analytics where process data, collated from different IT systems typically available in organisations, are analysed to uncover the real behaviour and performance of business operations [1]. Process mining was successfully applied in the healthcare domain and helped to uncover insights for improving operational efficiency of healthcare processes and evidence-informed decision making [4, 6,

⋆ Corresponding author

11, 12, 14]. A recent literature review [6] discovered 172 articles which report applications of various process mining techniques in the healthcare domain.

While the potential benefits of data analytics are widely acknowledged, many people have grave concerns about irresponsible use of their data. An increased concern of society with protecting the privacy of personal data is reflected in the growing number of privacy regulations that have been recently introduced or updated by governments around the world. Healthcare data can include highly sensitive attributes (e.g., patient health outcomes/diagnoses, the type of treatments being undertaken), and hence privacy of such data needs to be protected.

The need to consider data privacy in process mining and develop privacy-aware tools was raised at an early stage in the Process Mining Manifesto [3]. However, the process mining community has, until recently, largely overlooked the problem. A few recent articles highlight "a clear gap in the research on privacy in the field of process mining" [10] and make first attempts to address some privacy-related challenges [5, 7, 9, 10, 13] yet, significant challenges remain.

Privacy considerations are quite well-known in the field of data mining and a number of privacy-preserving data transformation techniques have been proposed [2, 17] (e.g., data swapping, generalisation or noise addition). *Although there are many similarities between data mining and process mining, some key differences exist that make some of the well-known privacy-preserving data mining techniques unsuitable to transform process data.* For example, the addition of noise to a data set may have an unpredictable impact on the accuracy of all kinds of process mining analyses.

In this article, we present related work (Section 2), analyse data privacy and utility requirements for process data typically recorded in the healthcare domain (Section 3) and then assess the suitability of privacy-preserving data transformation methods proposed in the data mining and process mining fields to anonymise healthcare process data (Section 4). We show that the problem of privacy protection for healthcare data while preserving data utility for process mining analyses is challenging and we propose a privacy-preserving process mining framework as a possible solution to address this problem in Section 5. Section 6 concludes the paper.

## 2 Related Work

**Privacy-preserving data mining.** Privacy, security, and access control considerations are quite well-known in the general field of data mining. A number of data transformation techniques, access control mechanisms and frameworks to preserve data privacy have been proposed [2, 8, 17]. In order to preserve data privacy, privacy-preserving methods usually reduce the representation accuracy of the data [2]. Such data modifications can affect the quality of analyses results. The effectiveness of the transformed data for analyses is often quantified explicitly as its *utility* and the goal of privacy-preserving methods is to *"maximize utility at a fixed level of privacy"* [2]. For example, privacy guarantees can be

specified in terms of *k-anonymity*: each record in a data set is indistinguishable from at least *k-1* other records.

Privacy-preserving data mining techniques can be generic or specific [17]. *Generic* approaches modify data in such a way that "the transformed data can be used as input to perform any data mining task" [17]. These approaches can provide anonymisation[4] by modifying records without introducing new values (e.g., data swapping) or they can modify original values (e.g., by adding noise). In *specific* approaches privacy preservation is embedded in specific data mining algorithms (e.g., privacy-preserving decision tree classification) [17]. Furthermore, *outputs* of some data mining algorithms can also be sensitive and methods that anonymise such outputs have been proposed (e.g., association rule hiding) [2]. Finally, *distributed* privacy-preserving methods are proposed for scenarios in which multiple data owners wish to derive insights from combined data without compromising privacy of their portions of the data [2]. Such methods often use cryptographic protocols for secure multi-party computations (SMC) [2].

Below, we describe traditional generic privacy-preserving data transformation approaches, such as data swapping, suppression, generalisation and noise addition [2]. *Data swapping* involves enacting privacy to a dataset by the existence of uncertainty. Uncertainty is introduced into individual records by swapping the true values of sensitive attributes between subsets of records [8]. *Suppression* anonymises data by omission. Values can be removed under three types of data suppression [2]. The most common type is column suppression which targets the presence of highly sensitive attributes whose values directly identify an individual (e.g., patient names). Alternatively, row suppression is used when outlier records are infrequent and difficult to anonymise. Value suppression omits selected sensitive attribute values. *Generalisation* methods define values approximately making it difficult for adversaries to identify records with full confidence [2]. The process of generalising usually includes the construction of a generalisation hierarchy, which is a predefined classification of values at decreasing levels of granularity. For numeric data, values are sorted into numerical ranges. For categorical data, a domain expert creates semantically meaningful generalisations using a tree structure. *Noise addition* can be used for both numerical and categorical data [17]. Numerical values are often anonymised by factoring randomly and independently generated "white noise" into the original data [2]. White noise is generated using a random distribution, often either uniform or Gaussian. Adding noise to categorical values is more complex, and can be achieved, for example, using clustering-based techniques [17].

**Privacy-preserving process mining.** A few recent articles made first attempts to address some privacy-related process mining challenges [5, 7, 9, 10, 13, 15, 16]. Mannhardt et al. [10] analysed privacy challenges in human-centered industrial environments and provided some generic guidelines for privacy in process mining. Liu et al. [9] presented a privacy-preserving cross-organisation process discovery framework based on access control. Tillem et al. [15, 16] presented interactive two-party protocols for discovery of process models from encrypted data,

[4] In this article, *anonymisation* refers to any method that can protect data privacy.

which are based on multiple communication rounds (and have high computation costs). The first privacy-preserving data transformation approach presented in the process mining community [5] proposes to use deterministic encryption methods for anonymisation of event log attribute values. (Such methods are also a part of the confidentiality framework proposed by Rafiei et al. [13].) Timestamps are treated as numeric values and are encrypted in a way that preserves the order of events. Deterministic encryption methods produce "the same ciphertext for a given plaintext" and preserve differences between values, which is important for process mining [13]. Encryption only provides weak data privacy protection and "could be prone to advanced de-anonymization techniques" [5]. More advanced privacy-preserving process mining approaches proposed by Rafiei et al. [13] and Fahrenkrog-Peterse et al. [7] will be discussed in detail in Section 4.

In this article, we focus on protecting privacy of process data in a healthcare organisation. Distributed privacy scenarios are not considered in this work.

## 3 Data Privacy and Utility Requirements: Healthcare

In order to realise our objective of privacy-preserving process mining for the healthcare domain, we first analyse privacy requirements for process data typically recorded in the healthcare domain, which is then followed by a discussion of data requirements of process mining approaches to analyse healthcare processes.

**Healthcare process data.** Process mining uses process data in the form of an event log, which represents collated and aggregated data from IT systems available in organisations. An event log contains events where each event refers to a case, an activity, a point in time, transaction type (e.g., *start* or *complete*) and (optionally) a resource and data attributes. An event log can be seen as a collection of cases and a case can be seen as a sequence of events.

Cases in healthcare processes typically refer to patients receiving treatments in a healthcare setting (e.g., a patient's pathway) and resources refer to medical personnel involved in the process. Figure 1 depicts an example event log which contains six events (represented by rows) related to two cases (*1* and *2*) where patient identifiers are already hidden. For example, we can see that case *1* refers to a patient whose age is *56*, who speaks English and was diagnosed with pancreatitis; activity *Register* is completed in this case; activity *Blood test* was started on *13/01/2019* at *17:01* by *Robert*; and treatment code *3456* is associated with activity *Triage* in case *1*. Data attributes can refer to cases (e.g., age, language and diagnosis) or to events (e.g., treatment codes are recorded for events associated with activity *Triage*). Data attributes used in this example are recorded in two publicly available healthcare logs. The healthcare MIMIC data set[5] contains information about language and diagnosis (as well as ethnicity, religion, marital status and insurance). The Dutch academic hospital event log[6] contains information about age, diagnosis and treatment codes.

[5] https://mimic.physionet.org/mimicdata/
[6] https://data.4tu.nl/repository/uuid:d9769f3d-0ab0-4fb8-803b-0d1120ffcf54

| Case ID | Activity | Type | Time | Resource | Age | Language | Diagnosis | Treatment Code |
|---|---|---|---|---|---|---|---|---|
| 1 | Register | complete | 12/01/2019 11:03 | Ann | 56 | EN | Pancreatitis | - |
| 1 | Triage | start | 12/01/2019 14:55 | Michael | 56 | EN | Pancreatitis | 3456 |
| 1 | Blood test | start | 13/01/2019 17:01 | Robert | 56 | EN | Pancreatitis | - |
| 2 | Register | complete | 14/01/2019 9:30 | Ann | 44 | IT | Pneumonia | - |
| 2 | X-ray | complete | 14/01/2019 11:00 | Mary | 44 | IT | Pneumonia | - |
| 2 | Triage | start | 14/01/2019 11:37 | Michael | 44 | IT | Pneumonia | 6543 |

**Fig. 1.** Example of an event log with typical healthcare data attributes.

**Legislative requirements.** An increased concern of people with protecting the privacy of their data is reflected in the growing number of privacy regulations that have been recently introduced (e.g., the EU General Data Protection Regulation (GDPR) 2018, the California Consumer Privacy Act of 2018) or updated by governments around the world (e.g., Australian Privacy Regulation 2013 under the Privacy Act 1988). In addition, data privacy requirements are often included in legislation governing specific sectors, e.g., Australian Healthcare Identifiers Act 2010.

Guidance for de-identification of protected health information in the US is provided in the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. For example, the "safe harbor" de-identification method of the HIPPA Privacy Rule prescribes removal of all elements of dates (except year) related to an individual (e.g., admission or discharge dates)[7]. In Australia, the Office of Australian Information Commissioner provides guidelines for the use of health information for research. The guidelines prescribe de-identification of personal information by "removing personal identifiers, such as name, address, d.o.b. or other identifying information" and "removing or altering other information that may allow an individual to be identified, for example, because of a rare characteristic of the individual, or a combination of unique or remarkable characteristics"[8]. Furthermore, the recently introduced My Health Records Amendment (Strengthening Privacy) Bill 2018 allows Australians to opt out of having an electronic health record and allows the deletion of their records permanently at any time. Whilst providing strong privacy protections for Australians; for analysis purposes, they also introduce data quality issues such as missing and incomplete data; thus reducing the utility of data and the accuracy of results.

Privacy of public healthcare data is typically protected by replacing sensitive attribute values with anonymised values (e.g., treatment codes are used in a publicly available Dutch academic hospital event log and subject IDs are used in the healthcare MIMIC data set) or by removing sensitive attributes from data (e.g., employee information is removed from both Dutch hospital and MIMIC data sets). All timestamps in the MIMIC data set were shifted to protect privacy: dates are randomly distributed, but consistent for each patient. The former

[7] https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#protected

[8] https://www.oaic.gov.au/engage-with-us/consultations/health-privacy-guidance/business-resource-collecting-using-and-disclosing-health-information-for-research

method only provides weak privacy protection while the latter methods can significantly decrease data utility.

**Privacy requirements for healthcare process data.** Healthcare process data can contain sensitive information such as patient or employee names or identifiers. Other attributes in the event log can also reveal patient or employee identities when combined with background knowledge about the process. For example, accident or admission time, a rare diagnosis or treatment, or a combination of age and language could potentially identify a patient. An employee could be identified by a combination of an activity name and execution time (e.g., when a blood test is always performed by the same employee during a shift). Hence, typical event log attributes such as *case ID, activity, time, resource* and many data attributes (e.g., a patient's personal and treatment information) can contribute to identity disclosure.

Furthermore, relations between events in a log can contribute to identity disclosure and this is especially pertinent for a healthcare event log due to the high variability of process paths typical for the sector [4]. Consider, for example, the Dutch hospital event log where 82% of cases follow unique process paths. Hence, someone with knowledge of the process could link these cases to individual patients. Moreover, cases which follow the same process path can include other atypical behaviors. In the Dutch hospital log, the fifth most frequent process variant is followed by 8 cases: 7 cases are related to only one organisational group ("Obstetrics and Gynecology clinic") and only one case is also related to the "Radiotherapy" group. Although the case does not follow a unique process path, the relation to the "Radiotherapy" group is unique and could be used by someone with knowledge of the process to identify the patient. Other examples of *atypical process behaviour* which could contribute to a patient's identity disclosure include abnormally short or long execution times of activities or cases, or an abnormally low or high number of resources involved in a case.

**Data requirements for process mining approaches.** All process mining algorithms require case IDs and activities to be recorded accurately in the log and most algorithms also require (accurate) timestamps. A recent literature review [6] discovered that the following types of process mining analyses were frequently used in healthcare: discovery techniques (which include process discovery as well as organisational mining approaches such as social network mining), conformance checking, process variant analysis and performance analysis.

- Process discovery techniques usually take as input a multi-set of traces (i.e., ordered sequences of activity labels) and do not require timestamps; however, timestamps are typically used to order events.
- Most academic process conformance and performance analysis techniques (e.g., alignment-based approaches) use formal models and require that complete traces are recorded in the log. Most commercial process mining tools (as well as some ProM plugins) convert the log to Directly Follows Graphs (DFG) annotated with frequencies and times, which show how frequently different activities follow each other and average times between them. DFG-based

tools do not require complete traces and only require that "directly-follows" relations between activities are preserved in the log.
- Organisational mining techniques require resource information to be recorded in the log (in addition to case IDs, activities and timestamps). Moreover, resource and data attributes can also be required by conformance checking approaches that consider different process perspectives.
- Process variant analysis, which is concerned with comparing process behaviour and performance of different cohorts, often uses case data attributes to distinguish between cohorts.

In order to comply with strict privacy requirements for healthcare data, one would need to consider *anonymising 1) event log attribute values and 2) atypical process behaviour*. However, many process mining techniques require that healthcare process data is accurate and representative. That is: *1) all events belong to a particular case*; *2) attributes that represent case identifiers and activity labels are accurate; and 3) timestamps are reliable and accurate*. Thus, the need to balance the privacy requirements of healthcare data and the utility requirements of process mining techniques is paramount. In the following section, we assess whether existing privacy-preserving data transformation approaches can preserve the attribute values and relations between events discussed above.

## 4 Anonymising Healthcare Process Data

### 4.1 Anonymising Sensitive Attribute Values

As discussed in Section 3, typical event log attributes such as *case, activity, time, resource* and many data attributes could contribute to identity disclosure. Below, we discuss how these attributes could be anonymised using generic data transformation approaches described in Section 2. We evaluate the suitability of deterministic encryption (referred to here as encryption), which was used to anonymise event log data [5, 13], and other traditional data transformation approaches proposed in the data mining community such as data swapping, value suppression, generalisation and noise addition (which, to the best of our knowledge, have not been applied to event logs). Figure 2 depicts how some of these techniques can be applied to the event log in Figure 1.

***Case*** identifiers can be encrypted (as well as other event log attributes); however, encryption does not provide strong data privacy protection (and may not be suitable to protect sensitive healthcare data). An underlying assumption of all process mining algorithms is that case identifiers are unique, which makes the application of value suppression and generalisation not suitable (these methods are used to hide infrequent attribute values). Adding noise to case identifiers can yield values that are no longer unique, which can decrease the accuracy of all process mining algorithms. Data swapping can be applied to case IDs without impact on process mining results.

***Activity*** labels can be encrypted; however, encrypted labels can be identified by someone with knowledge of the process (e.g., most or least frequent activities [13]). Moreover, encryption makes it difficult to interpret analysis results.

| Case ID | Activity | Type | Time | Resource | Age | Language | Diagnosis | Treatment Code |
|---|---|---|---|---|---|---|---|---|
| 2 | Register | complete | 12/01/2019 11:20 | Team A | 56 | EN | - | - |
| 2 | Triage | start | 12/01/2019 15:00 | Team B | 56 | EN | - | 3456 |
| 2 | Blood test | start | 13/01/2019 17:03 | Team C | 56 | EN | - | - |
| 1 | Register | complete | 14/01/2019 9:35 | Team A | 44 | IT | - | - |
| 1 | X-ray | complete | 14/01/2019 11:10 | Team C | 44 | IT | - | - |
| 1 | Triage | start | 14/01/2019 11:48 | Team B | 44 | IT | - | 6543 |

**Fig. 2.** Application of data transformation techniques to the event log in Figure 1: Case ID: swapping; Time: noise addition; Resource: generalisation; Diagnosis: suppression.

In addition, one must also encrypt process model labels when applying process mining algorithms that use process models as input (e.g., many process performance and conformance analysis approaches). Application of value suppression and generalisation to activity labels may affect the accuracy of process mining results where the utility loss depends on the process mining algorithm used. For example, removing infrequent activity labels may not have a significant effect on process discovery results (as process models often capture mainstream process behavior); however, process conformance analysis results may become invalid. One can use generalisation to hide some sensitive activities (e.g., replace activities "HIV test" and "Hepatitis C test" with activity "Blood test"). The result of process discovery performed on such logs will be correct; however, the discovered process model will be on a higher level of granularity. Noise addition and swapping activity labels will invalidate the results of all process mining algorithms. For example, if activity labels in a log are swapped, the resulting traces will consist of random activity sequences; hence, discovered process models will be incorrect, as well as other process mining results.

***Timestamps*** can be treated as numerical values and encrypted using methods which preserve the order of events. Such encryption will not affect the results of process mining algorithms that work with ordered events and do not require timestamps (such as many process discovery algorithms). On the other hand, an event log with encrypted timestamps will not be suitable for performance analysis. Value suppression and generalisation can be used to anonymise sensitive timestamps (e.g., as discussed in Section 3, according to the HIPAA Privacy Rule admission and discharge times must be anonymised). This will affect the accuracy of most process mining algorithms. For example, if value suppression is applied to admission times, the discovered process model will not include activity "Admission". On the other hand, if generalisation is applied to admission times (by only leaving year as prescribed by the HIPAA Privacy Rule), process discovery may not be affected; however, process performance analysis results may become invalid (as time between admission and other activities in the process will no longer be correct). Adding noise to timestamps or swapping their values will yield incorrect process mining results (as the order of events in the transformed log is no longer preserved).

***Resource*** information can be encrypted without impacting organisational mining results, while noise addition and swapping will invalidate such results (as resources will no longer be related to correct events and cases). One can

apply generalisation to resource information (e.g., by replacing individual identifiers with team identifiers), which will yield the analysis on a team level. Value suppression can affect the accuracy of organisational mining techniques (e.g., a discovered social network may have fewer nodes).

***Data*** attributes can be encrypted, though encryption of numerical values can make it difficult to conduct some analyses. For example, if *age* is encrypted, one can no longer compare process variants for different age cohorts. Value suppression can decrease the accuracy of process mining algorithms that use data (e.g., when infrequent age values are removed, the corresponding cases will not be included in process variant analysis). Using generalisation may decrease the accuracy of conformance analysis that consider data; however, it may not have any impact on variant analysis (e.g., when comparing different age groups). Noise addition and data swapping will nullify results of the methods that use data.

Table 1 summarises the suitability of different data transformation approaches to anonymising event log attribute values. Encryption has a minimal effect on data utility for most process mining algorithms; however, it may not provide a required level of privacy protection. Data swapping can be used to anonymise case IDs; however, application of this method to other event log attributes will invalidate process mining results. Noise addition will nullify all process mining results. Value suppression and generalisation are not suitable for case IDs (as they have unique values), these methods can be applied to other attributes; however, the accuracy of process mining results may be affected.

**Table 1.** Suitability of privacy-preserving data transformation approaches to anonymising event log attributes: NA: not applicable; '+': does not affect process mining results; '-': can be used to anonymise an attribute, however invalidates process mining results; '+/-': can decrease the accuracy of some process mining methods.

| | **Case ID** | **Activity** | **Time** | **Resource** | **Data** |
|---|---|---|---|---|---|
| Encryption (deterministic) | + | + | +/- | + | +/- |
| Swapping | + | - | - | - | - |
| Noise addition | - | - | - | - | - |
| Value suppression | NA | +/- | +/- | +/- | +/- |
| Generalisation | NA | +/- | +/- | +/- | +/- |

### 4.2 Anonymising Atypical Process Behaviour

As discussed in Section 3, relations between events in the log (such as event order or grouping of events by case identifiers) can be used to identify atypical process behaviour (which could be linked to individuals). There could be many different types of atypical process behaviour (e.g., infrequent activity sequences, abnormal number of resources or atypical durations). Below, we evaluate two approaches which target anonymisation of atypical process behaviour: a confidentiality framework [13] and PRETSA [7].

The ***confidentiality framework*** for process mining [13] combines a few data transformation techniques. The first step of the framework is filtering out all cases "that do not reach the minimal frequencies" [13]. The framework changes the structure of an event log: a new attribute "previous activity" is added (which specifies for each event the preceding activity in a case) and case IDs are removed. Since events in the transformed log are no longer related to cases, it is impossible to identify traces (and atypical process behaviour). However, the transformed log can no longer be used by process mining algorithms that require complete traces; it is only suitable for DFG-based tools (e.g., commercial process mining tools). Moreover, as discussed in Section 3, healthcare processes are often highly variable and in some processes all traces in the log may be unique. The confidentiality framework (which proposes to filter out traces with infrequent process behaviour) may not be suitable to anonymise event log data from such healthcare processes.

***PRETSA*** [7] is a log sanitisation algorithm, which represents a log as a prefix tree and then transforms the tree until given privacy guarantees are met while striving to preserve directly follows relations. The approach allows to anonymise two types of atypical process behaviour: infrequent traces and atypical activity execution times. The article [7] evaluates the impact of the log transformation on the results of process discovery and performance analysis algorithms using three real-life logs. It also compares the performance of PRETSA with a "baseline" approach which filters out infrequent traces. The evaluation showed that PRETSA outperforms the baseline approach on all logs and data utility losses are minimal for event logs which do not have many unique traces. However, for a log in which most traces are unique the utility of the transformed log is significantly decreased, even more so for stricter privacy requirements (which means that the algorithm may not be suitable for healthcare process data).

## 5 Privacy-Preserving Process Mining Framework

On the one hand, the healthcare sector needs to comply with strict data privacy requirements; on the other hand, healthcare process data often contain many sensitive attributes and highly variable process behaviour that presents additional threats to privacy. Ensuring high levels of privacy protection for such data while also preserving data utility for process mining purposes remains an open challenge for the healthcare domain.

The analysis of the suitability of existing data transformation approaches to anonymise healthcare process data (presented in Section 4) highlighted the trade-off between data privacy and utility. The methods that preserve higher data utility for process mining purposes (e.g., encryption) do not provide strong privacy protection. On the other hand, the methods that can satisfy stricter privacy requirements (e.g., value suppression and generalisation) can decrease the accuracy of results. The magnitude of the data utility loss depends on characteristics of a particular log and varies for different process mining algorithms. Furthermore, performing analyses on anonymised process data without understanding how the data was transformed can yield unpredictable results.

We propose a privacy-preserving process mining framework (Figure 3) which uses a history of privacy-preserving data transformations to quantify their impact and improve the accuracy of process mining results. The proposed framework can be applied to the healthcare domain as well as other domains with high privacy needs. The first two steps of the framework (i.e., data anonymisation and creation of privacy metadata) are performed by the data owner or a trusted representative. The third step (i.e., conducting privacy-preserving process mining analysis) can be performed by (not trusted) third parties.

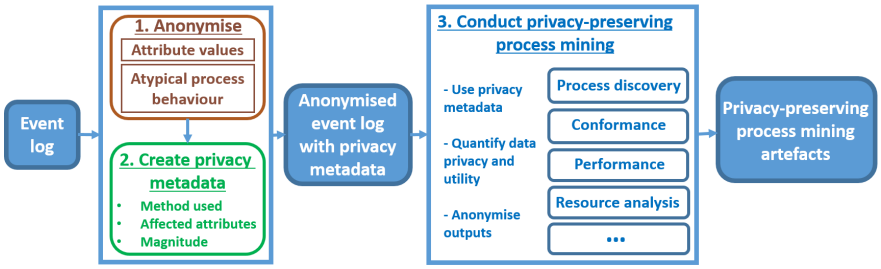


**Fig. 3.** Privacy-preserving process mining framework.

The first step of the framework is ***anonymising*** sensitive information such as sensitive attribute values and atypical process behavior. Anonymisation of sensitive attribute values could be achieved using data transformation approaches discussed in Section 4.1. Some atypical process behaviours can be anonymised using approaches discussed in Section 4.2; however, methods which could anonymise different types of atypical process behaviour in highly variable processes while preserving data utility for different algorithms are yet to be developed.

The second step of the framework is ***creating privacy metadata***, which maintains the history of privacy-preserving data transformations in a standardised and machine readable way. Such metadata can be stored in a privacy extension to the IEEE XES log format used for process mining. This privacy metadata will also assist in formally capturing the log characteristics that influence the anonymisation efforts for various forms of process mining.

The third step of the framework is ***conducting privacy-preserving process mining*** analysis of the anonymised event log with privacy metadata. The privacy metadata can be exploited by new "privacy-aware" process mining techniques to improve mining results. Privacy-aware process mining methods could also quantify data privacy and utility (e.g., by providing confidence measures). Finally, results of process mining techniques could also threaten privacy (by identifying patterns which are linked to individuals). To the best of our knowledge, anonymisation methods for process mining outputs are yet to be developed.

## 6 Conclusion

Keeping healthcare process data private while preserving data utility for process mining presents a challenge for the healthcare domain. In this article, we analysed data privacy and utility requirements for healthcare process data, assessed the suitability of existing privacy-preserving data transformation approaches and proposed a privacy-preserving process mining framework that can support process mining analyses of healthcare processes. A few directions for future work include: an empirical evaluation of the effects of privacy-preserving data transformation methods on healthcare logs, the development of privacy extensions for logs and the development of privacy-aware process mining algorithms.

## References

1. van der Aalst, W.: Process Mining: Data Science in Action. Springer-Verlag, Berlin (2016), http://www.springer.com/978-3-662-49850-7
2. Aggarwal, C.C.: Data mining: the textbook. Springer (2015)
3. van der Aalst et al., W.: Process mining manifesto. In: BPM 2011 Workshops proceedings. LNBIP, Springer-Verlag, Berlin (2011)
4. Andrews, R., Suriadi, S., Wynn, M., ter Hofstede, A.: Healthcare process analysis. Process Modelling and Management for HealthCare; CRC Press, USA (2017)
5. Burattin, A., Conti, M., Turato, D.: Toward an anonymous process mining. In: FiCloud 2015. pp. 58–63. IEEE (2015)
6. Erdogan, T.G., Tarhan, A.: Systematic mapping of process mining studies in healthcare. IEEE Access **6**, 24543–24567 (2018)
7. Fahrenkrog-Petersen, S.A., van der Aa, H., Weidlich, M.: PRETSA: Event log sanitization for privacy-aware process discovery. ICPM (accepted) (2019)
8. Fienberg, S.E., McIntyre, J.: Data Swapping: Variations on a Theme by Dalenius and Reiss. In: Int. Workshop on PSD. pp. 14–29. Springer (2004)
9. Liu, C., Duan, H., Qingtian, Z., Zhou, M., Lu, F., Cheng, J.: Towards comprehensive support for privacy preservation cross-organization business process mining. IEEE Transactions on Services Computing (2016)
10. Mannhardt, F., Petersen, S.A., Oliveira, M.F.: Privacy challenges for process mining in human-centered industrial environments. In: IE 2018. pp. 64–71. IEEE (2018)
11. Mans, R.S., Van der Aalst, W.M., Vanwersch, R.J.: Process mining in healthcare: evaluating and exploiting operational healthcare processes. Springer (2015)
12. Partington, A., et al.: Process mining for clinical processes: a comparative analysis of four Australian hospitals. ACM (TMIS) **5**(4), 19 (2015)
13. Rafiei, M., von Waldthausen, L., van der Aalst, W.: Ensuring confidentiality in process mining. In: SIMPDA 2018 (2018)
14. Rojas, E., Sepúlveda, M., Munoz-Gama, J., Capurro, D., Traver, V., Fernandez-Llatas, C.: Question-driven methodology for analyzing emergency room processes using process mining. Applied Sciences **7**(3), 302 (2017)
15. Tillem, G., Erkin, Z., Lagendijk, R.L.: Privacy-preserving alpha algorithm for software analysis. In: SITB 2016 (2016)
16. Tillem, G., Erkin, Z., Lagendijk, R.L.: Mining sequential patterns from outsourced data via encryption switching. In: PST 2018. pp. 1–10. IEEE (2018)
17. Toshniwal, D.: Privacy preserving data mining techniques for hiding sensitive data: A step towards open data. In: Data Science Landscape, pp. 205–212. Springer (2018)