Optimizing Privacy-Utility Trade-Off in Healthcare Processes: Simulation, Anonymization, and Evaluation (Using Process Mining) of Event Logs

Omar Samy Kamal[a], Syeda Amna Sohail[b] and Faiza Allah Bukhsh[c]
University of Twente, 7500 AE, Enschede, The Netherlands

Keywords: Simulation, Synthetic Data, Anonymization, K-Anonymity, Event Logs, Process Mining, Privacy-Utility Trade-Off.

Abstract: In healthcare, big data analytics involve balancing patients' privacy and data utility. Optimizing healthcare data utility often includes limited access to sensitive data by trusted onsite entities. This potentially hinders broader-scale data utilization by third-party data analysts. As a solution, this research simulates a healthcare process-based event log, inspired by a local hospital's radiology department. The simulated event log is anonymized using k-anonymity. The anonymized and un-anonymized event logs are evaluated, through process discovery techniques, using the process mining tool, ProM 6.11, for Privacy-utility trade-off assessment. Results indicate successful privacy preservation with a distinct loss in utility in the anonymized healthcare process model, which was not visible otherwise. Therefore, to ensure the efficacy of healthcare process analysis on anonymized sensitive event logs, the utilization of process mining techniques is beneficial for process utility and privacy protection evaluation.

1 INTRODUCTION

In healthcare, big data analytics involve balancing between the privacy of patients' personally identifiable information (PII) and optimal data utility. Optimal data utility entails the maximum usefulness of healthcare datasets (Mivule, 2013). Whereas, privacy involves patients' direct or indirect control over their PII in addition to their explicit informed consent. To avoid potential privacy threats, healthcare providers limit the access of their healthcare data to onsite data analysts only. This hampers the large-scale usefulness of the healthcare datasets. Recognizing the sensitivity of using real data belonging to individuals and its privacy concerns, the utilization of synthetic data becomes imperative and hence contributes immensely to the scientific community at large.

This research work ensures the preservation of the privacy rights of data subjects and compliance with data privacy regulations as per GDPR (2018). Simulation and synthetic data generation are considered essential in accurately mimicking a real-world system with private patients' data. The generated synthetic model allows the generation of the most suitable event logs. Event logs are the datasets that depict an end-to-end process of all the activities/treatments of all the patients involved. The extracted event logs can then serve as a valuable tool to assist healthcare providers in making informed decisions that enhance the quality of healthcare processes. The discrete event simulation technique was chosen to generate this model mainly because this technique guarantees the accurate representation of healthcare processes and events (being carried out in the model). The simulated healthcare event logs are later anonymized to ensure privacy. Afterward, the data utility of both, the anonymized and unanonymized event logs, is evaluated using ProM for process mining. ProM (ProM) is a one-stop shop for the evaluation of end-to-end processes using process mining.

The research objective of this paper is to report a healthcare event log simulation, its anonymization, and evaluation using Process Mining (PM). The evaluation will include the assessment of the privacy-utility-tradeoff of healthcare process workflow. For the latter, the un-anonymized event log will be compared against the anonymized event log using the inductive visual miner plugin in ProM. This compar-

[a] https://orcid.org/0009-0004-8836-2520
[b] https://orcid.org/0000-0001-8078-0411
[c] https://orcid.org/0000-0001-5978-2754

Kamal, O., Sohail, S. and Bukhsh, F.
Optimizing Privacy-Utility Trade-Off in Healthcare Processes: Simulation, Anonymization, and Evaluation (Using Process Mining) of Event Logs.
DOI: 10.5220/0012766800003758

In *Proceedings of the 14th International Conference on Simulation and Modeling Methodologies, Technologies and Applications (SIMULTECH 2024)*, pages 289-296
ISBN: 978-989-758-708-5; ISSN: 2184-2841

ison will reveal whether privacy is preserved while maintaining optimal data utility? If not, the underlying reasons will be provided on why the utility might have been affected. This research objective yields the research question *"How was the utility of event logs affected by the anonymization process when comparing un-anonymized event logs with anonymized event logs using ProM?"*

The structure of the paper is as mentioned ahead: The section 2 will explore the relevant literature. The section 3 will depict the employed overarching methodology and employed tools in this research work. The section 4 will exhibit the artifact design that includes simulation modeling, data set generation, even log simulation, and its anonymization using K-anonymity. The section 5 assesses the effectiveness of the design through evaluation and discovers their contribution to fulfilling the research objectives by answering the research question. The paper will discuss limitations encountered during the study and propose suggestions for future work in section 6.

2 RELATED WORK

The **simulation model** is a reasonable approximation to a real system. The aim of the simulation modeling process or a system is to predict what impact the modifications may have on the system (Maria, 1997). Simulation is used to validate, modify, and experiment in ways that are usually expensive and unrealistic using a real system. This simulated model can then be used to analyze the model's behavior to make judgments/assumptions about the actual processes and system. Additionally, simulation is a critical tool for evaluating the performance of a process within a defined period while under numerous configurations. Simulation is also generally used to lower the chances of not meeting the required standards, optimize the utilization of resources to avoid shortage or wastage, prevent unexpected inefficiencies, and ensure maximum system efficiency by optimizing the performance of a system before applying any changes to the existing system or the new system being built (Maria, 1997). The simulation technique used in this paper is discrete event simulation, due to its focus on healthcare operations which correlates to the radiology department model being simulated in this paper.

Discrete Event Simulation is a modeling technique that is focused more on how individual events in a system behave, rather than at a collective level. This technique models the systems which involve discrete events, examples of such events include a queuing, manufacturing process, or patients' care pathways. Discrete event simulation gives freedom in testing different scenarios and later evaluates the effect of these different factors on the system's performance.

Process Mining (PM) is a collection of techniques that enables the discovery, analysis, and enhancement of business processes using event logs, which provide valuable insights (Van Dongen et al., 2005). PM is utilized by organizations to identify relationships, visualize workflows, and uncover hidden dynamics within a process (Leemans and Fahland, 2014), (Leemans, 2017) & (Schrijver, n.d.). PM exposes inefficiencies and bottlenecks in the process being analyzed. Moreover, these analysis methods help organizations identify limitations within their processes, leading to informed decisions that reduce costs, increase efficiency, and maximize overall performance.

Event logs are end-to-end process-based datasets that are typically extracted from an organization's information systems. In process mining, event logs represent a sequential record of events that occur within a business process (Nogueira, 2021). Each event in the log represents an action, interaction, or incident. The events are timestamped with their start and stop dates of execution and may contain other associated data. Examples of data that event logs capture include the agents involved in a process, resources utilized, and outcomes of a process (Van Dongen et al., 2005).

Noise Addition/Anonymization Large data collected by organizations such as the Population and Housing Census in the EU usually release statistical databases (Eurostat, n.d.). However, before this is done, sensitive information such as personal identifying information (PII) is removed. Although PII is removed, researchers agree that if these databases are combined with extra data, malicious actors could succeed in identifying an individual. *Data Privacy & Confidentiality* are the measures taken to protect individuals from any unauthorized sharing of information. *Data Security* ensures that this private data is only accessible to authorized parties. *Data de-identification* involves a process where PII is first removed from a dataset. Moreover, the *Data de-identification process* which is also interchangeably named as *data anonymization* can remove or modify PII attributes in such a way that it makes the retrieval of an individual identity difficult upon the data release to the public. Furthermore, *Data Utility versus privacy* also referred to as *Privacy-Utility Trade-Off* (PUT) is the balance between how useful a dataset is to the user and the crucial need to safeguard privacy. Usually, before datasets are published to the public, publishers follow procedures to remove PII and apply noise-adding techniques to distort the data. However, these measures may result in the original data to suffer from losing its statistical properties. Consequently, achieving PUT is always a desired goal for researchers (Sramka et al., 2010), (Rastogi et al., n.d.), and (Sankar et al., 2010). Unfortunately, researchers in data privacy agreed that achieving data privacy without reducing data utility is a challenging task (Mivule, 2013).

3 METHODOLOGY AND TOOLS

For a systematic methodological approach, the paper is based on the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology (Hotz, 2018). The CRISP-DM is a framework, used in data mining projects, made up of six phases. The six main phases of the CRISP-DM methodology are: *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation*, *Deployment* see (Hotz, 2018) for details. Some of the publicly available tools used in this research work include:

AnyLogic (AnyLogic, n.d.) is a simulation software tool that is used to create models to gain insights into the workflows of several systems, enabling optimization. The AnyLogic software offers an easy-to-use user interface, that allows the modeling of healthcare processes as well. The software tool offers different simulation techniques such as Discrete-Event simulation, Agent-Based Modeling, and System dynamics, which help discover complexities found within real-world systems.

ARX (ARX, n.d.), also known as Anonymization Toolbox, is a software tool used for incorporating privacy-preserving techniques. This tool helps tackle concerns regarding data privacy by implementing techniques that aim to protect sensitive information while still preserving data utility (Kadampur, 2010). By utilizing this tool datasets are anonymized before being shared with third parties, which results in compliance with privacy regulations. Anonymization techniques that the tool offers include k-anonymity, t-closeness, differential privacy, and l-diversity (ARX, n.d.).

ProM (ProM, n.d.), is a tool that allows analysis and refinement of a business process by using event logs (Nogueira, 2021), (Leemans and Fahland, 2014), and (Leemans, 2017). The tool assists organizations in gaining valuable insights from their data. Process mining techniques ProM supports include process discovery, conformance checking, and performance analysis. ProM can effectively present process models, showcase bottlenecks, and detect unexpected behavior. Additionally, ProM includes a library that has an extensive collection of plugins, enhancing the capability and functionality of the software, making it an invaluable tool for process mining research projects.

4 DESIGN

4.1 AnyLogic: Radiology Department

AnyLogic is used for the simulated Radiology Department Model, modified from a pre-existing example model named “Emergency Department” in AnyLogic (AnyLogic, n.d.).This simulation model recreates the workflow and process of radiology services offered to patients. The model is made up of entities that fall under different resource categories. These entities provide utility and execute actions within the model. The three resource categories are: moving, static, and portable. The moving resources are entities within the model that can move freely. Examples include Nurses, Physician Assistants (PA’s), and Technicians. Static resources refer to the entities that remain in fixed positions, they are represented by a location or a physical piece of medical equipment. Examples are Waiting Rooms, Triage Rooms, EC Rooms, and the X-RAY device. The last resource type, portable, are entities within the model that can be moved around, however, movement is not possible on its own. Thus, only personnel within the hospital are allowed to carry these resources. So, the resources are carried and moved by specific hospital staff, who then use them for a particular task. The MRI and Ultrasound devices are examples of portable resources.

The radiology department model is made of various agents (Entities & Resources)namely: 2 Triage Room, 2 ECG Rooms, 1 Waiting Room, 1 Medical Device Storage Room, 1 X-Ray Room, 5 PA’s, 5 Nurses, 3 Technicians, 2 MRI Devices, 2 Ultrasound Devices, and 1 X-Ray Device. The simulation model can be viewed in 3D, 2D, and Logic 1 views. The 3D format displays a three-dimensional environment of the model, along with other 3D icons for resources. The 2D format displays a two-dimensional view, and the logic format shows the flowchart with blocks for building the model.

The healthcare process starts with patients’ arrival as agents in the simulation model. After that, the patients register at the front desk, waiting at the queue to get registered. After registration, the Patients wait to be allocated to a nurse to get checked on in a waiting room. Afterwards, the nurse escorts the patient into a triage room where the severity of the patient’s condition is examined. Later the patient returns to the

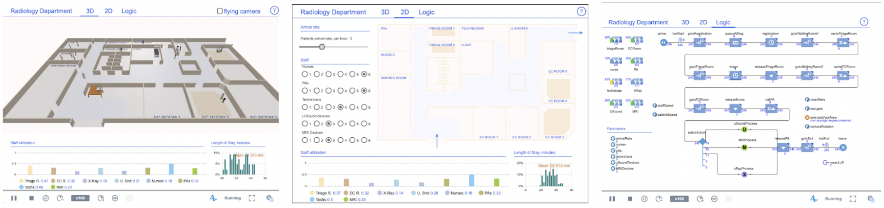


Figure 1: 3D, 2D, and Logic models of Radiology Department (Left to Right).

waiting room, and the nurse then decides whether the patient should be treated first. The patient then returns to the waiting room and when it's time for treatment a Physician Assistant (PA) and technician are called up to treat the patient. Then the patient is assigned a medical treatment process based on three different outputs namely: Ultra Sound , MRI , and X-Ray. After the process is completed, the patient is discharged and leaves the department.

4.2 Synthetic Data Generation

A synthetic dataset is generated and imported into the built simulation model. The website "Fake Name Generator" (Generator, 2006) is a platform for generating bulk fake data and allows the customization of many different attributes. The data is generated for around 500 patients with the following attributes: Given name, Surname, Gender, Birthday (m/d/yyyy), Telephone number, and Blood type. Afterwards, the data set is imported into the built-in AnyLogic database. This database mapped the attributes of the patients entering the radiology department. Lastly, the simulation model was run to generate required event logs.

4.3 Event Logs

Model execution recorded different event logs, capturing the behavior of diverse agents and resources within the radiology department, into the log folder inside the AnyLogic database. Since the focus is on testing the PUT using ProM, the emphasis is on evaluating event logs that offer information on patient IDs, timestamps of activities, resource pools, and processes patients undergo for valuable insight. The event logs providing these insights are the “agent parameter log” (APL) and “flowchart process states log” (FPSL). The APL stores parameter values of individual agents in the simulation model, and the FPSL records the timestamps individual agents spent in different states of the flowchart blocks.

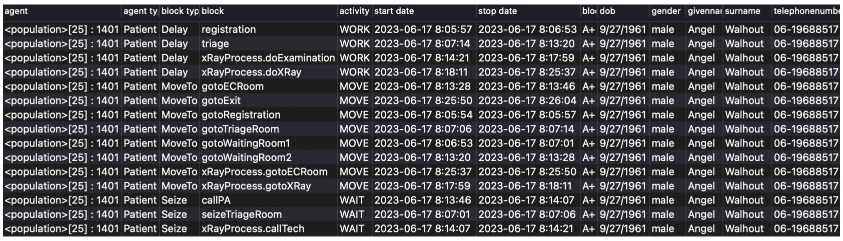
4.4 Anonymization Using ARX

After exporting the unanonymized event log from AnyLogic, it is imported into ARX for anonymization see 2. Before anonymization using ARX, attributes are labeled either under: Identifying, quasi-identifying, sensitive, or insensitive. The identifying attributes include given name, surname, and telephone number. Quasi-identifying attributes include blood type, date of birth, and gender. Sensitive attributes does not have any specific columns associated. Insensitive attributes include agent, agent type, block type, block, activity type, start date, and stop date.

The columns listed in the identifying attributes include personally identifiable information (PII). PII can directly identify a specific individual, thus. needs to be removed from the dataset to preserve the privacy of the patient. However, quasi-identifying attributes can indirectly be linked to individuals. The more quasi-identifier attributes are present in a dataset the easier it becomes for the risk of re-identification to increase. Thus, it is essential to safeguard these attributes during the anonymization process appropriately. Sensitive attributes are attributes that patients do not necessarily like being associated with, these may include symptoms, diagnosis, and health conditions. It is important to note that the dataset being anonymized does not include any columns of the sensitive attribute type. Finally, the last attribute type, insensitive, which includes most of the columns in figure 2, does not pose any privacy risks. Moreover, the inclusion or exclusion of these columns within the dataset does not affect the anonymization process.

4.5 Transforming Quasi-Identifiers

Once all columns are classified according to their attribute type, the anonymization process begins with a focus on transforming quasi-identifying attributes. Since quasi-identifiers when combined, can potentially lead to patient identification. For attribute transformation, a method known as generalization is used. This method aids in reducing the possibility of re-identification by broadening the data and making it



agent	agent ty	block typ	block	activity	start date	stop date	blo	dob	gender	givenna	surname	telephonenumbe
<population>[25] : 1401	Patient	Delay	registration	WORK	2023-06-17 8:05:57	2023-06-17 8:06:53	A+	9/27/1961	male	Angel	Walhout	06-19688517
<population>[25] : 1401	Patient	Delay	triage	WORK	2023-06-17 8:07:14	2023-06-17 8:13:20	A+	9/27/1961	male	Angel	Walhout	06-19688517
<population>[25] : 1401	Patient	Delay	xRayProcess.doExamination	WORK	2023-06-17 8:14:21	2023-06-17 8:17:59	A+	9/27/1961	male	Angel	Walhout	06-19688517
<population>[25] : 1401	Patient	Delay	xRayProcess.doXRay	WORK	2023-06-17 8:18:11	2023-06-17 8:25:37	A+	9/27/1961	male	Angel	Walhout	06-19688517
<population>[25] : 1401	Patient	MoveTo	gotoECRoom	MOVE	2023-06-17 8:13:28	2023-06-17 8:13:46	A+	9/27/1961	male	Angel	Walhout	06-19688517
<population>[25] : 1401	Patient	MoveTo	gotoExit	MOVE	2023-06-17 8:25:50	2023-06-17 8:26:04	A+	9/27/1961	male	Angel	Walhout	06-19688517
<population>[25] : 1401	Patient	MoveTo	gotoRegistration	MOVE	2023-06-17 8:05:54	2023-06-17 8:05:57	A+	9/27/1961	male	Angel	Walhout	06-19688517
<population>[25] : 1401	Patient	MoveTo	gotoTriageRoom	MOVE	2023-06-17 8:07:06	2023-06-17 8:07:14	A+	9/27/1961	male	Angel	Walhout	06-19688517
<population>[25] : 1401	Patient	MoveTo	gotoWaitingRoom1	MOVE	2023-06-17 8:06:53	2023-06-17 8:07:01	A+	9/27/1961	male	Angel	Walhout	06-19688517
<population>[25] : 1401	Patient	MoveTo	gotoWaitingRoom2	MOVE	2023-06-17 8:13:20	2023-06-17 8:13:28	A+	9/27/1961	male	Angel	Walhout	06-19688517
<population>[25] : 1401	Patient	MoveTo	xRayProcess.gotoECRoom	MOVE	2023-06-17 8:25:37	2023-06-17 8:25:50	A+	9/27/1961	male	Angel	Walhout	06-19688517
<population>[25] : 1401	Patient	MoveTo	xRayProcess.gotoXRay	MOVE	2023-06-17 8:17:59	2023-06-17 8:18:11	A+	9/27/1961	male	Angel	Walhout	06-19688517
<population>[25] : 1401	Patient	Seize	callPA	WAIT	2023-06-17 8:13:46	2023-06-17 8:14:07	A+	9/27/1961	male	Angel	Walhout	06-19688517
<population>[25] : 1401	Patient	Seize	seizeTriageRoom	WAIT	2023-06-17 8:07:01	2023-06-17 8:07:06	A+	9/27/1961	male	Angel	Walhout	06-19688517
<population>[25] : 1401	Patient	Seize	xRayProcess.callTech	WAIT	2023-06-17 8:14:07	2023-06-17 8:14:21	A+	9/27/1961	male	Angel	Walhout	06-19688517

Figure 2: Final version of imported event log.

less precise to a specific individual (Marques and Bernardino, 2020). The first quasi-identifying attribute, blood type, is transformed using the generalization method known as masking by adding an asterisk for each level, making it more privacy-preserving. As shown in figure 3, there are three different levels, level-0 is the raw data from the data set, level-1 is the masking of either the plus or minus sign, and level-2 is a fully suppressed data. The second quasi-identifying attribute being transformed is the date of birth. The first level, level 0, shows the raw data from the dataset. The second level, level 1, generalizes the date more by only using the birth year. The last level, level 2, shows the patient's decade intervals. Finally, the third quasi-identifying attribute , gender, has only two levels one being the raw data and the other includes a universal bracket with both genders.

4.6 Anonymization Using K-Anonymity

Following the transformation of quasi-identifiers using generalization, a privacy model in ARX also known as an anonymization technique is applied. In ARX various anonymization techniques are available, however, for this paper, the k-anonymity technique is used. K-anonymity ensures that each group of quasi-identifiers within a dataset is identical to at least k-1 other patients (Prasser et al., 2016). This means that the groups of attributes that could identify a patient are combined with other groups making it difficult for anyone to separate patients in and follow each other and re-identify them. This technique provides a higher level of privacy protection within a dataset and helps maintain confidentiality for sensitive information. An example of K-anonymity is shown in figure 4.

4.7 Anonymized Event Log Attributes

Finally, after the use of the k-anonymity, ARX generates an anonymized dataset. ARX generates this anonymized dataset by taking into consideration specific weight settings chosen for each quasi-identifying attribute to prioritize their importance. These weight settings ensure that information loss is minimized while preserving privacy. A subset of the dataset containing the anonymized attributes is shown in figure 5.

5 EVALUATION

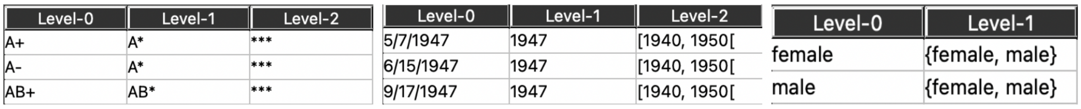
5.1 Utility and Re-Identification Risk Analysis Using ARX

ARX includes a feature that allows the analysis of the datasets' utility post-anonymization, this enables the evaluation of the success of the anonymization process. The insights provided concern the granularity and precision (generalization intensity) percentages of quasi-identifiers, which help assess the level of detail and accuracy preserved in the anonymized event log. ARX also offers a risk analysis feature that evaluates the potential re-identification risk associated with the event log before and after anonymization. As shown in table 1, the re-identification risk for the unanonymized log was 24.347% but after anonymization, this number decreased to 0.232%. This indicates the successful anonymization of the dataset.

5.2 Utility Analysis Using ProM

To import the anonymized and anonymized files into ProM, first, they had to be converted from CSV format into the eXtensible Event Stream (XES) file format, which is the format compatible with ProM. To successfully achieve this conversion, the "Convert CSV to XES" plugin by F. Mannhardt was used (Mannhardt et al., n.d.). This process involved mapping the columns of the event log to the case of the relevant field, event, start time, and completion time (Nogueira, 2021). The column "agent" is mapped to "case column", "block" to "event column", "start time" to "start date", and "completion time" to "stop date". Following the completion of this conversion, the event log is then suitable for analysis.

For measuring the utility difference between the unanonymized and anonymized event logs in ProM, a comparison was made using the process discov-



Level-0	Level-1	Level-2
A+	A*	***
A-	A*	***
AB+	AB*	***

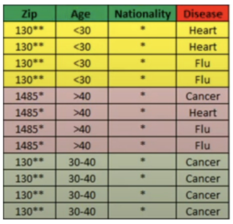
Level-0	Level-1	Level-2
5/7/1947	1947	[1940, 1950[
6/15/1947	1947	[1940, 1950[
9/17/1947	1947	[1940, 1950[

Level-0	Level-1
female	{female, male}
male	{female, male}

Figure 3: Levels of generalization transformations for blood type, dob, and gender attributes (left to right).

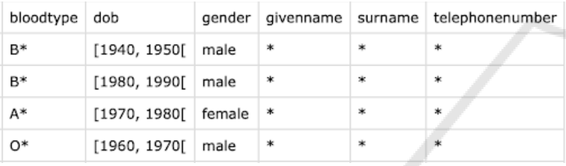
Table 1: Re-identification risk of un-anonymized log vs anonymized log.

	Unanonymized Log	*Anonymized Log*
Re-identification Risk (%)	**24.347%**	**0.232%**



Zip	Age	Nationality	Disease
130**	<30	*	Heart
130**	<30	*	Heart
130**	<30	*	Flu
130**	<30	*	Flu
1485*	>40	*	Cancer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer

Figure 4: Example of a k-anonymity dataset (k=4).



bloodtype	dob	gender	givenname	surname	telephonenumber
B*	[1940, 1950[	male	*	*	*
B*	[1980, 1990[	male	*	*	*
A*	[1970, 1980[	female	*	*	*
O*	[1960, 1970[	male	*	*	*

Figure 5: Snapshot of a subset of the anonymized event log.

ery technique namely, “inductive visual miner plugin” (Leemans and Fahland, 2014)(Leemans, 2017). This plugin extracts behavioral patterns and process flow from the events to then provide a visual animation of the process model (Schrijver, n.d.). This animated process model then enables understanding the relationship between events and gaining insights into the whole process.

Figure 6 displays both the unanonymized and anonymized event logs of the two different paths being compared using the IVM plugin in ProM 6.11. The paths in the unanonymized and anonymized logs exhibit distinct differences, indicating a potential impact on utility. In the left figure, there are fewer occurrences of paths, displaying only true paths without redundancy. However, the right figure shows a significantly higher number of path occurrences. This result suggests that the utility of the observed anonymized event log is distinctively affected by anonymization. Additionally, the observed differences in the paths demonstrate a potential trade-off between privacy and utility for the event logs, particularly in terms of privacy preservation for patients and the level of data granularity.

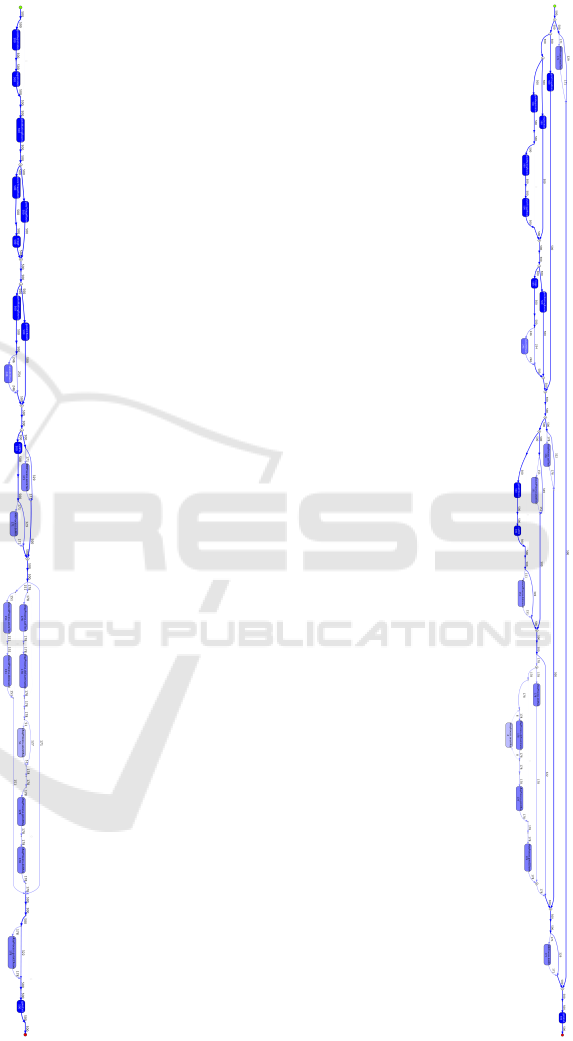


Figure 6: Process model of un-anonymized (left) and anonymized (right) event logs.

5.3 Discussion

The k-anonymity technique is crucial for preserving privacy, it also presents challenges that impact the data utility. Three reasons that might lead to this utility loss are information loss, indirect effects, and randomness.

The first reason for information loss is due to the reduction in information. Since attributes are either generalized or suppressed to achieve anonymity and avoid re-identification, the level of detail within the data decreases. This loss of precise detail in the event logs makes it harder when conduct a detailed analysis of the process model. Furthermore, the rich insights that could be previously extracted from the original data now become restricted, leading to a lack of in-depth information. The second critical reason for experiencing this utility loss is that k-anonymity may have indirect effects that result in additional changes to the data. Although the focus may be on preserving privacy for certain columns, the column related to activities in the event log may be accidentally modified during anonymization. This would then lead to consequences for the data such as reduced reliability. The third and final factor is the introduction of randomness to the process model due to anonymization. Since the k-anonymity technique groups different patients into equivalence classes, this can affect the ordering of events or activities and may introduce deviations to the process model. Therefore, randomness spreads in the event log, resulting in a less accurate and consistent representation compared to the error-less process model of the unanonymized event log. Furthermore, one notable thing that was observed is that process mining was able to reveal hidden insights from the simulation model that were not apparent at first. During the examination of the unanonymized event log, differences in the process flow were noticed, which were not configured in the model settings. This evidence proves the significance of process mining in uncovering valuable insights that would usually remain unnoticed. In a nutshell, with process mining, healthcare organizations can gain a deeper understanding of their processes, identify bottlenecks, and make informed decisions resulting in the optimization of their workflow performance, and can extend the evaluation of privacy-utility trade-off as well.

5.4 Limitations Concerning Process Mining

The limitations included the limited available literature and plugins on process mining applications for privacy-utility trade-off evaluation. Another sub-area limitation was the non-working condition of noise-adding plugins within the ProM toolkit. The noise-adding plugins would have allowed anonymization using ProM itself and would have prevented the addition of another tool usage outside the domain of ProM and process Mining. Another limitation is the usage of compliance-checking plugins within ProM and the calculation of precision, fitness, and generalization. We intentionally avoided the aforementioned quality assurance parameters to keep the focus on the healthcare process workflow comparisons of both the event logs and not an event log and its comparison to the discovered process model.

6 CONCLUSION & FUTURE WORK

6.1 Conclusion

Initially, this research work included the simulation of a model that reflected well the real-life hospital settings of a local radiology department with a wide range of customizable parameters using AnyLogic. The model allowed us to simulate a synthetic dataset, comprising 500 patients with varied attributes, respective agents, and resources, to be imported into the simulation model for event log creation. The simulation run created several event logs. However, the most suitable event logs were selected that could provide the most useful insight concerning healthcare process workflow using ProM for process mining (PM). The event logs are then anonymized using K-anonymity with the ARX tool. ARX also performed the re-identification risk and utility assessments of anonymized event logs to assess privacy and utility preservation respectively. ARX showed 24% more privacy and 40% less utility in the anonymized event logs. The decrease in utility resulted because of the shuffling in patients' physical attributes, which was required to apply the k-anonymity. We wanted to dig deeper to evaluate the privacy-utility-tradeoff for the events-based process workflows (namely patients' clinical pathways), their relationship to one another, and the whole process. To identify the latter, the event logs are evaluated using ProM 6.11. Process Mining (PM) offered valuable insights into process workflows. The utility difference between the unanonymized and anonymized event logs is evaluated using the process discovery technique namely the inductive visual miner plugin in ProM 6.11. This plugin extracts the process workflow from the events (i.e., treatments) and provides a visual animation of the process model. The animated process models enabled an understanding of the relationship between events and gaining insights into the whole process by replaying the log. Additionally, the PM highlighted privacy preservation in the anonymized event log but a distinct utility loss in the process model and hence potentially reduced the utility of the anonymized event log. Therefore, ensuring the efficacy of process analysis on anonymized sensitive event logs is imperative for progress in both process utility analysis and privacy protection through the utilization of process mining.

6.2 Future Work

The future work can include refining the simulation model by adding patients' symptoms to the selected output block for the three imaging processes to avoid patients being randomly assigned using probabilities. Additionally, other synthetic data attributes (such as symptoms, BSN, occupation, etc) inclusion will improve the quality of the dataset and enrich its information. Furthermore, other anonymization techniques such as l-diversity and t-closeness, can be built upon k-anonymity after the inclusion of sensitive attributes as they are a prerequisite for the use of other anonymization techniques. Lastly, the use of newly available plugins for process discovery and performance analysis would be beneficial in evaluating the utility of event logs.

REFERENCES

Hotz, N. (2018). What is CRISP DM? Data Science Process Alliance. https://www.datascience-pm.com/crisp-dm-2/.

Maria, A. (1997). Introduction to modeling and simulation. Proceedings of the 29th Conference on Winter Simulation - WSC '97, 7–13. https://doi.org/10.1145/268437.268440

Van Dongen, B. F., De Medeiros, A. K. A., Verbeek, H. M. W., Weijters, A. J. M. M., & Van Der Aalst, W. M. P. (2005). The ProM Framework: A New Era in Process Mining Tool Support. In G. Ciardo & P. Darondeau (Eds.), Applications and Theory of Petri Nets 2005 (Vol. 3536, pp. 444–454). Springer Berlin Heidelberg. https://doi.org/10.1007/11494744_25

Population and housing censuses—Population and demography—Eurostat. (n.d.). Retrieved December 8, 2023, from https://ec.europa.eu/eurostat/web/population-demography/population-housing-censuses

Mivule, K. (n.d.). Utilizing Noise Addition for Data Privacy, an Overview.

Prasser, F., Kohlmayer, F., & Kuhn, K. (2016). The Importance of Context: Risk-based De-identification of Biomedical Data. Methods of Information in Medicine, 55(04), 347–355. https://doi.org/10.3414/ME16-01-0012

Mannhardt, F., Tax, N., Schunselaar, D., & Verbeek, E. (n.d.). Den Dolech 2, 5612 AZ Eindhoven P.O. Box 513, 5600 MB Eindhoven The Netherlands www.tue.nl.

Nogueira, F. (2021). Hands-on Process Mining: Event Visualisation with ProM. Laredoute.Io. https://laredoute.io/blog/hands-on-process-mining-event-visualisation-with-prom/

Leemans, S. J. J., & Fahland, D. (2014). Process and Deviation Exploration with Inductive visual Miner.

Leemans, S. (2017). Inductive visual Miner manual. https://www.semanticscholar.org/paper/Inductive-visual-Miner-manual-Leemans-Prom/c09fd03d82bea9d2df50682ed4c220df21648005

Schrijver, G. (n.d.). Using process mining to compare different variants of the same reimbursement process: A case study.

Radiologie. (n.d.). Medisch Spectrum Twente. Retrieved July 21, 2023, from https://www.mst.nl/p/specialismen/radiologie/

Sramka, M., Safavi-Naini, R., Denzinger, J., & Askari, M. (2010). A practice-oriented framework for measuring privacy and utility in data sanitization systems. Proceedings of the 2010 EDBT/ICDT Workshops, 1–10. https://doi.org/10.1145/1754239.1754270

Rastogi, V., Suciu, D., & Hong, S. (n.d.). The Boundary Between Privacy and Utility in Data Publishing.

Sankar, L., Rajagopalan, S. R., & Poor, H. V. (2010). Utility and privacy of data sources: Can Shannon help conceal and reveal information? 2010 Information Theory and Applications Workshop (ITA), 1–7. https://doi.org/10.1109/ITA.2010.5454092

AnyLogic: Simulation Modeling Software Tools & Solutions for Business. (n.d.). Retrieved July 21, 2023, from https://www.anylogic.com/

Kadampur, M. A. (2010). A Noise Addition Scheme in Decision Tree for Privacy Preserving Data Mining. 2(1).

ARX - Data Anonymization Tool — A comprehensive software for privacy-preserving microdata publishing. (n.d.). Retrieved July 21, 2023, from https://arx.deidentifier.org/

ProM Tools – The Process Mining Framework. (n.d.). Retrieved July 21, 2023, from https://promtools.org/

Order Free Random Names—Fake Name Generator. (n.d.). Retrieved July 21, 2023, from https://www.fakenamegenerator.com/order.php

Marques, J., & Bernardino, J. (2020). Analysis of Data Anonymization Techniques: Proceedings of the 12th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, 235–241. https://doi.org/10.5220/0010142302350241