



# UNIVERSITY OF TWENTE.

Faculty of Electrical Engineering,  
Mathematics & Computer Science



## Tracing Data Utility for Noise Adding Plugins in Process Mining

Omar Ahmed Samy  
B.Sc. Thesis  
July 2023

---

**Supervisor:**  
Dr. Faiza Bukhsh  
**Critical Observer:**  
Syeda Sohail

Data Management & Biometrics Group  
Faculty of Electrical Engineering,  
Mathematics and Computer Science  
University of Twente  
P.O. Box 217  
7522 NB Enschede  
The Netherlands

---

# Abstract

This bachelor thesis project explores the implementation of simulation, anonymization techniques, and process mining tools in the healthcare sector for event log analysis. The main objective is to gain insights into the Privacy-Utility Trade-off by evaluating the impact of anonymization on the utility of simulated event logs. The simulation part utilizes discrete event simulation to model a radiology department using AnyLogic. Afterwards an event log with synthetic patient data is extracted from the simulated model which will then goes through an anonymization process. For anonymization, the k-anonymity technique in the ARX tool is applied to the event logs, ensuring the protection of private patient data while balancing data utility. Process mining, specifically process discovery using ProM, is then used for further analysis to compare the unanonymized and anonymized event logs. This analysis discover insights and assesses whether data utility was impacted by anonymization. The results in ProM reveal a loss of utility in the anonymized log, due to the presence of unnecessary paths. This finding highlights the importance of using process mining for analysis of sensitive healthcare data and proves that process mining is an effective approach in understanding event log utility and uncovering hidden insights. Insights gained from process mining contribute to more informed decisions and overall improved patient care.

**Keywords:** *Discrete Event Simulation, Synthetic Data Generation, Anonymization, K-anonymity, Event Logs, Process Mining, Privacy-Utility Trade-off.*

# Acknowledgements

I would like to express my sincere gratitude to my supervisor Dr. Faiza Bukhsh for her guidance, support, and insightful feedback throughout the project. Her expertise and encouragement have been helpful in delivering quality work, and i am very thankful for her mentorship.

I am also deeply grateful to my critical observer Syeda Sohail for her constructive feedback, and thoughtful suggestions. Her insights and attention to detail played an important role in refining my project work. Additionally, I am thankful for her in helping me find relevant research papers and for her assistance in operating software tools.

I would also like to thank my family members for their constant encouragement during this journey. Their support and belief in my abilities have been a constant source of motivation. Lastly, I would like to thank all my friends and colleagues who have provided support, encouragement, and assistance. Your contributions have been very much appreciated and have made a meaningful impact on the completion of this project.

# List of Tables

2.1 Industries of Simulation and Applications. . . . .	8
2.2 Continued-1 Industries of Simulation and Applications. . . . .	8
2.3 Continued-2 Industries of Simulation and Applications. . . . .	9
4.1 Re-identification Risk of Unanonymized Log vs Anonymized Log. . . . .	37

# List of Figures

1.1 The Six Phases of the CRISP-DM Methodology [1]. . . . .	5
2.1 Standardized Data Privacy Procedure. . . . .	13
4.1 3D Model of Radiology Department . . . . .	20
4.2 2D Model of Radiology Department . . . . .	20
4.3 Logic Model of Radiology Department. . . . .	21
4.4 Name set, Country, Gender, and Age of Synthetic Dataset. . . . .	23
4.5 Final Synthetic Attributes Included into the Dataset. . . . .	25
4.6 First 10 Rows of the Generated Synthetic Data. . . . .	26
4.7 Imported Dataset into the AnyLogic Database. . . . .	26
4.8 The Agent Parameters Log. . . . .	27
4.9 The Flowchart Process States Log. . . . .	27
4.10 Merged Event Logs. . . . .	29
4.11 Data Type and Format of Attributes in ARX. . . . .	29
4.12 Final Version of Imported Event Log. . . . .	30
4.13 Plugin Used for Conversion. . . . .	31
4.14 Mapping Event Log Columns to XES Attributes. . . . .	31
4.15 Three levels of Generalization Transformation for Blood Type Attribute. . . . .	34
4.16 Three levels of Generalization Transformation for dob Attribute. . . . .	34
4.17 Two levels of Generalization Transformation for Gender Attribute. . . . .	35
4.18 Example of a K-anonymity Dataset (K=4). . . . .	36
4.19 Snapshot of a Subset of the Anonymized Event Log Dataset. . . . .	36
4.20 Separate Granularity & Precision Percentages of Quasi-identifying Attributes. . . . .	37
4.21 Combined Granularity & Precision Percentages of Quasi-identifying Attributes. . . . .	37
4.22 Process Model of Unanonymized Event Log. . . . .	39
4.23 Process Model of Anonymized Event Log. . . . .	40

C.1 Logic format of USound Process.	53
C.2 Logic format of MRI Process.	54
C.3 Logic format of X-Ray Process.	54
C.4 Columns overview of Anonymized Event Log.	55
C.5 Remaining 26 Patients with full Suppression for all Attributes.	55
C.6 Fitness Score of Unanonymized Event Log.	56
C.7 Fitness Score of Anonymized Event Log.	56
C.8 Event Log Dashboard in ProM.	56
C.9 Log Summary in ProM.	57
C.10 Start & End Events of Log Summary in ProM.	57
C.11 Waiting Times of Unanonymized Event Log.	58
C.12 Waiting Times of Anonymized Event Log.	59
C.13 Service Times of Unanonymized Event Log.	60
C.14 Service Times of Anonymized Event Log.	61
C.15 Petri Net of Unanonymized Event Log.	62
C.16 Petri Net of Anonymized Event Log.	63

# Acronyms

<b>ABS</b>	Agent-Based Simulation
<b>CSV</b>	Comma-Separated Values
<b>DES</b>	Discrete Event Simulation
<b>DOB</b>	Date of Birth
<b>EDPB</b>	European Data Protection Board
<b>GDPR</b>	General Data Protection Regulation
<b>ISO</b>	International Organization for Standardization
<b>MCS</b>	Monte Carlo Simulation
<b>MRI</b>	Magnetic Resonance Imaging
<b>MST</b>	Medisch Spectrum Twente
<b>NEN</b>	Netherlands Standardization Institute
<b>PA</b>	Physician Assistant
<b>PII</b>	Personally Identifiable Information
<b>PLE</b>	Personal Learning Edition
<b>PUT</b>	Privacy-Utility Trade-Off
<b>SDC</b>	Statistical Disclosure Control
<b>SD</b>	System Dynamics
<b>XES</b>	eXtensible Event Stream

**ZGT**      Ziekenhuisgroep Twente



# Contents

<b>Abstract</b>	ii
<b>Acknowledgements</b>	iii
<b>Acronyms</b>	vii
<b>1 Introduction</b>	1
1.1 Research questions	2
1.2 Methodology	4
<b>2 Background research</b>	7
2.1 Modeling	7
2.2 Simulation	7
2.2.1 Contexts of Simulation Industries	8
2.2.2 Simulation Techniques and their Applications in Healthcare	9
2.2.3 Discrete Event simulation software tools	11
2.3 Process Mining	12
2.4 Event Logs	12
2.5 Anonymization/Noise Addition	13
2.6 Software Tools Used	14
2.6.1 AnyLogic	14
2.6.2 ARX	15
2.6.3 ProM	15
<b>3 Stakeholder Identification &amp; Analysis</b>	16
<b>4 Approach and Implementation</b>	17
4.1 Radiology Department Model	17

4.2 Synthetic Data Generation . . . . .	22
4.3 Data Pre-Processing . . . . .	24
4.3.1 Synthetic Data . . . . .	24
4.3.2 Event Logs . . . . .	26
4.3.3 ARX . . . . .	29
4.3.4 ProM . . . . .	30
4.4 Anonymization using ARX . . . . .	32
4.4.1 Transforming quasi-identifiers . . . . .	33
4.4.2 Anonymization using K-anonymity . . . . .	35
4.4.3 Anonymized event log attributes . . . . .	35
4.5 Utility analysis in ARX . . . . .	37
4.6 Risk analysis in ARX . . . . .	37
4.7 Utility analysis using ProM . . . . .	38
<b>5 Discussion and Limitations</b> . . . . .	<b>41</b>
5.1 Discussion . . . . .	41
5.2 Limitations . . . . .	42
<b>6 Conclusion and Future work</b> . . . . .	<b>43</b>
6.1 Conclusion . . . . .	43
6.2 Future Work . . . . .	44
<b>References</b> . . . . .	<b>45</b>
<b>Appendices</b>	
<b>A Patient Dataset and Event Logs Used</b> . . . . .	<b>51</b>
<b>B Inductive Visual Miner Animation</b> . . . . .	<b>52</b>
<b>C Additional Figures</b> . . . . .	<b>53</b>
<b>D Python Code</b> . . . . .	<b>64</b>
<b>E SQL Query</b> . . . . .	<b>65</b>
I . . . . .	66

# Introduction

Simulation and synthetic data generation are regarded as essential steps in developing a model that accurately mimics a real-world system with private patient data, specifically in the healthcare sector. The use of real sensitive data which belongs to actual individuals is considered a breach in data privacy. Therefore, synthetic data must be generated instead to respect the privacy rights of these individuals. A model can then be simulated which would contain the synthetic patient data and later useful event logs can be exported from the simulated model. This model is also considered to be beneficial in assisting healthcare providers in making thoughtful decisions that will enhance the quality of a patient's care. To generate this simulation model various techniques existing must first be researched before deciding on one, as the most appropriate technique would depend on specific domain and context where it is used. The goal of this paper is to explore different simulation techniques, anonymization techniques, and process mining tools to help gain insights on the event logs from the simulated model. The first part of this paper identifies the different techniques available and focuses on finding the most suitable one for use in the healthcare sector. The second part will focus on the researching and implementing a suitable anonymization technique that would then be used on an event log to prevent the re-identification of sensitive patient data. Finally, the third part will focus on implementing process mining techniques using ProM tool on the event log (anonymized & unanonymized) to analyze the Privacy-Utility Trade-off.

## 1.1 Research questions

1. *How can end-to-end process-based healthcare event log simulation be used to improve performance in healthcare system operations?*
  - (a) *What are the different data simulation techniques that can be used in healthcare?*
  - (b) *What are the different software tools for discrete event simulation that can be used for healthcare?*
2. *What are the various anonymization techniques available for preserving privacy in sensitive data sets, and does the use of the K-anonymity technique using the ARX tool preserve patient privacy?*
  - (a) *What are the different anonymization techniques utilized for preserving privacy in sensitive data sets?*
  - (b) *How does the K-anonymity technique, implemented through the ARX tool, contribute to the preservation of patient privacy?*
3. *What preprocessing techniques can be employed on the synthetic dataset and event logs before importing them into the software tools utilized?*
4. *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 four main Research Questions of this report are mentioned above. Research questions one and two each have two sub-questions, as opposed to the third and fourth research questions. Each of these sub-questions break down the main questions to then help form an answer for them.

The first main research question this report will aim to answer is “How can end-to-end process-based healthcare event log simulation be used to improve performance in healthcare system operations?” To answer this research question, an analysis of the various data simulation techniques applied in healthcare will be done. Examples include the Monte Carlo Simulation, Discrete Event Simulation, Agent-based simulation, and system dynamics simulation. Moreover, different commercial and open-source discrete event simulation software tools in healthcare will be explored. The sub-questions of the first research question are:

- (a) *What are the different data simulation techniques that can be used in healthcare?*
- (b) *What are the different software tools for discrete event simulation that can be used for healthcare?*

The second main research questions will answer the question “What are the various anonymization techniques available for preserving privacy in sensitive data sets, and does the use of the K-anonymity technique using the ARX tool preserve patient privacy”. To answer this question various anonymization techniques used to protect individuals sensitive information will be researched. Furthermore, the ARX tool ,for data anonymization, will be used to measure the level of preserved patient privacy in event logs using the technique know as K-anonymity. The sub-questions of the second research question are:

- (a) *What are the different anonymization techniques utilized for preserving privacy in sensitive data sets?*
- (b) *How does the K-anonymity technique, implemented through the ARX tool, contribute to the preservation of patient privacy?*

The third main research question which is “What pre-processing techniques can be employed on the synthetic dataset and event logs before importing them into the software tools utilized?” will go over the various pre-processing techniques used in order to clean up datasets and event logs prior to their use in the three software tools used in this paper.

Doing so helps ensure there is sufficient data quality and the exported datasets/event logs are compatible with software tools where they will be imported. The third main research question is as follows:

- *What pre-processing techniques can be employed on the synthetic dataset and event logs before importing them into the software tools utilized?*

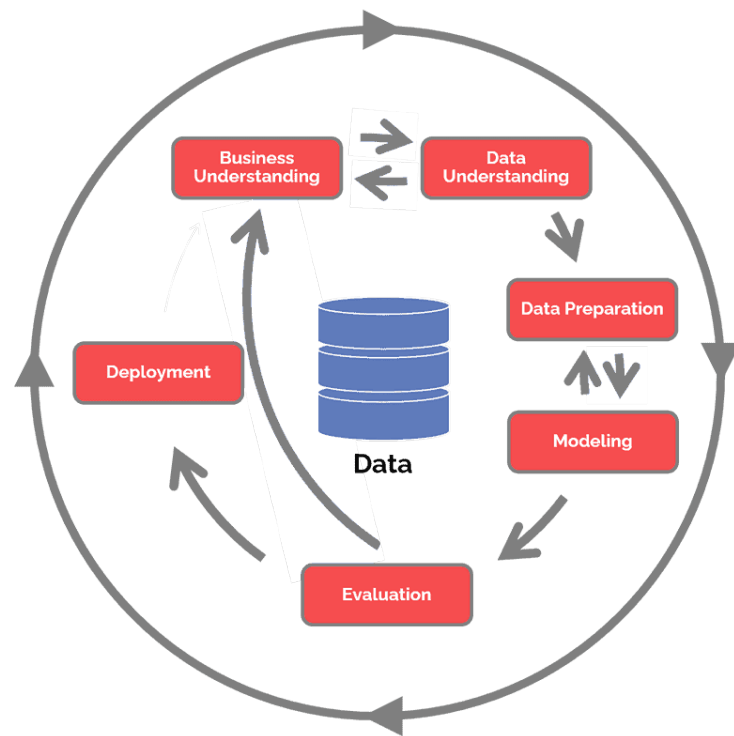
The fourth main research question aims to answer, “How was the utility of event logs affected by the anonymization process when comparing un-anonymized event logs with anonymized event logs using ProM?” In order to answer this research question, the un-anonymized event log will be compared to the anonymized event log using the inductive visual miner in ProM. This comparison will aid in finding out whether Privacy-utility-tradeoff can be achieved between the two event logs, and if not assumptions will be made on why the utility might have been affected. The fourth main research question is:

- *How was the utility of event logs affected by the anonymization process when comparing un-anonymized event logs with anonymized event logs using ProM?*

## **1.2 Methodology**

The methodology followed in this paper is based on the **CRISP-DM** (Cross-Industry Standard Process for Data Mining) methodology. The CRISP-DM is a framework used in data mining projects that is made up of six phases. By following these six phases, the project will follow a structured approach from start to finish. The six main phases of the CRISP-DM methodology are:

1. ***Business Understanding***
2. ***Data Understanding***
3. ***Data Preparation***
4. ***Modeling***
5. ***Evaluation***
6. ***Deployment***



**Figure 1.1:** The Six Phases of the CRISP-DM Methodology [1].

## 1- Business Understanding

The Business Understanding phase focuses on understanding the context and objectives of the project. During this phase collaboration with stakeholders is essential to help identify their expectations and requirements. While this project may not be directly linked to specific stakeholders, it may seem less relevant however it is still considered important. Therefore, inspiration is drawn from local healthcare departments in Enschede, which can be considered indirect stakeholders. By including these healthcare departments, the project aims to fulfill the needs that the department might have in terms of optimizing operational processes.

## 2- Data Understanding

During the Data Understanding phase the goal is to gain in-depth understanding of the data being worked with. This involves exploring data, collecting data, explaining it, and verifying its relevance and quality. Furthermore, any limitations or challenges faced can also be documented in this phase.

### **3- Data Preparation**

The Data Preparation phase focuses on modifying the original unprocessed data into a format that is appropriate for import and analysis. This may include selecting data to be used, data cleaning, combining data from different sources, and making conversions. Moreover, ensuring the quality and consistency of the data structure is important to meet the specifications required for modeling and analysis.

### **4- Modeling**

In the Modeling phase, various simulation modeling and process mining techniques are examined and the chosen technique is then used to build the model and extract insights from it. Thus this phase involves various steps depending on what is being modelled, such as selecting a simulation model or technique, developing or analyzing it, and for the model linking with the generated dataset. Throughout the process of the simulation model, the models workflow process and parameters is continuously refined to help optimize the models performance and accuracy.

### **5- Evaluation**

During the Evaluation phase, results of the developed models are evaluated based on their effectiveness. The Specific metrics used to measure this aids during this evaluation. Additionally, the results and conclusion regarding insights are reported and future steps are also discussed.

### **6- Deployment**

The Deployment phase, reviews the project where a final report is made summarizing all project findings and recommendations. In this last phase the outcomes are presented to the stakeholders in a transparent and comprehensible manner. Furthermore, if possible the developed models become integrated into existing systems and these systems can also be monitored for further improvements. However since this project do not have any direct stakeholders the deployment phase will be skipped as it is less relevant.



# Background research

## 2.1 Modeling

A model entails the creation of a simple version of a process or system that is less complex than the actual real-life system. The aim of Modeling process or a system is to predict what impact the modifications made have on the system. The model must be a reasonable approximation to a real system, however not so complex that it becomes impractical for use. Experts in the field of simulation recommend gradually adding more complexity to the model over time. This helps have a better representation of the actual processes or system and the interactions between its elements. Moreover, by gradually adding detail, the model becomes more realistic and errorless without being too difficult to understand. Model validity is a critical concern in modeling, and techniques to validate it include comparing the model output to the actual process or system output. Typically, simulation involves the use of mathematical models which are created with the help of a simulation software tool. These mathematical models can either be stochastic (a probabilistic value for at least one of the input or output variables) or deterministic (where input & output variables have fixed values) and dynamic (time-changing interactions taken into account) or static (time isn't taken into account). In simulation models are usually dynamic and stochastic [2].

## 2.2 Simulation

Simulation is regarded as a powerful tool when used properly. When simulating a process or system it involves running a model. 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 in order to make judgments and assumptions about the actual processes and system. Additionally, simulation is an critical tool for evaluating the performance of process within a defined time 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 [2].

### 2.2.1 Contexts of Simulation Industries

Simulation has a wide range of applications in various industries. Simulation can be applied to multiple fields such as engineering, government, healthcare, business, ecology, and much more. Tables 2.1, 2.2, and 2.3 below show some of the diverse industries where simulation has been used along with the respective contexts for each industry [3] [4].

Industry	<i>Manufacturing</i>	<i>Environmental</i>
<b>Contexts</b>	Plant layout	Water purification/pollution
	Machine design	Waste control
	Material handling systems	Air pollution
	Assembly lines	Pest control
	Production facilities	storm/earthquake analysis
	Automated storage facilities	Crop production

**Table 2.1:** Industries of Simulation and Applications.

<i>Government</i>	<i>Healthcare</i>	<i>Social/Behavioral</i>
Military tactics	Healthcare interventions	Educational policies
Roadway design	Health policy	University administration
Traffic control	Disease transmission and control	Organizational structure
Police services	Patient flow (admission/scheduling)	Welfare systems
Population forecasting	Epidemics	Social systems
Sanitation services		

**Table 2.2:** Continued-1 Industries of Simulation and Applications.

<i>Computer Systems</i>	<i>Business</i>	<i>Biosciences</i>
Database structure/management	Stock analysis	Biomedical studies
Reliability of soft/hard-ware	Marketing strategies	Sports performance
Hardware components	Cash flow analysis	Disease control
Software systems	Pricing policy	Biological life cycles
Information processing		

**Table 2.3:** Continued-2 Industries of Simulation and Applications.

## 2.2.2 Simulation Techniques and their Applications in Healthcare

Since this project specifically focuses on simulation within the healthcare industry, techniques relevant to only this industry were researched. The techniques found to be widely used in healthcare include Monte Carlo simulation (MCS), Discrete Event Simulation (DES), Agent-Based Simulation (ABS), System Dynamics (SD), and Hybrid simulation [4] [5] [6] [7] [8] [9] [10] [11]. Each of these 5 simulation techniques has its own weaknesses and strengths. Therefore, choosing the right simulation technique depends on various factors such as the research problem, the context of where its be used, and the process or system being modeled.

### Monte Carlo simulation

The Monte Carlo simulation is a statistical technique used to analyze complex systems. This simulation technique makes use of probability distributions and random sampling to be able to simulate the behavior of the respective system. When running the simulation multiple times, Monte Carlo is able to give approximations to the probability of uncertain outcomes that occur. This helps in identifying the benefits, risks, and inefficiencies of different choices in a system and results in more informed decisions being taken. The different applications in healthcare where Monte Carlo simulation is used include patient flows through a hospital and the effectiveness of medical treatments.

### Discrete Event Simulation

The 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 or manufacturing process. Discrete event simulation gives freedom in testing different scenarios and later evaluates the effect of these different factors on the system's performance, examples include changes in customer arrival rates and processing time of a machine. Discrete Event simulation has various modeling applications in healthcare such as resource allocation, patient admission/scheduling, and new healthcare interventions.

### **Agent-Based Simulation**

Agent-Based simulation is a simulation technique that models the behavior of individual agents in a system, with each agent possessing its own objectives, behaviors, and rules while interacting with other agents within the system. This technique helps in studying the general behavior of the system, by examining how the decisions and interactions of agents with each other influence the outcome of a system. Furthermore, although ABS appears to have been rarely used in recent years it has started to gain more attention and is being utilized more. In healthcare, ABS is utilized to model healthcare providers, health behaviors that increase the risk of diseases, disease transmission in populations, and infection control for interaction between individuals. Thus, more insights can be obtained from a system and bottleneck and improvements can be made by simulating interactions within a system.

### **System Dynamics**

The System Dynamics simulation is generally used to model complex systems over a period of time. This is done through simulation technology by analyzing how different resources, information, and feedback loops move in the system. The technique helps understand how changes in one part of a system can affect other parts of it over time. Using this technique models that reveal bottlenecks in a system can be created and improvement in system outcomes can be achieved through informed decision making. Therefore, using this technique the complexity of a system can become more clear to understand. Application areas of system dynamics in healthcare are healthcare policy-making services, health improvement, and evaluation of economic models such as health insurance strategies.

### **Hybrid Simulation**

Hybrid simulation is a powerful technique, which is known to combine two or more simulation techniques. This technique helps model complex systems which cannot be entirely

represented using only one technique. By using this technique, it is possible to have a much more complete understanding of the complex system. The traditional simulation techniques which are generally combined include Discrete event simulation, Agent-based simulation, and System dynamics simulation. In the healthcare field Hybrid simulation can be very useful especially when modeling complex interactions between a healthcare system, healthcare staff, and patients. The applications of a hybrid simulation are the same as the traditional simulation techniques used, such as the impact of changes in health policy, optimization of resource allocation, and patient flow through hospitals.

### **2.2.3 Discrete Event simulation software tools**

After choosing the Discrete event simulation technique over other simulation techniques, due to having an interest in hospital operations. the process of exploring various software tools that support Discrete event simulation began. This list of software tools includes both commercial and open source software tools:

Commercial DES tools:

- ExtendSim
- Simul8
- AnyLogic
- Arena
- FlexSim
- ExtendSim
- Plant Simulation
- Witness
- Simcad
- Simio
- NetLogo

Open source DES tools:

- Python (SimPy package)
- R (Simmer Package)

From all the researched commercial tools which are mentioned, all require a license subscription to use. However, some tools offered free plans such as Plant Simulation, Witness, Anylogic, and Arena. Due to most free plans of the simulation tools only supporting the Windows operating system, a decision was made to use Anylogic. Since it offered a free Personal Learning Edition (PLE) and also supported MacOs, which will be the operating system to be used in making the simulation model.

Furthermore, although there are free open-source software tools that could be used like Python and R, which offer packages for discrete event simulation such as SimPy and simmer. The decision to use a commercial software tool was due to the fact that these tools have a user-friendly interface with drag-and-drop blocks. This makes it less time-consuming and more efficient when creating the simulation model.

## **2.3 Process Mining**

Process mining is a technique that enables the discovery, analysis, and enhancement of business processes based on event logs, which provide valuable insights. This technique can be utilized by organizations to identify relationships, visualize workflows, and uncover hidden dynamics within a process [12] [13].

By analyzing event logs, process mining techniques can uncover insights into the actual execution of a business process. This analysis exposes inefficiencies or 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.

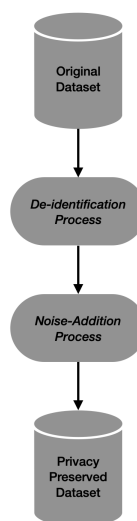
## **2.4 Event Logs**

Event logs are typically extracted from information systems. In process mining, event logs represent a sequential record of events that occur within a business process. During the execution of a specific process, these logs capture and record what happens. 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 may include are the agents involved in a process, resources utilized, and outcomes of a process [14].

## 2.5 Anonymization/Noise Addition

Large data collected by organizations such as the Population and Housing Census in the EU usually release statistical databases out to the public [15]. However, before this is done sensitive information such as personal identifying information (PII) is removed. Although PII is removed, researchers agreed that if these databases are combined with extra data malicious actors could then conduct privacy attacks and succeed in uncovering an individual's identity resulting in their private information being revealed. Thus, although data de-identification is a critical step it is only the first stage in preserving privacy. To ensure an even higher level of confidentiality after the removal of PII, methods such as noise addition are recommended [16]. Figure 2.1 below shows the standardized data privacy procedure.



**Figure 2.1:** Standardized Data Privacy Procedure.

In the following section relevant key terms used in noise addition procedure are explained. The terms *Data Privacy & Confidentiality* are the measures taken in order to protect individuals from any unauthorized sharing of information. Which aids in guaranteeing that sensitive data remains properly secured and private. In addition, the term *Data*

*Security* ensures that this private data is only accessible to authorized parties. *Data de-identification* involves a process where PII are first removed from a dataset. Moreover, the *Data de-identification process* which is also known as *data sanitization*, *data anonymization*, or *statistical disclosure control (SDC)* can remove or modify PII attributes in such a way that it makes the retrieval of an individual's 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 in removing PII and apply noise adding techniques to distort the data. However, while this can address the confidentiality issue, these measures may result in the original data to suffer from losing its statistical properties. Thus, causing the dataset to be less meaningful to the user. Consequently, achieving PUT is always a desired goal for researchers [17] [18] [19]. Unfortunately, researchers in data privacy agreed that achieving data privacy without reducing data utility is a difficult and challenging task [16] [20].

Even after the removal of PII attributes from datasets using data de-identification, researchers agreed on data de-identification as being an insufficient technique for protecting patient data. As the remaining sanitized data is still prone to being compromised and utilized for the reconstruction of an individual's identity [16]. Thus, it is essential for other measures to be applied in order to protect this sensitive information. Moreover, when inference attacks occur, the attributes are transformed significantly which then help prevent any connection with external sources. In order to address this issue, noise addition is used as a technique of distortion. This technique aims to modify numerical attributes to guarantee confidentiality, the confidential numerical attributes are modified by either multiplying or adding them with a random number. This random value is chosen from a normal distribution with a small standard deviation and a zero mean [16] [21].

## **2.6 Software Tools Used**

### **2.6.1 AnyLogic**

AnyLogic [22] is a simulation software tool that is used to create models to gain insights into workflows of several processes and systems, enabling optimization. The AnyLogic software



offers an easy to use user interface, that allows the modeling of several processes, including logistics, manufacturing, transportation, and healthcare. The software tool offers different sets simulation techniques such as discrete event simulation, Agent Based modeling, and system dynamics, which help discover complexities found within real-world systems. other than the modeling capabilities the tool provides, AnyLogic also offers visualization tools to help analyze simulation results and gain deeper insights of a systems behavior.

### **2.6.2 ARX**

The ARX software [23], also known as Anonymization Toolbox, is a software tool used for privacy preserving techniques. This tool helps tackle concerns regarding data privacy by implementing techniques that aim to protect private sensitive information while still preserving data utility. By utilizing this tool datasets can be anonymized before being shared to third parties, which results in compliance with privacy regulations. Anonymization techniques which the tool offers include k-anonymity, t-closeness, and differential privacy, and l-diversity [24]. Additionally, ARX provides features for exploring anonymized datasets, analyzing a datasets utility, and evaluating any linked risks for these anonymized datasets [25].

### **2.6.3 ProM**

The software tool ProM [26], is a tool that allows for the analysis and refinement of a business process by using event log data. 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 is able to effectively present process models, showcase bottlenecks, and detect unexpected behavior. Additionally, it includes a library which has an extensive collection of plugins, enhancing the capability and functionality of the software, making it an invaluable tool for process mining projects in different domains.

# Stakeholder Identification & Analysis

Identifying and analyzing stakeholders is an essential aspect of the project. Identifying and analyzing these stakeholders helps in figuring out the affected stakeholders and their interests. The stakeholders in this project include patients, hospital administrators, researchers, healthcare providers, and regulatory authorities. Patients are considered to be stakeholders due to their data being impacted by the use of anonymization and noise techniques that protect it. Patients want to ensure that their confidential health data is kept private and used in the way they expect it to be used. Hospital administrators are stakeholders due to their responsibility for making sure that hospital operations efficiently and effectively follow all rules and are compliant with regulations. Additionally, researchers and healthcare providers are also stakeholders because they use patient data in order to discover new breakthroughs and make informed decisions regarding patient treatment. Finally, the common stakeholders are regulatory authorities, the European Data Protection Board (EDPB) [27], the International Organization for Standardization (ISO) [28], the General Data Protection Regulation (GDPR) [29], and the Netherlands Standardization Institute (NEN) [30]. since all of them play a vital role in endorsing data privacy and protection, and ensuring healthcare organizations are effectively following all guidelines and regulations [31].

# Approach and Implementation

## 4.1 Radiology Department Model

The simulation model created using AnyLogic is a Radiology Department Model, this model is based on an existing model named “Emergency Department” from the example models by AnyLogic. However, modifications were made to it, in order to have it aligned with the objectives and goals of this project. The choice of using a radiology department was to make the research more interesting due to there being some relevance to real-life examples of other radiology departments within the region of Twente. Therefore, the research is focused on the radiology services offered in Enschede which are provided by the Medisch Spectrum Twente (MST) hospital in Enschede and the Ziekenhuisgroep Twente (ZGT) hospital in Hengelo [32] [33].

The simulation model of the Emergency Department model in AnyLogic recreates the workflow and process of a hospital emergency department. This model has a wide range of parameters which could be changed to better reflect the characteristics of a specific hospital setting.

The model is made up of entities which fall under different resource categories. These entities are then used to provide utility and execute actions within the model. The three resource categories are:

- Moving
- Static
- Portable

The moving resources are entities within the model that can move freely. In the Radiol-

ogy department examples of these moving resources 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 of the static resources used in the radiology department include 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 their own. Thus, only allowed personnel within the hospital are able to carry these resources. So the resources are being carried and moved by specific hospital staff where they are then later used for a particular task. The magnetic resonance imaging (MRI) and Ultrasound devices are examples of portable resources used in the Radiology department.

By using these three different resource types in a model one can easily and correctly represent behaviors of any real-life healthcare system. Moreover, choosing a resource depends on what its aim will be in the model since each resource has unique behaviors and features. Typically Moving resources are used when assigning schedules, while static resources are used during availability. Lastly, portable resources are assigned to hospital staff and are moved through locations within the model by them.

The radiology department model is made of various agents (Entities & Resources). A full list of the agents and the allocation number for each of them is listed below:

- Triage Room: Amount 2
- EC Room: Amount 2
- Waiting Room: Amount 1
- Medical Devices Storage Room: Amount 1
- X-Ray Room: Amount 1
- PA's: Amount 5
- Nurses: Amount 5
- Technicians: Amount 3
- MRI Device: Amount 2
- Ultrasound Device: Amount 2
- X-Ray Device: Amount 1

The simulation model can be viewed in three different formats: 3D, 2D, and Logic. The 3D format displays a three-dimensional environment of the radiology department model with also other 3D icons for the resources. When the simulation model is running, a representation of patients flowing through the department can be visualized. Furthermore, under the running model, the Resource utilization of the radiology department can be viewed. Which is measured using a value from 0 to 1, 0 meaning the resource has been used yet and 1 the resource is at full capacity and there is no availability for it. Additionally, at the bottom right of figure [4.1](#), an average for the length of stay between patients is calculated.

The second format is 2D, which displays a two-dimensional view of the model. The 2D view also has measurements of the resource utilization like the 3D environment, which constantly changes throughout the running of the model. However, there is a new window where the number of patients arriving and resources can be adjusted. These adjustments can be made by using the Circular horizontal scroll button and Radio buttons on the left side of the model

Furthermore, the third format is Logic, which shows a flowchart with blocks of how the model was initially built. The different blocks that make up this model are from a library called the process modeling library. This library is known to be used when modeling discrete event systems in AnyLogic. Moreover, on the top left of figure [4.3](#), the percentages of the utilization of different resources can be viewed.

Additionally, for all formats a status bar at the bottom can be seen where the simulation mode can be paused, stopped, sped up, slowed down, or put into full screen.

Using the Logic format in figure [4.3](#), the model built will be explained. The simulation model starts with the arrival of patients, who are represented as agents in the simulation. After that the patients register at the front desk where they queue in line waiting to register. Then the Patients enter a waiting room where they wait to be allocated a nurse to get checked on. The nurse then escorts the patient into a triage room where the severity of the patient's condition is examined. Following that the patient returns to the waiting room, and the nurse then decides whether the patient should be seen and 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. Furthermore, the patient is assigned a medical treatment process based on three different outputs. These three treatment process outputs are Usound Process, MRI Process, and X-Ray Process. After the process is done the patient is discharged and exits the department.

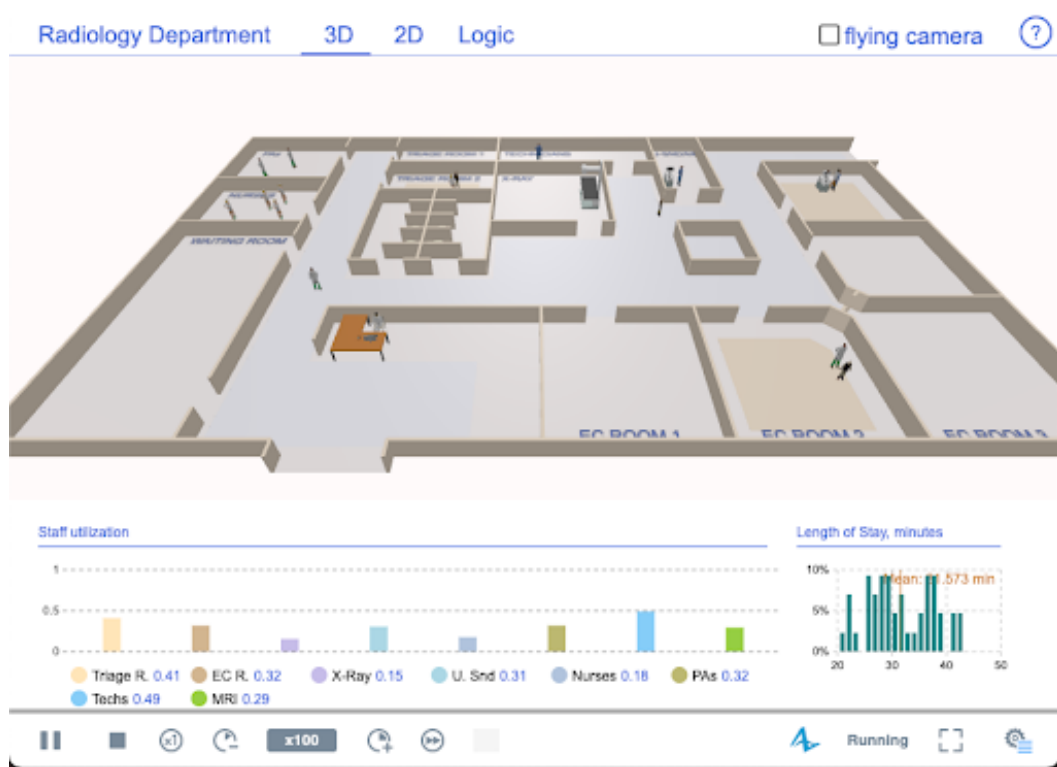
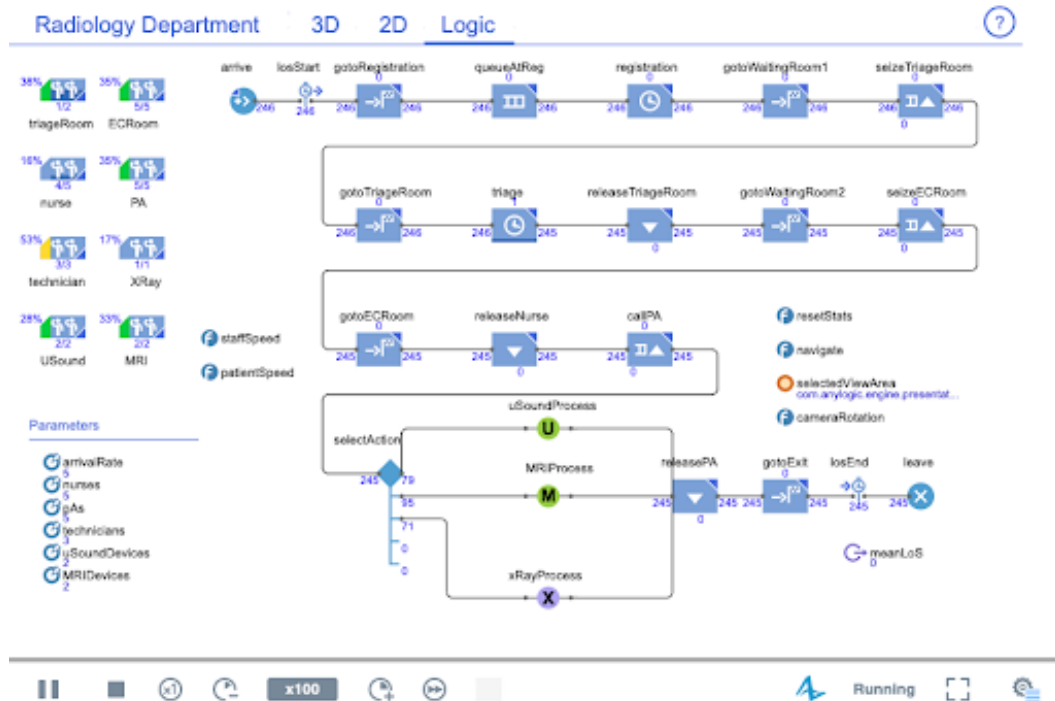


Figure 4.1: 3D Model of Radiology Department



Figure 4.2: 2D Model of Radiology Department



**Figure 4.3:** Logic Model of Radiology Department.

The previous paragraph explains how a patient flows through the model, however the flowchart elements and blocks from the process modeling library that model these different behaviors are not explicitly mentioned. Therefore, A description of them is mentioned and briefly explained below:

Space Markup (Elements):

- Path: Defines which path an entity can take in a simulation space.
- Point Node: A location (point) where entities can stop or move to.
- Rectangular Node: An area where entities can stop or move to.
- Attractor: Used to draw entities towards a certain location.

Blocks:

- Source: Represents the starting point of a model by generating entities that flow in the simulation.
- Sink: Represents the end point of an entity flow by removing entities from the simulation.

- Delay: introduces a waiting time in the simulation.
- Release: Releases a resource after its usage by an entity.
- Seize: Seizes a resource through an entity.
- Resource Pool: a group of similar resources that could be used by entities.
- Resource Attach: Attaches a resource pool to an entity/Agent.
- Queue: A buffer for entities waiting for a service or resource.
- Select Output 5: Randomly selects one of 5 output paths based on a chosen probability for each path.
- Move To: Moves an entity/Agent to a specified location.
- Time Measure Start: Indicates the start time of a process.
- Time Measure End: Indicates the end time of a process.

## 4.2 Synthetic Data Generation

The next phase of the project was to generate a synthetic dataset which then be imported into the simulation model for use. This section delves into the considerations involved and attributes chosen in generating this synthetic dataset. To generate a synthetic dataset of around 500 patients the a website called Fake Name Generator was used [34]. This website is a platform for generating bulk fake data by allowing the customisation of many different attributes. Using this tool certain settings can be configured in the generated dataset such as specifying name sets, countries, and the required percentage between male and female genders. Having offered these different settings, the website allows to create large amount of synthetic data all tailor-made particular requirements which can then be used further for testing, research, and analysis.

Figure 4.4 below shows the characteristics chosen for the synthetic dataset which include name set, countries, gender and, age. Since the model is based on an existing radiology department at MST, located in Enschede, the Netherlands, ensuring the dataset represents a realistic patient population is a must. Which is why correct and exact characteristics need to be selected carefully. The generated dataset for the model covers several considerations. Firstly, due to the research project representing a dutch population it is unquestionable that



the name set would be Dutch and the country would be the Netherlands as shown in figure 4.4 below.

**Step 3 - Choose name sets, countries, gender, and age**

**Name set**

- Chinese
- Chinese (Traditional)
- Croatian
- Czech
- Danish
- Dutch**

**Country**

- Germany
- Greenland
- Hungary
- Iceland
- Italy
- Netherlands**

**Gender**

Male: **51%** Female: **49%**

**Age**

**17 - 75 years old**

**Figure 4.4:** Name set, Country, Gender, and Age of Synthetic Dataset.

Regarding gender distribution as shown in the slider in figure 4.4, the decision to allocate 51% male and 49% female distribution was based on statistical data specific to Enschede [35]. Despite the general population of the Netherlands having the opposite distribution of 51% female and 49% male, this choice was made to align with local demographics of Enschede.

Furthermore, for the age range selection next to the gender distribution slider, certain practical considerations led the decision of including individuals aged 17 to 75. In order to determine the appropriate age range, various factors were taken into account. Similarly to the last example of choosing an appropriate gender distribution, the same website provided statistical insight into the age distribution of people living in Enschede [36]. Additionally, to ensure a diverse representation of patient demographic there is a mixed inclusion of younger and older age groups. Moreover, the choice not going lower than the age of 17 was due to different age groups having varying healthcare needs. To clarify, elderly individuals maybe require more frequent radiology services due to their age related health condition. Thus, starting the age of the dataset from age 17 ensures various age groups are included.

The next steps after that is choosing the desired attributes for the data set yet to be generated. Initially many attributes were selected to be part of the dataset such as:

- Title
- Given name
- Surname

- Name set
- Gender
- Street address
- Postal code
- City
- Birthday/Date of birth (dob)
- Age
- Occupation
- Blood type
- Weight (Kg)
- Height (ft/in)

## **4.3 Data Pre-Processing**

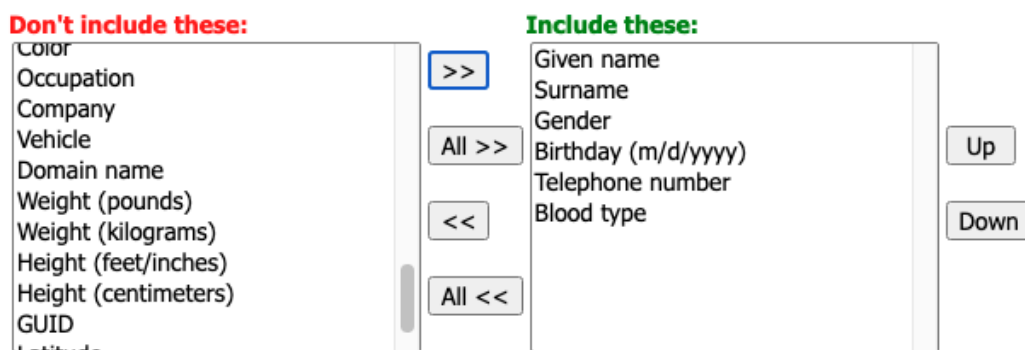
### **4.3.1 Synthetic Data**

However after the dataset was made some changes were made and a new dataset was generated as the inclusion of some attributes were found to be unnecessary. The first attribute that was removed was the name set, as all generated patients had a dutch name set so it was unnecessary to leave it in there. Secondly, the street address, postal code, and attributes were all removed because the dataset included addresses for patients from all over the Netherlands. This problem causes the dataset to not be realistic since its fails in representing the demographic of Enschede only without including other irrelevant cities. Furthermore, the Age attribute was removed as it would have been redundant to have with the inclusion of the Birthday attribute. Not only that, but another reason to include the Birthday instead of the age is its more formal and precise format, which follows the (m/d/yyyy) structure. Lastly, the same issue was encountered with the attributes Occupation, weight, and height. The problem occurred during the import process of the entire dataset into the AnyLogic database. The iterative approach of importing the dataset multiple times, each time with different attributes, allowed for the identification of the specific attributes causing the issue. Although the exact reason for the failure of these attributes to import into the

database remains unknown, it is possible that the data format did not comply with AnyLogic's database specifications. After these changes the final list for the attributes includes the following:

- Given name
- Surname
- Gender
- Birthday (m/d/yyyy)
- Telephone number
- Blood type

Refer to Figure 4.5 below for a visual representation of these attributes.



**Figure 4.5:** Final Synthetic Attributes Included into the Dataset.

After configuring and selecting all the attributes, a dataset of 500 patients was generated. The dataset was downloaded as an Excel workbook (.xlsx) format, which was also compressed into a zip file. To ensure the dataset is accurate, the file was imported into google sheets for verification. Other minor adjustments like the renaming of columns were made. A glimpse of the data in figure 4.6 below, which shows the first 10 rows of the dataset.

Following that the dataset was then imported into the AnyLogic builtin AnyLogic database

After checking that the dataset is accurate and reliable, the next step was to import it into the built-in AnyLogic database. This database served as a mean to map the parameter attributes to the patients entering the radiology department of the simulation model. By utilizing the AnyLogic database functionality, a connection between the generated dataset and the simulation model was established, allowing for seamless integration of patient attributes.

	A	B	C	D	E	F
1	GivenName	Surname	Gender	DOB	TelephoneNumber	BloodType
2	Viënne	van Sluijs	female	4/15/1995	06-23499151	B+
3	Leonardus	Vossebeld	male	12/12/1951	06-25233988	O+
4	Iefke	van Iterson	female	2/10/2004	06-35323834	O+
5	Natalya	Steijn	female	11/10/1972	06-57547041	A+
6	Jing	Braak	female	12/16/1972	06-38815514	B+
7	Ashley	Windt	female	1/21/1959	06-76949781	O+
8	Eyüp	Wilschut	male	11/13/1969	06-62920934	B+
9	Kerwin	Spijker	male	5/8/1993	06-80533025	A+
10	Aliye	Banning	female	12/23/1955	06-83067572	B+

**Figure 4.6:** First 10 Rows of the Generated Synthetic Data.

Furthermore, the AnyLogic database offered a convenient method for storing and retrieving patient information during simulation runs. The dataset imported into the AnyLogic database is illustrated in figure 4.7 below.”

	given_name	surname	gender	dob	telephone_number	blood_type
1	Viënne	van Sluijs	female	4/15/1995	06-23499151	B+
2	Leonardus	Vossebeld	male	12/12/1951	06-25233988	O+
3	Iefke	van Iterson	female	2/10/2004	06-35323834	O+
4	Natalya	Steijn	female	11/10/1972	06-57547041	A+
5	Jing	Braak	female	12/16/1972	06-38815514	B+
6	Ashley	Windt	female	1/21/1959	06-76949781	O+
7	Eyüp	Wilschut	male	11/13/1969	06-62920934	B+
8	Kerwin	Spijker	male	5/8/1993	06-80533025	A+
9	Aliye	Banning	female	12/23/1955	06-83067572	B+
10	Ramón	Piels	male	12/13/1982	06-20754413	O+
11	Nermin	Gierveld	female	6/14/1974	06-69393610	B+
12	Liana	Worm	female	9/29/2002	06-96784769	O+
13	Palma	Meesters	female	8/1/1965	06-21832800	B+

**Figure 4.7:** Imported Dataset into the AnyLogic Database.

### 4.3.2 Event Logs

The next steps following the import of the dataset into the AnyLogic database is to generate the event logs. To generate the event logs, it is necessary to run the simulation model. During the model execution, different event logs are being recorded, which capture the behavior of different agents and resources within the radiology department. After the completion of the simulation run, several event logs are saved into the log folder inside the AnyLogic database. However, most of these logs were filtered out and specific event logs were selected as they will provide the most relevant insight during future analysis in ProM.

Moreover, given the focus on testing the PUT in ProM, emphasis is on analyzing event logs that represent the processes patients undergo in the radiology department. These event logs offer important information on patient ids, timestamps of activities, resource pools, and processes patients undergo. Thus, using these event logs will provide valuable insights into the interplay between privacy and utility will be gained. These event logs provide valuable insights into the relationship between privacy and utility. The event logs which will provide these insights are "agent\_parameter\_log" and "flowchart\_process\_states\_log". The agent\_parameters\_log (see figure 4.8) stores the parameter values of individual agents in the simulation model. While the "flowchart\_process\_states\_log" (see figure 4.9) records the timestamps individual agents spent in different states of flowchart blocks.

	agent_type	agent	parameter_name	parameter_value
1	Patient	<population>[25] : 1401	givenname	Angel
2	Patient	<population>[25] : 1401	surname	Walhout
3	Patient	<population>[25] : 1401	dob	9/27/1961
4	Patient	<population>[25] : 1401	bloodtype	A+
5	Patient	<population>[25] : 1401	gender	male
6	Patient	<population>[25] : 1401	telephonenumber	06-19688517
7	Patient	<population>[25] : 962	telephonenumber	06-23499151

**Figure 4.8:** The Agent Parameters Log.

	agent_type	agent	block_type	block	activity_type	start_date	stop_date
1	Patient	<population>[25] : 962	MoveTo	gotoRegistration	MOVE	15-06-2023 00:07:54	15-06-2023 00:07:56
2	Patient	<population>[25] : 962	Delay	registration	WORK	15-06-2023 00:07:56	15-06-2023 00:08:39
3	Patient	<population>[26] : 963	MoveTo	gotoRegistration	MOVE	15-06-2023 00:08:24	15-06-2023 00:08:27
4	Patient	<population>[26] : 963	Queue	queueAtReg	WAIT	15-06-2023 00:08:27	15-06-2023 00:08:39
5	Patient	<population>[25] : 962	MoveTo	gotoWaitingRoom1	MOVE	15-06-2023 00:08:39	15-06-2023 00:08:45

**Figure 4.9:** The Flowchart Process States Log.

The agent parameters log in figure 4.8 includes columns:

- agent\_type: refers to the entity or object involved in the simulation.
- agent: refers to the entity or object itself that interacts with the simulation and is recorded by it.
- parameter\_name: refers to the label assigned to a parameter.

- `parameter_value`: refers to the categorical, symbolic, or numerical attribute assigned to the parameter name

The flowchart process states log in figure [4.9](#) includes columns:

- `agent_type`: refers to the entity or object involved in the simulation.
- `agent`: refers to the entity or object itself that interacts with the simulation and is recorded by it.
- `block_type`: refers to the type of block used (from process modeling library) in the simulation model.
- `block`: refers to the block name in the simulation model.
- `activity_type`: refers to the type of task being performed for a specific process.
- `start_date`: Timestamp indicating when a process began.
- `stop_date`: Timestamp indicating when a process has concluded.

However before the event logs are exported for analysis, it is essential to filter only the required columns. Therefore, in the agent parameters log of [4.8](#), `agent_type` will be filtered to only display patient agents. As a result, the second agent column will automatically be filtered to also show patient agents, meaning no filtering is required for that column. As for the last two columns, `parameter_name` and `parameter_value`, they will remain unfiltered, allowing the display of attributes for all patients. For the second event log, namely flow chart process in Figure [4.9](#), the first two columns, `agent_type` and `agent`, will be filtered similarly to the last event log. The third column and fourth column, `block_type` and `block`, will remain unfiltered to include all block types and processes to show the whole process a patient undergoes. The `activity_type` column will remain unchanged, as it just explains the activity type of the specified process in the block column. Similarly, the last two columns, `start_date` and `stop_date`, will remain unchanged, as the timestamps for patients is a requirement for analysis in ProM.

Finally after these two event logs have been modified they can be merged together, resulting in the unanonymized event log shown in figure [4.10](#). In order to merge these two event logs, the SQL query code for each log was used. The SQL query code used to merge these two logs can be found in the appendix at [E.1](#).

”The SQL query code used to merge these event logs can be found in the provided code snippet below.”

	agent_type	agent	block_type	block	activity_type	start_date	stop_date	parameter_name	parameter_value
1	Patient	<population>[25] : 962	MoveTo	gotoRegistration	MOVE	15-06-2023 00:07:54	15-06-2023 00:07:56	givenname	Vienne
2	Patient	<population>[25] : 962	MoveTo	gotoRegistration	MOVE	15-06-2023 00:07:54	15-06-2023 00:07:56	surname	van Sluijs
3	Patient	<population>[25] : 962	MoveTo	gotoRegistration	MOVE	15-06-2023 00:07:54	15-06-2023 00:07:56	dob	4/15/1995
4	Patient	<population>[25] : 962	MoveTo	gotoRegistration	MOVE	15-06-2023 00:07:54	15-06-2023 00:07:56	bloodtype	B+
5	Patient	<population>[25] : 962	MoveTo	gotoRegistration	MOVE	15-06-2023 00:07:54	15-06-2023 00:07:56	gender	female

Figure 4.10: Merged Event Logs.

### 4.3.3 ARX

After exporting the unanonymized event log as a comma-separated values (CSV) file from AnyLogic, the next step is to import the CSV file into ARX and begin with the anonymization. However to ensure accurate analysis is done, the data types and format for each attribute were carefully reviewed first. All data types were correct but two formats of the data type:date/time required slight adjustments as they were incorrect. The format for the "start date" and "stop date" attributes had to be changed from (yyyy-MM-dd hh:mm:ss) to (yyyy-MM-dd HH:mm:ss) as the timestamps used in AnyLogic utilize the 24-hour format and not the 12-hour format. This slight modification for the hour component ensures the date data type is accurate and consistent within the entire dataset resulting in the event log being effectively processed without any errors by ARX. The format for the second attribute ,dob, also had to be modified from (dd/MM/yyyy) to (MM/dd/yyyy).

Selected	Name	Data type	Format
✓	agent	String	
✓	agent type	String	
✓	block type	String	
✓	block	String	
✓	activity type	String	
✓	start date	Date/Time	yyyy-MM-dd hh:mm:ss
✓	stop date	Date/Time	yyyy-MM-dd hh:mm:ss
✓	bloodtype	String	
✓	dob	Date/Time	dd/MM/yyyy
✓	gender	String	
✓	givenname	String	
✓	surname	String	
✓	telephonenumber	String	

Figure 4.11: Data Type and Format of Attributes in ARX.

Furthermore after these corrections were made the following task was to transform the layout of the attributes from a vertical listing to being a horizontal arrangement to be able to continue with anonymizing the attributes. Since there is no way around this in AnyLogic as the event log will always display the columns "parameter\_name" and "parameter\_value" vertically, python was used to pivot the attributes as needed. This allowed for the continuation of the anonymization process. The code used for pivoting these two column can be found in the appendix at [D.1](#).

Finally after making these changes the imported event log in ARX is shown in figure [4.12](#) below.

agent	agent by block typ	block	activity	start date	stop date	blo: dob	gender	givenna: surname	telephonenumb
<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
<population>[25] : 962	Patient	Delay	MRIProcess.doMRI	WORK 2023-06-15 0:16:42	2023-06-15 0:35:01	B+ 4/15/1995	female	Viënne van Stuijs	06-23499151
<population>[25] : 962	Patient	Delay	registration	WORK 2023-06-15 0:07:56	2023-06-15 0:08:39	B+ 4/15/1995	female	Viënne van Stuijs	06-23499151
<population>[25] : 962	Patient	Delay	triage	WORK 2023-06-15 0:09:04	2023-06-15 0:15:41	B+ 4/15/1995	female	Viënne van Stuijs	06-23499151
<population>[25] : 962	Patient	MoveTo	gotoECRoom	MOVE 2023-06-15 0:15:53	2023-06-15 0:16:08	B+ 4/15/1995	female	Viënne van Stuijs	06-23499151
<population>[25] : 962	Patient	MoveTo	gotoExit	MOVE 2023-06-15 0:35:01	2023-06-15 0:35:17	B+ 4/15/1995	female	Viënne van Stuijs	06-23499151
<population>[25] : 962	Patient	MoveTo	gotoRegistration	MOVE 2023-06-15 0:07:54	2023-06-15 0:07:56	B+ 4/15/1995	female	Viënne van Stuijs	06-23499151
<population>[25] : 962	Patient	MoveTo	gotoTriageRoom	MOVE 2023-06-15 0:08:53	2023-06-15 0:09:04	B+ 4/15/1995	female	Viënne van Stuijs	06-23499151
<population>[25] : 962	Patient	MoveTo	gotoWaitingRoom1	MOVE 2023-06-15 0:08:39	2023-06-15 0:08:45	B+ 4/15/1995	female	Viënne van Stuijs	06-23499151
<population>[25] : 962	Patient	MoveTo	gotoWaitingRoom2	MOVE 2023-06-15 0:15:41	2023-06-15 0:15:53	B+ 4/15/1995	female	Viënne van Stuijs	06-23499151
<population>[25] : 962	Patient	Seize	MRIProcess.seizeTech	WAIT 2023-06-15 0:16:25	2023-06-15 0:16:42	B+ 4/15/1995	female	Viënne van Stuijs	06-23499151

Figure 4.12: Final Version of Imported Event Log.

#### 4.3.4 ProM

During the data pre-processing for ProM, minimal time was needed since it mainly focused on converting the CSV of the event logs 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 [\[37\]](#). First the CSV file was imported into ProM, then the plugin was used to transform the CSV file to the desired XES format. This process involved mapping the columns of the event log to the relevant fields case, event, start time, and completion time [\[38\]](#). As shown in figure [4.14](#), "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 becomes suitable for analysis.



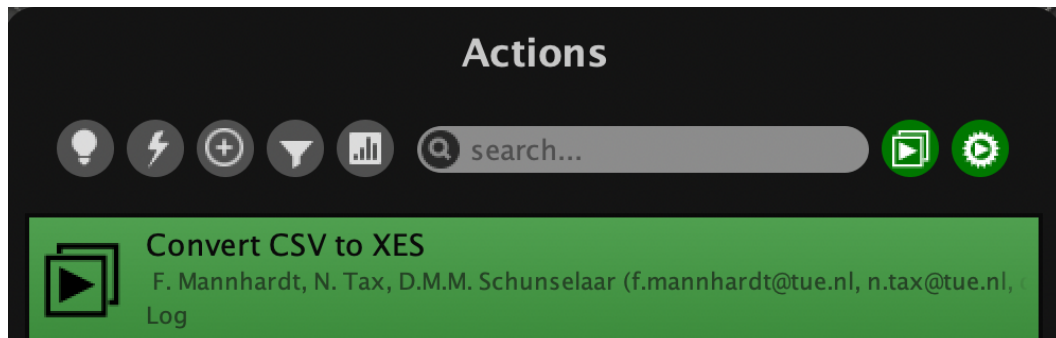


Figure 4.13: Plugin Used for Conversion.

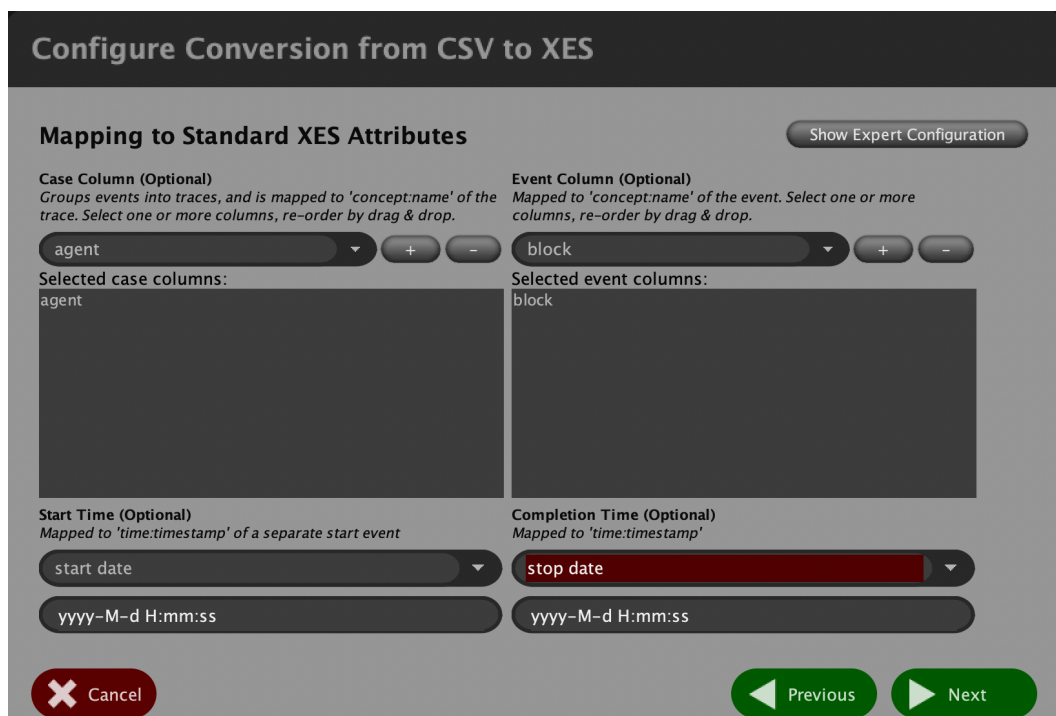


Figure 4.14: Mapping Event Log Columns to XES Attributes.

## 4.4 Anonymization using ARX

Before anonymizing the data in ARX, an important step is to label the attributes appropriately. There are four different categorizations for the types of attributes: Identifying, quasi-identifying, sensitive, and insensitive. Each attribute in the dataset should be assigned a type depending on its role.

1. Identifying: attributes with a high risk of re-identification
2. Quasi-identifying: attributes that when combined can potentially lead to re-identification
3. Sensitive: attributes which contain information that an individual is not willing to be linked with. This attribute type may cause harm to individuals and is usually of interest to potential attackers
4. Insensitive: attributes which do not pose any privacy risks.

The categorization of the attributes/columns included in the event log is as follows:

1. Identifying attributes
  - given name
  - surname
  - telephone number
2. Quasi-identifying attributes
  - blood type
  - dob
  - gender
3. Sensitive attributes
  - no specific columns associated
4. Insensitive attributes
  - agent
  - agent type
  - block type

- block
- activity type
- start date
- stop date

The columns listed in the identifying attributes are classified as such due to being personally identifiable information (PII). These PII attributes can be used directly to identify a specific individual. Therefore they should be removed from the dataset to preserve the privacy of the patient's personal information. However, quasi-identifying attributes can indirectly be linked back 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 appropriately safeguard these attributes during the anonymization process. Sensitive attributes are attributes which 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 the list above, does not pose any privacy risks. Moreover, the inclusion or exclusion of these columns within the dataset does not affect the anonymization process.

#### **4.4.1 Transforming quasi-identifiers**

Once all columns are classified based on their attribute type, the anonymization process can begin. For the anonymization process, focus is on transforming quasi-identifying attributes. Since quasi-identifiers when combined contain information that leads to identifying a patient, a method known as generalization is used. This method aids in reducing the possibility of re-identification by broadening the data and making it less precise to a specific individual [24].

The first quasi-identifying attribute which was transformed is the blood type. For this attribute, a masking method was used. This masking method gradually generalizes the data by adding an asterisk for each level, making it more privacy preserved. As shown in figure 4.15 below, there are three different levels: level-0 is the raw data from the data set, level-1 is the only the plus or minus sign of the blood types, and level-2 is a fully suppressed data.

The second quasi-identifying attribute being transformed is the date of birth. This attribute also has three different levels. 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 decade intervals of the patient.

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

**Figure 4.15:** Three levels of Generalization Transformation for Blood Type Attribute.

Level-0	Level-1	Level-2
5/7/1947	1947	[1940, 1950[
6/15/1947	1947	[1940, 1950[
9/17/1947	1947	[1940, 1950[
10/26/1947	1947	[1940, 1950[
2/5/1948	1948	[1940, 1950[
4/29/1948	1948	[1940, 1950[
5/25/1948	1948	[1940, 1950[
7/20/1948	1948	[1940, 1950[
7/21/1948	1948	[1940, 1950[

**Figure 4.16:** Three levels of Generalization Transformation for dob Attribute.

The third and last quasi-identifying attribute is the gender, this attribute only has two levels. The first level , level-0, shows the exact gender from the raw dataset so either male or female. The second level ,level-1, combines the two genders into one grouped category.

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

**Figure 4.17:** Two levels of Generalization Transformation for Gender Attribute.

#### 4.4.2 Anonymization using K-anonymity

Following the transformation of quasi-identifiers using generalization, a privacy model which is also known as an anonymization technique is applied. In ARX various anonymization techniques are available, however for this project 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 [39]. What this means is that the groups of attributes which could possibly identify a patient are combined with other groups making it more difficult for anyone to separate patients in the dataset from each other and identify them [40]. This technique provides a higher level of privacy protection within a dataset and helps maintain confidentiality for sensitive information.

#### 4.4.3 Anonymized event log attributes

Finally, after the addition of the k-anonymity privacy model, ARX generates an anonymized dataset. ARX generates this anonymized dataset by taking into consideration specific weights settings chosen for each quasi-identifying attribute to prioritize their importance. These weight settings ensure that information loss is minimized while preserving privacy. Furthermore, following the process of anonymizing the event log, further analysis can be done using ProM. The dataset containing the anonymized attributes is shown in figure 4.19 below.

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

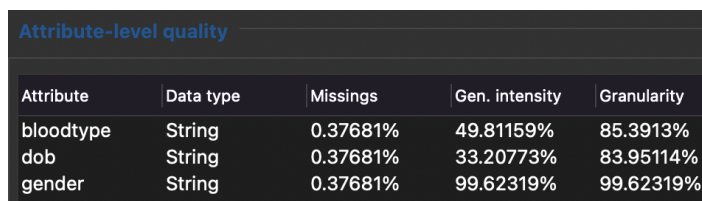
Figure 4.18: Example of a K-anonymity Dataset (K=4).

bloodtype	dob	gender	givenname	surname	telephonenumber
B*	[1940, 1950[	male	*	*	*
B*	[1980, 1990[	male	*	*	*
A*	[1970, 1980[	female	*	*	*
O*	[1960, 1970[	male	*	*	*
B*	[1960, 1970[	male	*	*	*
A*	[1970, 1980[	female	*	*	*
B*	[1970, 1980[	female	*	*	*
B*	[1960, 1970[	female	*	*	*
A*	[1950, 1960[	male	*	*	*
B*	[2000, 2010[	male	*	*	*
A*	[1950, 1960[	female	*	*	*
A*	[1980, 1990[	male	*	*	*
B*	[1990, 2000[	male	*	*	*
B*	[1970, 1980[	male	*	*	*
O*	[1990, 2000[	male	*	*	*
B*	[1950, 1960[	male	*	*	*
A*	[1950, 1960[	female	*	*	*

Figure 4.19: Snapshot of a Subset of the Anonymized Event Log Dataset.

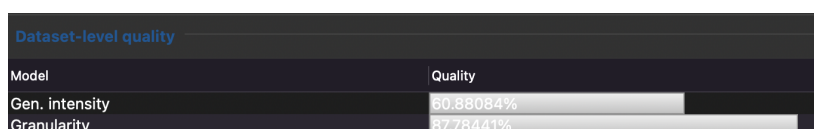
## 4.5 Utility analysis in ARX

ARX includes a feature that allows the analysis of the datasets utility post anonymization, this enables the evaluation of the usefulness 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.



Attribute-level quality				
Attribute	Data type	Missings	Gen. intensity	Granularity
bloodtype	String	0.37681%	49.81159%	85.3913%
dob	String	0.37681%	33.20773%	83.95114%
gender	String	0.37681%	99.62319%	99.62319%

**Figure 4.20:** Separate Granularity & Precision Percentages of Quasi-identifying Attributes.



Dataset-level quality	
Model	Quality
Gen. intensity	60.88084%
Granularity	87.78441%

**Figure 4.21:** Combined Granularity & Precision Percentages of Quasi-identifying Attributes.

## 4.6 Risk analysis in ARX

ARX also offers a risk analysis feature, that evaluates the potential re-identification risk associated with the dataset before and after anonymization. As shown in table [4.1](#) below, the re-identification risk for the unanonymized log was 24.347% but after anonymization this number decreases to 0.232%. This indicates the successful anonymization of the dataset.

	<i>Unanonymized Log</i>	<i>Anonymized Log</i>
<b>Re-identification Risk (%)</b>	<b>24.347%</b>	<b>0.232%</b>

**Table 4.1:** Re-identification Risk of Unanonymized Log vs Anonymized Log.

## 4.7 Utility analysis using ProM

For measuring the utility difference between the unanonymized and anonymized event logs in ProM, a comparison was made using the technique known as process discovery, specifically the "inductive visual miner plugin" [41] [42]. This plugin extracts behavioral patterns and process flow from the events in to then provide a visual animation of the process model [43] [44]. This animated process model then enables in understanding the relationship between events and in gaining insights of the whole process by replaying over the log [45].

Figures 4.22 and 4.23 below display the unanonymized and anonymized event logs of the two different paths being compared. It is evident that the paths in the unanonymized and anonymized logs exhibit distinct differences, indicating a potential impact on utility. In Figure 1, there are fewer occurrences of paths, displaying only true paths without redundancy. However, Figure 2 shows a significantly higher number of path occurrences, for possible reasons explored in the discussion chapter. In addition to these two figures, Appendix-C includes other figures comparing various activities, such as waiting times, service times, and Petri nets, for both the unanonymized and anonymized event logs.

This result suggests that the utility of the observed anonymized event log may have been affected due to 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 maintained.



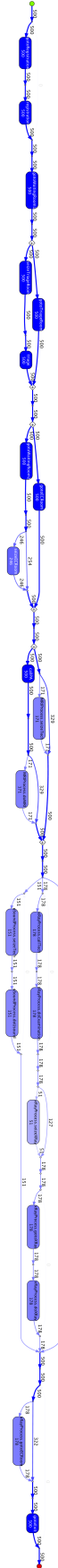


Figure 4.22: Process Model of Unanonymized Event Log.

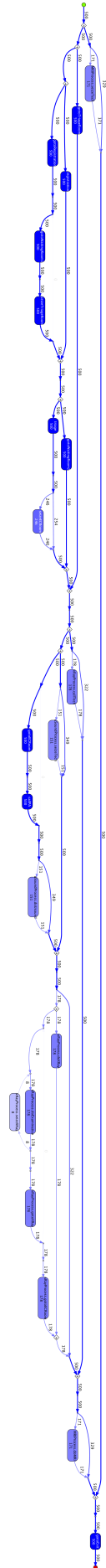


Figure 4.23: Process Model of Anonymized Event Log.

# Discussion and Limitations

## 5.1 Discussion

There are several interesting reasons contributing to the observed utility loss in the anonymized event log. While the k-anonymity technique is crucial for preserving privacy, it also presents challenges which impact the data utility. Three reasons that might lead to this utility loss include 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 conducting 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/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.

## 5.2 Limitations

The project encountered several limitations which should be considered. Firstly, conducting research in a relatively new field like process mining was challenging. The available literature and plugins on process mining applications for privacy-utility trade-off were limited, which affected the depth of analysis done on the event logs. Secondly, although the simulation model could have been more realistic, the free (PLE) version by AnyLogic enforced restrictions due to being constraint to only 10 agents in the model. Moreover, although the model was already quite complex, as visualized in ProM by the process model, it could've had an even more complex representation. Thirdly, due to the indirect contact with real world radiology departments in Enschede the model did not fully represent exact details of their operations. Although this limitation did not affect the results, it is worth noting that creating a more precise model, which exactly mimics real world operations, would have been a valuable addition because it improves the models real-world applicability and accuracy. The fourth limitation was being only limited using the k-anonymity technique and not exploring and comparing it to other anonymization techniques such as t-closeness and l-diversity. This limitation was due to the dataset lacking sensitive attribute types which is required for implementing these alternative techniques in ARX.

# Conclusion and Future work

## 6.1 Conclusion

In conclusion process mining offers valuable insights into process workflows, especially when analyzing the utility of anonymized event logs. This allows organizations and third parties to evaluate the usefulness of their anonymized datasets prior to sharing them with other parties. Additionally, anonymization plays an important role in preserving patient information, however it is important to also address concerns surrounding the utility of anonymized datasets and be transparent about the used anonymization techniques. Nevertheless, more research is still required to better understand how to strike a balance between data privacy and data utility. As ensuring the effectiveness of process analysis on anonymized sensitive datasets is important for developments in data analysis and privacy protection.

Furthermore, one notable thing that was observed is that process mining was able to reveal hidden insights from the simulation model which were not apparent at first. During the examination of the unanonymized event log, small unexpected differences in the process flow were noticed, which were not configured in the model's settings. This evidence proves the significance of process mining in uncovering valuable insights which would usually remain unnoticed. Although the impact was obvious with the simulation model, the true potential of process mining stands out when applied to actual real life event logs. Since analyzing real data enables organizations to discover highly valuable insights with interesting new findings. In summary, through the use of 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.

## 6.2 Future Work

There are several components that can be considered for future work to contribute to the improvement of this project. The first is refining the simulation model by adding patient symptoms to the select output block for the three imaging processes. This ensures a more realistic model and avoids in patients being randomly assigned using probabilities. In addition, the inclusion of other synthetic data attributes will improve the quality of the dataset and enrich its information. Examples of attributes that could be added include symptoms, BSN (Burger Service Nummer or the citizen service number), occupation, and salary. A third improvement would be to directly collaborate with the radiology departments like MST in Enschede, this will help in adding more specific operational details resulting a better representation model. Furthermore, research shows that k-anonymity alone is not sufficient and the use of different anonymization techniques which build upon k-anonymity are required, such as l-diversity and t-closeness [46] [47] [48]. However, it is required for sensitive attributes to be present in the synthetic dataset to use these techniques. Additionally, the use of newly available plugins for process discovery and performance analysis would be beneficial in evaluating the utility of anonymized event logs. A final recommendation ,that is not directly related to the utility analysis of anonymized log, is to use publicly available real-life healthcare event logs and maybe even see how they compare with the simulated event logs. These opportunities for improvement will significantly advance the project and result in more complete outcomes within simulation, anonymization, and process mining.

# Bibliography

- [1] C. Schröer, F. Kruse, and J. M. Gómez, “A systematic literature review on applying crisp-dm process model,” *Procedia Computer Science*, vol. 181, p. 526–534, 2021. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S1877050921002416>
- [2] A. Maria, “Introduction to modeling and simulation,” in *Proceedings of the 29th conference on Winter simulation - WSC '97*. Atlanta, Georgia, United States: ACM Press, 1997, p. 7–13. [Online]. Available: <http://portal.acm.org/citation.cfm?doid=268437.268440>
- [3] R. E. Shannon, “Introduction to simulation,” in *Proceedings of the 24th conference on Winter simulation - WSC '92*. Arlington, Virginia, United States: ACM Press, 1992, p. 65–73. [Online]. Available: <http://portal.acm.org/citation.cfm?doid=167293.167302>
- [4] K. Katsaliaki and N. Mustafee, “Applications of simulation within the healthcare context,” *The Journal of the Operational Research Society*, vol. 62, no. 8, p. 1431–1451, 2011. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7099916/>
- [5] N. Mustafee, K. Katsaliaki, and S. J. Taylor, “Profiling literature in healthcare simulation,” *SIMULATION*, vol. 86, no. 8–9, p. 543–558, Aug 2010. [Online]. Available: <https://doi.org/10.1177/0037549709359090>
- [6] S. Salleh, P. Thokala, A. Brennan, R. Hughes, and A. Booth, “Simulation modelling in healthcare: An umbrella review of systematic literature reviews,” *PharmacoEconomics*, vol. 35, no. 9, p. 937–949, Sep 2017. [Online]. Available: <https://doi.org/10.1007/s40273-017-0523-3>
- [7] J. Viana, “Reflections on two approaches to hybrid simulation in healthcare,” in *Proceedings of the Winter Simulation Conference 2014*, Dec 2014, p. 1585–1596.

- [8] S. C. Brailsford, "Hybrid simulation in healthcare: New concepts and new tools," in *2015 Winter Simulation Conference (WSC)*, Dec 2015, p. 1645–1653.
- [9] B. Mielczarek, "Review of modelling approaches for healthcare simulation," *Oper Res Decis*, vol. 26, Jan 2016.
- [10] J. I. Vázquez-Serrano, R. E. Peimbert-García, and L. E. Cárdenas-Barrón, "Discrete-event simulation modeling in healthcare: A comprehensive review," *International Journal of Environmental Research and Public Health*, vol. 18, no. 22, p. 12262, Nov 2021. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8625660/>
- [11] S. Almagoshi, "Simulation modelling in healthcare: Challenges and trends," *Procedia Manufacturing*, vol. 3, p. 301–307, Jan 2015. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2351978915001560>
- [12] E. Rojas, J. Munoz-Gama, M. Sepúlveda, and D. Capurro, "Process mining in healthcare: A literature review," *Journal of Biomedical Informatics*, vol. 61, p. 224–236, Jun 2016. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S1532046416300296>
- [13] W. Van Der Aalst, "Process mining: Overview and opportunities," *ACM Transactions on Management Information Systems*, vol. 3, no. 2, p. 1–17, Jul 2012. [Online]. Available: <https://dl.acm.org/doi/10.1145/2229156.2229157>
- [14] B. F. Van Dongen, A. K. A. De Medeiros, H. M. W. Verbeek, A. J. M. M. Weijters, and W. M. P. Van Der Aalst, *The ProM Framework: A New Era in Process Mining Tool Support*, ser. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, vol. 3536, p. 444–454. [Online]. Available: [http://link.springer.com/10.1007/11494744\\_25](http://link.springer.com/10.1007/11494744_25)
- [15] [Online]. Available: <https://ec.europa.eu/eurostat/web/population-demography/population-housing-censuses>
- [16] K. Mivule, "Utilizing noise addition for data privacy, an overview."
- [17] M. Sramka, R. Safavi-Naini, J. Denzinger, and M. Askari, "A practice-oriented framework for measuring privacy and utility in data sanitization systems," in *Proceedings of the 2010 EDBT/ICDT Workshops*. Lausanne Switzerland: ACM, Mar 2010, p. 1–10. [Online]. Available: <https://dl.acm.org/doi/10.1145/1754239.1754270>



- [18] V. Rastogi, D. Suciú, and S. Hong, "The boundary between privacy and utility in data publishing."
- [19] L. Sankar, S. R. Rajagopalan, and H. V. Poor, "Utility and privacy of data sources: Can Shannon help conceal and reveal information?" in *2010 Information Theory and Applications Workshop (ITA)*. La Jolla, CA, USA: IEEE, Jan 2010, p. 1–7. [Online]. Available: <http://ieeexplore.ieee.org/document/5454092/>
- [20] S. A. Sohail, "Normative and empirical evaluation of privacy utility trade-off in health-care."
- [21] M. A. Kadampur, "A noise addition scheme in decision tree for privacy preserving data mining," vol. 2, no. 1, 2010.
- [22] [Online]. Available: <https://www.anylogic.com/>
- [23] [Online]. Available: <https://arx.deidentifier.org/>
- [24] J. Marques and J. Bernardino, "Analysis of data anonymization techniques:," in *Proceedings of the 12th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*. Budapest, Hungary: SCITEPRESS - Science and Technology Publications, 2020, p. 235–241. [Online]. Available: <https://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0010142302350241>
- [25] J. Tomás, D. Rasteiro, and J. Bernardino, "Data anonymization: An experimental evaluation using open-source tools," *Future Internet*, vol. 14, no. 6, p. 167, May 2022. [Online]. Available: <https://www.mdpi.com/1999-5903/14/6/167>
- [26] [Online]. Available: <https://promtools.org/>
- [27] [Online]. Available: [https://edpb.europa.eu/edpb\\_en](https://edpb.europa.eu/edpb_en)
- [28] [Online]. Available: <https://www.iso.org/standards.html>
- [29] [Online]. Available: <https://gdpr.eu/>
- [30] [Online]. Available: <https://www.nen.nl/en/>
- [31] S. A. Sohail, J. Krabbe, P. de, A. Silva, and F. A. Bukhsh, "Privacy value modeling: A gateway to ethical big data handling."
- [32] [Online]. Available: <https://www.mst.nl/p/specialismen/radiologie/>

- [33] [Online]. Available: <https://www.zgt.nl/aandoening-en-behandeling/onze-specialismen/radiologie/>
- [34] [Online]. Available: <https://www.fakenamegenerator.com/>
- [35] [Online]. Available: <https://ugeo.urbistat.com/AdminStat/en/nl/demografia/dati-sintesi/enschede/23055684/4>
- [36] [Online]. Available: <https://ugeo.urbistat.com/AdminStat/en/nl/demografia/eta/enschede/23055684/4>
- [37] F. Mannhardt, N. Tax, D. Schunselaar, and E. Verbeek, “Den dolech 2, 5612 az eindhoven p.o. box 513, 5600 mb eindhoven the netherlands www.tue.nl.”
- [38] F. Nogueira, “Hands-on process mining: Event visualisation with prom,” Jan 2021. [Online]. Available: <https://laredoute.io/blog/hands-on-process-mining-event-visualisation-with-prom/>
- [39] F. Prasser, F. Kohlmayer, and K. Kuhn, “The importance of context: Risk-based de-identification of biomedical data,” *Methods of Information in Medicine*, vol. 55, no. 04, p. 347–355, 2016. [Online]. Available: <http://www.thieme-connect.de/DOI/DOI?10.3414/ME16-01-0012>
- [40] N. Mohammed, B. C. Fung, P. C. Hung, and C.-k. Lee, “Anonymizing healthcare data: a case study on the blood transfusion service,” in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. Paris France: ACM, Jun 2009, p. 1285–1294. [Online]. Available: <https://dl.acm.org/doi/10.1145/1557019.1557157>
- [41] S. J. J. Leemans and D. Fahland, “Process and deviation exploration with inductive visual miner,” Jan 2014.
- [42] S. Leemans, “Inductive visual miner manual,” 2017. [Online]. Available: <https://www.semanticscholar.org/paper/Inductive-visual-Miner-manual-Leemans-Prom/c09fd03d82bea9d2df50682ed4c220df21648005>
- [43] G. Schrijver, “Using process mining to compare different variants of the same reimbursement process: a case study.”

- [44] F. A. Yasmin, R. Bemthuis, M. Elhagaly, and F. A. Bukhsh, "A process mining starting guideline for process analysts and process owners: A practical process analytics guide using prom."
- [45] J. C. A. M. Buijs, B. F. Van Dongen, and W. M. P. Van Der Aalst, *On the Role of Fitness, Precision, Generalization and Simplicity in Process Discovery*, ser. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, vol. 7565, p. 305–322. [Online]. Available: [http://link.springer.com/10.1007/978-3-642-33606-5\\_19](http://link.springer.com/10.1007/978-3-642-33606-5_19)
- [46] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkatasubramanian, "L - diversity: Privacy beyond k -anonymity," *ACM Transactions on Knowledge Discovery from Data*, vol. 1, no. 1, p. 3, Mar 2007. [Online]. Available: <https://dl.acm.org/doi/10.1145/1217299.1217302>
- [47] N. Li, T. Li, S. Venkatasubramanian, and T. Labs, "t-closeness: Privacy beyond k-anonymity and -diversity."
- [48] M. E. Nergiz, M. Atzori, and C. Clifton, "Hiding the presence of individuals from shared databases," in *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*. Beijing China: ACM, Jun 2007, p. 665–676. [Online]. Available: <https://dl.acm.org/doi/10.1145/1247480.1247554>
- [49] A. Pika, M. T. Wynn, S. Budiono, A. H. Ter Hofstede, W. M. Van Der Aalst, and H. A. Reijers, "Privacy-preserving process mining in healthcare," *International Journal of Environmental Research and Public Health*, vol. 17, no. 5, p. 1612, Mar 2020. [Online]. Available: <https://www.mdpi.com/1660-4601/17/5/1612>
- [50] L. Sankar, S. R. Rajagopalan, and H. V. Poor, "Utility and privacy of data sources: Can shannon help conceal and reveal information?" in *2010 Information Theory and Applications Workshop (ITA)*. La Jolla, CA, USA: IEEE, Jan 2010, p. 1–7. [Online]. Available: <http://ieeexplore.ieee.org/document/5454092/>
- [51] S. N. von Voigt, S. A. Fahrenkrog-Petersen, D. Janssen, A. Koschmider, F. Tschorsch, F. Mannhardt, O. Landsiedel, and M. Weidlich, *Quantifying the Re-identification Risk of Event Logs for Process Mining*, 2020, vol. 12127, p. 252–267, arXiv:2003.10707 [cs]. [Online]. Available: <http://arxiv.org/abs/2003.10707>

- [52] S. A. Sohail, F. A. Bukhsh, and M. Van Keulen, "Multilevel privacy assurance evaluation of healthcare metadata," *Applied Sciences*, vol. 11, no. 22, p. 10686, Nov 2021. [Online]. Available: <https://www.mdpi.com/2076-3417/11/22/10686>
- [53] M. A. Kadampur, "A noise addition scheme in decision tree for privacy preserving data mining," vol. 2, no. 1, 2010.
- [54] H. Kupwade Patil and R. Seshadri, "Big data security and privacy issues in healthcare," in *2014 IEEE International Congress on Big Data*. Anchorage, AK: IEEE, Jun 2014, p. 762–765. [Online]. Available: <https://ieeexplore.ieee.org/document/6906856/>

# Patient Dataset and Event Logs Used

[\\*Click here to view the repository\\*](#)

# **Inductive Visual Miner Animation**

[\\*Click here to view the repository\\*](#)

# Additional Figures

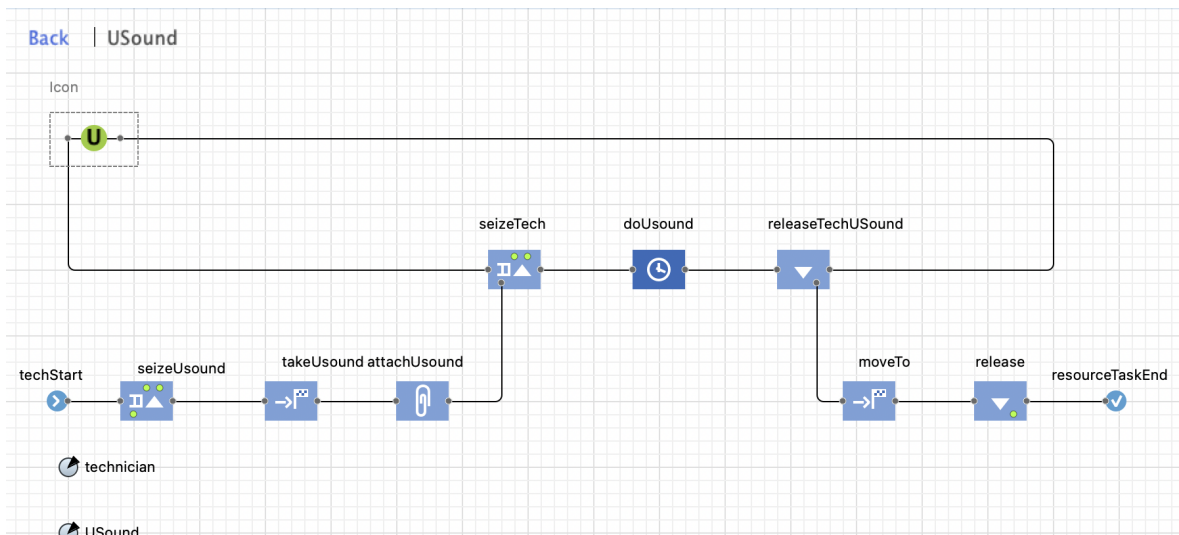


Figure C.1: Logic format of USound Process.

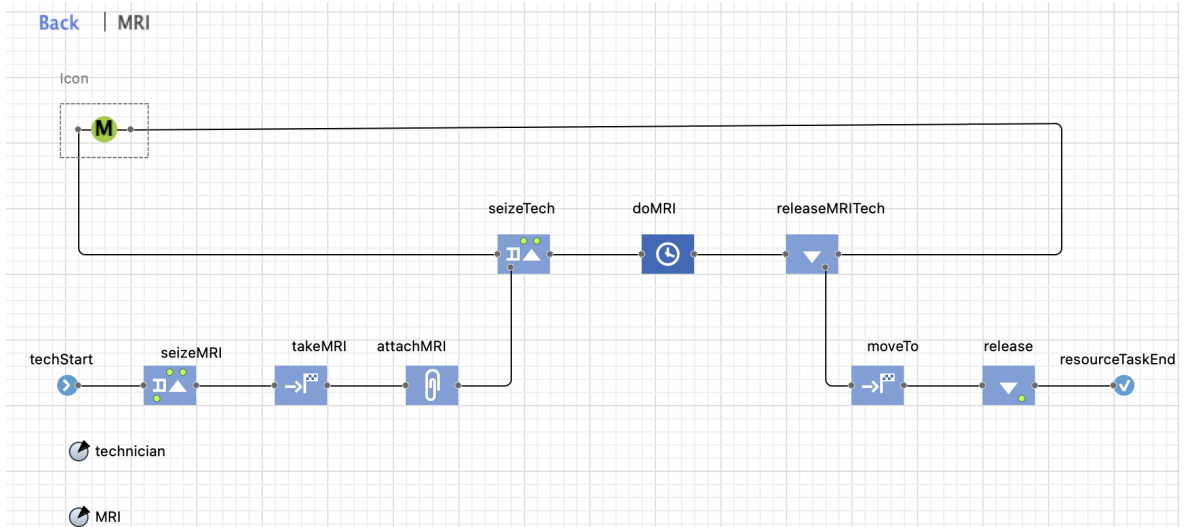


Figure C.2: Logic format of MRI Process.

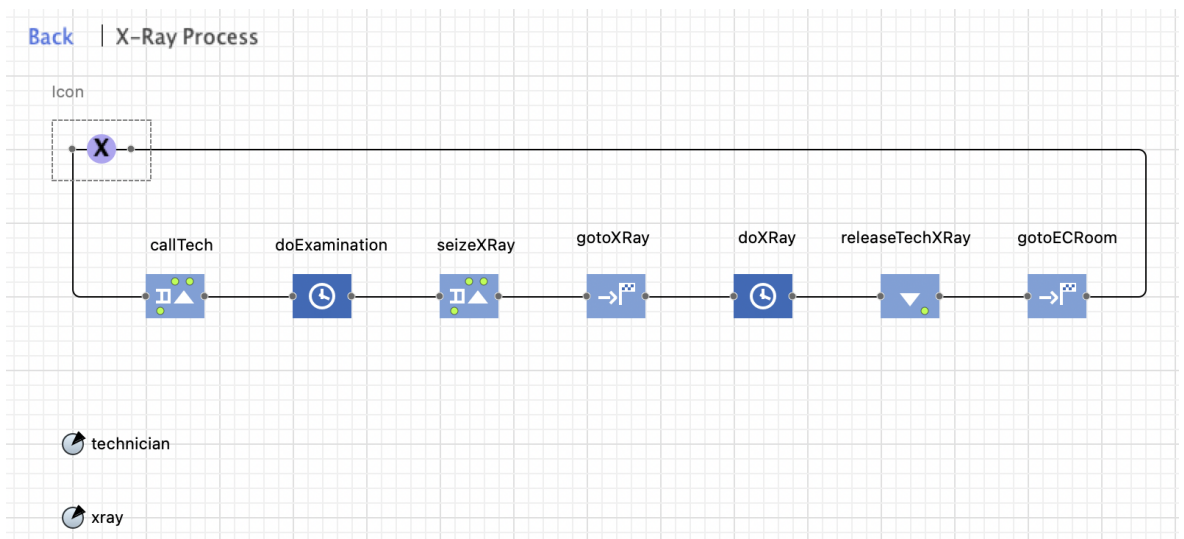


Figure C.3: Logic format of X-Ray Process.



agent	agent type	block type	block	activity type	start date	stop date	bloodtype	dob	gender	givenname	surname	telephonenumber
<population>[38] : 1193	Patient	Delay	registration	WORK	2023-06-16 5:09:57	2023-06-16 5:10:54	A*	[1940, 1950[	female	*	*	*
<population>[38] : 1193	Patient	Delay	triage	WORK	2023-06-16 5:36:42	2023-06-16 5:45:30	A*	[1940, 1950[	female	*	*	*
<population>[38] : 1193	Patient	Delay	xRayProcess.doExamination	WORK	2023-06-16 6:27:48	2023-06-16 6:31:59	A*	[1940, 1950[	female	*	*	*
<population>[38] : 1193	Patient	Delay	xRayProcess.doXRay	WORK	2023-06-16 6:39:30	2023-06-16 6:45:41	A*	[1940, 1950[	female	*	*	*
<population>[38] : 1193	Patient	MoveTo	gotoECRoom	MOVE	2023-06-16 6:17:50	2023-06-16 6:18:12	A*	[1940, 1950[	female	*	*	*
<population>[38] : 1193	Patient	MoveTo	gotoExit	MOVE	2023-06-16 6:45:55	2023-06-16 6:46:09	A*	[1940, 1950[	female	*	*	*
<population>[38] : 1193	Patient	MoveTo	gotoRegistration	MOVE	2023-06-16 5:09:54	2023-06-16 5:09:57	A*	[1940, 1950[	female	*	*	*
<population>[38] : 1193	Patient	MoveTo	gotoTriageRoom	MOVE	2023-06-16 5:36:31	2023-06-16 5:36:42	A*	[1940, 1950[	female	*	*	*
<population>[38] : 1193	Patient	MoveTo	gotoWaitingRoom1	MOVE	2023-06-16 5:10:54	2023-06-16 5:10:59	A*	[1940, 1950[	female	*	*	*
<population>[38] : 1193	Patient	MoveTo	gotoWaitingRoom2	MOVE	2023-06-16 5:45:30	2023-06-16 5:45:42	A*	[1940, 1950[	female	*	*	*
<population>[38] : 1193	Patient	MoveTo	xRayProcess.gotoECRoom	MOVE	2023-06-16 6:45:41	2023-06-16 6:45:55	A*	[1940, 1950[	female	*	*	*
<population>[38] : 1193	Patient	MoveTo	xRayProcess.gotoXRay	MOVE	2023-06-16 6:39:16	2023-06-16 6:39:30	A*	[1940, 1950[	female	*	*	*
<population>[38] : 1193	Patient	Seize	callIPA	WAIT	2023-06-16 6:18:12	2023-06-16 6:18:33	A*	[1940, 1950[	female	*	*	*
<population>[38] : 1193	Patient	Seize	seizeECRoom	WAIT	2023-06-16 5:45:42	2023-06-16 6:17:50	A*	[1940, 1950[	female	*	*	*
<population>[38] : 1193	Patient	Seize	seizeTriageRoom	WAIT	2023-06-16 5:10:59	2023-06-16 5:36:31	A*	[1940, 1950[	female	*	*	*
<population>[38] : 1193	Patient	Seize	xRayProcess.callTech	WAIT	2023-06-16 6:18:33	2023-06-16 6:27:48	A*	[1940, 1950[	female	*	*	*
<population>[38] : 1193	Patient	Seize	xRayProcess.seizeXRay	WAIT	2023-06-16 6:31:59	2023-06-16 6:39:16	A*	[1940, 1950[	female	*	*	*
<population>[32] : 1010	Patient	Delay	registration	WORK	2023-06-15 6:10:06	2023-06-15 6:11:06	A*	[1950, 1960[	female	*	*	*
<population>[32] : 1010	Patient	Delay	triage	WORK	2023-06-15 6:14:00	2023-06-15 6:22:17	A*	[1950, 1960[	female	*	*	*
<population>[32] : 1010	Patient	Delay	uSoundProcess.doUsound	WORK	2023-06-15 6:47:21	2023-06-15 7:15:31	A*	[1950, 1960[	female	*	*	*
<population>[32] : 1010	Patient	MoveTo	gotoECRoom	MOVE	2023-06-15 6:27:03	2023-06-15 6:27:28	A*	[1950, 1960[	female	*	*	*
<population>[32] : 1010	Patient	MoveTo	gotoExit	MOVE	2023-06-15 7:15:31	2023-06-15 7:15:45	A*	[1950, 1960[	female	*	*	*
<population>[32] : 1010	Patient	MoveTo	gotoRegistration	MOVE	2023-06-15 6:10:03	2023-06-15 6:10:06	A*	[1950, 1960[	female	*	*	*
<population>[32] : 1010	Patient	MoveTo	gotoTriageRoom	MOVE	2023-06-15 6:13:48	2023-06-15 6:14:00	A*	[1950, 1960[	female	*	*	*
<population>[32] : 1010	Patient	MoveTo	gotoWaitingRoom1	MOVE	2023-06-15 6:11:06	2023-06-15 6:11:12	A*	[1950, 1960[	female	*	*	*
<population>[32] : 1010	Patient	MoveTo	gotoWaitingRoom2	MOVE	2023-06-15 6:22:17	2023-06-15 6:22:29	A*	[1950, 1960[	female	*	*	*
<population>[32] : 1010	Patient	Seize	callIPA	WAIT	2023-06-15 6:27:28	2023-06-15 6:27:51	A*	[1950, 1960[	female	*	*	*
<population>[32] : 1010	Patient	Seize	seizeECRoom	WAIT	2023-06-15 6:22:29	2023-06-15 6:27:03	A*	[1950, 1960[	female	*	*	*
<population>[32] : 1010	Patient	Seize	seizeTriageRoom	WAIT	2023-06-15 6:11:12	2023-06-15 6:13:48	A*	[1950, 1960[	female	*	*	*
<population>[32] : 1010	Patient	Seize	uSoundProcess.seizeTech	WAIT	2023-06-15 6:27:51	2023-06-15 6:47:21	A*	[1950, 1960[	female	*	*	*

Figure C.4: Columns overview of Anonymized Event Log.

<population>[33] : 1306	Patient	Delay	registration	WORK	2023-06-16 19:28:07	2023-06-16 19:29:04	*	*	*	*	*	*
<population>[33] : 1306	Patient	Delay	triage	WORK	2023-06-16 19:29:28	2023-06-16 19:39:14	*	*	*	*	*	*
<population>[33] : 1306	Patient	Delay	uSoundProcess.doUsound	WORK	2023-06-16 20:13:56	2023-06-16 20:35:21	*	*	*	*	*	*
<population>[33] : 1306	Patient	MoveTo	gotoECRoom	MOVE	2023-06-16 19:53:48	2023-06-16 19:54:07	*	*	*	*	*	*
<population>[33] : 1306	Patient	MoveTo	gotoExit	MOVE	2023-06-16 20:35:21	2023-06-16 20:35:34	*	*	*	*	*	*
<population>[33] : 1306	Patient	MoveTo	gotoRegistration	MOVE	2023-06-16 19:28:04	2023-06-16 19:28:07	*	*	*	*	*	*
<population>[33] : 1306	Patient	MoveTo	gotoTriageRoom	MOVE	2023-06-16 19:29:18	2023-06-16 19:29:28	*	*	*	*	*	*
<population>[33] : 1306	Patient	MoveTo	gotoWaitingRoom1	MOVE	2023-06-16 19:29:04	2023-06-16 19:29:12	*	*	*	*	*	*
<population>[33] : 1306	Patient	MoveTo	gotoWaitingRoom2	MOVE	2023-06-16 19:39:14	2023-06-16 19:39:24	*	*	*	*	*	*
<population>[33] : 1306	Patient	Seize	callIPA	WAIT	2023-06-16 19:54:07	2023-06-16 19:54:30	*	*	*	*	*	*
<population>[33] : 1306	Patient	Seize	seizeECRoom	WAIT	2023-06-16 19:39:24	2023-06-16 19:53:48	*	*	*	*	*	*
<population>[33] : 1306	Patient	Seize	seizeTriageRoom	WAIT	2023-06-16 19:29:12	2023-06-16 19:29:18	*	*	*	*	*	*
<population>[33] : 1306	Patient	Seize	uSoundProcess.seizeTech	WAIT	2023-06-16 19:54:30	2023-06-16 20:13:56	*	*	*	*	*	*
<population>[31] : 1022	Patient	Delay	registration	WORK	2023-06-15 7:43:44	2023-06-15 7:44:41	*	*	*	*	*	*
<population>[31] : 1022	Patient	Delay	triage	WORK	2023-06-15 7:45:08	2023-06-15 7:54:33	*	*	*	*	*	*
<population>[31] : 1022	Patient	Delay	uSoundProcess.doUsound	WORK	2023-06-15 8:07:43	2023-06-15 8:32:54	*	*	*	*	*	*
<population>[31] : 1022	Patient	MoveTo	gotoECRoom	MOVE	2023-06-15 7:55:02	2023-06-15 7:55:26	*	*	*	*	*	*
<population>[31] : 1022	Patient	MoveTo	gotoExit	MOVE	2023-06-15 8:32:54	2023-06-15 8:33:09	*	*	*	*	*	*
<population>[31] : 1022	Patient	MoveTo	gotoRegistration	MOVE	2023-06-15 7:43:42	2023-06-15 7:43:44	*	*	*	*	*	*
<population>[31] : 1022	Patient	MoveTo	gotoTriageRoom	MOVE	2023-06-15 7:44:55	2023-06-15 7:45:08	*	*	*	*	*	*
<population>[31] : 1022	Patient	MoveTo	gotoWaitingRoom1	MOVE	2023-06-15 7:44:41	2023-06-15 7:44:46	*	*	*	*	*	*
<population>[31] : 1022	Patient	MoveTo	gotoWaitingRoom2	MOVE	2023-06-15 7:54:33	2023-06-15 7:54:46	*	*	*	*	*	*
<population>[31] : 1022	Patient	Seize	callIPA	WAIT	2023-06-15 7:55:26	2023-06-15 7:55:48	*	*	*	*	*	*
<population>[31] : 1022	Patient	Seize	seizeECRoom	WAIT	2023-06-15 7:54:46	2023-06-15 7:55:02	*	*	*	*	*	*
<population>[31] : 1022	Patient	Seize	seizeTriageRoom	WAIT	2023-06-15 7:44:46	2023-06-15 7:44:55	*	*	*	*	*	*
<population>[31] : 1022	Patient	Seize	uSoundProcess.seizeTech	WAIT	2023-06-15 7:55:48	2023-06-15 8:07:43	*	*	*	*	*	*

Figure C.5: Remaining 26 Patients with full Suppression for all Attributes.

Log attributes		Trace attributes
Attributes at the log level.		
Attribute ▲	value	
concept:name	GroupedbyEl.csv	
fitness	0.993	
number of events	13800	
number of traces	500	

**Figure C.6:** Fitness Score of Unanonymized Event Log.

Attributes at the log level.		
Attribute ▲	value	
concept:name	k-anonymity-anonymized...	
fitness	0.989	
number of events	13800	
number of traces	500	

**Figure C.7:** Fitness Score of Anonymized Event Log.



**Figure C.8:** Event Log Dashboard in ProM.

Log Summary		
Total number of process instances: 500		
Total number of events: 13800		
Event Name		
Event classes defined by Event Name		
All events		
Total number of classes: 22		
Class	Occurrences (absolute)	Occurrences (relative)
gotoWaitingRoom1	1000	7.246%
gotoWaitingRoom2	1000	7.246%
seizeTriageRoom	1000	7.246%
gotoTriageRoom	1000	7.246%
gotoRegistration	1000	7.246%
callIPA	1000	7.246%
gotoECRoom	1000	7.246%
registration	1000	7.246%
gotoExit	1000	7.246%

Figure C.9: Log Summary in ProM.

Start events		
Total number of classes: 1		
Class	Occurrences (absolute)	Occurrences (relative)
gotoRegistration+start	500	100.0%
End events		
Total number of classes: 1		
Class	Occurrences (absolute)	Occurrences (relative)
gotoExit+complete	500	100.0%

Figure C.10: Start & End Events of Log Summary in ProM.

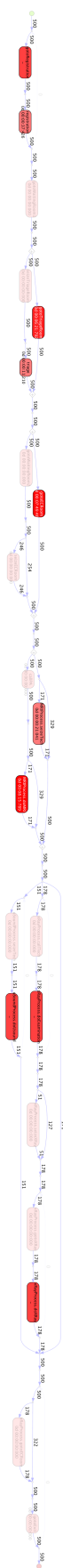


Figure C.11: Waiting Times of Unanonymized Event Log.

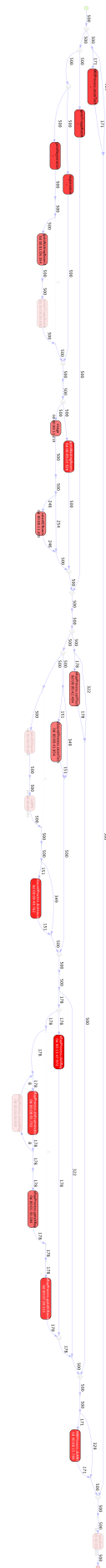


Figure C.12: Waiting Times of Anonymized Event Log.

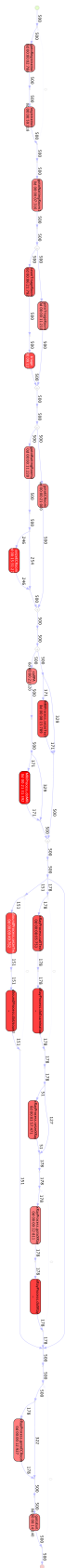


Figure C.13: Service Times of Unanonymized Event Log.

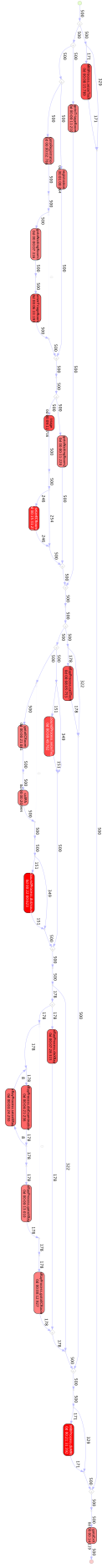
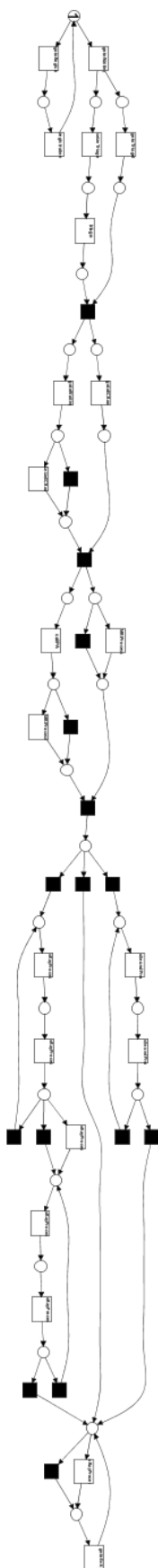
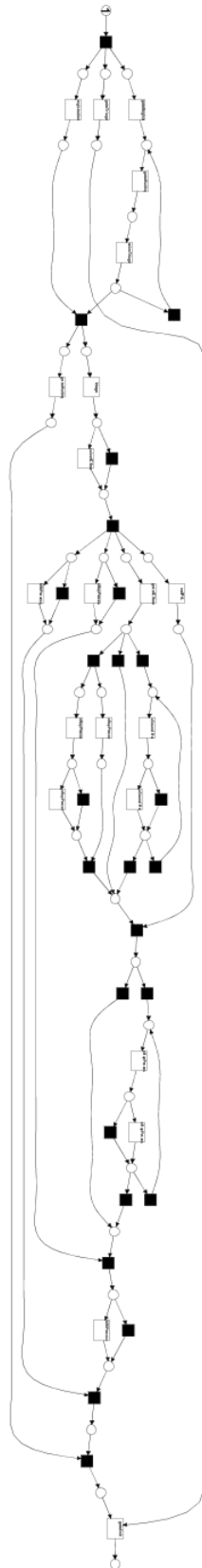


Figure C.14: Service Times of Anonymized Event Log.



**Figure C.15:** Petri Net of Unanonymized Event Log.





**Figure C.16:** Petri Net of Anonymized Event Log.

# Python Code

```
1 import pandas as pd
2
3 path = '/content/drive/MyDrive/Filter_EL.csv'
4
5 df = pd.read_csv(path)
6
7 # Pivot the dataset
8 pivoted_df = df.pivot(index='agent', columns='parameter_name', values=
    'parameter_value')
9
10 # Reset the index
11 pivoted_df.reset_index(inplace=True)
12
13 # Merge with the other columns
14 merged_df = pd.merge(df.drop(['parameter_name', 'parameter_value'],
    axis=1), pivoted_df, on='agent')
15
16 merged_df.drop_duplicates(subset='agent', keep='first', inplace=True)
17
18 # Save the reformatted dataset to a CSV file
19 merged_df.to_csv('reformatted_dataset.csv', index=False)
20
21 # Print the reformatted dataset
22 print(pivoted_df)
```

**Listing D.1:** Python Code Example.

# SQL Query

```
1 SELECT
2     PUBLIC.AL_FORMAT_AGENT_TYPE_NAME_LOG(AGENT_TYPES.NAME) AS
3     AGENT_TYPE ,
4     PUBLIC.AL_FORMAT_AGENT_NAME_LOG(AGENTS.NAME , AGENTS.ID) AS AGENT ,
5     PUBLIC.AL_FORMAT_AGENT_TYPE_NAME_LOG(BLOCK_TYPES.NAME) AS
6     BLOCK_TYPE ,
7     PUBLIC.AL_FORMAT_AGENT_NAME_LOG(BLOCKS.NAME , BLOCKS.ID) AS BLOCK ,
8     FLOWCHART_PROCESS_STATES.ACTIVITY_TYPE ,
9     FLOWCHART_PROCESS_STATES.START_DATE ,
10    FLOWCHART_PROCESS_STATES.STOP_DATE ,
11    PARAMETERS.PARAMETER_NAME AS PARAMETER_NAME ,
12    PARAMETERS.PARAMETER_VALUE AS PARAMETER_VALUE
13 FROM
14    PUBLIC.FLOWCHART_PROCESS_STATES_RAW_LOG FLOWCHART_PROCESS_STATES
15    INNER JOIN PUBLIC.AGENTS_RAW_LOG AGENTS ON
16        FLOWCHART_PROCESS_STATES.AGENT_ID = AGENTS.ID
17    INNER JOIN PUBLIC.AGENT_TYPES_RAW_LOG AGENT_TYPES ON AGENTS.
18        AGENT_TYPE_ID = AGENT_TYPES.ID
19    INNER JOIN PUBLIC.AGENTS_RAW_LOG BLOCKS ON
20        FLOWCHART_PROCESS_STATES.BLOCK_ID = BLOCKS.ID
21    INNER JOIN PUBLIC.AGENT_TYPES_RAW_LOG BLOCK_TYPES ON BLOCKS.
22        AGENT_TYPE_ID = BLOCK_TYPES.ID
23    INNER JOIN PUBLIC.AGENT_PARAMETERS_RAW_LOG PARAMETERS ON
24        PARAMETERS.AGENT_ID = AGENTS.ID
25 WHERE
26    AGENT_TYPES.NAME NOT LIKE 'com.anylogic.libraries.%'
```

```
20 ORDER BY
21 FLOWCHART_PROCESS_STATES . START_DATE ;
```

**Listing E.1:** SQL Query.