p–ISSN: 2723 - 6609 e-ISSN: 2745-5254
Vol. 6, No. 1 January 2025 http://jist.publikasiindonesia.id/

Analysis of the Quality of Patient Treatment Data in MIMIC-IV Using Alpha, Heuristic, and Inductive Miner

Muhammad Reynaldi Mujantara[1*], Angelina Prima Kurniati[2]
Universitas Telkom Bandung, Indonesia[1,2]
Email: reynaldimujantara@student.telkomuniversity.ac.id[1], angelina@telkomuniversity.ac.id[2]

*Correspondence

Keywords: Alpha Miner; Conformance Checking; Data Quality; Healthcare Heuristic Miner; Inductive Miner; Process Discovery; Process Mining

ABSTRACT

MIMIC-IV (Medical Information Mart for Intensive Care IV), is a medical dataset used for research in the fields of medicine and health computer science. This dataset contains health information collected from intensive care units (ICUs) at the Beth Israel Deaconess Medical Center in Boston, MA. This research uses process mining to analyze data quality of patient treatment in the MIMIC-IV dataset using the Alpha Miner, Heuristic Miner, and Inductive Miner algorithms. It begins with planning and justification, followed by database reconstruction, data quality assessment, and data extraction, leading to the development of a control flow model. Subsequently, conformance checking is performed, and the study concludes with an evaluation of the results. It is expected that the results of this study will provide a better understanding of the quality of patient care process data in the MIMIC-IV dataset and a positive contribution to developing more effective health services.

Introduction

Optimization of business processes in healthcare organizations is a crucial need to respond to economic pressures, improve service quality, and reduce the risk of technical errors that potentially threaten patient safety. (Alter, 2015). This study aims to improve the efficiency of health data analysis by implementing process mining. This research focuses on checking the data quality of the MIMIC-IV dataset and proceeding to the Process Mining stage. The approach used is the Process Mining Project Methodology (*PM2*) using the inductive miner algorithm (Garg & Agarwal, 2016), alpha miner (Sundari & Nayak, 2020), and heuristic miner (Rebuge & Ferreira, 2012). This study highlights the evaluation of process mining algorithms. Each process mining algorithm has specific advantages in handling its tasks. Therefore, choosing the correct algorithm is very important. This study uses a new framework to observe the results of Alpha Miner, Heuristic, and Inductive Miner algorithms (Fox et al., 2018).

The application of process mining techniques in business analytics, particularly in the healthcare sector, has gained increasing significance due to the inherent complexity, variability, and patient-centered nature of healthcare processes. Process-Oriented Data Science for Healthcare (PODS4H) Alliance was established to promote the research and implementation of process mining methodologies aimed at driving data-driven improvements in healthcare workflows (Fox et al., 2018). Process mining, a rapidly growing discipline, focuses on extracting insights from data stored in information systems, particularly event records, to improve process understanding and efficiency. In healthcare, processes' dynamic and multidisciplinary nature provides unique opportunities for process mining applications. These applications, from process discovery to conformance checking and improvement, enable healthcare institutions to improve service quality, reduce costs, and enhance overall process management. (Perimal-Lewis et al.,

2016). This study aims to identify and analyze the results of three process mining algorithms: Alpha Miner, Heuristic Miner, and Inductive Miner. This study utilizes the ProM tool to run the three algorithms so that the results obtained can be evaluated and analyzed in depth. The results of these three methods are expected to provide meaningful insights into the dataset (Johnson et al., 2023).

In the contet of healthcare, an accurate understanding of workflows or care flows is essential. This study applies process mining techniques using Petri nets to real-time data from a private community hospital. This study aims to obtain information and knowledge about the flow through control flow analysis, including finding the path of certain patient groups. This approach uses the alpha miner, heuristic miner (Rebuge & Ferreira, 2012), and inductive miner (Garg & Agarwal, 2016) methods. By comparing the results of these three algorithms, it can add insight into data quality and determine the process model that best suits the needs. The study aims to provide insights that can improve hospital performance and reduce unnecessary medical costs (Rojas et al., 2016). This study applies process mining techniques to evaluate the quality of time-based hospital performance metrics data from ED electronic health records. By building a patient journey model, the study aims to identify flow patterns and improve the quality of care. The survey results indicate that process mining can be a viable methodology for evaluating hospital performance metrics and data quality, allowing for appropriate corrective action (De Weerdt et al., 2013). A structured approach is needed to implement process mining research on electronic health records (EHR). This study utilizes the Care Pathway Data Quality Framework (CP-DQF) for data quality analysis. This framework manages data quality within process mining, with a case study focused on medical records. It offers solutions to identify and address data quality issues, making it a valuable tool for researchers in the healthcare sector (Bolt et al., 2016) The application of process mining in the healthcare sector, using CP-DQF, not only aids in understanding optimal procedures but also enhances service efficiency, improves medical facility management, identifies resource and patient behavior patterns, provides recommendations for design changes, facilitates performance analysis, and reduces service waiting times (Perimal-Lewis et al., 2016).

In electronic health records (EHR), data quality is a very important aspect to obtain research results (Bolt et al., 2016). Handling complex data quality issues in the context of EHR is often not clearly understood. To overcome this challenge, the use of new technologies such as process mining is relevant to improve understanding of the care pathway. However, the successful implementation of this technology requires special attention to strategies and methods to improve data quality (Bolt et al., 2016). This study proves that Process Mining is a feasible methodology to be used to assess data quality in hospitals. The results of the study also provide insights that allow for appropriate actions to address data quality issues (De Weerdt et al., 2013).

In this study, researchers attempted to conduct an in-depth analysis of data quality in the patient care process using the MIMICIV dataset. MIMIC-IV, as a credible source of health data, provides invaluable information related to patient medical records in clinical settings. To comprehensively uncover and understand the documented workflow, researchers chose to apply three methods at once, namely Alpha Miner, Heuristic Miner, and Inductive Miner. The use of three methods, Alpha Miner, Heuristic Miner, and Inductive Miner, aims to achieve a holistic view of the patient care process while identifying potential discrepancies or anomalies in the data that could affect its quality. Alpha Miner detects control patterns underlying the sequence of activities, Heuristic Miner provides heuristic rule-based insights, and Inductive Miner explores process structures that may not be directly visible. Through this in-depth data quality analysis, this study aims to contribute significantly to a deeper understanding of the dynamics of the patient

care process in the healthcare environment. The findings of the study also provide insights that enable appropriate actions to address data quality issues.

Related Work

A. Mimic-IV

MIMIC-IV is a publicly released EHR dataset, offering several advantages over previous datasets. First, MIMIC-IV includes contemporary information reflecting the evolution of clinical practice over the last decade (Johnson et al., 2023). Second, this dataset integrates new digital information sources, such as electronic medication administration records, increasing the accuracy and completeness of the data. MIMIC-IV results from a collaboration between Beth Israel Deaconess Medical Center (BIDMC) and Massachusetts Institute of Technology (MIT), with data de-identified and made available for research following ethics approval and a data use agreement. The process involves the stages of acquisition, transformation, and deidentification (Johnson et al., 2023). The dataset structure can be seen in Figure 1.

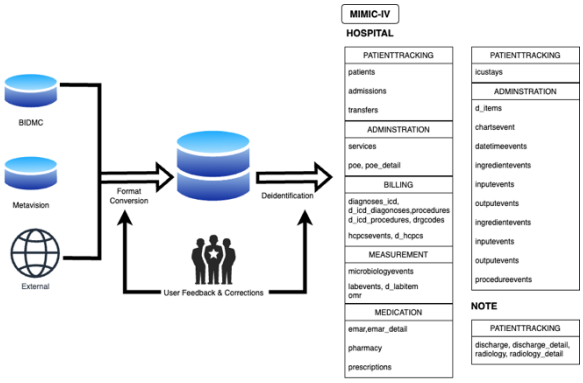

Figure 1. Dataset Structure

The MIMIC development process went through several stages. First, data were obtained from the BIDMC data warehouse, the ICU information system (MetaVision), and external sources of data collection (Johnson et al., 2023). Next, using Structured Query Language (SQL), multiple data sources are combined into a single schema (transformation). Then, a deidentification algorithm is applied to remove protected health information from the reformatted schema selectively. The MIMIC Code Repository evaluates and improves the development process as needed (Johnson et al., 2023). The table relationship structure can be seen in Figure 2.

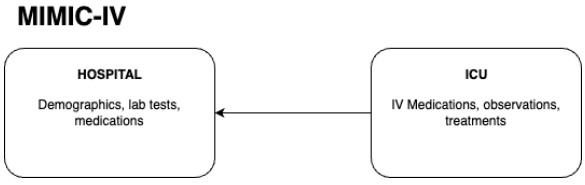

Figure 2. Table Relationship Structure

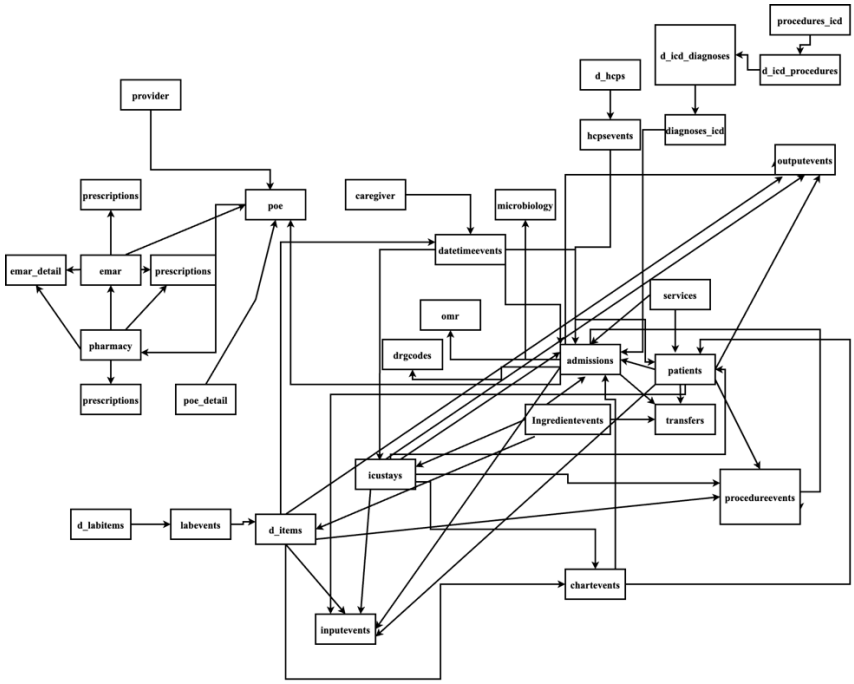

Figure 3. Relationship diagram of MIMIC-IV

MIMIC-IV follows a relationship structure between tables. This structure can be linked through identifiers such as subject_id, hadm_id, and deidentified date and time (Johnson et al., 2023). The relationship diagram of MIMIC-IV can be seen in Figure 3.

B. Process Mining

The main goal of process mining is to improve the efficiency of data-driven operational processes by providing analysis techniques for event data generated during process execution. Information on the results of the process implementation is stored in the relevant organizational information system (Rojas et al., 2016). Process mining is a technique that bridges the gap between data analysis and business process management by analyzing event logs generated by information systems. It typically involves three main stages: discovery, where the actual process model is automatically constructed from event logs; conformance checking, which compares the discovered process with the intended model to identify deviations; and enhancement, where the process model is refined to improve performance or align more closely with reality. Process mining is widely used to gain insights into how business processes are carried out, identify inefficiencies, ensure compliance with regulations, and support continuous improvement initiatives (De Weerdt et al., 2013). An illustration of process mining is provided in Figure 4.

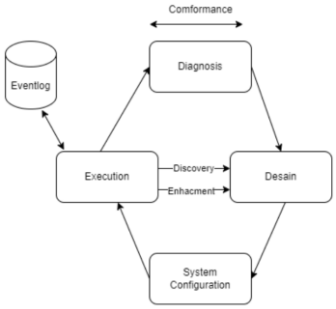

Figure 3. Process Mining (De Weerdt et al., 2013)

C. Data Quality

Data quality (DQ) plays a vital role in health research, especially in using electronic health records (EHRs), because it can influence the evaluation and improvement of hospital health processes. DQ covers several key aspects, such as accuracy, completeness, consistency, integrity, standardization, availability, usability, and timeliness of health data (Perimal-Lewis et al., 2016). For example, the timeliness of health data ensures that relevant information is available in sufficient time to support appropriate decision-making in clinical practice. Meanwhile, data completeness and consistency are essential in providing accurate and representative analysis of patient populations. At the same time, the integrity and accuracy of health data avoid errors that can influence research results and clinical decisions. Therefore, efforts to ensure good DQ in the use of EHRs not only increase confidence in research results and support the development of more effective and efficient clinical practice. The Comprehensive Data Quality Framework (CP-DQF) approach using the Plan-Do-Study-Act (PDSA) cycle has proven effective in ensuring high data quality in this research. Data quality can be improved through this continuous cycle so that research results become more accurate and reliable.

Methods

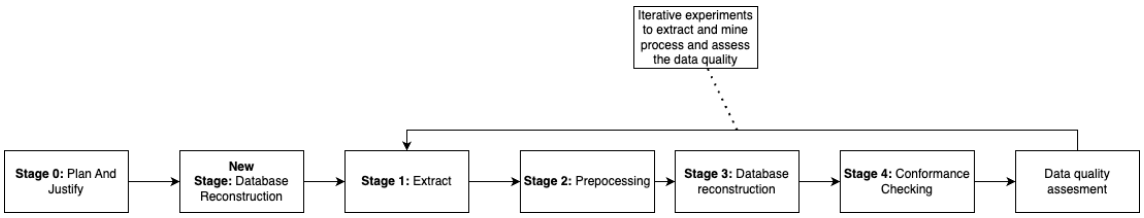

Figure 4. Modeling Flow

Figure 5 illustrates the design and modeling process, beginning with planning and justification, followed by database reconstruction, data quality assessment, and data extraction, Resulting in the development of a control flow model. Subsequently, conformance checking is conducted (Suriadi et al., 2017). This process involves interrelated steps, from planning to data quality assessment (Kurniati et al., 2019).

Plan and Justify

In the plan and justify phase, the rationale for this research is based on the recognition of a need for a high-quality, publicly accessible research dataset that can be leveraged internationally to advance process mining methodology development. (Alter, 2017b). After the category identification process is complete, the next step is to identify research questions, where we need to formulate questions that can be answered using log data (Alter, 2017a)

Database Reconstruction

Database reconstruction involves rebuilding the structure and content of a database through analysis of business process data (Schuster et al., 2022). In modern business, where transactions are executed automatically, and systems are integrated, this type of data becomes the basis for reporting information. Database reconstruction focuses on reorganizing, cleaning, and consolidating data from various sources to ensure consistency and accuracy. This process includes correcting data errors, eliminating redundancies, filling in missing information, and restructuring the data to better reflect the actual processes. The objective is to create a well-structured and reliable dataset that enables more accurate analysis and supports decision-making based on solid data foundations (Schuster et al., 2022).

1. Data Quality

Data quality (DQ) is crucial in health research, especially in the use of electronic health records (EHRs) to assess and improve hospital health processes. DQ includes the accuracy, completeness, consistency, integrity, standardization, availability, usability, and timeliness of health data (Perimal-Lewis et al., 2016).

- Accuracy: The level of conformity of the data representation to the actual situation.
- Completeness: The extent to which the data includes the expected values.
- Consistency: Uniformity of data format, structure, and values.
- Integrity: Sustainability and integrity of data.
- Standardization: Harmonization of data formats and structures.
- Availability: Availability of data for decision-making.
- Usability: Ease of using and interpreting data.
- Timeliness: Availability of data at a suitable time.

Data quality (DQ) issues in process mining often involve missing, incorrect, inaccurate, or irrelevant data (Bolt et al., 2016). To address these challenges, the Comprehensive Data Quality Framework (CP-DQF) employs the Plan-Do-Study-Act (PDSA) cycle to ensure data quality in research. This process begins with planning quality-related questions and identifying relevant data quality dimensions. The plan is then implemented through data collection and processing. The results are subsequently analyzed to assess whether the data meets the expected quality standards. Finally, corrective actions are taken to resolve any identified data quality issues.

Preprocessing

Preprocessing is a critical stage that aims to improve the quality and representation of data in the context of data mining and process mining [22]. Essential steps involve data cleaning to handle missing or incomplete values and abnormal values, dimensionality reduction to reduce irrelevant attributes, and normalization or standardization to ensure consistency of attribute scales. This process is crucial in preparing data to support the effectiveness and reliability of modelling algorithms, thus facilitating success in knowledge process mining [22].

Create Control Flow Model

The next step is to develop a control flow model and link it to the event log. In this example, three main plugins are used: first, ‘Convert CSV to XES’ to convert the event log into the XES format required by ProM; second, ‘Add Artificial Events >> START and END Events’ to address the absence of explicit start and end events in the log; and third, ‘Perform Analysis’ using three main methods Alpha Miner, Heuristic Miner, and Inductive Miner—as commonly used discovery algorithms. The analysis is applied to data from the four units with the highest patient volumes, starting with the initial room entry, followed by the procedures performed, and ending with the final room from which the patient was transferred. In the initial phase, the data is filtered to focus on the four units with the highest patient admissions: the Cardiovascular Intensive Care Unit, the Medical Intensive Care Unit, the Surgical Intensive Care Unit, and the Trauma Surgical Intensive Care Unit (Trauma SICU). The entire process is designed to ensure a comprehensive and accurate analysis of patient flow through these critical units (Alter, 2017b)

Process Mining

Process mining is a very new and developing technology that provides deep insight into the business processes occurring within an organization. Process mining is based on the use of data mining terms and technologies to analyze processes (Bolt et al., 2016). Process mining is a

technology that allows us to extract information or knowledge about processes from data sets and develop them for further analysis. Process mining begins by collecting data from information systems and automatically recording various organizational transactions and activities. These information systems log each activity as event logs containing digital traces such as timestamps, actors, and activities associated with specific processes. Once the event logs are generated, they serve as input for the process mining analysis. The first step in process mining is to extract and process the event logs to build process models. These models depict how workflows or business processes unfold within the organization. Organizations can effectively analyze, optimize, and improve their workflows by utilizing the process models derived from event logs. Figure 6 illustrates how event logs are created and used in the process mining procedure (Garg & Agarwal, 2016).

Figure 5. Process Mining

2. Process Discovery

Process discovery is a central element in process mining, starting from event data, where the discovery algorithm examines a process model that reflects the sequence of events as recorded in existing event data (Mayr et al., 2022). Figure 7, illustrates how this process works. This is a significant first step in process mining, where algorithms automatically understand and reconstruct process paths based on information in event data. Data-driven process discovery assumes that event data is the most objective representation of a process (Suriadi et al., 2017).

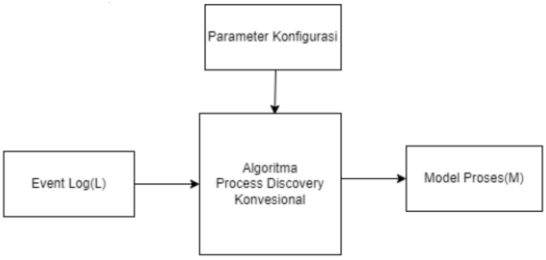

Figure 6. Process Discovery

2.1 Alpha Miner

Alpha Miner is a Process Mining algorithm that extracts process models from event logs by identifying sequences of activities and relationships, constructing Petri net models that represent healthcare workflows based on patient event logs, which is crucial for visualizing process flow and improving the quality of healthcare delivery, as illustrated in Figure 8, where the Alpha Miner methodology is detailed step by step, from event log input to Petri net model construction (Sundari & Nayak, 2020).

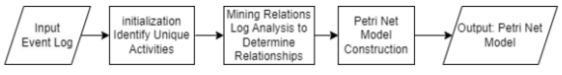

Figure 7. Alpha Miner

2.2 Inductive Miner

The Inductive Miner is a Process Mining algorithm that focuses on producing highly accurate and complete process models by identifying structured patterns within event logs. Figure 9 illustrates how this algorithm works. Its primary advantage lies in its ability to handle complex healthcare processes, ensuring that even the most intricate care pathways are modeled accurately (Jans et al., 2021). The Inductive Miner utilizes a block structure to enhance model accuracy, making it especially beneficial for healthcare settings where precise process representation is crucial for improving patient care. By filtering out noise and focusing on the most relevant activities, it ensures that healthcare workflows are both comprehensive and easy to analyze, allowing for better decision-making and process optimization (Bogarín et al., 2018).

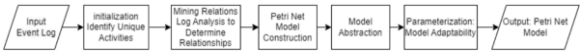

Figure 8. Inductive Miner

2.3 Heuristic Miner

A Heuristic Miner is an algorithm in Process Mining that overcomes obstacles in obtaining process models from activity records. Figure 10 provides a visual representation of how this algorithm works. It is effective for complex processes, low structure, and the presence of noise. This algorithm uses a split/joins frequency table to ensure an easy-to-understand model. It is implemented in the ProM framework and has proven successful for low-noise structured processes (Evermann et al., 2016). Heuristic Miner infers process structures based on event sequences, prioritizes frequently occurring patterns, is suitable for noisy data, and captures implicit knowledge in business processes (Nuritha & Mahendrawathi, 2017). This algorithm balances simplicity and effectiveness, making it a valuable tool for modeling complex and varied business processes (Bogarín et al., 2018).

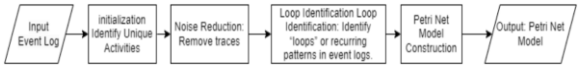

Figure 9. Heuristic Miner

3. Conformance Checking

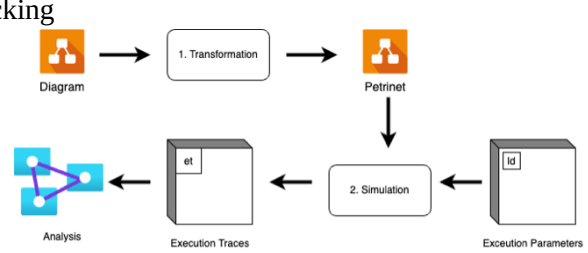

Figure 10. Conformance Checking

Conformance checking in Process Mining is an evaluation of the extent to which the actual behavior of a process conforms to the desired model or standard (Rubin et al., 2014). Figure 11 illustrates this comparison. This involves comparing the observed process execution from activity logs with the corresponding process model to identify deviations or nonconformities (Alter, 2017a). The aim is to assess the extent of correspondence between modeled and observed behavior, providing insight into the effectiveness and efficiency of the implemented processes. Conformance checking is important in various aspects, including

process diagnostics, compliance verification, process model improvement, and prediction-based business monitoring. Research in conformance checking explores new techniques and methods to improve the accuracy, efficiency, and applicability of process conformance assessment in various domains (Alter, 2017b).

Conformance checking can use various tools, including ProM. ProM allows the comparison of multiple model size evaluations based on four aspects of quality strength (Kurniati et al., 2019). An illustrative diagram depicting this comparison is presented in Figure 12.

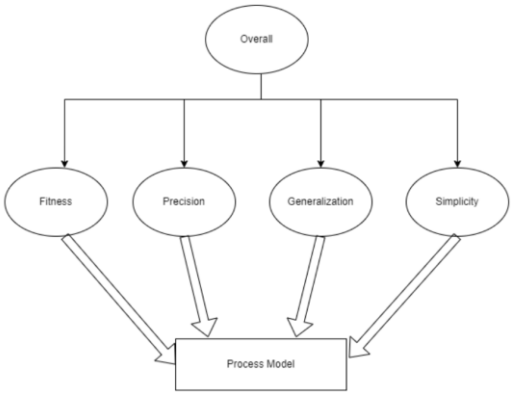

Figure 11. Conformance Checking

3.1 Fitness

Fitness in conformance checking measures the extent to which the model can reproduce the behaviour observed in the event log. This indicates the extent to which the model agrees with the execution traces recorded in the logs (Burattin et al., 2016). The extent to which the model that has been formed can accurately reproduce the cases recorded in the log can be measured (Kurniati et al., 2019).

$$\text{fitness}(\text{L}, \text{M}) = \frac{\text{fcost}(\text{L}, \text{M})}{\text{moveL} + |\text{L}| \,.\, \text{moveM}(\text{M})}$$

- fcost(L,M): The total cost of adjustment for the L and M model event logs.
- moveL(L): The total cost of moving through the entire log without moving with the model.
- |L|: The number of agencies in the log.
- moveM(M): The total cost of performing the movement only on the model.

Fitness scores range from 0 (poor fitness) to (perfect fitness).

3.2 Precision

Precision in conformance checking measures the extent to which the model can avoid "underfitting" by checking the extent to which the model allows behavior that matches the observed data (Rubin et al., 2014). It can show the proportion of behavior represented by the model that cannot be seen in the activity log (Kurniati et al., 2019).

$$\text{precision}(\text{L}, \text{M}) = 1 - \frac{1}{|\text{E}|} \sum \text{e E} \frac{|\text{enL}(\text{e})|}{|\text{enM}(\text{e})|}$$

|E|: The total number of events in the log.

enL(e): A set of activities executed in the same context.

enM(e): The set of activities that are activated in the model at a specific point.

If all the behaviors allowed by the model are also observed, then precision(L,M)=1.

3.3 Generalization

Generalization in conformance checking measures the extent to which the model can avoid "overfitting" by checking the extent to which the model can describe more general behavior than the examples in the event log [26]. It can assess the extent to which the model that has been formed can reproduce the behavior of the process in the future and can be considered a measure of confidence in precision (Kurniati et al., 2019).

$$\text{generalization}(L, M) - 1 \frac{1}{|E|} \Sigma e \in E \, \text{Pnew}(|\text{diff}(e)|, |\text{sim}(e)|)$$

$$\text{Pnew}(w, n) = \frac{w(w+1)}{\{n(n-1)} \; \text{otherwise}$$

- |E|: The total number of events in the log.
- |diff(e)|: The number of different activities in a state.
- |sim(e)|: The number of events that occurred under the same circumstances.

Generalization approaches 0 if it is very likely that a new event will exhibit behavior that has never been seen before. Generalization approaches 1 if it is doubtful that subsequent events will reveal new behavior.

3.4 Simplicity

Simplicity in conformance checking emphasizes that the process model should be no more complex than necessary to explain the data in the event log. The goal is to find the "simplest process model" to explain what is observed (Rubin et al., 2014). It can capture the complexity of the process model in terms of readability (Kurniati et al., 2019).

$$S(M) = f(N, A)$$

- N is the number of nodes in the model.
- A is the number of arcs in the model.
- F is a function that describes the relationship between the number of nodes and the ARC to produce the value of simplicity.

We can specify f based on special considerations for special cases. For example, f can be simply the sum of N and A:

$$f(N, A) = N + A$$

If we want to consider other factors, such as structuralist or entropy, we may need to replace f according to specific metrics relevant to the context at hand. The choice of function f will depend mainly on the specific needs and characteristics of the model and the data to be handled.

Results and Discussion

The MIMIC-IV dataset's pre-process begins with planning and justifying the data import to ensure all relevant data is available for the study. The dataset then undergoes reconstruction, combining various subsets into a unified dataset, maintaining data integrity through completeness and consistency across variables. Data quality is rigorously assessed, focusing on accuracy,

completeness, consistency, integrity, standardization, availability, usability, and timeliness. Next, relevant data is extracted, emphasizing critical information for analysis, which is then used to construct a control flow model. Finally, conformance checks ensure the data aligns with established standards. The results of the filtered dataset, which focuses on the four First Care Units with the highest patient numbers, are detailed in Table 1, describing the filtered dataset outcome.

Table 1. Description of Dataset

Diagnosis Name	Total Record
Cardiac Vascular Intensive Care Unit	155461
Medical Intensive Care Unit	127208
Surgical Intensive Care Unit	109692
Trauma SICU	94249

Data Quality

Completeness: To measure the completeness of the data, we calculate the proportion of available values compared to the total values that should be in the dataset. We count the missing values (missing values) in each column and subtract that ratio from 1, which produces a completeness value. Completeness values range from 0 to 1, where 1 indicates that the column has complete data with no missing values, while values closer to 0 indicate that the column is almost completely unfilled.

The completeness of the ICU and Hospital datasets was evaluated based on the average percentage of data filled in various data types. The ICU dataset showed high completeness, with a total average of 95.79%, indicating that most of the required data was well represented. Meanwhile, the Hospital dataset has an overall completeness of 90.88%, although several areas, such as microbiology events (59.28%) and pharmacy (66.86%), show lower completeness, which could indicate areas that require data quality improvement. Table 2 Completeness of Hospital Dataset and Table 3 Completeness ICU Dataset, provides further details on these value

Table 2. Completness Hospital

Completeness Hospital			
Table Name	**Average Score**	**Table Name**	**Average Score**
Admissions	88.18%	omr	100%
d_hcpcs	96.17%	patients	83.5%
d_icd_diagnoses	100%	pharmacy	66.86%
d_icd_procedures	100%	poe_detail	100%
d_labitems	99.95%	poe	81.49%
diagnoses_icd	100%	prescriptions	89.38%
drgcodes	85.25%	procedures_icd	100%
Emar	95.04%	provider	100%
hcpcsevents	100%	services	96.31%
labevents	76.17%	transfers	90.98%
microbiologyevents	59.28%	Average	90.88%

Table 3. Completeness ICU

Completeness ICU			
Table Name	**Average Score**	**Table Name**	**Average Score**
Caregiver	100%	Ingredientevents	95.63%
Chartevents	86.83%	Inputevents	93.92%
Datetimeevents	100%	Outputevents	100%
Icustays	100%	Procedureevents	89.97%
		Average	95.79%

Consistency: Consistency checking is carried out to ensure that each column in the dataset has the same data type and is as expected. This is important so that the data can be analyzed correctly and that there is no mixture of data types, which could cause errors in further processing. For example, a column that should contain numbers should not contain text. The results of this check help identify whether the data type in each column is consistent so that data quality is better maintained.

The consistency of data types in the hospital and ICU datasets was examined in the analysis of the MIMIC-IV dataset. This dataset consists of three main data types: integer, float, and object. After going through the verification process, it was found that all data types in the dataset were consistent, thus supporting the validity of further analysis.

Integrity: Data integrity verification ensures that all values in the primary key (ID) column are unique and accurately represent distinct entities. This process is crucial for maintaining data quality and reliability, as duplicate entries can lead to distortions in analysis and decision-making. If duplicates are detected in the primary key column, corrective actions such as removing or merging duplicate records are undertaken to restore data integrity. The effectiveness of these actions is assessed by calculating the ratio of unique records to the total number of records, providing a clear indication of how well data integrity is upheld. This step is critical for ensuring high-quality data for further analysis in both research and operational contexts.

The data integrity analysis of the MIMIC-IV dataset reveals that the Hospital dataset demonstrates a high level of integrity, with an average of 92.24%, as presented in Table 4. In contrast, the ICU dataset exhibits considerable variability in data integrity, with an average of only 45.49%, as illustrated in Table 5.

Table 4 Hospital Integrity

Hospital Integrity			
Table Name	**Average Score**	**Table Name**	**Average Score**
Admissions	41.91%	omr	38.10%
d_hcpcs	100%	patients	100%
d_icd_diagnoses	99.55%	pharmacy	100%
d_icd_procedures	99.99%	poe_detail	70.16%
d_labitems	100%	poe	100%
diagnoses_icd	99.99%	prescriptions	87.78%
drgcodes	99.75%	procedures_icd	99.99%
Emar	100%	provider	100%
hcpcsevents	100%	services	99.99%
labevents	100%	transfers	100%
microbiologyevents	100%	Average	92.24%

Table 5. ICU Integrity

ICU Integrity			
Table Name	**Average Score**	**Table Name**	**Average Score**
Caregiver	100%	Ingredientevents	0.62%
Chartevents	0.15%	Inputevents	0.8%
Datetimeevents	100%	Outputevents	1.68%%
Icustays	100%	Procedureevents	10.44%
		Average	45.49%

Standardization: The MIMIC-IV dataset, while rich in medical information, does not always adhere to strict data standardization practices across all columns. This can lead to inconsistencies, particularly in data types, missing entries, and non-standardized values, which pose challenges for accurate analysis. In such cases, several corrective measures are essential. First, preprocessing techniques should be applied to identify and resolve missing or improperly formatted data. Cleaning and reformatting inconsistent entries can help ensure uniformity. Second, robust validation rules should be introduced to enforce compliance with standardization criteria, ensuring that future data entries adhere to the expected format. Lastly, periodic audits of the dataset are recommended to continually assess and correct standardization issues, ultimately improving the reliability and quality of the data for meaningful analysis.

Availability: Despite being comprehensive, the MIMIC-IV dataset only sometimes guarantees that all critical data elements are consistently available. Some key columns may need to be included or completed, leading to gaps in the analysis and limiting the potential to draw accurate conclusions. This issue can stem from factors such as incomplete data entry, extraction errors, or inconsistencies in the data logging process.

To address these shortcomings, researchers should employ a systematic approach to fill in missing columns or rectify incomplete data. Solutions include re-extracting the data from the source, ensuring proper data collection methods, and implementing automated checks during the data import to flag missing columns. Additionally, if critical data is not retrievable, alternative methods such as imputing missing values based on available data trends or consulting with domain experts for possible manual corrections may be necessary. These measures will ensure that the dataset is well-prepared for accurate and comprehensive analysis, ultimately improving the quality of research outcomes.

Usability: The usability of data in the MIMIC-IV dataset is evaluated by ensuring that values in critical columns adhere to valid formats and are correctly represented for accurate analysis. This process involves validating data types, confirming that numerical fields contain only integers or digit-based strings, and ensuring that text fields include only printable characters. These checks are crucial to maintaining data accuracy, consistency, and reliability, minimizing errors caused by outliers, invalid entries, or improperly formatted values. Ensuring high usability of the dataset enhances the quality of the analysis and leads to more trustworthy results.

Despite the efforts to maintain usability, some columns in the hospital and ICU sections of the MIMIC-IV dataset do not fully meet the required standards. These usability issues may stem from the presence of invalid data types, non-digit values in fields expected to be numerical, or unreadable characters in text fields. To address these challenges, it is necessary to identify and correct the problematic data points, implement robust validation mechanisms during data collection, and ensure adherence to data standards. By proactively addressing these issues, the dataset's overall usability will improve, supporting more effective and accurate analysis in healthcare research.

Timelines: In evaluating the timeliness of the MIMIC-IV dataset, the goal is to assess whether the sequence of time-related data is accurate and logically ordered. This includes ensuring that recorded events, such as patient admission or discharge times, are chronologically correct. For example, discharge times should not occur before admission times. Timeliness plays a critical role in maintaining the reliability and relevance of the dataset, as it ensures that all time-based data aligns with real-world sequences, providing a solid foundation for further analysis and research.

However, during the evaluation of the timeliness of the MIMIC-IV dataset for the hospital and ICU departments, it was found that several tables lacked the necessary time-stamped columns required to effectively assess timeliness. The absence of these time-related columns poses a challenge as it limits the ability to fully verify the chronological accuracy of the data. To address this issue, the inclusion of timestamped columns should be prioritized within the dataset. If feasible, additional data collection efforts should focus on recording time-related information to ensure that the dataset meets the required criteria for timeliness, thereby enhancing the overall integrity and usability of the data for future analysis.

Process Mining

Event Log The processed dataset is converted into an event log, which represents a series of events related to the clinical process in MIMIC-IV. This event log is then used as a basis for further analysis in Process Mining.

Process Models The analysis in this study begins with extracting process models from event logs generated using process mining methods. The resulting process models detail specific workflows and sequences of events within the clinical environment, starting from the first care unit, the procedure phase, and the last care unit. Each phase is associated with a timestamp that records when the patient enters and exits the hospital. This model provides a comprehensive overview of the patient's care pathway from start to finish within the hospital system, supporting an in-depth analysis of the efficiency and quality of clinical processes. An example of a process model can be found in Figure 13, which utilizes the Heuristic Miner method.

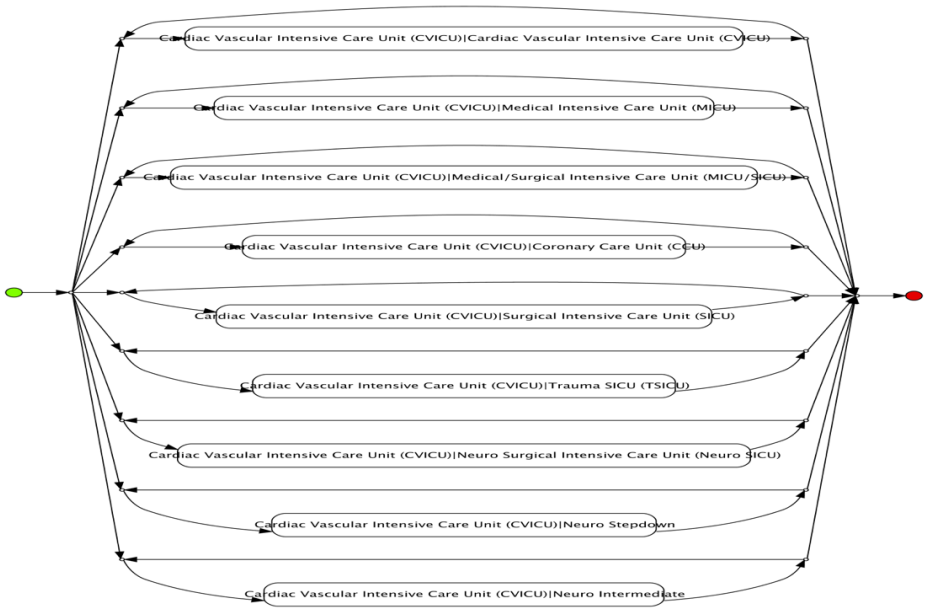

Figure 12 Process Model CVICU Heuristic

Inductive Miner: In the Inductive Miner methodology, the generated process models automatically detail specific workflows and the sequence of events within the clinical environment. This process begins by extracting patterns from event logs, which are then used to construct a process model that outlines the patient's journey through various stages of care. The

model includes the first care unit, the procedure stage, and the last care unit. Each stage in the model is linked to a timestamp that records the times the patient enters and exits the hospital. The Inductive Miner methodology allows for the identification of recurring workflow patterns. It ensures that the generated model accurately reflects the reality of clinical processes by considering all events recorded in the logs.

Alpha Miner: Alpha Miner extracts process models from event logs. Although this algorithm is basic and more straightforward, it provides a helpful initial understanding of the structure of the existing process.

Heuristic Miner: In the Heuristic Miner methodology, the generated process model details specific workflows and the sequence of events within the clinical environment. This process begins by analyzing event logs to identify causal relationships and the frequency of occurrences. The resulting model outlines the patient's journey through various stages of care, starting from the first care unit, then the procedure stage, and finally, the last care unit. Each stage in the model is linked to a timestamp that records when the patient enters and exits the hospital. The Heuristic Miner methodology enables the identification of solid relationships between events. It ensures that the resulting model accurately represents the clinical processes based on the observed frequency and correlation of events. The process model is illustrated in Figure 14.

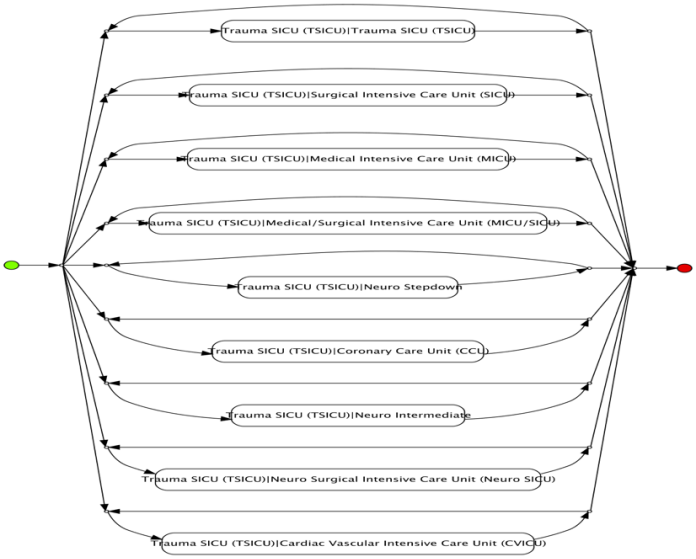

Figure 14 Process Model TSICU Heuristic

Conformance Checking: The models produced from various algorithms are then tested using the Conformance Checking method to evaluate how well they fit the existing data. This evaluation includes measurements of fitness, generalization, precision, and simplicity. Table 6, Conformance Checking, provides detailed results of this evaluation.

Table 6. Conformance Checking

CVICU				
Type	**Fitness**	**Generalization**	**Precision**	**Simplicity**
Inductive	0.999	0.763	0.384	0.604
Alpha	0.134	0.722	0.443	0.736
Heuristic	0.989	0.797	0.475	0.450
MICU				
Type	**Fitness**	**Generalization**	**Precision**	**Simplicity**
Inductive	0.999	0.795	0.531	0.610
Alpha	0.232	0.771	0.545	0.670
Heuristic	0.991	0.782	0.637	0.458

SICU				
Type	**Fitness**	**Generalization**	**Precision**	**Simplicity**
Inductive	0.998	0.800	0.478	0.617
Alpha	0.180	0.796	0.590	0.614
Heuristic	0.990	0.827	0.490	0.451
TSICU				
Type	**Fitness**	**Generalization**	**Precision**	**Simplicity**
Inductive	0.999	0.792	0.455	0.614
Alpha	0.155	0.817	0.644	0.582
Heuristic	0.989	0.792	0.457	0.452

The models produced by the different mining algorithms—Inductive Miner, Heuristic Miner, and Alpha Miner—each exhibit distinct characteristics in terms of fitness, generalization, precision, and simplicity when applied to various clinical units, such as CVICU, MICU, SICU, and TSICU. Inductive Miner consistently achieves very high fitness values across all units, indicating an excellent fit to the event log data. However, this high fitness comes at the cost of lower precision, suggesting that while the model can replay the event logs accurately, it may also allow for behaviors not strongly supported by the data. Despite this trade-off, the simplicity of the models produced by Inductive Miner remains moderate, making it a robust choice when fitness is prioritized over precision.

In contrast, Heuristic Miner offers a better balance between fitness, precision, and generalization, albeit at the expense of simplicity. While slightly more complex, the models generated by Heuristic Miner demonstrate improved precision and generalization across the clinical units, indicating that these models are better at capturing the actual underlying processes with less allowance for unobserved behaviors. On the other hand, Alpha Miner produces simpler models with lower fitness values but maintains relatively good precision in some cases, particularly in the TSICU unit. This suggests that Alpha Miner might be more suitable for scenarios where simplicity and precision are valued over fitness and generalization. Overall, the choice of mining algorithm depends on the specific requirements of the analysis, such as whether the focus is on model accuracy, the ability to generalize, or the simplicity of the model representation.

Conclusion

The conclusion of this research demonstrates that in the Process Mining analysis using the MIMIC-IV dataset, the model generated by the Inductive Miner algorithm shows a very high fitness level across all ICU units despite its lower precision. The Heuristic Miner algorithm produces a more balanced model with a better mix of fitness, accuracy, and generalization, though it comes with increased complexity. Meanwhile, while more straightforward, the Alpha Miner algorithm displays variability in its outcomes, with commendable precision in certain instances.

From a data quality perspective, the Hospital dataset generally shows higher quality than the ICU dataset, with completeness levels of 90.88% for the Hospital and 95.79% for the ICU. However, there are variations in aspects such as data integrity, standardization, availability, usability, and timeliness between the two datasets. These variations reveal that data quality significantly impacts the effectiveness of process mining. High-quality data enhances the process models' accuracy, precision, and reliability, leading to more actionable insights. Conversely, lower-quality data can introduce noise and inconsistencies, affecting the precision and generalization of the models.

The findings underscore the critical role of data quality in process mining, particularly in healthcare settings, where accurate and reliable analysis is essential. Improving data quality in specific dataset areas will refine the process models and ensure that the insights derived from these models are both valid and valuable for clinical decision-making and process optimization.

Bibliography

Alter, S. (2015). Work system theory: A bridge between business and IT views of systems. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 9097, 520–521. https://doi.org/10.1007/978-3-319-19069-3

Alter, S. (2017a). Work System Theory and Work System Method. Proceedings of the 10th Innovations in Software Engineering Conference, 211–211. https://doi.org/10.1145/3021460.3021488

Alter, S. (2017b). Work system theory and work system method: A bridge between business and IT views of IT-reliant systems in organizations. ACM International Conference Proceeding Series, 211. https://doi.org/10.1145/3021460.3021488

Bogarín, A., Cerezo, R., & Romero, C. (2018). Discovering learning processes using Inductive Miner: A case study with Learning Management Systems (LMSs). Psicothema, 3(30), 322–329. https://doi.org/10.7334/psicothema2018.116

Bolt, A., de Leoni, M., & van der Aalst, W. M. P. (2016). Scientific workflows for process mining: building blocks, scenarios, and implementation. International Journal on Software Tools for Technology Transfer, 18(6), 607–628. https://doi.org/10.1007/s10009-015-0399-5

Burattin, A., Maggi, F. M., & Sperduti, A. (2016). Conformance checking based on multi-perspective declarative process models. Expert Systems with Applications, 65, 194–211. https://doi.org/10.1016/j.eswa.2016.08.040

De Weerdt, J., Schupp, A., Vanderloock, A., & Baesens, B. (2013). Process Mining for the multi-faceted analysis of business processes - A case study in a financial services organization. Computers in Industry, 64(1), 57–67. https://doi.org/10.1016/j.compind.2012.09.010

Evermann, J., Rehse, J.-R., & Fettke, P. (2016). Process Discovery from Event Stream Data in the Cloud - A Scalable, Distributed Implementation of the Flexible Heuristics Miner on the Amazon Kinesis Cloud Infrastructure. 2016 IEEE International Conference on Cloud Computing Technology and Science (CloudCom), 645–652. https://doi.org/10.1109/CloudCom.2016.0111

Fox, F., Aggarwal, V. R., Whelton, H., & Johnson, O. (2018). A data quality framework for process mining of electronic health record data. Proceedings - 2018 IEEE International Conference on Healthcare Informatics, ICHI 2018, 12–21. https://doi.org/10.1109/ICHI.2018.00009

Garg, N., & Agarwal, S. (2016). Process mining for clinical workflows. ACM International Conference Proceeding Series, 12-13-August-2016. https://doi.org/10.1145/2979779.2979784

Jans, M., Weerdt, J. De, Depaire, B., Dumas, M., & Janssenswillen, G. (2021). Conformance Checking in Process Mining. Information Systems, 102, 101851. https://doi.org/10.1016/j.is.2021.101851

Johnson, A. E. W., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Moody, B., Gow, B., Lehman, L. wei H., Celi, L. A., & Mark, R. G. (2023). MIMIC-IV, a freely accessible electronic health record dataset. Scientific Data, 10(1). https://doi.org/10.1038/s41597-022-01899-x

Kurniati, A. P., Rojas, E., Hogg, D., Hall, G., & Johnson, O. A. (2019). The assessment of data quality issues for process mining in healthcare using Medical Information Mart for Intensive Care III, a freely available e-health record database. Health Informatics Journal, 25(4), 1878–1893. https://doi.org/10.1177/1460458218810760

Mayr, M., Luftensteiner, S., & Chasparis, G. C. (2022). Abstracting Process Mining Event Logs From Process-State Data To Monitor Control-Flow Of Industrial Manufacturing Processes. Procedia Computer Science, 200, 1442–1450. https://doi.org/10.1016/j.procs.2022.01.345

Nuritha, I., & Mahendrawathi, E. R. (2017). Structural Similarity Measurement of Business Process Model to Compare Heuristic and Inductive Miner Algorithms Performance in Dealing with Noise. Procedia Computer Science, 124, 255–263. https://doi.org/10.1016/j.procs.2017.12.154

Perimal-Lewis, L., Teubner, D., Hakendorf, P., & Horwood, C. (2016). Application of process mining to assess the data quality of routinely collected time-based performance data sourced from electronic health records by validating process conformance. Health Informatics Journal, 22(4), 1017–1029. https://doi.org/10.1177/1460458215604348

Rebuge, Á., & Ferreira, D. R. (2012). Business process analysis in healthcare environments: A methodology based on process mining. Information Systems, 37(2), 99–116. https://doi.org/10.1016/j.is.2011.01.003

Rojas, E., Munoz-Gama, J., Sepúlveda, M., & Capurro, D. (2016). Process mining in healthcare: A literature review. In Journal of Biomedical Informatics (Vol. 61, pp. 224–236). Academic Press Inc. https://doi.org/10.1016/j.jbi.2016.04.007

Rubin, V. A., Mitsyuk, A. A., Lomazova, I. A., & van der Aalst, W. M. P. (2014). Process mining can be applied to software too! Proceedings of the 8th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 1–8. https://doi.org/10.1145/2652524.2652583

Schuster, D., van Zelst, S. J., & van der Aalst, W. M. P. (2022). Utilizing domain knowledge in data-driven process discovery: A literature review. Computers in Industry, 137, 103612. https://doi.org/10.1016/j.compind.2022.103612

Sundari, M. S., & Nayak, R. K. (2020). Process Mining in Healthcare Systems: A Critical Review and its Future. International Journal of Emerging Trends in Engineering Research, 8(9), 5197–5208. https://doi.org/10.30534/ijeter/2020/50892020

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