UHASSELT

KNOWLEDGE IN ACTION

Faculty of Business Economics

Master of Management

Master's thesis

GM4M-IV: a generic method for extracting an event log from the MIMIC-IV database

Teresa Nalikka

Thesis presented in fulfillment of the requirements for the degree of Master of Management, specialization Business Process Management

SUPERVISOR :
Prof. dr. Niels MARTIN

2021
2022

UHASSELT

KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

UHASSELT

KNOWLEDGE IN ACTION

Faculty of Business Economics

Master of Management

Master's thesis

GM4M-IV: a generic method for extracting an event log from the MIMIC-IV database

Teresa Nalikka

Thesis presented in fulfillment of the requirements for the degree of Master of Management, specialization Business Process Management

SUPERVISOR :

Prof. dr. Niels MARTIN

DECLARATION

I certify that all the material in this dissertation that is not my work has been identified and that no material is included for which a degree has previously been conferred on me. The contents of this dissertation reflect my own personal views and are not necessarily endorsed by the University.

Signature

Teresa Gladys Nalikka

Date: August 12th, 2022

Supervisor: Prof. dr. Niels MARTIN

Hasselt University

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my supervisor Prof. dr. Niels MARTIN, for the constructive feedback, support, and patience throughout authoring this thesis. The feedback guided me to improve the content of the thesis and become more knowledgeable about the topic of process mining.

To my family, thanks to mama and papa, Christine, and Marco Methorst for taking care of Noah while I wrote my thesis. And aunt Pauline Linnenbank, thanks for the continuous support and for pushing me not to give up.

Lastly, I would like to thank my friends for being there.

Teresa

ABSTRACT

Process mining is a method for discovering, analysing, and improving organisational processes by extracting knowledge from an event log. An event log is simply a list of activities in their order of occurrence. To build an event log, data is extracted from a database, one of which is MIMIC-IV, a database containing de-identified medical records of patients admitted to critical care units of Beth Israel Deaconess Medical Centre in Boston, Massachusetts, in the United States. Unlike existing event logs, an extra step must be taken to use the MIMIC-IV database for process mining: extracting an event log from the database. This thesis proposes a generic method for extracting an event log from the MIMIC-IV database. The thesis successfully applies the generic method GM4M-IV when discovering the process model for a patient's stay in the ICU.

Keywords - process mining, event log, MIMIC-IV, GM4M-IV, generic method.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Case Id	Case Identification
CITI	Collaborative Institute Training Initiative
CPT	Current Procedural Terminology
Cxr	Chest x-ray
DRG codes	Diagnosis-related groups codes
DS	Design Science
DUA	Data User Agreement
ed	Emergency department
EHR	Electronic Health Record
ERP	Enterprise Resource Planning
ETL	Extraction, Transformation and Loading
EVS	Enterprise Visualisation suite
HIS	Health Information System
Hcpcs	Healthcare Common Procedural Coding System.
ICD	International Classification of Diseases
ICU	Intensive Care Unit
iDHM	The Interactive Data-aware Heuristics Miner
IT	Information Technology
Medrecon	Medical reconciliation
MIMIC	Medical Information Mart for Intensive Care
MIT	Massachusetts Institute of Technology
PM2	Process Mining Methodology
SAP	Systems Applications and Products
SQL	Structured Query Language
SMART	Specific, Measurable, Attainable, Realistic and Time-bond

1. Introduction

Process mining is discovering and analysing processes based on knowledge extracted from an event log (Kurniati et al., 2018; Jans, 2010). An event log is a list of activities showing the execution of an action (van der Aalst, 2016; Rule et al., 2019). Several attributes describe an action, such as the activity it is associated with, when it was performed, who performed it and for whom (Martin et al., 2020). Traditionally in process mining, an event log is extracted from an organisation's database (Bano et al., 2021), most of which store process execution data in a non-process-centric way hindering immediate process mining (Diba et al., 2020). Process execution data is detailed information about what happens during each execution of each task (De Smedt et al., 2019; Liu et al., 2016). The non-process-centric data is usually in different formats, which, requires substantial effort to locate and transform into an event log for process mining (Diba et al., 2020); an example of a database with non-process-centric data is MIMIC-IV.

MIMIC-IV is a freely accessible database containing de-identified medical records of patients admitted to Beth Israel Deaconess Medical Centre in Boston, Massachusetts, United States of America (Johnson et al., 2021). The medical records in MIMIC-IV correspond to patients admitted to the hospital's intensive care unit (ICU) and emergency department (ed) between 2008 and 2019 (Johnson et al., 2021). MIMIC is unique because it is the only database of its kind with open-source critical care data about individual patients that spans over a decade (Johnson et al., 2016). The availability of a vast amount of detailed critical care data in MIMIC-IV boots the database's potential for process mining. Cremerius and Weske (2021, use process mining on the MIMIC-IV database to introduce a visualisation technique for enhancing discovered process models with domain data, allowing process exploration based on data. Lichtenstein et al. (2021) develop an attribute driven case notion discovery approach for unlabelled event logs, which detects cyclic behaviour correlates the events closer to the original process instances without additional input using the MIMIC-IV database.

The MIMIC-IV database does not automatically generate event logs; to use the database for process mining, an additional step must be taken: extracting an event log. The aim of this thesis is to introduce GM4M-IV, a generic method for extracting an event log from the MIMIC-IV database. A general method will ensure maximum utilisation of the MIMIC-IV database to its full potential for process mining. Using GM4M-IV, the process mining analysts can customise an event log by selecting and extracting only the variables they need for a particular process mining project.

The rest of this thesis is structured as follows. Section 2 introduces preliminary concepts about process mining, what an event log is, process mining in healthcare and the MIMIC-IV database. Section 3 presents the research methodology that the thesis follows. Section 4, the literature review examines the literature on process mining using the MIMIC database in section 4.1 and on extracting an event log in section 4.2. Section 5 presents the GM4M-IV, a generic method for extracting an event log from the MIMIC-IV database. Section 6 demonstrates GM4M-IV, and section 7 evaluates the method. Section 8 summarises the contribution of this thesis and provides a future recommendation.

2. Preliminaries

2.1 Process mining

Process mining is a technique for discovering, monitoring, and improving actual processes by extracting knowledge from an event log (Rojas et al., 2016), which supports three distinct types of process mining.

The first type of process mining is discovery. Discovery involves developing a process model without prior knowledge based on behaviour observed in an event log (van der Aalst, 2016). Several organisations are surprised to learn that exiting techniques can detect real processes based solely on examples of behaviours stored in an event log (van der Aalst, 2012). Researchers and analysts use discovery as a starting point to conduct other types of process mining (D'Castro et al., 2018).

The second type of process mining is conformance checking. According to Ghahfarokhi et al. (2021), conformance checking involves using algorithms to compare an existing process model to an event log. The goal of this comparison is to verify compliance. Compliance entails checking whether the process model differs from the event logs and vice versa (Verbeek et al., 2010). One example is checking the compliance between an existing process model for a clinical-surgical pathway and its event log. If the model runs thirty cases, they should all be able to play through beginning to end without getting stuck. The model needs to be improved if one or more cases become stuck.

The third type of process mining is enhancement. Enhancement involves improving or extending an existing process model based on insights from an event log (van der Aalst, 2016). For example, the event log can add information about service times to an existing process model (van der Aalst, 2011).

2.1.1 Event log

An event log is a combination of cases that represent distinct process instances (Munoz-Gama et al., 2022). A process instance is a sequence of events that occur in a case (Suriadi et al., 2017). Events are the ordered activities in an event log, each with a timestamp indicating when they occurred. For

instance, the activities carried out during a patient's visit to the emergency room. Such as patient registration, examination, and ordering tests the patient needs like blood analysis.

A case is an occurrence of a series of events in a process (Pourbafrani & van der Aalst, 2021). The standard attributes an event log records from every event include: (i) 'CaseID', this is the unique case identifier for every case. (ii) 'Activity' shows what happened. (iii) 'Timestamp' indicates when the event took place, events never have duration but happen at a certain point. And (iv) 'Resource' shows who performed the activity. The first event in table1 shows the admission of a patient with caseID 5671 at 7:50:26 on January 26th, 2022, by receptionist Simon. Event two shows the patient getting her blood pressure taken by nurse Yves on the same date.

Table 1: An example of an event log

CaseID	**Activity**	**Timestamp**	**Status**	**Resource**
5671	Admission	2022-2-04 13:50:26	Complete	Receptionist Simon
5671	Blood pressure	2022-21-04 13:56:10	Start	Nurse Yves
5671	Blood pressure	2022-21-04 14:14:00	Complete	Nurse Yves
5671	Sugar test	2022-21-04 14:15:00	Start	Midwife Janna
5671	Sugar test	2022-21-04 14:18:00	Complete	Midwife Janna
5671	Draw blood	2022-21-04 14:20:00	Start	Nurse Piet
5671	Draw blood	2022-21-04 14:22:00	Complete	Nurse Piet
5671	Induce delivery	2022-22-04 06:00:00	Start	Midwife Hedwij
5671	Induce delivery	2022-22-04 6:10:00	Complete	Midwife Hedwij
5671	Delivery	2022-22-04 12:06:00	Start	Midwife Kato
5671	Discharge	2022-25-04 10:35:01	complete	Gynaecologist Jenifer

2.2 Process mining in healthcare

Healthcare refers to all services medical professionals provide to maintain people's physical and mental well-being (Agarwal et al., 2010; Kraus et al., 2021). A literature review by Batista and Solanas (2018) identifies process discovery as the most popular type of process mining in the healthcare sector. Process discovery is popular because it reconstructs processes from their executions and allows for discovering complex processes in healthcare (Batista & Solanas, 2018). According to Rojas et al. (2016), process mining research in healthcare is divided into four categories. The first perspective is control-flow, which aims at discovering the execution order of process activities (Rojas et al., 2016). An example is research

by Bos et al. (2011), showing the steps of treating a patient diagnosed with cancer. The second perspective is performance, which analyses the execution time of activities and identifies bottlenecks (Rojas et al., 2016). For example, Mans et al. (2012) use process mining to obtain details on the execution of dental processes. The third perspective is conformance checking, which detects process deviations regarding an existing model (Rojas et al., 2016). For example, Bouarfa and Dankelman (2012) use process mining to detect workflow outliers from surgical activity logs automatically. The fourth perspective is the organisation, which analyses resource collaboration (Rojas et al., 2016). For example, Rattanavayakorn and Premchaiswadi (2015) use the process mining social network miner to investigate the relationship between staff and resources in a hospital, by tracking and tracing the behaviour of doctors during treatment processes of patients.

2.2.1 Hospital processes

There are two types of hospital processes in which process mining is effective (Rojas et al. (2016). These two processes are the medical treatment and organisational processes. The medical treatment processes are the clinical processes that manage patients, such as a pathway describing a patient's oncology diagnosis (Rojas et al., 2016). Organisational processes manage organisational knowledge, categorise resources, and establish relationships between healthcare professionals and organisational units (Kaymak et al., 2012).

2.2.2 Health data

The data for healthcare process mining is mostly gathered from hospital emergency units (Batista & Solanas, 2018). The data types mainly gathered are from treatments, followed by clinical pathways and diagnostics (Batista & Solanas, 2018). Oncology is the most widely researched medical field that uses process mining in healthcare (Rojas et al., 2016). Examples of research in this area include Mans et al. (2008), Fei & Mensken (2010) and Perimal-Lewis et al. (2014).

2.3 The MIMIC-IV database

As mentioned in section 1, MIMIC is a freely accessible database that contains de-identified health-related data from patients admitted to the Beth Israel Deaconess Medical Canter's critical care units. The MIMIC database is a product of combining medical data by the Massachusetts Institute of Technology (MIT) Laboratory for computational physiology. The MIMIC database currently has four versions, three of which are hosted by Physionet[1], a website containing documentation on getting started using the MIMIC database. The three versions of MIMIC that Physionet hosts are MIMIC-II 2001-2008, MIMIC-III 2001-2012 and MIMIC-IV 2008-2019. The MIMIC-IV database has six modules to reflect on the origin of the data:(i) core contains patient stay information (i.e. admissions and transfers), (ii) hosp contains hospital-level data for patients: labs, micro, and electronic medication administration. (iii) ICU contains ICU level data. (iv) ed contains data from the emergency department. (v) cxr contains lookup tables and meta-data from MIMIC-CXR, allowing linking to MIMIC-IV. (vi) note contains de-identified free-text clinical notes (Johnson et al., 2021).The six modules of the MIMIC-IV database all together contain forty-three tables i.e. Core has three tables, Hosp has seventeen tables, ICU has seven tables, ed has seven tables, cxr has one table, and note contains eight tables. Tables 2, 3, 4 and 5 summarise the content of each table as per the respective MIMIC-IV database module. These forty-three tables contain data from which process mining researchers extract an event log. The detailed description and data per table are available on the MIMIC website, which hosts documentation of MIMIC[2] and how to use the database.

The core concepts of the MIMIC-IV are patient identifiers, which are unique identification numbers assigned to every patient in the database. Patient identifiers include subject_id which refers to a unique patient number or identifier, hadm_id - refers to a unique admission number or identifier and stay_id - refers to a unique stay identifier per patient. The three patient identifiers make it easy to trace the

[1] https://mimic.physionet.org/
[2] https://mimic.mit.edu/docs/iv/

patient's path during their stay at the hospital. When the patient registers at the emergency department, she gets a subject_id, on admission, she is given a hadm_id and a stay_id.

Table 2:Brief description of tables from core and cxr modules

Table	**Description**
Admissions	Information regarding a patient's admission to the hospital.
Patients	Information that is consistent for a patient's lifetime is stored in this table.
Transfers	Physical locations for patients throughout their hospital stay.

Table 3: Brief description of cxr module

Table	**Description**
cxr_record_list	Lists all records in the MIMIC-CXR database

Table 4:Brief description of tables from hosp module

Table	**Description**
D_hcpcs	Dimension table for hcpcs events; provides a description of Current Procedural Terminology codes
D_icd_diagnoses	Describes icd-9/icd-10 billed diagnoses.
D_icd_procedures	Describes icd-9/icd-10 billed procedures.
D_labitems	Dimension table for lab events; describes all lab items.
Diagnoses_icd	Billed icd-9/icd-10 diagnoses for hospitalizations.
Drgcodes	Billed DRG codes for hospitalisations.
Emar	The Electronic Medicine Administration Record (eMAR); barcode scanning of medications at the time of administration
emar_detail	Supplementary information for electronic administrations recorded in emar.
Hpcsevents	Billed events occurring during the hospitalization. Includes CPT codes.
Labevents	Laboratory measurements sourced from patient-derived specimens.
Microbiology events	Microbiology cultures
Pharmacy	Formulary, dosing, and other information for prescribed medications.
Poe	Orders made by providers relating to patient care.
Poe_detail	Supplementary information for orders made by providers in the hospital.
Prescriptions	Prescribed medications.
Procedures_icd	Billed procedures for patients during their hospital stay.
Services	The hospital service(s) cared for the patient during hospitalisation.

Table 5:Brief description of tables from ICU module

Table	**Description**
D_items	Dimension table, describing itemid. Defines concepts recorded in the events table in the ICU module.
Chart events	Charted items occurring during the ICU stay.
Datetimeevents	Documented information is in a date format (e.g., date of the last dialysis).
Icustays	Tracking information for ICU stays, including admission and discharge times.
Input events	The information documented regarding continuous infusions or Intermittent administrations.
Output events	Information regarding patient outputs, including urine, and drainage among others.
Procedural events	Procedures documented during the ICU stay (e.g., Ventilation), though not necessarily conducted within the ICU (e.g., X-ray imaging).

Table 6:Brief description of tables from ed module

Table	**Description**
Diagnosis	Provides billed diagnoses for patients
Edstays	This is the primary tracking table for emergency department visits. It provides the time the patient entered the emergency department and the time they left the emergency department.
Medrecon	This process is called medicine reconciliation, and the medrecon table stores the findings of the care providers.
Pyxis	The pyxis table provides information for medicine dispensations made via the Pyxis system.
Triage	The triage table contains information about the patient when they were first triaged in the emergency department.
Vital sign	Patients admitted to the emergency department have routine vital signs every 1-4 hours.
vitalsign_hl7 table	Patients admitted to the emergency department may be monitored by telemetry.

3. Research Methodology.

Section 3 discusses how the thesis uses the Design Science Research (DSR) method to develop GM4M-IV. DSR is a method presented by Peffers et al. (2007), with a focus on developing and accessing problem-solving artifacts in information technology. The DSR method has six steps that this thesis uses to develop GM4M-IV, the first step is Problem identification and motivation. This involves defining the specific research problem and justifying the value of a solution (Peffers et al., 2007). The second step is defining the objective of the solution. This involves determining the ways in which the solution is better than the existing ones. (Peffers et al., 2007). The third step is design and development which involves the creation of an artifact (i.e., method, model) (Peffers et al., 2007). The fourth step is a demonstration of how the artifact solves the identified problem (Peffers et al., 2007). The fifth step is evaluation. This involves comparing the objectives of a solution to actual results from the use of the artifact during a demonstration (Peffers et al., 2007). The sixth step is communication which communicates the problem and its importance, the artifact, its utility and novelty, the rigour of its design, and its effectiveness to researchers and other relevant audiences such as practising professionals, when appropriate (Peffers et al., 2007). Figure 1 visualizes how the GM4M-IV implements the six steps of the DSRM as follows.

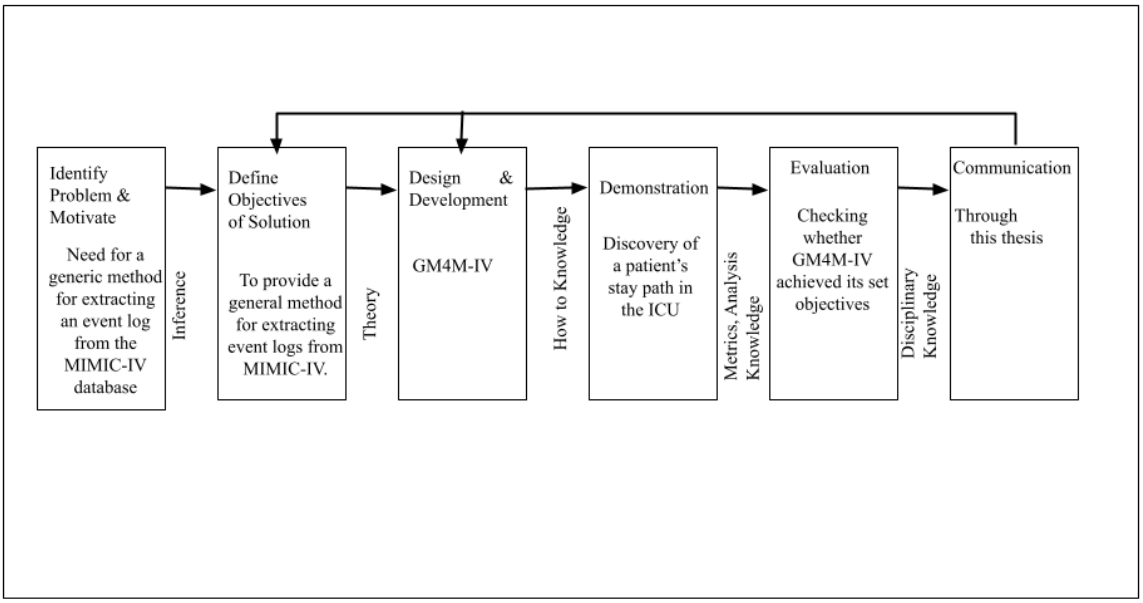

Figure 1:Brief description of tables from ed module

3.1 Problem identification and motivation

Modern information systems generate event logs automatically, but some explicitly store their process activities and require a method to extract their event logs (Alharbi et al., 2017). A process mining analyst must manually extract the event log from the MIMIC-IV database because it is not generated automatically. Based on section 4.1, there is no general method for extracting an event log from MIMIC-IV, hence the need for one. This thesis presents a solution GM4M-IV, a generic method for extracting an event log from the MIMIC-IV database. With a general method, process mining analysts will have a systematic approach of extracting event logs the database, which will enable the full exploitation of MIMIC-IV. Section 1 discusses the problem identification and motivation in detail.

3.2 Defining the objective of the solution

The objective of the thesis is to develop a generic method that guides a process mining analyst through the process of extracting an event log from the MIMIC-IV database. The generic method is to provide a step-by-step approach creating a simple sequence that both experienced and novice analysts can use when extracting event logs from MIMIC-IV. GM4M-IV will enable the user.

- Gain access to the MIMIC-IV database.
- Select and join tables from the MIMIC-IV database.
- Identify of events and process instance for the process mining project.
- Build queries to extract data from the MIMIC-IV to PostgreSQL and create an event log using bupaR.

3.3 Design and development

The artifact is GM4M-IV, a generic method for extracting an event log from the MIMIC-IV database. The thesis develops GM4M-IV based on existing literature on extracting an event log from a database. GM4M-IV starts with the user gaining access to the MIMIC-IV database, she then selects the required tables for her process mining project. The user thereafter selects a process instance and events that will make up the process instance. Next is the selection of the event attributes, building queries to extract data from the MIMIC-IV database and lastly is creating an event log using bupaR. Section 5 explains the generic method in detail.

3.4 Demonstration

As a proof of concept, the demonstration involves using the generic method in section 5 to extract an event log from the ICU module of MIMIC-IV. Using the extracted event log, a path for a patient's stay in the ICU is discovered. The details of the demonstration are in section 6.

3.5 Evaluation

There is an evaluation of the generic method by comparing the objectives of the method with the GM4M-IV method, this is to check whether the method meets the set objective. The evaluation of the generic method is in section 7.

3.6 Communication

This thesis communicates the generic guidelines for building an event log using MIMIC-IV.

4. Literature review

For the literature review, the thesis uses four of the most relevant search engines for information technology, namely, google scholar, PubMed, research gate and IEEE Xplore. In addition to the search engines, the thesis also utilises the library of Hasselt University for process mining textbooks. The thesis uses both MIMIC-III and MIMIC-IV databases for the literature review section. The inclusion of research on the MIMIC-III database is due to its considerable number of papers that provide insight into the overall potential of the MIMIC database. To narrow the search to relevant papers, the search engine filters are set to papers between 2020 - 2022 for MIMIC-IV and 2012-2022 for MIMIC-III.

4.1 Using the MIMIC-IV database for process mining

Section 4.1 discusses three research areas that the MIMIC-IV database for process mining, it should be noted that some papers are from MIMIC-III as little research is available using MIMIC-IV. The three research areas are sections 4.1.1 data, 4.1.2 diseases and 4.1.3 clinical pathways, each of the sections explains breakthroughs in process mining using data from the MIMIC. Section 4.1.1 discusses data-focused research from Cremerius and Weske (2021), Cremerius and Weske (2022) and Kurniati et al. (2018). Section 4.1.2 examines disease - research by Kurniati et al. (2018) and Kusuma et al. (2020), section 4.1.3 examines clinical pathways- focused research by Alharbi et al. (2017). Table 7 summarises the papers discussed in section 4.1 by indicating their focus area, MIMIC database version, and process mining technique used by the researchers. Table 7 also shows the contributions of the papers to process mining.

Table 7: Summary of papers using the MIMIC database for process mining

Focus	Paper	MIMIC version	Event log extraction method	Contribution
Data	Cremerius and Weske (2021)	MIMIC-IV	Not mentioned	A visualisation technique for enhancing discovered process data with domain data
	Cremerius and Weske (2022)	MIMIC-IV	Not mentioned	A method to classify domain data
	Kurniati et al. (2018)	MIMIC-III	L*lifecycle model	Data quality assessment
Diseases	Kurniati et al. (2018)	MIMIC-III	L*lifecycle model	Understanding cancer treatment pathways
	Kusuma et al. (2020)	MIMIC-III	PM^2	A disease trajectory mining method
Clinical pathways	Alharbi et al. (2017)	MIMIC-III	Manual extraction using PostgreSQL	An approach for detecting variations in clinical pathways

4.1.1 Data

Cremerius and Weske (2021) address the need for an adequate representation of domain data in process models. Domain data refers to additional event attributes collected during the execution of a process besides the essential ones (Cremerius & Weske, 2021). In their research, Cremerius and Weske (2021) introduce a visualisation technique for enhancing discovered process models with domain data. The technique is demonstrated using an event log extracted from the MIMIC-IV database for heart failure patients. Cremerius and Weske (2022) propose a method to classify domain data according to its process characteristics. The method measures the degree of variability among the process characteristics and uses the variability to filter event attributes. Cremerius and Weske (2022) apply their method to the attributes of an event log created from the MIMIC-IV database. Kurniati et al. (2018) demonstrate the applicability of process mining using the MIMIC III database by performing a data quality assessment, part of the assessment is to understand data quality issues. Kurniati et al. (2018) use the heuristics miner to discover the most followed admission path for cancer patients and analyse it to find a quality issue. The data quality issue found was that discharge took place after death.

4.1.2 Diseases

Kurniati et al. (2018) use the L*lifecycle model to discuss the potential of using MIMIC-III for process mining in oncology. Kurniati et al. (2018) gain insight into cancer treatment pathways by analysing an event log they built from cancer patient treatment records stored in MIMIC-III. The L*lifecycle model is a method for conducting process mining projects (van der Aalst et al., 2011). The standard L* model consists of five stages: plan and justify (Stage 0), extract (Stage 1), create a control-flow model and connect it to the event log (Stage 2), create an integrated process model (Stage 3), and provide operational support (Stage 4). Kusuma et al. (2020) present a novel method for mining disease trajectories. Disease trajectories describe the progression of chronic disease over time (Henly et al., 2011). Following the steps of PM^2, Kusuma et al. (2020) use the MIMIC III database to assess the directionality of a method to recognise the unique nature of disease trajectory models. The disease process models are discovered using the Interactive Data-aware Heuristics Miner (iDHM) plug-in (Kusuma et al., 2020). The iDHM is a process discovery tool that uses data attributes to improve the discovery procedure and provides built-in conformance checking to get direct feedback on the quality of the model (Mannhardt et al., 2017). PM^2 is a methodology that guides the execution of a process mining project.

4.1.3 Clinical pathways

Alharbi et al. (2017) develop an approach for detecting variations in clinical pathways. Alharbi et al. (2017), test the approach on clinical pathways data of diabetes patients with congestive heart failure. Alharbi et al. (2017) manually extract data from MIMIC-III and analyse it using the ProM process mining tool. ProM is a tool that converts data by providing a programming framework with supporting functions and a user-friendly interface (Gunter, 2009; Gunter et al., 2006; Buijs, 2010).

The six papers reviewed in section 4.1 do not provide a structured method for extracting an event log from the MIMIC database, hence the need for section 4.2 to examine research on extracting an event log from a database.

4.2 Extracting an event log

Section 4.2 is the primary source of literature for the development of the GM4M-IV, because the research reviewed in this section contains knowledge on event log extraction from different databases. Sections 4.2.1 examines approaches by Mahendrawathi et al. (2015), Ingvaldsen and Gulla (2007), Jans et al. (2019) for extracting an event log from ERP databases. Section 4.2.2 discuses research by Calvanese et al. (2016) on how to extract an event log from an ontology database, and section 4.2.3 examines literature by de Murillas et al. (2016) and Bano et al. (2021) about extracting an event log from redo logs.

4.2.1 ERP databases.

Mahendrawathi et al. (2015) discuss extracting an event log from an ERP system using the Extraction, Transformation and Loading (ETL) process. ERP systems are core software programs companies use to integrate and coordinate information in every area of the business. (Jans et al. 2019). Extraction involves writing Structured Query Language (SQL) queries to extract data from the ERP database. Transformation involves filtering the data to remain with relevant information for the process mining project. And Loading is uploading the data into a process mining application. Ingvaldsen and Gulla (2007) describe the Enterprise Visualisation suite (EVS), a log analysis system that supports pre-processing phase of process mining in Enterprise Resource Planning (ERP) systems. The EVS enables users to define at a high level how to store and transform events, resources, and their interactions for use in process mining. Jans et al. (2019) provide a step-by-step guideline for novice process mining analysts to follow when extracting an event log from an ERP database. When implementing the guideline, the user must first establish a primary business goal for the project and identify the key project

activities. Following that, the user identifies the key tables and their relationships, and then she selects a process instance document on which she bases her process instance level selection. Finally, the user selects attributes and associates them with activities.

4.2.2 Ontology database

Calvanese et al. (2016) presents a two-phase approach for extracting an event log from a legacy ontology relational database. The first phase is design, where the user describes the type of ontology data she will extract from the database and the second phase is design access, where the user extracts the data described in the first phase to create an event log. A relational database is a collection of information that organises data in predefined relationships and stores data in one or more tables (relations), making it simple to see and understand how different data structures relate to one another (Jatana et al., 2012).

4.2.3 Redo logs

De Murillas et al. (2016) present an approach that generates an event log from database redo logs. Redo logs are a list of changes made to a database as they occur (de Murillas et al., 2016). The approach has three steps: (i) Extract events from the redo log. (ii) Obtain the data model from the database (i.e., involves querying the tables, columns, and keys defined in the database schema) (de Murillas et al., 2016). (iii) Process instance identification; involves selecting which events go into what traces (de Murillas et al., 2016). Bano et al. (2021) propose database-less event log extraction from redo log, where the database schema is inferred automatically from the redo log. A domain expert evaluates the inferred schema, which is then used to extract an event log based on a selected case notion (Bano et al., 2021).

5. GM4M-IV: Using DSRM to develop the method.

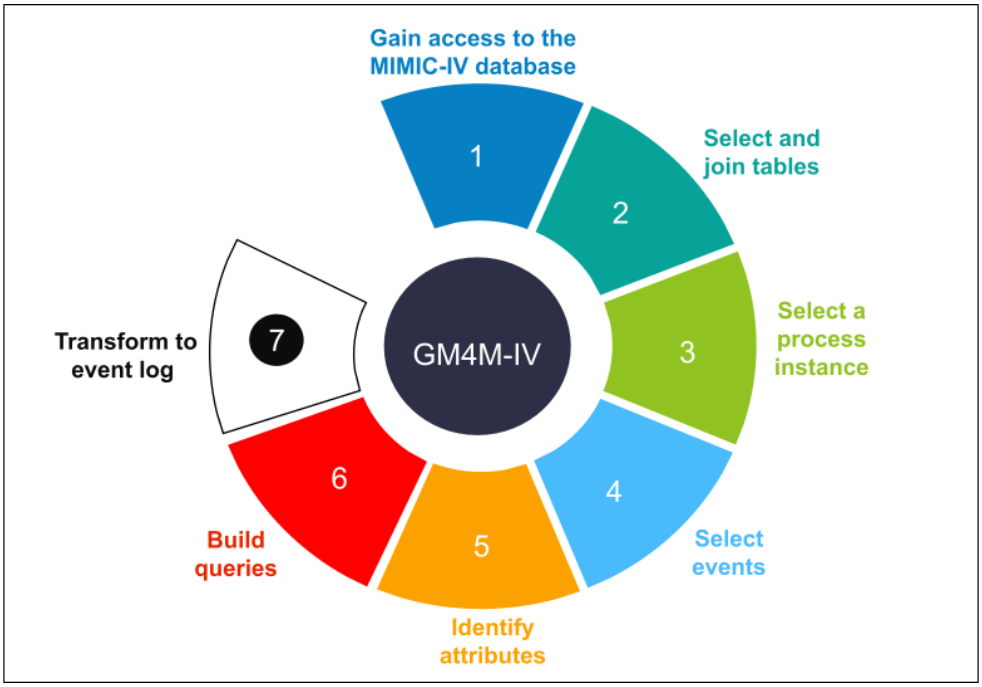

Figure 2: GM4M-IV

The GM4M-IV is a seven-step general method for extracting an event log from the MIMIC-IV database. When using GM4M-IV, the user starts with gaining access to the MIMIC-IV database. After successfully gaining access to MIMIC-IV, the user selects tables from the database that are required to meet the project's goal. Next, the user selects a process instance and events from the required tables. Finally, the user selects event log attributes, starts building queries and uses bupaR create an event log.

Five of the six steps of the GM4M-IV are developed based on knowledge from previous literature. Step 2 is based on Jans et al. (2019) and de Murillas et al. (2019), step 3 is inspired by Buijs (2010), Jans (2017) and Wei et al. (2022), step 4 (van der Aalst, 2016; Jans, 2017) step 5 (Buijs, 2010; Jans, 2017; Jans et al., 2019; Wei et al., 2022) and step 6 (Ingvaldsen et al. (2008).

GM4M-IV method assumes the user already has a holistic understanding of the process for the project she or he is doing. For instance, if the project is to discover an oncology treatment path, the user should at least have elementary knowledge about the activities that are performed during oncology treatment.

Examples of the activities of oncology treatment could include admitting the patient to the treatment unit, where his medical chart is checked to see which type of cancer and treatment the patient should receive. After which the right cancer treatment is administered to the patient. Below is a detailed explanation of the GM4M-IV method.

5.1 Step 1. Gain access to the MIMIC-IV database

The goal of step 1 is to gain access to the MIMIC-IV database, there are requirements that a user needs to fulfil to gain access to MIMIC-IV. First, the user must become a credentialed user of Physionet, the website that hosts the MIMIC-IV database, by opening an account. Secondly, the user must undertake a training course in Human Research and Data, or Specimen Only Research offered by the Collaborative Institute Training Initiative (CITI) program. The training course ensures the user knows and understands the ethics of conducting human research. The most important ethical principles are described in the Belmond Report, added to this thesis as appendix 1. Third, the user must sign a Data User Agreement (DUA). A DUA is a binding agreement in which a user pledges to use the data ethically. The user gains full access to the MIMIC database upon successfully meeting the above requirements.

5.2 Step 2. Select and join tables

The goal of step 2 is to select tables from the MIMIC-IV database. This step involves selecting tables that align with the process mining project. Jans et al. (2019) believe that only a specific subset of tables is relevant for a log to serve its purpose. A holistic understanding of the project acts as a guide toward selecting the relevant tables. For instance, the tables of admissions, patients, drgcodes, procedural events, input events, and output events would be some of the most relevant in an oncology treatment path because they contain information about the patient admission, the treatment procedures, and the infusions into and out of the patient's body, respectively. Please note, that the number of tables is not limited, the tables to use entirely depend on the user’s choice for her project. Section 2.3 contains a brief overview of the MIMIC-IV tables.

After selecting the relevant tables for the project, the user must choose a unique identifier which joins the tables (de Murillas et al., 2019). As mentioned in section 2.3, MIMIC–IV has three unique identifiers (ids) which are subject_id, hadm_id and stay_id. First, each patient in the MIMIC-IV database is assigned a subject_id specific to him. Second, an hadm_id is given to a patient for follow-up during their stay at the hospital. Third, is the stay_id which is assigned to a patient staying in the ICU.

5.3 Step 3. Select a process instance

The goal of step 3 is to select a process instance. Selecting a process instance involves identifying the sequence of events to include in the event log (Buijs, 2010). The user must determine the boundary of the process instance by identifying what triggers its start and end (Jans et al., 2019). Wei et al. (2022) emphasizes that most of the tables in the MIMIC-IV database have timestamps that aid the user in identifying the beginning and end of an activity instance. A patient's ICU stay might include, for instance, placing the patient on an oxygen ventilator at 15:00 on July 27, collecting blood samples at 15:20, and sponge bathing the patient at 15:30 on the same day.

5.4 Step 4. Select events

The goal of step 4 is to select events for the process instance. The user should decide which events to include in the process instance using a list of crucial SMART questions to which she anticipates receiving answers at the completion of the process mining project (van der Aalst, 2016; Jans, 2017). SMART stands for specific, measurable, attainable, relevant and time bound. An example of a SMART question could be “what happens during a patient’s stay in the ICU from admission to discharge?” With such a question, the user identifies events that happen during the patient’s stay in the ICU. The user identifies activities from the events and uses timestamps to know the start and end of the activity. An example of an event from MIMIC-IV is the admission (activity) of subject_id 10000032 (case) to the ICU on the 27th of July 2180 at 15:00 (timestamp). This step is followed by selecting the attributes of an event log.

5.5 Step 5. Identify attributes

This step focuses on identifying all relevant attributes from the selected MIMIC-IV in step 2 tables in addition to the three main attributes of an event log (caseID, activity and timestamp) (Wei et al., 2022; Jans et al., 2019). According to Buijs (2010), the user should carefully select the required attributes for the event log by avoiding the selection of too many attributes making the event log unnecessarily large. Or very few attributes which limit several analyses that would have been done on the event log (Buijs, 2010). MIMIC attributes may include subject_id, which is recorded in all the tables, procedures from procedures_icd table, chart time from chart events table and stay_id recorded in the ICU tables. Depending on the user's preference, the caseID might be the subject id or stay id, procedures would be an activity, chart time would be the timestamp.

5.6 Step 6. Build queries

After completing the earlier steps 1-5, the user imports data from MIMIC-IV to PostgreSQL (Ingvaldsen et al., 2008) to start writing queries. PostgreSQL is an open-source relational database system[3]. The queries extract relevant events by calling the attributes needed to build an event log (de Murillas et al., 2019). There user should focus on the following functions of PostgreSQL in table 8.

Table 8:PostgreSQL functions for extracting events from MIMIC-IV

Function	**Description**
CREATE TABLE	Creates a new table
SELECT	Extracts the specific attribute from a given table into the event log
DISTINCT	Ensures the attributes selected have no duplicated records
AS	Renames the attribute and records its values into another attribute the user chooses.
UNION	Unites the selected events or tables or attributes into one.
FROM	Indicates the table from which the user extracts events and attributes
ORDER BY	Orders the attributes in the sequence the user desires
NOT NULL	Does not return empty cells

To illustrating the functions in table 8, CREATE TABLE mimic. stayIDs AS (SELECT DISTINCT subject_id, hadm_id, stay_id, intime AS timestamp NOT NULL, FROM icutstays table UNION

[3] https://www.postgresql.org/about/

SELECT subject_id, stay_id, hadm_id, intime AS timestamp FROM edstays table). ORDER BY subject_id, icuID, hadm_id;

To export data from the MIMIC-IV database, the user must create tables in PostgreSQL with the same number of columns as those of the database using CREATE TABLE function. The user should then use SELECT DISTINCT to select column names with unique values for the created tables. Please keep in mind that the table names should be identical to those of the MIMIC-IV database tables from which the data is exported. To join tables, the user should use UNION and ORDER BY to arrange the columns in a preferred order. The symbol (;) alerts PostgreSQL that the query has ended, next the user pressures the run button on the upper right side of query tool to execute the query. The executed query returns a unified table which the user stores as a comma-separated values (CSV) file.

5.7 Step 7. Create event log

The goal of step is creating an event log using bupaR, is an open-source integrated suite of R-packages for handling the analysis of business process data (Janssenswillen et al., 2019). To create an event log using bupaR, the user extracts data from the CSV file by using the BupaR function. Thereafter, the user specifies the event log attributes then runs the code in R-studio[4], which is an integrated environment for bupaR programming. An event log is discovered summarising the number of cases, events, process instances and activities of the event log. Using the write_XES function, the event log is saved in the standard format for event logs. The eXtensible Event Stream (XES)[5] is a standard language that transports, stores, and exchanges event data.

[4] http://mercury.webster.edu/aleshunas/R_learning_infrastructure/Introduction_to_R_and_RStudio.html
[5] https://www.tf-pm.org/resources/xes-standard

6. Demonstration

In Section 6, there are two sub sections: Section 6.1, states the case and shows how to use GM4M-IV to produce an event log. Section 6.2 explains how to use an event log for process mining. Section 6 uses tables from the ICU module of the MIMIC-IV database as a demonstration case.

6.1 Using GM4M-IV to create an event log using ICU tables

6.1.1 Case. Discovery of a patient's stay path in the ICU

As mentioned in section 2.3, the MIMIC-IV database has six modules, the ICU module being one of them. The ICU module stores data about a patient's stay in seven tables, each containing specific details of what happened in the ICU. The ICU tables include D_items, chartevents, datetime events, Icustays, input events, output events and procedural events, these tables are displayed with their descriptions in Table 5 in section 2.3. The discovery of a patient's stay path in the ICU shows the various activities the patient experiences during their stay.

6.1.2 Step 1. Gain access to the MIMIC-IV database.

Access to the MIMIC-IV database is gained by opening an account on Physionet to become a credentialed user. Using the Physionet, we successfully completed the training course in Human Research and Data, or Specimen Only Research and signed a DUA giving us full access to the MIMIC-IV database.

6.1.3 Step 2. Select and join tables

Tables that contain data pertaining a patient's stay in the ICU are selected, these include, ICU stays, input events, output events and procedure events. ICU stays tables contains information regarding a patient's admission and discharge to the ICU, input events table contains information about continuous infusions or intermittent administrations into the patient, output events table contains information

regarding a patient's outputs like urine. And the procedure events table contains procedures documented during the stay in the ICU, although the procedures do not necessarily happen in the ICU.

6.1.4 Step 3. Select a process instance

From the selected tables, the process instance starts when the patient enters the ICU, followed by receiving inputs into his body, then output from the patient's body, and finally performing a procedure at the end of the process.

6.1.5 Step 4. Select events

Three questions have been identified to serve as a guide for selecting events, the first question is "What path does the patient follow during a stay in the ICU?" The second question is "What is the least followed stay path in the ICU?" and the third question is, "What is the most followed stay path in the ICU?" The selected events show the patient entering the ICU, infusions received, output from patient's body and the performed procedures.

Timestamps from when an activity began and ended are used to identify events from the selected tables. In the Icustsay tables, intime is used to determine when the patient enters the ICU, hence the activity name "Enter the ICU." In the input events tables, columns of start time and end time are used to determine infusions received by the patient, in the output events table hence the activity "Input infusion." The chart time column is used to determine when an output like urine was removed from the patient's body hence the activity "Output from patient." In the procedures table, start time and end time columns are considered to determine when the procedure started and ended hence the activity "Perform procedure."

6.1.6 Step 5. Identify attributes

The attributes are subject_id selected as the caseID, stay_id, timestamp, and activity. The subject_id identifies the patient, stay_id shows different stays of the patient in the ICU, timestamp shows when different activities took place and activities are what happened in the ICU.

6.1.7 Step 6. Build queries

In PostgreSQL, three queries are created, in the first query, there is creation of tables of four tables using the CREATE TABLE AS i.e., Icustays, input evets, output events and procedures events, data are imported into these tables from the selected ICU tables in the MIMIC-IV database. After importing the data to PostgreSQL, in the second query, four new tables are created using data from tables in the first query. The newly created tables are "Icu_icustays_activity_enter," "Icu_icustays_activity_input", "Icu_icustays_activity_output" and Icu_icustays_activity_procedures". Each of the tables contains four attributes which are subject_id, stay_id, timestamp, and activity. In the third query, the four tables are merged into one using the UNION function, to eliminate repeated data, DISTINCT function is applied to the new table "mimic_insights.icu_icutables." After merging the tables, the new table is saved as a CSV file and exported to R.

6.1.8 Step 7. Transform to event log

The CSV file is uploaded to R and transformed into an event log using bupaR. The script used is; Icutable %>% event log (case_id = "subject_id", activity_id = "activity", timestamp = "timestamp", activity_instance_id = "row_activityinstance", lifecycle transition= "lifecycle", resource_id = "resource") The script has three main attributes and four other attributes that bupaR suggests, without them the script fails to run. Figure 3 visualises the event log created using bupaR.

```
# A tibble: 9,252,765 × 8
    stay_id subject_id timestamp           activity           row_activityinstance lifecycle resource .order
      <int>      <int> <dttm>              <chr>                             <int>     <int>    <int>  <int>
 1 10000032   29079034 2180-07-23 15:00:00 Output from patient                1         1        1      1
 2 10000032   29079034 2180-07-23 17:00:00 Input infusion                      2         2        2      2
 3 10000032   29079034 2180-07-23 17:33:00 Input infusion                      3         3        3      3
 4 10000032   29079034 2180-07-23 18:56:00 Input infusion                      4         4        4      4
 5 10000032   29079034 2180-07-23 21:10:00 Input infusion                      5         5        5      5
 6 10000032   39553978 2180-07-23 14:00:00 Enter the ICU                       6         6        6      6
 7 10000980   26913865 2189-06-27 07:40:00 Input infusion                      7         7        7      7
 8 10000980   26913865 2189-06-27 09:08:00 Output from patient                 8         8        8      8
 9 10000980   26913865 2189-06-27 10:51:00 Input infusion                      9         9        9      9
10 10000980   26913865 2189-06-27 11:00:00 Output from patient                10        10       10     10
# … with 9,252,755 more rows
```

Figure 3: Event log of a patient's stay in the ICU

Figure 3 is a summary of the event log generated using the GM4M-IV. Figure 3 has the main attributes of an event log i.e., subject_id which is the caseID, timestamp and activity. Figure 3 consists of 73492 cases and 9252765 instances of four activities. The event log shows that events occur from 2110-01-11 12:30:00 until 2211-11-11 15:32:00. The event log activities answer the questions used when choosing the events in line with the project (a patient's stay in the ICU).

The process model shows the paths a patient takes once they enter the ICU. The first path is input infusion then output from the patient's body and the process ends. The second path is the patient enters the ICU, then input infusion, output from the patient's body and end. The third path is output from the patient's body and end. Lastly, the fourth is to perform the procedure and end. To answer questions, the most followed stay path in the ICU is the second path which starts with 23,542 patients entering the ICU and 27,205 exiting. The least followed path is the fourth where 19,882 procedures are performed first before exiting the ICU.

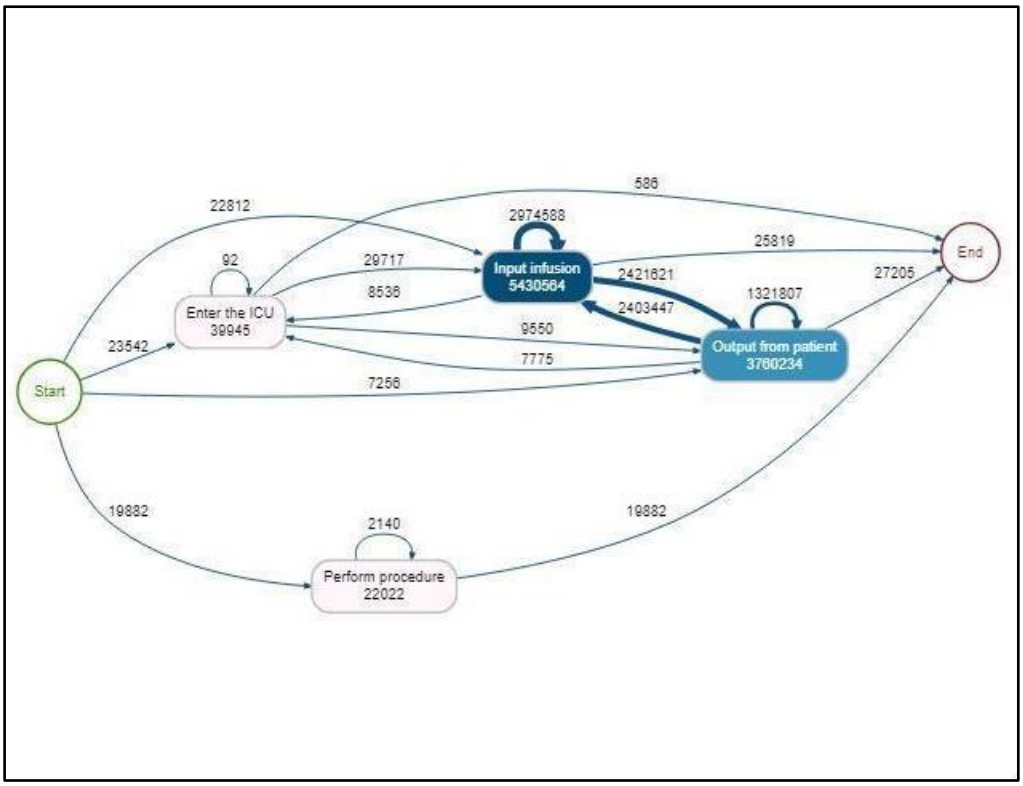

Figure 4: A patient's stay in the ICU

7. Evaluation

To meet the objective of the solution in section 3.2, GM4M-IV has seven easy to follow steps that guide both experienced and inexperienced process mining analyst to extract an event log from the MIMIC-IV database. GM4M-IV ensures the user gains access to the database by specifying the website that hosts the MIMIC-IV database and what training course should be taken. GM4M-IV highlights how the user can select and join MIMIC-IV tables that are relevant to the user's project, the method also explains how to identify a process instance and to select events that make up the process instance. GM4M-IV explains to the user the functions to use when building queries in PostgreSQL for extracting data from the MIMIC-IV database and transforming that data into a CSV file. GM4M-IV shows its user how to convert a CSV file to an event log using bupaR. Detailed descriptions are in section 5 and the demonstration in section 6.

8. Conclusion

The MIMIC-IV database is a freely accessible database containing de-identified medical data from patients admitted to critical care units from 2008-2019. Given its richness in critical care data, the MIMIC-IV database has great potential for process mining, the database can be used to discover disease trajectories, discover methods of improving data quality, discover clinical paths among others. In section 4.1, researchers use different process mining methods; PM2 (van Eck et al., 2015), L*lifecycle model (van der Aalst) and manual extraction (Alharbi et al., 2017) to extract event logs from MIMIC. This indicates a lack of a general method for extracting an event log from the MIMIC database. This thesis sought to provide a generic method and template for extracting an event log from MIMIC-IV, which is GM4M-IV, a method using existing literature on building event logs.

GM4M-IV has seven steps that a user follows to extract an event log from MIMIC-IV. To summarise the steps, the user starts with gaining access to MIMIC-IV. After successfully gaining access to MIMIC-IV, the user selects tables from the database that are required to meet the project's goal. Next, the user selects a process instance and events from the required tables. Finally, the user selects event log attributes, starts building queries and uses bupaR to create an event log. The feasibility of GM4M-IV is tested by extracting an event log from MIMIC-IV and using it to discover a process model for a patient's stay in the ICU.

Future research. The aim is to improve the usability of GM4M-IV. There is also a consideration of developing a tool based on the GM4M-IV which automatically extracts an event log from the MIMIC-IV database.

9. References

Agarwal, R., Gao, G. G., DesRoches, C., & Jha, A. K. (2010). Research Commentary. The Digital Transformation of Healthcare: current status and the road ahead. *Information Systems Research*.

Alharbi, A., Bulpitt, A., Johnson, O. (2017). Improving Pattern Detection in Healthcare Process Mining Using an Interval-Based Event Selection Method. *Business Information Processing.*

Archer, L. B. (1964). A systematic method for designers. *Design*

Bano, D., Lichtenstein, T., Klessascheck, F., & Weske, M. (2021, July). Database-less Extraction of Event Logs from Redo Logs. *Business Information Systems*

Batista, E., & Solanas, A. (2018). Process Mining in Healthcare: a systematic review. *9th International Conference on Information, Intelligence, Systems and Applications.*

Batyuk, A., & Voityshyn, V. (2018). Streaming Process Discovery for Lambda Architecture-Based Process Monitoring Platform. *IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies*

Buijs, J.C.A.M. (2010). Mapping data sources to xes in a generic way. *Student thesis: Master*

Bose, R. J. C., & van der Aalst, W. M. P. (2011). Analysis of patient treatment procedures: the BPI challenge case study

Bouarfa, L., & Dankelman, J. (2012). Workflow mining and outlier detection from clinical activity logs. *Journal of Biomedical Informatic*

Calvanese, D., Montali, M., Syamsiyah, A., & van der Aalst, W. M. (2016, September). Ontology-driven extraction of event logs from relational databases. In *International Conference on Business Process Management*

Calvert, J. S., Price, D. A., Chettipally, U. K., Barton, C. W., Feldman, M. D., Hoffman, J. L., Jay, M., & Das, R. (2016). A computational approach to early sepsis detection. *Computers in Biology and Medicine*

Caron, F., Vanthienen, J., De Weerdt, J., Baesens, B., De Weerdt, J., & Baesens, B. (2011). Beyond x-raying a care-flow: adopting different focuses on care-flow mining. *Business Process Intelligence*

Chan, B. K. C. (2018). Data Analysis using R Programming. *Advances in Experimental Medicine and Biology*

Chen, Y., Patel, M. B., McNaughton, C. D., & Malin, B. A. (2018). Interaction patterns of trauma providers are associated with length of stay. *American Medical Informatics Association*

Cremerius, J. & Weske, M. (2021). Data Enhanced Process Models in Process Mining. *ZEUS*

Cremerius, J., & Weske, M. (2022). Supporting Domain Data Selection in Data-Enhanced Process Models. *arXiv preprint arXiv*

Cremerius, J., König, M., Warmuth, C., & Weske, M. (2022). Patient Discharge Classification Based on the Hospital Treatment Process. *International Conference on Process Mining*

Dallagassa, M. R., dos Santos Garcia, C., Scalabrin, E. E., Ioshii, S. O., & Carvalho, D. R. (2021). Opportunities and Challenges for Applying Process Mining in Healthcare: a systematic mapping study. *Ambient Intelligence and Humanized Computing*

D'Castro, R. J., Oliveira, A. L., & Terra, A. H. (2018, October). Process Mining Discovery Techniques in a Low-Structured Process Works? *Brazilian Conference on Intelligent Systems*

de Murillas, E. G. L., van der Aalst, W. M., & Reijers, H. A. (2016, September). Process mining on databases: Unearthing historical data from redo logs. In *International Conference on Business Process Management*

De Smedt, J., Hasić, F., vanden Broucke, S. K., & Vanthienen, J. (2019). Holistic discovery of decision models from process execution data. *Knowledge-Based Systems*

Diba, K., Batoulis, K., Weidlich, M., & Weske, M. (2020). Extraction, correlation, and abstraction of event data for process mining. *Wiley Interdisciplinary Reviews: data mining and knowledge discovery*

Eekels, J., & Roozenburg, N. F. (1991). A Methodological Comparison of the Structures of Scientific Research and Engineering Design: their similarities and differences. *Design studies*

Erdogan, T. G., & Tarhan, A. (2018). Systematic mapping of process mining studies in healthcare. *IEEE Access*

Fei, H., & Meskens, N. (2010). Discovering patient care process models from event logs. *International conference of modeling*

Ghahfarokhi, A. F., Park, G., Berti, A., & van der Aalst, W. M. P. (2021). OCEL: a standard for object-centric event logs. In *Communications in Computer and Information Science*

Ghasemi, M., & Amyot, D. (2016). Process Mining in Healthcare: a systematised literature review. *International Journal of Electronic Healthcare*

González López De Murillas, E., Reijers, H. A., & van der Aalst, W. M. P. (2019). Case notion discovery and recommendation: automated event log building on databases. *Knowledge and Information Systems*

Gonzalez Lopez de Murillas, E. (2019). Process mining on databases: extracting event data from real-life data sources. *Technische Universiteit Eindhoven*.

González López De Murillas, E., Reijers, H. A., & van der Aalst, W. M. P. (2018). Connecting databases with process mining: a meta model and toolset. *Software & Systems Modeling*

G¨unther, C. (2009). Process Mining in Flexible Environments. PhD thesis, *Eindhoven University of Technology*.

G¨unther, C.W., & van der Aalst, W.M.P. (2006). A generic import framework for process event logs. *Business Process Intelligence*

Harutyunyan, H., Khachatrian, H., Kale, D. C., ver Steeg, G., & Galstyan, A. (2019). Multitask Learning and Benchmarking with Clinical Time Series Data. *Scientific Data*

Haux, R. (2006). Health information systems–past, present, future. *International journal of medical informatics*

Henly, S. J., Wyman, J. F., & Findorff, M. J. (2011). Health and illness over time: the trajectory perspective in nursing science. *Nursing research*

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS quarterly*

IEEE Standard for eXtensible Event Stream (XES) for Achieving Interoperability in Event Logs and Event Streams (2016)

Ingvaldsen, J. E., & Gulla, J. A. (2007). Pre-processing Support for Large Scale Process Mining of SAP transactions. *International Conference on Business process management*

Jans, M. J., Alles, M., & Vasarhelyi, M. A. (2010). Process Mining of Event Logs in Auditing: opportunities and challenges. *SSRN Electronic Journal*

Jans, M. (2017). From Relational Database to Valuable Event Logs for Process Mining Purposes: a procedure. *Technical report, Hasselt University.*

Jans, M., Soffer, P. (2018). From Relational Database to Event Log: decisions with quality impact. *Business Process Management Workshop*

Jans, M., Soffer, P., & Jouck, T. (2019). Building a Valuable Event Log for Process Mining: an experimental exploration of a guided process. *Enterprise Information System*

Janssenswillen, G., Depaire, B., Swennen, M., Jans, M., & Vanhoof, K. (2019). bupaR: enabling reproducible business process analysis. *Knowledge-Based System*

Jansen-Vullers, M. H., & Reijers, H. A. (2005). Business Process Redesign in Healthcare: towards a structured approach

Jatana, N., Puri, S., Ahuja, M., Kathuria, I., & Gosain, D. (2012). A Survey and Comparison of Relational and Non-relational Database. *International Journal of Engineering Research & Technology*

Johnson, S. B. (1996). Generic Data Modeling for Clinical Repositories. *Journal of the American Medical Informatics Association*

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L. W. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., & Mark, R. G. (2016). MIMIC-III, a Freely Accessible Critical Care Database. *Scientific Data*

Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L. A., & Mark, R. (2021). *MIMIC-IVv1.0*

Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L. A., & Mark, R. (2022). MIMIC-IV (version 2.0). *PhysioNet*

Kalenkova, A. A., van der Aalst, W. M. P., Lomazova, I. A., & Rubin, V. A. (2017). Process Mining Using Bpmn: relating event logs and process models. *Software and Systems Modeling*

Kaymak, U., Mans, R., Steeg v. d. T., & Dierks, M. (2012). On Process Mining in Health Care. *IEEE International Conference on Systems, Man, and Cybernetics*

Kempa-Liehr, A. W., Lin, C. Y. C., Britten, R., Armstrong, D., Wallace, J., Mordaunt, D., & O'Sullivan, M. (2020). Healthcare Pathway Discovery and Probabilistic Machine Learning. *International Journal of Medical Informatics*

Kia, A. S., Beheshti, M., & Shahmoradi, L. (2022). Health Information Systems Evaluation Criteria: overview of systematic reviews. *Frontiers in Health Informatics*

Kim, E., Kim, S., Song, M., Kim, S., Yoo, D., Hwang, H., & Yoo, S. (2013). Discovery of Outpatient Care Process of a Tertiary University Hospital Using Process Mining. *Healthcare Informatics Research*

Kurniati, A. P., Hall, G., Hogg, D., & Johnson, O. (2018). Process Mining in Oncology Using The MIMIC-III Dataset. *Journal of Physics: Conference Series*

Kurniati, A. P., Rojas, E., Hogg, D., Hall, G., & Johnson, O. A. (2018). 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*

Kusuma, G., Kurniati, A., McInerney, C. D., Hall, M., Gale, C. P., & Johnson, O. (2021, October). Process Mining of Disease Trajectories in MIMIC-III: a case study. *International Conference on Process Mining*

Lenz, R., & Reichert, M. (2007). IT Support for Healthcare Processes Premises, Challenges, Perspectives. *Data & Knowledge Engineering*

Li, Y., Bai, C., & Reddy, C. K. (2016). A Distributed Ensemble Approach for Mining Healthcare Data Under Privacy Constraints. *Information Sciences*

Lichtenstein, T., Bano, D., & Weske, M. (2021). Attribute-Driven Case Notion Discovery for Unlabelled Event Logs. *International Conference on Business Process Management*

Liu, C., van Dongen, B., Assy, N., & van der Aalst, W. M. (2016). Component behavior discovery from software execution data. *IEEE symposium series on computational intelligence*

Lippeveld, T., Sauerborn, R., & Bodart, C. (2000). Design And Implementation of Health Information Systems. *World Health Organization*

Mahendrawathi, E., Astuti, H.M., Wardhani, I.R.K. (2015). Material Movement Analysis for Warehouse Business Process Improvement with Process Mining: a case study. *Pacific Business Process Management*

Mannhardt, F., De Leoni, M., & Reijers, H. A. (2017, September). Heuristic Mining Revamped: an interactive, data-aware, and conformance-aware miner. *Bpm (demos)*.

Mans, R. S., Schonenberg, M. H., Song, M., van der Aalst, W. M., & Bakker, P. J. (2008). Application of Process Mining in Healthcare: a case study in a Dutch hospital. *International Joint Conference on Biomedical Engineering Systems and Technologies*

Mans, R. S., Van der Aalst, W. M., Vanwersch, R. J., & Moleman, A. J. (2012). Process Mining in Healthcare: data challenges when answering frequently posed questions. *Process Support and Knowledge Representation in Health Care*

Mans, R., Reijers, H., van Genuchten, M., & Wismeijer, D. (2012, January). Mining Processes in Dentistry. *ACM SIGHIT International Health Informatics Symposium*

Marques, I. C. P., & Ferreira, J. J. M. (2019). Digital Transformation in the Area Of Health: systematic review of 45 years of evolution. *Health and Technology*

Martin, N., de Weerdt, J., Fernández-Llatas, C., Gal, A., Gatta, R., Ibáñez, G., Johnson, O., Mannhardt, F., Marco-Ruiz, L., Mertens, S., Munoz-Gama, J., Seoane, F., Vanthienen, J., Wynn, M. T., Boilève, D. B., Bergs, J., Joosten-Melis, M., Schretlen, S., & van Acker, B. (2020). Recommendations For Enhancing the Usability and Understandability of Process Mining in Healthcare. *Artificial Intelligence in Medicine*

Maruster, L., van der Aalst, W., Weijters, T., van den Bosch, A., & Daelemans, W. (2001). Automated Discovery of Workflow Models from Hospital Data.

Munoz-Gama, J., Martin, N., Fernandez-Llatas, C., Johnson, O. A., Sepúlveda, M., Helm, E., Galvez-Yanjari, V., Rojas, E., Martinez-Millana, A., Aloini, D., Amantea, I. A., Andrews, R., Arias, M., Beerepoot, I., Benevento, E., Burattin, A., Capurro, D., Carmona, J., Comuzzi, M., . . . Zerbato, F. (2022). Process mining for healthcare: characteristics and challenges. *Journal of Biomedical Informatic*

Nunamaker Jr, J. F., Chen, M., & Purdin, T. D. (1990). Systems Development in Information Systems Research. *Journal of management information systems*

Pakbin, A., Wang, X., Mortazavi, B. J., & Lee, D. K. (2021). BoXHED2. 0: scalable boosting of dynamic survival analysis.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*

Poelmans, G. Dedene, G. Verheyden, H. Van Der Mussele, S. Viaene, and E. Peters. (2010). Combining Business Process and Data Discovery Techniques for Analysing and Improving Integrated Care Pathways. *Conference of Advanced data Mining Application*

Pourbafrani, M., & van der Aalst, W. M. P. (2021). Extracting Process Features from Event Logs to Learn Coarse-Grained Simulation Models. *Lecture Notes in Computer Science*

Perimal-Lewis, L., De Vries, D., & Thompson, C. H. (2014). Health Intelligence: discovering the process model using process mining by constructing Start-to-End patient journeys. *Proceedings of the Seventh Australasian Workshop on Health Informatics and Knowledge Management*

Rattanavayakorn, P., & Premchaiswadi, W. (2015). Analysis of the Social Network Miner (working together) of Physicians. *International Conference on ICT and Knowledge Engineering*

Reddy, C. K., & Aggarwal, C. C. (2015). *Healthcare data analytics*. Amsterdam University Press

Rojas, E., Arias, M., & Sepúlveda, M. (2015). Clinical Processes and its Data, What Can We Do with Them. In *Proceedings of the International Conference on Health Informatics*

Rojas, E., Munoz-Gama, J., Sepúlveda, M., & Capurro, D. (2016). Process Mining in Healthcare: a literature review. *Journal of Biomedical Informatics*

Rojas, E., Sepúlveda, M., Munoz-Gama, J., Capurro, D., Traver, V., & Fernandez-Llatas, C. (2017). Question-Driven Methodology for Analysing Emergency Room Processes Using Process Mining. *Applied Sciences*

Roock, E. De, & Martin, N. (2022). Process Mining in Healthcare: An Updated Perspective on the State of the Art. *Journal of Biomedical Informatics*

Rossi, M., & Sein, M. K. (2003). Design Research Workshop: a proactive research approach. *Presentation delivered at IRIS*

Rule, A., Chiang, M. F., & Hribar, M. R. (2019). Using Electronic Health Record Audit Logs to Study Clinical Activity: a systematic review of aims, measures, and methods. *Journal of the American Medical Informatics Associatio*n

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*

van Der Aalst, W. M. (2011, April). Process Mining: discovering and improving Spaghetti and lasagna processes. *IEEE Symposium on Computational Intelligence and Data Mining*

van Der Aalst, W., Adriansyah, A., De Medeiros, A. K. A., Arcieri, F., Baier, T., Blickle, T., & Wynn, M. (2011). Process mining manifesto. *International Conference on Business Process Management*

van Der Aalst, W. (2012). Process Mining: overview and opportunities. *ACM transactions on management information systems*

van der Aalst, W. (2016). Process *Mining: data science in action*

van Eck, M. L., Lu, X., Leemans, S. J. J., & van der Aalst, W. M. P. (2015). PM^{2}: A process mining project methodology. *Advanced Information Systems Engineering*

Vanhaecht K, De Witte K, Sermeus W. (2007). The Impact of Clinical Pathways on the Organisation of Care Processes. *PhD dissertation, Belgium: KU Leuven*

Verbeek, H. M. W., Buijs, J. C. A. M., van Dongen, B. F., & van der Aalst, W. M. P. (2011). XES, XESame, and ProM 6. *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*

Walls, J. G., Widmeyer, G. R., & El Sawy, O. A. (1992). Building an Information System Design Theory for Vigilant EIS. *Information systems research*,

Wang, S., McDermott, M. B., Chauhan, G., Ghassemi, M., Hughes, M. C., & Naumann, T. (2020). MIMIC-Extract: a data extraction, pre-processing, and representation pipeline for MIMIC-III. In *Proceedings of the ACM conference on health, inference, and learning*

Wei, J., He, Z., Ouyang, C., & Moreira, C. (2022). MIMICEL: MIMIC-IV event log for emergency department (version 1.0.0). *PhysioNet*

Weijters, A., & Ribeiro, J. (2011). Flexible Heuristics Miner (FHM). *IEEE Symposium on Computational Intelligence and Data Mining*

Pulsanong, W., Porouhan, P., Tumswadi, S., & Premchaiswadi, W. (*2017)*. Using Inductive Miner to Find the Most Optimised Path of Workflow Process. *International Conference on ICT and Knowledge Engineering*

Yang, W., & Su, Q. (2014). Process Mining for Clinical Pathway: literature review and future directions. *International Conference on Service Systems and Service Management*

10. Appendices

The Belmont Report

Office of the Secretary

Ethical Principles and Guidelines for the Protection of Human Subjects of Research

The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research

April 18, 1979

AGENCY: Department of Health, Education, and Welfare.

ACTION: Notice of Report for Public Comment.

SUMMARY: On July 12, 1974, the National Research Act (Pub. L. 93-348) was signed into law, there-by creating the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. One of the charges to the Commission was to identify the basic ethical principles that should underlie the conduct of biomedical and behavioural research involving human subjects and to develop guidelines which should be followed to assure that such research is conducted in accordance with those principles. In carrying out the above, the Commission was directed to consider: **(i)** the boundaries between biomedical and behavioural research and the accepted and routine practice of medicine, **(ii)** the role of assessment of risk-benefit criteria in the determination of the appropriateness of research involving human subjects, **(iii)** appropriate guidelines for the selection of human subjects for participation in such research and **(iv)** the nature and definition of informed consent in various research settings.

The Belmont Report attempts to summarize the basic ethical principles identified by the Commission in the course of its deliberations. It is the outgrowth of an intensive four-day period of discussions that

were held in February 1976 at the Smithsonian Institution's Belmont Conference Canter supplemented by the monthly deliberations of the Commission that were held over a period of nearly four years. It is a statement of basic ethical principles and guidelines that should assist in resolving the ethical problems that surround the conduct of research with human subjects. By publishing the Report in the Federal Register, and providing reprints upon request, the Secretary intends that it may be made readily available to scientists, members of Institutional Review Boards, and Federal employees. The two-volume Appendix, containing the lengthy reports of experts and specialists who assisted the Commission in fulfilling this part of its charge, is available as DHEW Publication No. (OS) 78-0013 and No. (OS) 78-0014, for sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402.

Unlike most other reports of the Commission, the Belmont Report does not make specific recommendations for administrative action by the Secretary of Health, Education, and Welfare. Rather, the Commission recommended that the Belmont Report be adopted in its entirety, as a statement of the Department's policy. The Department requests public comment on this recommendation.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research

Members of the Commission

Kenneth John Ryan, M.D., Chairman, Chief of Staff, Boston Hospital for Women.

Joseph V. Brady, Ph.D., Professor of Behavioral Biology, Johns Hopkins University.

Robert E. Cooke, M.D., President, Medical College of Pennsylvania.

Dorothy I. Height, President, National Council of Negro Women, Inc.

Albert R. Jonsen, Ph.D., Associate Professor of Bioethics, University of California at San Francisco. Patricia King, J.D., Associate Professor of Law, Georgetown University Law Center.

*Karen Lebacqz, Ph.D., Associate Professor of Christian Ethics, Pacific School of Religion. *** David W. Louisell, J.D., Professor of Law, University of California at Berkeley.*

Donald W. Seldin, M.D., Professor and Chairman, Department of Internal Medicine, University of Texas at Dallas.

*** *Eliot Stellar, Ph.D., Provost of the University and Professor of Physiological Psychology, University of Pennsylvania.*

*** *Robert H. Turtle, LL.B., Attorney, VomBaur, Coburn, Simmons & Turtle, Washington, D.C.* *** Deceased.

Table of Contents

Ethical Principles & Guidelines for Research Involving Human Subjects

Scientific research has produced substantial social benefits. It has also posed some troubling

ethical questions. Public attention was drawn to these questions by reported abuses of human subjects in biomedical experiments, especially during the Second World War. During the Nuremberg War Crime Trials, the Nuremberg code was drafted as a set of standards for judging physicians and scientists who had conducted biomedical experiments on concentration camp prisoners. This code became the prototype of many later codes [1] intended to assure that research involving human subjects would be carried out in an ethical manner.

The codes consist of rules, some general, others specific, that guide the investigators or the reviewers of research in their work. Such rules often are inadequate to cover complex situations; at times they come into conflict, and they are frequently difficult to interpret or apply. Broader ethical principles will provide a basis on which specific rules may be formulated, criticized and interpreted.

Three principles, or general prescriptive judgments, which are relevant to research involving human subjects are identified in this statement. Other principles may also be relevant. These three are comprehensive, however, and are stated at a level of generalization that should assist scientists, subjects, reviewers and interested citizens to understand the ethical issues inherent in research involving human subjects. These principles cannot always be applied so as to resolve beyond dispute particular ethical problems. The objective is to provide an analytical framework that will guide the resolution of ethical problems arising from research involving human subjects.

This statement consists of a distinction between research and practice, a discussion of the three basic ethical principles, and remarks about the application of these principles.

Part A: Boundaries Between Practice & Research

A. Boundaries Between Practice and Research

It is important to distinguish between biomedical and behavioral research, on the one hand, and the practice of accepted therapy on the other, in order to know what activities ought to undergo review for the protection of human subjects of research. The distinction between research and practice is blurred partly because both often occur together (as in research designed to evaluate a therapy) and partly because notable departures from standard practice are often called "experimental" when the terms "experimental" and "research" are not carefully defined.

For the most part, the term "practice" refers to interventions that are designed solely to enhance the wellbeing of an individual patient or client and that have a reasonable expectation of success. The purpose of medical or behavioral practice is to provide diagnosis, preventive treatment or therapy to particular individuals [2]. By contrast, the term "research' designates an activity designed to test a hypothesis, permit conclusions to be drawn, and thereby to develop or contribute to generalizable knowledge (expressed, for example, in theories, principles, and statements of relationships). Research is usually described in a formal protocol that sets forth an objective and a set of procedures designed to reach that objective.

When a clinician departs in a significant way from standard or accepted practice, the innovation does not, in and of itself, constitute research. The fact that a procedure is "experimental," in the sense of new, untested or different, does not automatically place it in the category of research. Radically new procedures of this
description should, however, be made the object of formal research at an early stage in order to determine whether they are safe and effective. Thus, it is the responsibility of medical practice committees, for example, to insist that a major innovation be incorporated into a formal research project [3].

Research and practice may be carried on together when research is designed to evaluate the safety and efficacy of a therapy. This need not cause any confusion regarding whether or not the activity requires review; the general rule is that if there is any element of research in an activity, that activity should undergo review for the protection of human subjects.

Part B: Basic Ethical Principles

B. Basic Ethical Principles

The expression "basic ethical principles" refers to those general judgments that serve as a basic justification for the many particular ethical prescriptions and evaluations of human actions. Three basic principles, among those generally accepted in our cultural tradition, are particularly relevant to the ethics of research involving human subjects: the principles of respect of persons, beneficence and justice.

1. Respect for Persons. — Respect for persons incorporates at least two ethical convictions: first, that individuals should be treated as autonomous agents, and second, that persons with diminished autonomy are entitled to protection. The principle of respect for persons thus divides into two separate moral requirements: the requirement to acknowledge autonomy and the requirement to protect those with diminished autonomy.

An autonomous person is an individual capable of deliberation about personal goals and of acting under the direction of such deliberation. To respect autonomy is to give weight to autonomous persons' considered opinions and choices while refraining from obstructing their actions unless they are clearly detrimental to others. To show lack of respect for an autonomous agent is to repudiate that person's considered judgments, to deny an individual the freedom to act on those considered judgments, or to withhold information necessary to make a considered judgment, when there are no compelling reasons to do so.

However, not every human being is capable of self-determination. The capacity for self-determination matures during an individual's life, and some individuals lose this capacity wholly or in part because of illness, mental disability, or circumstances that severely restrict liberty. Respect for the immature and the incapacitated may require protecting them as they mature or while they are incapacitated.

Some persons are in need of extensive protection, even to the point of excluding them from activities which may harm them; other persons require little protection beyond making sure they undertake activities freely and with awareness of possible adverse consequence. The extent of protection afforded should depend upon the risk of harm and the likelihood of benefit. The judgment that any individual lacks autonomy should be periodically reevaluated and will vary in different situations.

In most cases of research involving human subjects, respect for persons demands that subjects enter into the research voluntarily and with adequate information. In some situations, however, application of the principle is not obvious. The involvement of prisoners as subjects of research provides an instructive example. On the one hand, it would seem that the principle of respect for persons requires that prisoners not be deprived of the opportunity to volunteer for research. On the other hand, under prison conditions they may be subtly coerced or unduly influenced to engage in research activities for which they would not otherwise volunteer. Respect for persons would then dictate that prisoners be protected. Whether to allow
prisoners to "volunteer" or to "protect" them presents a dilemma. Respecting persons, in most hard cases, is often a matter of balancing competing claims urged by the principle of respect itself.

2. Beneficence. — Persons are treated in an ethical manner not only by respecting their decisions and protecting them from harm, but also by making efforts to secure their well-being. Such treatment falls under the principle of beneficence. The term "beneficence" is often

understood to cover acts of kindness or charity

that go beyond strict obligation. In this document, beneficence is understood in a stronger sense, as an obligation. Two general rules have been formulated as complementary expressions of beneficent actions in this sense: **(1)** do not harm and **(2)** maximize possible benefits and minimize possible harms.

The Hippocratic maxim "do no harm" has long been a fundamental principle of medical ethics. Claude Bernard extended it to the realm of research, saying that one should not injure one person regardless of the benefits that might come to others. However, even avoiding harm requires learning what is harmful; and, in the process of obtaining this information, persons may be exposed to risk of harm. Further, the Hippocratic Oath requires physicians to benefit their patients "according to their best judgment." Learning what will in fact benefit may require exposing persons to risk. The problem posed by these imperatives is to decide when it is justifiable to seek certain benefits despite the risks involved, and when the benefits should be foregone because of the risks.

The obligations of beneficence affect both individual investigators and society at large, because they extend both to particular research projects and to the entire enterprise of research. In the case of particular projects, investigators and members of their institutions are obliged to give forethought to the maximization of benefits and the reduction of risk that might occur from the research investigation. In the case of scientific research in

general, members of the larger society are obliged to recognize the longer-term benefits and risks that may result from the improvement of knowledge and from the development of novel medical, psychotherapeutic, and social procedures.

The principle of beneficence often occupies a well-defined justifying role in many areas of research involving human subjects. An example is found in research involving children. Effective ways of treating childhood diseases and fostering healthy development are benefits

that serve to justify research involving children -- even when individual research subjects are not direct beneficiaries. Research also makes it possible to avoid the harm that may result from the application of previously accepted routine practices that on closer investigation turn out to be dangerous. But the role of the principle of beneficence is not always so unambiguous. A difficult ethical problem remains, for example, about research that presents more than minimal risk without immediate prospect of direct benefit to the children involved. Some have argued that such research is inadmissible, while others have pointed out that this limit would rule out much research promising great benefit to children in the future. Here again, as with all hard cases, the different claims covered by the principle of beneficence may come into conflict and force difficult choices.

3. Justice. — Who ought to receive the benefits of research and bear its burdens? This is a question of justice, in the sense of "fairness in distribution" or "what is deserved." An injustice occurs when some benefit to which a person is entitled is denied without good reason or when some burden is imposed unduly. Another way of conceiving the principle of justice is that equals ought to be treated equally. However, this statement requires explication. Who is equal and who is unequal? What considerations justify departure from equal distribution? Almost all commentators allow that distinctions based on experience, age, deprivation, competence, merit and position do sometimes constitute criteria justifying differential treatment for certain purposes. It is necessary, then, to explain in what respects people should be treated equally. There are several widely accepted formulations of just ways to distribute burdens and benefits. Each formulation mentions some relevant property on the basis of which burdens and benefits should be distributed. These formulations are **(1)** to each person an equal share, **(2)** to each person according to individual need, **(3)** to each person according to individual effort, **(4)** to each person according to societal contribution, and **(5)** to each person according to merit.

Questions of justice have long been associated with social practices such as punishment, taxation and political representation. Until recently these questions have not generally been

associated with scientific research. However, they are foreshadowed even in the earliest reflections on the ethics of research involving human subjects. For example, during the 19th and early 20th centuries the burdens of serving as research subjects fell largely upon poor ward patients, while the benefits of improved medical care flowed primarily to private patients. Subsequently, the exploitation of unwilling prisoners as research subjects in Nazi concentration camps was condemned as a particularly flagrant injustice. In this country, in the 1940's, the Tuskegee syphilis study used disadvantaged, rural black men to study the untreated course of a disease that is by no means confined to that population. These subjects were deprived of demonstrably effective treatment in order not to interrupt the project, long after such treatment became generally available.

Against this historical background, it can be seen how conceptions of justice are relevant to research involving human subjects. For example, the selection of research subjects needs to be scrutinized in order to determine whether some classes (e.g., welfare patients, particular racial and ethnic minorities, or persons confined to institutions) are being systematically selected simply because of their easy availability, their compromised position, or their manipulability, rather than for reasons directly related to the problem being studied. Finally, whenever research supported by public funds leads to the development of therapeutic devices and procedures, justice demands both that these not provide advantages only to those who can afford them and that such research should not unduly involve persons from groups unlikely to be among the beneficiaries of subsequent applications of the research.

Part C: Applications

C. Applications

Applications of the general principles to the conduct of research leads to consideration of the following requirements: informed consent, risk/benefit assessment, and the selection of subjects of research.

1. Informed Consent. — Respect for persons requires that subjects, to the degree that they are capable, be given the opportunity to choose what shall or shall not happen to them. This opportunity is provided when adequate standards for informed consent are satisfied.

While the importance of informed consent is unquestioned, controversy prevails over the nature and possibility of an informed consent. Nonetheless, there is widespread agreement that the consent process can be analyzed as containing three elements: information, comprehension and voluntariness.

Information. Most codes of research establish specific items for disclosure intended to assure that subjects are given sufficient information. These items generally include: the research procedure, their purposes, risks and anticipated benefits, alternative procedures (where therapy is involved), and a statement offering the subject the opportunity to ask questions and to withdraw at any time from the research. Additional items have been proposed, including how subjects are selected, the person responsible for the research, etc.

However, a simple listing of items does not answer the question of what the standard should be for judging how much and what sort of information should be provided. One standard frequently invoked in medical practice, namely the information commonly provided by practitioners in the field or in the locale, is inadequate since research takes place precisely when a common understanding does not exist. Another standard, currently popular in malpractice law, requires the practitioner to reveal the information that reasonable persons would wish to know in order to make a decision regarding their care. This, too, seems insufficient since the research subject, being in essence a volunteer, may wish to know considerably more about risks gratuitously undertaken than do patients who deliver themselves into the hand of a clinician for

needed care. It may be that a standard of "the reasonable volunteer" should be proposed: the

extent and nature of information should be such that persons, knowing that the procedure is neither necessary for their care nor perhaps fully understood, can decide whether they wish to participate in the furthering of knowledge. Even when some direct benefit to them is anticipated, the subjects should understand clearly the range of risk and the voluntary nature of participation.

A special problem of consent arises where informing subjects of some pertinent aspect of the research is likely to impair the validity of the research. In many cases, it is sufficient to indicate to subjects that they are being invited to participate in research of which some features will not be revealed until the research is concluded. In all cases of research involving incomplete disclosure, such research is justified only if it is clear that **(1)** incomplete disclosure is truly necessary to accomplish the goals of the research, **(2)** there are no undisclosed risks to subjects that are more than minimal, and **(3)** there is an adequate plan for debriefing subjects, when appropriate, and for dissemination of research results to them. Information about risks should never be withheld for the purpose of eliciting the cooperation of subjects, and truthful answers should always be given to direct questions about the research. Care should be taken to distinguish cases in which disclosure would destroy or invalidate the research from cases in which disclosure would simply inconvenience the investigator.

Comprehension. The manner and context in which information is conveyed is as important as the information itself. For example, presenting information in a disorganized and rapid fashion, allowing too little time for consideration or curtailing opportunities for questioning, all may adversely affect a subject's ability to make an informed choice.

Because the subject's ability to understand is a function of intelligence, rationality, maturity and language, it is necessary to adapt the presentation of the information to the subject's capacities. Investigators are responsible for ascertaining that the subject has comprehended the information. While there is always an obligation to ascertain that the information about risk to

subjects is complete and adequately comprehended, when the risks are more serious, that obligation increases. On occasion, it may be suitable to give some oral or written tests of comprehension.

Special provision may need to be made when comprehension is severely limited -- for example, by conditions of immaturity or mental disability. Each class of subjects that one might consider as incompetent (e.g., infants and young children, mentally disable patients, the terminally ill and the comatose) should be considered on its own terms. Even for these persons, however, respect requires giving them the opportunity to choose to the extent they are able, whether or not to participate in research. The objections of these subjects to involvement should be honoured, unless the research entails providing them a therapy unavailable elsewhere. Respect for persons also requires seeking the permission of other parties in order to protect the subjects from harm. Such persons are thus respected both by acknowledging their own wishes and by the use of third parties to protect them from harm.

The third parties chosen should be those who are most likely to understand the incompetent subject’s situation and to act in that person's best interest. The person authorized to act on behalf of the subject should be given an opportunity to observe the research as it proceeds in order to be able to withdraw the subject from the research, if such action appears in the subject's best interest.

Voluntariness. An agreement to participate in research constitutes a valid consent only if voluntarily given. This element of informed consent requires conditions free of coercion and undue influence. Coercion occurs when an overt threat of harm is intentionally presented by one person to another in order to obtain compliance. Undue influence, by contrast, occurs through an offer of an excessive, unwarranted, inappropriate or improper reward or other overture in order to obtain compliance. Also, inducements that would ordinarily be acceptable may become undue influences if the subject is especially vulnerable.

Unjustifiable pressures usually occur when persons in positions of authority or commanding influence -- especially where possible sanctions are involved -- urge a course of action for a subject. A continuum of

such influencing factors exists, however, and it is impossible to state precisely where justifiable persuasion ends and undue influence begins. But undue influence would include actions such as manipulating a person's choice through the controlling influence of a close relative and threatening to withdraw health services to which an individual would otherwise be entitled.

2. Assessment of Risks and Benefits. — The assessment of risks and benefits requires a careful arrayal of relevant data, including, in some cases, alternative ways of obtaining the benefits sought in the research. Thus, the assessment presents both an opportunity and a responsibility to gather systematic and comprehensive information about proposed research. For the investigator, it is a means to examine whether the proposed research is properly designed. For a review committee, it is a method for determining whether the risks that will be presented to subjects are justified. For prospective subjects, the assessment will assist the determination whether or not to participate.

The Nature and Scope of Risks and Benefits. The requirement that researches be justified on the basis of a favourable risk/benefit assessment bears a close relation to the principle of beneficence, just as the moral requirement that informed consent be obtained is derived primarily from the principle of respect for persons. The term "risk" refers to a possibility that harm may occur. However, when expressions such as "small risk" or "high risk" are used, they usually refer (often ambiguously) both to the chance (probability) of experiencing a harm and the severity (magnitude) of the envisioned harm.

The term "benefit" is used in the research context to refer to something of positive value related to health or welfare. Unlike, "risk," "benefit" is not a term that expresses probabilities.

Risk is properly contrasted to probability of benefits, and benefits are properly contrasted with harms rather than risks of harm. Accordingly, so-called risk/benefit assessments are concerned with the probabilities and magnitudes of possible harm and anticipated benefits. Many kinds of possible harms and benefits need to be taken into account. There are, for example, risks of psychological harm, physical harm, legal harm, social harm and economic harm and the corresponding benefits. While the most likely types of harms to research subjects are those of psychological or physical pain or injury, other possible kinds should not be overlooked.

Risks and benefits of research may affect the individual subjects, the families of the individual subjects, and society at large (or special groups of subjects in society). Previous codes and Federal regulations have required that risks to subjects be outweighed by the sum of both the anticipated benefit to the subject, if any, and the anticipated benefit to society in the form of knowledge to be gained from the research. In balancing these different elements, the risks and benefits affecting the immediate research subject will normally carry special weight. On the other hand, interests other than those of the subject may on some occasions be sufficient by themselves to justify the risks involved in the research, so long as the subjects' rights have been protected. Beneficence thus requires that we protect against risk of harm to subjects and also that we be concerned about the loss of the substantial benefits that might be gained from research.

The Systematic Assessment of Risks and Benefits. It is commonly said that benefits and risks must be “balanced" and shown to be "in a favourable ratio." The metaphorical character of these terms draws attention to the difficulty of making precise judgments. Only on rare occasions will quantitative techniques be available for the scrutiny of research protocols. However, the idea of systematic, nonarbitrary analysis of risks and benefits should be emulated insofar as possible. This ideal requires those making decisions about the justifiability of

research to be thorough in the accumulation and assessment of information about all aspects of the research, and to consider alternatives systematically. This procedure renders the assessment of research more rigorous and precise, while making communication between review board members and investigators less subject to misinterpretation, misinformation and conflicting judgments. Thus, there should first be a determination of the validity of the presuppositions of the research; then the nature, probability and magnitude of risk should be distinguished with as much clarity as possible. The method of ascertaining risks should be explicit, especially where there is no alternative to the use of such vague categories as small or slight risk. It should also be determined whether an investigator's estimates of the probability of harm or benefits are reasonable, as judged by known facts or other available studies.

Finally, assessment of the justifiability of research should reflect at least the following considerations: **(i)**Brutal or inhumane treatment of human subjects is never morally justified. **(ii)** Risks should be reduced to those necessary to achieve the research objective. It should be determined whether it is in fact necessary to use human subjects at all. Risk can perhaps never be entirely eliminated, but it can often be reduced by careful attention to alternative procedures. **(iii)** When research involves significant risk of serious impairment, review committees should be extraordinarily insistent on the justification of the risk (looking usually to the likelihood of benefit to the subject -- or, in some rare cases, to the manifest voluntariness of the participation). **(iv)** When vulnerable populations are involved in research, the appropriateness of involving them should itself be demonstrated. A number of variables go into such judgments, including the nature and degree of risk, the condition of the particular population involved, and the nature and level of the anticipated benefits. **(v)** Relevant risks and benefits must be thoroughly arrayed in documents and procedures used in the informed consent process.

3. Selection of Subjects. — Just as the principle of respect for persons finds expression in the requirements for consent, and the principle of beneficence in risk/benefit assessment, the principle of justice gives rise to moral requirements that there be fair procedures and outcomes

in the selection of research subjects.

Justice is relevant to the selection of subjects of research at two levels: the social and the individual. Individual justice in the selection of subjects would require that researchers exhibit fairness: thus, they should not offer potentially beneficial research only to some patients who are in their favour or select only “undesirable" persons for risky research. Social justice requires that distinction be drawn between classes of subjects that ought, and ought not, to participate in any particular kind of research, based on the ability of members of that class to bear burdens and on the appropriateness of placing further burdens on already burdened persons. Thus, it can be considered a matter of social justice that there is an order of preference in the selection of classes of subjects (e.g., adults before children) and that some classes of potential subjects (e.g., the institutionalized mentally infirm or prisoners) may be involved as research subjects, if at all, only on certain conditions.

Injustice may appear in the selection of subjects, even if individual subjects are selected fairly by investigators and treated fairly in the course of research. Thus, injustice arises from social, racial, sexual and cultural biases institutionalized in society. Thus, even if individual researchers are treating their research subjects fairly, and even if IRBs are taking care to assure that subjects are selected fairly within a particular institution, unjust social patterns may nevertheless appear in the overall distribution of the burdens and benefits of research. Although individual institutions or investigators may not be able to resolve a problem that is pervasive in their social setting, they can consider distributive justice in selecting research subjects.

Some populations, especially institutionalized ones, are already burdened in many ways by their infirmities and environments. When research is proposed that involves risks and does not include a therapeutic component, other less burdened classes of persons should be called upon first to accept these risks of research, except where the research is directly related to the

specific conditions of the class involved. Also, even though public funds for research may often flow in the same directions as public funds for health care, it seems unfair that populations dependent on public health care constitute a pool of preferred research subjects if more advantaged populations are likely to be the recipients of the benefits.

One special instance of injustice results from the involvement of vulnerable subjects. Certain groups, such as racial minorities, the economically disadvantaged, the very sick, and the institutionalized may continually be sought as research subjects, owing to their ready availability in settings where research is conducted. Given their dependent status and their frequently compromised capacity for free consent, they should be protected against the danger of being involved in research solely for administrative convenience, or because they are easy to manipulate as a result of their illness or socioeconomic condition.

[1] Since 1945, various codes for the proper and responsible conduct of human experimentation in medical research have been adopted by different organizations. The best known of these codes are the Nuremberg Code of 1947, the Helsinki Declaration of 1964 (revised in 1975), and the 1971 Guidelines (codified into Federal Regulations in 1974) issued by the U.S. Department of Health, Education, and Welfare Codes for the conduct of social and behavioral research have also been adopted, the best known being that of the American Psychological Association, published in 1973.

[2] Although practice usually involves interventions designed solely to enhance the well-being of a particular individual, interventions are sometimes applied to one individual for the enhancement of the well-being of another (e.g., blood donation, skin grafts, organ transplants) or an intervention may have the dual purpose of enhancing the well-being of a particular individual, and, at the same time, providing some benefit to others (e.g., vaccination, which protects both the person who is vaccinated and society generally). The fact that some forms of practice have elements other than immediate benefit to the individual receiving an intervention, however, should not confuse the general distinction between research and practice. Even when a procedure applied in practice may benefit some other person, it

remains an intervention designed to enhance the well-being of a particular individual or groups of individuals; thus, it is practice and need not be reviewed as research.

[3] Because the problems related to social experimentation may differ substantially from those of biomedical and behavioral research, the Commission specifically declines to make any policy determination regarding such research at this time. Rather, the Commission believes that the problem ought to be addressed by one of its successor bodies.

Completion report

Collaborative Institutional Training Initiative (Citi Program) Completion Report- Part 1 Of 2

Coursework Requirements*

* NOTE: Scores on this Requirements Report reflect quiz completions at the time all requirements for the course were met. See list below for details.
See separate Transcript Report for more recent quiz scores, including those on optional (supplemental) course elements.

- **Name:** Teresa Nalikka (ID: 10663184)
- **Institution Affiliation:** Massachusetts Institute of Technology Affiliates (ID: 1912)
- **Institution Email:** gladysnalikka@gmail.com
- **Institution Unit:** Computer science
- **Curriculum Group:** Human Research
- **Course Learner Group:** Data or Specimens Only Research
- **Stage:** Stage 1 - Basic Course
- **Record ID:** 45860685
- **Completion Date:** 01-Nov-2021

- **Expiration Date:** 31-Oct-2024
- **Minimum Passing:** 90
- **Reported Score*:** 92

REQUIRED AND ELECTIVE MODULES ONLY	DATE COMPLETED	SCORE
Belmont Report and Its Principles (ID: 1127)	01-Nov-2021	3/3 (100%)
History and Ethics of Human Subjects Research (ID: 498)	01-Nov-2021	5/5 (100%)
Basic Institutional Review Board (IRB) Regulations and Review Process (ID: 2)	01-Nov-2021	5/5 (100%)
Records-Based Research (ID: 5)	01-Nov-2021	3/3 (100%)
Genetic Research in Human Populations (ID: 6)	01-Nov-2021	4/5 (80%)
Populations in Research Requiring Additional Considerations and/or Protections (ID: 16680)	01-Nov-2021	5/5 (100%)
Research and HIPAA Privacy Protections (ID: 14)	01-Nov-2021	4/5 (80%)
Conflicts of Interest in Human Subjects Research (ID: 17464)	01-Nov-2021	4/5 (80%)
Massachusetts Institute of Technology (ID: 1290)	01-Nov-2021	No Quiz

For this Report to be valid, the learner identified above must have had a valid affiliation with the CITI Program subscribing institution

identified above or have been a paid Independent Learner.

Verify at: www.citiprogram.org/verify/?k34972204-8978-4e0d-8d4b-ea36747a8502-45860685

Collaborative Institutional Training Initiative (CITI Program)

Email: support@citiprogram.org

Phone: 888-529-5929

Web: https://www.citiprogram.org

Collaborative Institutional Training Initiative (Citi Program) Completion Report - Part 2 Of 2

Coursework Transcript**

** NOTE: Scores on this Transcript Report reflect the most current quiz completions, including quizzes on optional (supplemental) elements of the course. See list below for details. See separate Requirements Report for the reported scores at the time all requirements for the course were met.

- **Name:** Teresa Nalikka (ID: 10663184)
- **Institution Affiliation:** Massachusetts Institute of Technology Affiliates (ID: 1912)
- **Institution Email:** gladysnalikka@gmail.com
- **Institution Unit:** Computer science
- **Curriculum Group:** Human Research
- **Course Learner Group:** Data or Specimens Only Research
- **Stage:** Stage 1 - Basic Course
- **Record ID:** 45860685
- **Report Date:** 01-Nov-2021
- **Current Score**:** 92

REQUIRED, ELECTIVE, AND SUPPLEMENTAL MODULES	MOST RECENT	SCORE
Basic Institutional Review Board (IRB) Regulations and Review Process (ID: 2)	01-Nov-2021	5/5 (100%)
Belmont Report and Its Principles (ID: 1127)	01-Nov-2021	3/3 (100%)
Records-Based Research (ID: 5)	01-Nov-2021	3/3 (100%)
Genetic Research in Human Populations (ID: 6)	01-Nov-2021	4/5 (80%)
Research and HIPAA Privacy Protections (ID: 14)	01-Nov-2021	4/5 (80%)
History and Ethics of Human Subjects Research (ID: 498)	01-Nov-	

2021 5/5 (100%) Populations in Research Requiring Additional Considerations and/or Protections (ID: 16680) 01-Nov-2021 5/5 (100%) Conflicts of Interest in Human Subjects Research (ID: 17464) 01-Nov-2021 4/5 (80%) Massachusetts Institute of Technology (ID: 1290) 01-Nov-2021 No Quiz

For this Report to be valid, the learner identified above must have had a valid affiliation wi th the CITI Program subscribing institution identified above or have been a paid Independent Learner.

Collaborative Institutional Training Initiative (CITI Program)

Email: support@citiprogram.org

Phone: 888-529-5929

Web: https://www.citiprogram.org

Queries

```
--tables into activities-----

--1-icustays----

CREATE TABLE mimic_insights.icu_icustays_activity_enter AS

 (SELECT

subject_id, stay_id, hadm_id, intime AS timestamp, 'Enter the ICU' AS activity

FROM mimiciv. icustays)

ORDER BY stay_id.

-----2---inputevents-----

CREATE TABLE mimic_insights.icu_icustays_activity_input AS

 (SELECT

 subject_id, stay_id, hadm_id, starttime AS timestamp, 'Input infusion' AS activity

FROM "mimiciv. inputevents)

ORDER BY stay_id.

3-----outputevents-------

CREATE TABLE mimic_insights.icu_icustays_activity_output AS

(SELECT

subject_id, stay_id, hadm_id, charttime AS timestamp, 'Output from patient' AS activity

FROM mimiciv. outputevents)

ORDER BY stay_id.

4-------procedureevents----

CREATE TABLE mimic_insights.icu_icustays_activity_procedures AS

 (SELECT

i.stay_id, i. subject_id, i. intime AS timestamp, 'Perform procedure' AS activity,

FROM mimiciv. procedureevents p
```

```
INNER JOIN mimiciv. icustays i)
    ON i. stay_id = p. stay_id;
```

4.

```
CREATE TABLE mimic_insights.icu_icutables AS (
SELECT DISTINCT
subject_id, hadm_id, stay_id, timestamp, activity
FROM
mimic_insights.icu_icustays_activity_enter
UNION
SELECT DISTINCT
subject_id, stay_id, hadm_id, timestamp, activity
FROM
mimic_insights.icu_icustays_activity_input
UNION
select DISTINCT
subject_id, stay_id, hadm_id, timestamp, activity
FROM
mimic_insights.icu_icustays_activity_output
UNION
SELECT DISTINCT
subject_id, stay_id, hadm_id, timestamp, activity
FROM
mimic_insights.icu_icustays_activity_procedures
)
```

```
ORDER BY stay_id, subject_id, hadm_id, timestamp.

 select * from mimic_insights.icu_icutables.

-----EXPORT IN CSV -----

COPY TO mimic_insights.icu_icutables TO

'C:\Users\glady\Desktop\thesis\3.0\RESULTS\QUERIES\icutables.csv' DELIMITER ',' CSV

HEADER;
```