Predicting Post-Procedural Complications on MIMC-III Dataset

****

Master of Science in Data Science

Faisal Maqbool

MSDS17027

Session: 2017 – 2019

DEPARTMENT OF COMPUTER SCIENCE

INFORMATION TECHNOLOGY UNIVERSITY

LAHORE, PAKISTAN

****

Predicting Post Procedural Complications on MIMIC-III Data

A thesis submitted in partial fulfillment of the requirements for the

Degree of Master of Science in

Data Science

Faisal Maqbool

Dr. Saeed Ul Hassan

Committee Member Name

Committee Member Name

**Declaration**

This thesis is a presentation of my original research work. Wherever contributions of others are involved, every effort is made to indicate this clearly, with due reference to the literature, and acknowledgement of collaborative research and discussions. I also declare that this work is the result of my own investigations, except where identified by references and free from plagiarism of the work of others.

Signature: ………….

Student Name

Date: ………………..

The undersigned hereby certify that they have read and recommend the thesis entitled “…………………………..” by ………………….. for the degree of Master of Science in Data Science.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Supervisor Name (ITU), Thesis Advisor

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Committee Member Name (ITU), Thesis Committee Member

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Committee Member Name, Thesis Committee Member

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Chairperson Name, Chairperson of the Department

**Acknowledgment**

First and foremost, I would like to express sincere gratitude to my advisor Dr. Saeed Ul Hassan for being the supportive advisor I could have asked for. His precious insights, guidance and support throughout the research journey has not only helped me become a better student but also examine data for research and contribute to previously done research. Though he is very busy, he always made me feel like a priority. He took time to review my findings, answer questions, and give thoughtful feedback for which I am extremely grateful. This thesis would not have been possible without them.

I would like to thank Dr. Mohsin Ali for his help, advice and support throughout my work and for agreeing to serve on my thesis committee.

Moreover, I would like to thank my family especially my parents and my friends for me letting my follow my dreams and for their encouragement which has been the vital part of the research.

**Table of Contents**

Table of Contents

Chapter: 1 11

[1.1 Introduction 11](#_Toc27025657)

[1.2 ICD, Hippa AND Complications 11](#_Toc27025658)

[1.3 Why COMPLICATIONS? 12](#_Toc27025659)

[1.4 APPROACH Followed 12](#_Toc27025660)

[1.5 Description of the Content 13](#_Toc27025661)

[Chapter: 2 MIMIC-III Critical Care Database 14](#_Toc27025662)

[2.1 About Mimic-III 14](#_Toc27025663)

[2.2 MIMIc-iii Tables 16](#_Toc27025664)

[2.3 MIMIC-III Construction 18](#_Toc27025665)

[2.4 MIMIC-III derived Concepts 18](#_Toc27025666)

[Chapter: 3 ETL and DataSet Building From DWH Mart 20](#_Toc27025667)

[3.1 Technical Process 20](#_Toc27025668)

[3.2 Relational Mapping and Batch Processing 21](#_Toc27025669)

**List of Tables**

[Table 1: Class Distribution of data for MIMIC-III Dataset 15](#_Toc26919824)

Table 2: MIMIC-III Tables Summary 16

[Table 3 : Derived Concepts 19](#_Toc26919826)

**List of Figures**

[Figure 1: Methodology 12](file:///D:\DataScience\Thesis-Bioinformatics-MIMICIII-\Initial%20Draft.docx#_Toc26919520)

[Figure 2: MIMIC-III Construction Model 18](file:///D:\DataScience\Thesis-Bioinformatics-MIMICIII-\Initial%20Draft.docx#_Toc26919521)

[Figure 3: Technical Chain of Steps 20](#_Toc26919522)

[Figure 4: Batch Processing for Huge Files Using Python 21](#_Toc26919523)

[Figure 5 : One Hot Encoding for Organ Failure 24](#_Toc26919524)

**List of Abbreviations**

MIMIC:

PHI:

ICU:

ETL:

ICD:

Hippa

BMI:

**Abstract**

The advancements in bioinformatics and health care sector has inspired researchers to develop systems that can mimic work of doctors because quantifying patient health and predicting outcomes is an important problem in critical care research. But, predicting outcomes for critically ill patients admitted in intensive care units requires specific characteristics of clinical data: quality, volume, access and dimensionality. In this Master Thesis, an analysis of the data from critical patients was carried out in order to study the influence of several factors for predicting the post-procedural complications and seeking if those complications can lead to mortality of patients. To derive insights for that, we used well-known clinical dataset named Medical Information Mart for Intensive Care III (MIMIC-III).

**.. Results Pending**

**Chapter: 1**

# Introduction

With the initiation of digital technology, advanced techniques are increasingly making it possible to utilize big data to more precisely risk evaluation and predict how an individual patient will behave based on a given diagnose or procedure. Intensive care unit **(ICU)** is a ward in hospital, where critically ill patients are admitted requires accurate predictors that can help doctors with the assessment of severity of illness.

In this thesis we investigate the different methodologies for extracting, transforming and loading **(ETL)** processes, which obtain data from original source to perform informative analysis and features extraction to aid model to predict post-procedural (diagnoses & procedure) complications of critically ill patients and investigating those complications if those can lead to mortality of patient or not. The methods use demographics, data from different hospital system, lab events, diagnoses, notes and other engineered information regarding each patient. The database used for the study is Medical Information Mart for Intensive Care **MIMIC-III [1]** which comes from health service with anonymized data for protecting health information **(PHI).**

Other researches about MIMIC-III data is also presented to motivate our problem, establish understanding of dataset, key findings and recommendations for future investigations. The question of predicting post-procedural complications from data science perspective and critical health perspective is not only important for doctors, administrators but also for the patient as well. For administrators this would help managing patients and required resources. Avoiding predicted complications can further be avoided if such information is known during the stay of patient at ICU. On MIMIC-III researchers widely contributed to support the cause. **[2]** Presented an evaluation of the influence of body mass index. They hypothesize that selected severity of illness scores would perform differently if body mass index categorization was incorporated and that the performance of these score models would improve after consideration of body mass index as an additional model feature. **[3]** Presented a study recently where they characterized the relationship of Body Mass Index **(BMI)** with survival and explored gender-based interactions with surrogates of body composition of nutrition in a real world setting. Another cohort study **[4]** examine the impact of overstays of patients and discharge delays on in-hospital morbidity and mortality.

Many other researcher have contributed in prediction of other critical factors related to medicine and the health of patient. **[5]** For different

# ICD, Hippa AND Complications

The International Classification of Diseases **(ICD) [10]** is the foundation for the identification of health trends and statistics globally, and the international standard for reporting diseases and health conditions. It is the diagnostic classification standard for all clinical and research purposes. Under revision of ICD9 codes, the code 996 defines complications particular to certain specified procedures and diagnoses. Most complications are caused due to cardiac, vascular or other used devices and some of them relates to reaction caused due to a procedure performed. In our work, we are focusing on such complications and investigating if those can lead to the mortality of patient. **HIPAA** (Health Insurance Portability and Accountability Act of 1996) is United States legislation that provides data privacy and security provisions for safeguarding medical information. To protect health information MIMIC-III provided anonymized data, still we need to make sure we are following HIPPA compliance rules so that our research does not conflict with any of the standards defined.

# Why COMPLICATIONS?

Extended diagnoses and patients stay at the hospital is associated with not only the health of patient, cost, increased number of deaths but also increased number of readmissions. Each of these parameters defines the hospital performance. So, our focus is to produce insights that can complement these parameters ranging from cost to patient health.

# APPROACH Followed

As open sourced MIMIC-III data is available for research purposes for free. Our main focus was to use tools and technologies in a way that can complement already done research. Hence, we used **Python**, **R** and **Postgres SQL** for ETL process, features selection and modeling our features to extract results which I will explain in respective chapters in details.

We have used the basic Data Science (mining) **[3]** model as full cycle of development in our research.



Figure 1: Methodology

The model divides several phases which are executed in iteration to get to goal which we can see in **Figure 1.1.** The ﬁrst Business Understanding phase focuses on understanding the objectives of the project. In the Data Understanding phase, the tasks related to the data collection, the exploratory analysis and the quality review of the data are included. During the Data Preparation phase, the data that will be used for the successive phases are selected, the data is cleaned if necessary, derived data are obtained from the sources and in the modeling phase we design and implement our models with appropriate parameters to extract results. In the end final deployment stage is carried as per requirement.

# Description of the Content

The thesis is organized as follows: In chapter **1** we introduce the problem, explains purpose of our work, methodologies followed and detail of contents. In Chapter **2**, we define the data source MIMIC-III in detail. In Chapter **3**, we explained the ETL process to obtain data and process it effectively for posterior use. In Chapter **4**, we explain the preprocessing and analysis of data which is used of feature engineering and model selection process. In Chapter **5**, results, conclusion and future work is been discussed.

# Chapter: 2 MIMIC-III Critical Care Database

This chapter explains the structure, context and development researchers have done on data source MIMIC-III.

# 2.1 About Mimic-III

Over the past decade, much have been written about the field of data science regarding the explosion of big data. In health care, every decision made for a critical patient requires precision by clinicians and doctors. To carry out research to aid clinicians and doctors to make better, reliable and quick decision using research and applications. This demands privacy existence of data and wide-ranging analysis.

To avoid these difﬁculties, it is available to the researcher and educational community the **MIMIC-III (Medical Information Mart for Intensive Care III)** database, openly accessible at <https://mimic.physionet.org/> This database is an evolution of the MIMIC-II database created by the Laboratory of Computational Physiology of The Massachusetts Institute for Technology (MIT) with the goal of providing tools for the creation of clinical knowledge through the application of data analysis techniques. MIMIC-III is a large, freely-available relational database comprising de-identiﬁed health related data associated with over forty thousand patients who stayed in Intensive Care Units at Beth Israel Deaconess Medical Center (Boston, Massachusetts). The data spans June 2001 October2012.

MIMIC-III is a comprehensive collection of de-identified data from **53,423** distinct critical care hospital admissions from **38,597** distinct adult patients. The data has been compiled into 26 tables which contain, for example, an average of **4579** charted observations and 380 laboratory measurements for each hospital admission as well as a total of **3.8** gigabytes of unstructured textual data from various healthcare provider notes and analyses. In addition to de-identifying patient data, MIT requires training in the protection of patient data for anyone requesting access to the MIMIC dataset. After completing the prescribed training, data can be downloaded as 26 comma separated values (csv) files representing the **26** tables in the MIMIC-III database. Sample SQL code can be acquired from GitHub (<https://github.com/MIT-LCP/mimic-code>) for establishing relationships between the tables. Additionally, there is a published data dictionary which can be found at <https://mimic.physionet.org/mimictables/admissions/>. There is variability in the usage of “unique” attributes and definition of primary keys between the sample SQL code and the published data dictionary. For example, every table has an attribute called “ROW\_ID”, and the sample SQL code consistently declares this attribute as “unique” and/or as a “primary key” for every table despite the fact that tables like the “PATIENTS” table have a unique identifier (SUBJECT\_ID) that is intended to be the primary key and serve as foreign key in child relations that refer to the “PATIENTS” table.

After downloading and analyzing the MIMIC source tables, implementation occurs in 5 additional steps:

* Create tables with attribute rules (data types) and identify the primary key for each table. 1. Load records from csv files into each table.
* Declare the indexes for each table.
* Define foreign keys in each table and establish table relationships.
* Implement user interface (with appropriately granted permissions) for the database.

This database includes information on demographic data of patients, laboratory test results, vital sign measurements, procedures, medications, caregiver notes, imaging reports, mortality (both in and out of the hospital), manual evolution annotations regarding events, discharge reports, prescription, and so on. The class distribution of data shown in table **1** (adopted from <https://www.nature.com/articles/sdata201635/tables/3>)

Table : Class Distribution of data for MIMIC-III Dataset

| **Class of data** | **Description** |
| --- | --- |
| Billing | Coded data recorded primarily for billing and administrative purposes. Includes Current Procedural Terminology (CPT) codes, Diagnosis-Related Group (DRG) codes, and International Classification of Diseases (ICD) codes. |
| Descriptive | Demographic detail, admission and discharge times, and dates of death. |
| Dictionary | Look-up tables for cross referencing concept identifiers (for example, International Classification of Diseases (ICD) codes) with associated labels. |
| Interventions | Procedures such as dialysis, imaging studies, and placement of lines. |
| Laboratory | Blood chemistry, hematology, urine analysis, and microbiology test results. |
| Medications | Administration records of intravenous medications and medication orders. |
| Notes | Free text notes such as provider progress notes and hospital discharge summaries. |
| Physiologic | Nurse-verified vital signs, approximately hourly (e.g., heart rate, blood pressure, respiratory rate). |
| Reports | Free text reports of electrocardiogram and imaging studies. |

# 2.2 MIMIc-iii Tables

MIMIC-III is structured in a relational manner containing 26 files from which we created following tables in PostgreSQL.

Table : MIMIC-III Tables Summary (Adopted from: <https://mimic.physionet.org/gettingstarted/access/>)

|  |  |  |
| --- | --- | --- |
| File Name | Dimension | Summary |
| ADMISSIONS | (58976, 19) | The ADMISSIONS table gives information regarding a patient’s admission to the hospital. |
| CALLOUT | (34499, 24) | The CALLOUT table provides information about ICU discharge planning |
| CAREGIVERS | (7567, 4) | This table provides information regarding care givers. For example, it would deﬁne if a caregiver is a research nurse (RN), medical doctor (MD), and so on. |
| CHARTEVENTS | (330712483, 15) | CHARTEVENTS contains all the charted data available for a patient. |
| CPTEVENTS | (573146, 12) | The CPTEVENTS table contains a list of which current procedural terminology codes were billed for which patients. This can be useful for determining if certain procedures have been performed (e.g. ventilation). |
| D\_CPT | (134, 9) | This table gives some high level information regarding current procedural terminology (CPT) codes. Unfortunately, detailed information for individual codes is unavailable. |
| D\_ICD\_DIAGNOSES | (14567, 4) | This table deﬁnes International Classiﬁcation of Diseases Version 9 (ICD-9) codes for diagnoses. These codes are assigned at the end of the patient’s stay and are used by the hospital to bill for care provided. |
| D\_ICD\_PROCEDURES | (3882, 4) | This table deﬁnes International Classiﬁcation of Diseases Version 9 (ICD-9) codes for procedures. These codes are assigned at the end of the patient’s stay and are used by the hospital to bill for care provided. |
| D\_ITEMS | (12487, 10) | The D\_ITEMS table deﬁnes ITEMID, which represents measurements in the database. |
| D\_LABITEMS | (753, 6) | D\_LABITEMS contains deﬁnitions for all ITEMID associated with lab measurements in the MIMIC database. |
| DATETIMEEVENTS | (4485937, 14) | DATETIMEEVENTS contains all date measurements about a patient in the ICU. |
| DIAGNOSES\_ICD | (651047, 5) | This table deﬁnes ICD-9 codes for diagnoses. The ICD codes are generated for billing purposes at the end of the hospital stay. |
| DRGCODES | (125557, 8) | This table deﬁnes HCFA-DRG and APR-DRG codes which provide information regarding Diagnosis-Related Group recorded primarily for billing and administrative purposes. |
| ICUSTAYS | (61532, 12) | This table gives information regarding ICU hospital stays. |
| INPUTEVENTS\_CV | (17527935, 22) | This table contains data of ﬂuid input events (serums, intravenous medication, insulin, etc.) regarding Carevue database source associated to ICU episodes. |
| INPUTEVENTS\_MV | (3618991, 31) | This table contains input data for patients. |
| LABEVENTS | (27854055, 9) | Contains all laboratory measurements for a given patient, including outpatient data. |
| MICROBIOLOGYEVENTS | (631726, 16) | Contains microbiology information, including tests performed and sensitivities. |
| NOTEEVENTS | (2083180, 9) | This table contains all notes for patients took in a manual way by their caregivers. |
| OUTPUTEVENTS | (4349218, 13) | This table contains output data for patients. |
| PATIENTS | (46520, 8) | This table contains hospitalization-independent data for all patients such as, gender, date of birth, etc. |
| PRESCRIPTIONS | (4156450, 19) | This table contains medication related order entries, i.e. prescriptions. |
| PROCEDUREEVENTS\_MV | (258066, 25) | This table contains procedures for patients |
| PROCEDURES\_ICD | (17527935, 22) | Contains ICD procedures for patients, most notably ICD-9 procedures. The ICD codes are generated for billing purposes at the end of the hospital stay and are recorded for all patient hospitalizations. |
| SERVICES | (73343, 6) | The SERVICES table describes the service that a patient was admitted under. This service admission can be elective or caused due to a number of reasons, including bed shortage. |
| TRANSFERS | (261897, 13) | This table contains physical locations for patients throughout their hospital stay. |

# 2.3 MIMIC-III Construction

MIMIC-III was constructed based upon hospital level, patient level, ICU level & used systems level. Furthermore it includes billing, notes and reports as shown in below figure.



Figure 2: MIMIC-III Construction Model

# 2.4 MIMIC-III derived Concepts

The active researchers have contributed to already given data with additional scripts to generate new concepts and insights at MIMIC code repository **[4]** which includes views and tables as well. They also encourage other researchers to contribute to derived insights which helps to distinct between the original data and derived data and one can use as per the problem they are solving and contribute as well.

Following are the major concepts that are being used frequently by researchers.

Table : Derived Concepts

|  |  |
| --- | --- |
| Class of Data | Summary |
| Comorbidity | These scripts derive binary ﬂags indicating the presence of various comorbidities using billing codes (ICD-9) assigned to the patient at hospital discharge. |
| First day | The ﬁrst day subfolder contains scripts used to calculate various clinical concepts on the ﬁrst day of a patient’s admission to the ICU, such as the highest blood pressure, lowest temperature, etc. This folder contains many useful scripts which can be adapted to capture data outside the ﬁrst day. |
| Sepsis | Deﬁnitions of sepsis, a common cause of mortality for intensive care unit patients. |
| Severity Scores | Severity of illness scores which summarize the acuity of a patient’s illness on admission to the intensive care unit (usually in the ﬁrst 24 hours). |
| Durations | Start and stop times for administration of various treatments or durations of various phenomena, including: medical agents which have a vasoactive effect on a patient’s circulatory system, continuous renal replacement therapy (CRRT), and mechanical ventilation. |
| Organ Failure | This script derives binary flags for major organ failures |

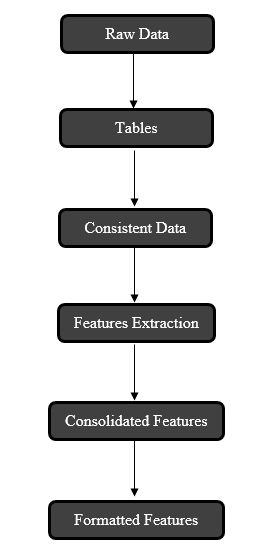
[[1]](#footnote-1)The tables are linked by identifiers which usually have the suffix “**ID**”. For example **HADM\_ID** refers to a unique hospital admission and **SUBJECT\_ID** refers to a unique patient. One exception is **ROW\_ID**, which is simply a row identifier unique to that table.

Tables pre-fixed with “**D\_**” are dictionaries and provide definitions for identifiers. For example, every row of **OUTPUTEVENTS** is associated with a single **ITEMID** which represents the concept measured, but it does not contain the actual name of the drug. By joining **OUTPUTEVENTS** and **D\_ITEMS** on **ITEMID**, it is possible to identify what concept a given **ITEMID** represents

# Chapter: 3 ETL and DataSet Building From DWH Mart

In this section, we would introduce the ETL followed by us to derive certain insights which will lead us to conclusion of stated problem. The section is divided into extraction, transformation and loading sections to reach to our features.

# 3.1 Technical Process



* Tables Creation
* Relationship Mapping (Indexes and Keys)
* Materialized views from already given tables
* Trim down values for ICD-9 Codes
* Filter rows with subject id lookup and pass it to items lookup for certain diagnoses and procedure
* ICD-9 Codes for class complications which is 996
* Making it to binary classes with 1 and all other classes to 0
* Extracting derived features from chart events and lab events with batch processing
* Consolidate all other features with derived concept
* Format all features, fill out invalid fields and normalize features for model training

Figure : Technical Chain of Steps

Given the size of data mart and the volume of raw data, we devoted most of time to extraction and transformation of data.

In the first step prior to requesting access to MIMIC, you will need to complete the CITI “Data or Specimens Only Research” course by registering yourself on CITI program. After getting data access we are provided links to the 26 comma separated file containing patient, hospital and ICU related data[[2]](#footnote-2).

# 3.2 Relational Mapping and Batch Processing

All these are relationally mapped with each other[[3]](#footnote-3). After getting these files we created a database of all those file and created respective tables. To improve the performance indexes and constraints were added. There were several of them that are huge and others are medium to tiny. The small files were dealt with PSQL but on the other hand, the big files caused a problem for not only creating table but also of processing those files in RAM. To handle such problems with huge files, we implemented Python script for asyn batch processing using Pandas[[4]](#footnote-4) which is an open sourced library to manipulate structured data and very highly efficient because of its reliable data frame objects along with transformation tools available with it.

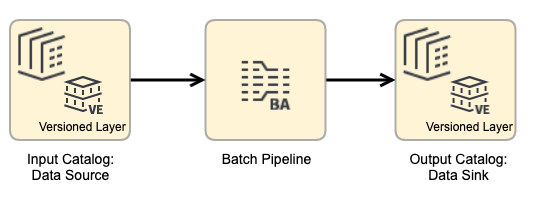


Figure : Batch Processing for Huge Files Using Python

To complete ETL process, PostgreSQL and Python played important role. Multiple SQL scripts were written for creation of tables, indexes, materialized views and derived tables. All of which are presented on a public repository.

<https://github.com/faisalmaqbool94/Thesis-Bioinformatics-MIMICIII->

Extraction of major chart events and lab events against each patient involved filtering of specific patients, lookup against particular diagnoses and procedures.



Figure : One Hot Encoding for Organ Failure

References

[1] Johnson, Alistair EW, Tom J. Pollard, Lu Shen, H. Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. "MIMIC-III, a freely accessible critical care database." *Scientific data* 3 (2016): 160035.

**[2]** Deliberato, Rodrigo Octavio, Ary Serpa Neto, Matthieu Komorowski, David J. Stone, Stephanie Q. Ko, Lucas Bulgarelli, Carolina Rodrigues Ponzoni, Renato Carneiro de Freitas Chaves, Leo Anthony Celi, and Alistair EW Johnson. "An evaluation of the influence of body mass index on severity scoring." *Critical care medicine* 47, no. 2 (2019): 247-253.

**[3]** Naik, Girish S., Sushrut S. Waikar, Alistair EW Johnson, Elizabeth I. Buchbinder, Rizwan Haq, F. Stephen Hodi, Jonathan D. Schoenfeld, and Patrick A. Ott. "Complex inter-relationship of body mass index, gender and serum creatinine on survival: exploring the obesity paradox in melanoma patients treated with checkpoint inhibition." *Journal for immunotherapy of cancer* 7, no. 1 (2019): 89.

**[4]** Bose, Somnath, Alistair EW Johnson, Ari Moskowitz, Leo Anthony Celi, and Jesse D. Raffa. "Impact of intensive care unit discharge delays on patient outcomes: a retrospective cohort study." *Journal of intensive care medicine* 34, no. 11-12 (2019): 924-929.

2) ICD

3) CrispDM

4) derived concepts

**THESIS GUIDELINES**

**Thesis Cover:**

For thesis, the following color scheme will be used:

* Bachelor Thesis: Blue
* Master Thesis: Dark Maroon
* Ph.D. Thesis: Black

**Format for Spine:**

* For spine of the thesis, the student will put ***Degree title i.e. MSCS***, ***Thesis title*** and mention ***Session***.

**Font Size and Style:**

* The ***Headings*** in the thesis will be in ***Times New Roman*** with font size ***16.***
* The **Body** of the thesis will be in ***Times New Roman*** with font size ***12.***
* The line spacing will be *1.5.*

**Referencing:**

* The student should follow ***IEEE*** referencing style.

**Appendices:**

* Provide appendices at the end of the thesis labeling as appendix A, appendix B and so on.

1. <https://mimic.physionet.org/gettingstarted/overview/> [↑](#footnote-ref-1)
2. <https://mimic.physionet.org/gettingstarted/access/> [↑](#footnote-ref-2)
3. <https://mit-lcp.github.io/mimic-schema-spy/> [↑](#footnote-ref-3)
4. <https://pandas.pydata.org/> [↑](#footnote-ref-4)