Archives of Proteomics and Bioinformatics

Commentary

Assessing Different Diagnoses in MIMIC-IV v2.2 and MIMIC-IV-ED Datasets

Muhammad Adib Uz Zaman[1,*]

[1]School of IT, University of Cincinnati, Cincinnati, Ohio, USA

[*]Correspondence should be addressed to Muhammad Adib Uz Zaman, a_u_z_ipe@yahoo.com

Received date: July 10, 2023, **Accepted date:** January 04, 2024

Citation: Zaman MAU. Assessing Different Diagnoses in MIMIC-IV v2.2 and MIMIC-IV-ED Datasets. Arch Proteom and Bioinform. 2024;4(1):1-5.

Abstract

This study aims to reveal some important insights into the different diagnoses that are listed in Medical Information Mart for Intensive Care (MIMIC) dataset. This dataset includes patients from diverse backgrounds, ethnicity, demographics, etc. The diagnosis records are stored electronically using ICD-09 and ICD-10 codes. It is found that most of the patients were diagnosed at least once for essential hypertension and other related diseases.

Keywords: MIMIC, ICU, Hypertension, Emergency Department, Data mining

Introduction

Since critical patients require constant monitoring, the intensive care unit (ICU) is a data-rich setting. Researchers are drawn to the ICU environment because of the often-acute nature of ICU patient sickness and the requirement of early intervention. There are a lot of publicly available critical care datasets that have enabled research in this domain, which is unique.

These initiatives are mostly based on MIMIC [1], and a waveform database including demographic information digitally transcribed from paper records for over 90 patients. MIMIC- IV v2.2 [2] and MIMIC-IV ED [3] have been investigated in this study.

MIMIC-IV originates from two in-hospital database systems: a customized hospital-wide Electronic Health Record (EHR) and a Clinical Information System specific to Intensive Care Units (ICUs). The development of MIMIC-IV involved a three-step process:

Acquisition: Data extraction was performed for patients admitted to the Beth Israel Deaconess Medical Center (BIDMC) emergency department or any of the intensive care units from the respective hospital databases. A comprehensive master patient list was established, encompassing all medical record numbers corresponding to ICU or emergency department admissions between 2008 and 2019. Source tables were then filtered to include only records related to patients in the master patient list [2].

Preparation: The data underwent reorganization to enhance retrospective data analysis. This involved denormalizing tables, removing audit traces, and restructuring into a more condensed set of tables. The primary objective of this procedure was to simplify retrospective analysis of the database. Notably, data cleaning procedures were intentionally omitted to maintain the representation of a real-world clinical dataset [2].

Deidentification: Patient identifiers, as mandated by the Health Insurance Portability and Accountability Act (HIPAA), were eliminated. Random ciphers were used to replace patient identifiers, resulting in deidentified integer identifiers for patients, hospitalizations, and ICU stays. Structured data underwent filtering using lookup tables and allow lists. If required, a free-text deidentification algorithm was applied to remove Personal Health Information (PHI) from free-text. Additionally, date and times were arbitrarily shifted into the future with an offset measured in days. Each patient ID (e.g., subject ID) was assigned a unique date shift, ensuring internal consistency for a single patient's data. For instance, if the time between two measures in the original data was 4 hours, the calculated time difference in MIMIC-IV would also be 4 hours. However, patients were not temporally comparable, meaning two patients hospitalized in 2130 were not necessarily admitted in the same year. Following these three stages, the database was exported to a character-based comma-delimited format.

MIMIC-IV-ED is a large, publicly accessible database of Beth Israel Deaconess Medical Center emergency department admissions from 2011 to 2019. In the database, there are 422,500 ED stays. Vital signs, triage information, medication reconciliation, medication delivery, and discharge diagnoses are accessible [3]. To comply with the Safe Harbor requirement of the Health Insurance Portability and Accountability Act (HIPAA), all data are deidentified. MIMIC-IV-ED is intended to facilitate a vast array of educational and research endeavors. Patients are evaluated and prioritized for further care in a congested emergency department (ED). The severity of the conditions of ED patients extends from minor cuts to potentially fatal heart conditions. The emergency department (ED) is, at its essence, a setting with limited resources in which the most valuable resource, human attention, is rationed to achieve positive patient outcomes. Recent advancements in algorithmic methods present a thrilling opportunity to improve emergency department care. Large datasets are required for data-driven studies, and open data access facilitates study replication. MIMIC-IV-ED, a large database of admissions to an ED at a Boston, Massachusetts academic medical center, is intended to facilitate data analysis in emergency care [3].

Background Studies

ICU stands for intensive care unit where many hypertensive patients are admitted. The combination of heart failure (HF) and hypertension is a leading cause of hospital mortality, particularly among intensive care unit (ICU) patients. However, under intensive work pressure, the large number of clinical signals generated in the ICU can easily overwhelm the medical staff, leading to treatment delays, suboptimal care, or even incorrect clinical decisions. Individual risk stratification is crucial for the management of ICU patients with HF and hypertension. Artificial intelligence, particularly machine learning (ML), can generate superior prognostic models for these patients [4].

Data Insights

Table 1 shows the number of patients who have been diagnosed at least once with specific ICD codes. It shows the first-time diagnosis that the patients received while they were admitted. Quite obviously, many of the patients have been diagnosed with different diseases throughout their follow-up period. But here, only the first-time diagnosis for patients is reported. Around 50,000 patients were diagnosed with unspecified essential hypertension at least once during their admission. **Figure 1** shows an infographic of the same.

Table 2 provides insights on the emergency department database (MIMIC-IV-ED). However, the diagnoses do not follow the diagnosis order of **Table 1** except for a few. Hypertension takes the first place in both tables since there appears to be a highly disproportionate number of patients being diagnosed with it.

Conclusion and Future Directions

This study reveals some insights into an ICU database that is widely used for research. Based on the findings, a disproportionate number of patients are associated with essential hypertension than other related diseases. There are many scopes to investigate hypertensive patients further like tracking the readmission rate, predicting mortality, vitals, etc.

MIMIC IV also contains many clinical notes where large language models can be implemented. Recent developments in scaling large language models (LLMs) have resulted in significant enhancements to several benchmarks for natural language processing [5]. These language models have been partially trained on clinical text. These studies demonstrate that training a language model with clinical notes using masked language modeling (MLM) is an effective method for improving performance on downstream tasks. All these previous works employ architectures with only decoders.

Table 1. Total patients for top 20 diagnoses.

ICD code	Long title	Total patients
4019	Unspecified essential hypertension	49741
2724	Other and unspecified hyperlipidemia	33448
I10	Essential (primary) hypertension	31521
E785	Hyperlipidemia, unspecified	27903
53081	Esophageal reflux	23955
Z87891	Personal history of nicotine dependence	21356

25000	Diabetes mellitus without mention of complication, type II or unspecified type, not stated as uncontrolled	19401
311	Depressive disorder, not elsewhere classified	19216
K219	Gastro-esophageal reflux disease without esophagitis	19067
41401	Coronary atherosclerosis of native coronary artery	17863
V1582	Personal history of tobacco use	17849
5849	Acute kidney failure, unspecified	17295
2859	Anemia, unspecified	16592
F329	Major depressive disorder, single episode, unspecified	16476
42731	Atrial fibrillation	16454
4280	Congestive heart failure, unspecified	14432
3051	Unspecified essential hypertension	14343
F419	Other and unspecified hyperlipidemia	14223
30000	Essential (primary) hypertension	13705

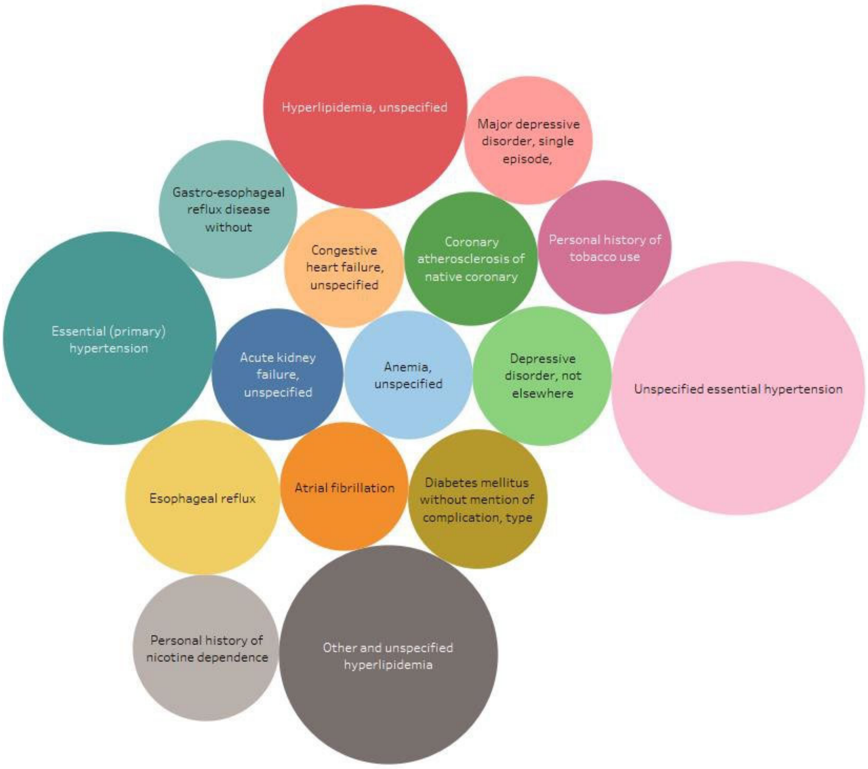

Figure 1. Infographic of the top 20 diagnoses across patients.

Table 2. Total patients for top 20 diagnosis (Emergency Department).

ICD code	Long title	Total patients
4019	Unspecified essential hypertension	18493
I10	Essential (primary) hypertension	15410
R079	Chest pain, unspecified	10499

78650	Chest pain, unspecified	9515
R109	Unspecified abdominal pain	8307
25000	Diabetes mellitus without mention of complication, type II or unspecified type, not stated as uncontrolled	7591
78909	Abdominal pain, other specified site	7174
W1830XA	Fall on same level, unspecified, initial encounter	7141
E8889	Unspecified fall	6431
E119	Type 2 diabetes mellitus without complications	6048
R51	Headache	5908
R0600	Dyspnea, unspecified	5812
2720	Pure hypercholesterolemia	5721
5990	Urinary tract infection, site not specified	5370
7840	Headache	5191
R531	Weakness	5164
R55	Syncope and collapse	5149
R42	Dizziness and giddiness	4959
R4182	Altered mental status, unspecified	4773
N390	Urinary tract infection, site not specified	4583

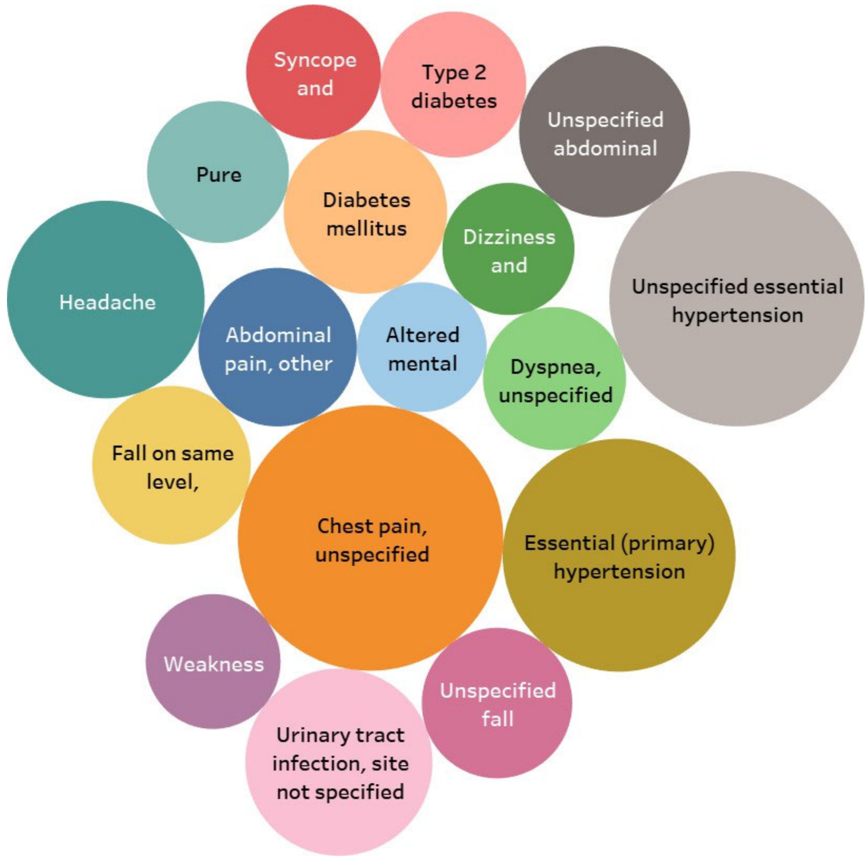

Figure 2. Infographic of the top 20 diagnoses across ED patients.

References

1. Zaman MA, Du D. A stochastic multivariate irregularly sampled time series imputation method for electronic health records. BioMedInformatics. 2021 Nov 16;1(3):166-81.

2. Johnson AE, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. Scientific Data. 2023 Jan 3;10(1):1.

3. Johnson A, Bulgarelli L, Pollard T, Celi LA, Mark R, Horng S. MIMIC-IV-ED (version 1.0). PhysioNet.

4. Peng S, Huang J, Liu X, Deng J, Sun C, Tang J, et al. Interpretable machine learning for 28-day all-cause in-hospital mortality prediction in critically ill patients with heart failure combined with hypertension: A retrospective cohort study based on medical information mart for intensive care database-IV and eICU databases. Frontiers in Cardiovascular Medicine. 2022 Oct 12; 9:994359.

5. Lehman E, Johnson A. Clinical-t5: Large language models built using mimic clinical text. PhysioNet.