Appendix for Whose Health Matters in Healthcare Models? Unmasking Data Bias for Data-Aware Modeling

Anonymous Author(s)
Affiliation
Address
email

A Datasheet for MIMIC Database

In the advancement of healthcare machine learning (HML) models, the quest for unbiased predictive outcomes is not just technical—it's fundamentally ethicalRajkomar et al. (2018). Despite substantial strides facilitated by HML, the specter of unfairness looms Buolamwini and Gebru (2018), Corbett-Davies and Goel (2018), propelled by the dual engines of data and algorithmic biasesMehrabi et al. (2021).The Impossibility Theorem in Fairness asserts that models with perfect performance and optimal fairness are simultaneously unattainable except special cases. Nevertheless, we can endeavor to find an optimal balance—a perfect trade-off tailored to specific domains. In the realm of healthcare machine learning (HML), this balance is not just about mitigating algorithmic bias; it's also about understanding and rectifying data biases. Researchers often face the colossal task of sifting through extensive data documentation to unearth task-specific anomalies. To streamline this process, we present the Datasheet for MIMIC IV v2.0. This resource empowers researchers to discern and address data inconsistencies, guides the selection of sensitive attributes for fairness assessments, and facilitates the creation of robust, just, and data-conscious HML models. More than a mere inventory, the datasheet provides comprehensive insights into the database's structure, data collection methodologies, management practices, and potential biases, ensuring researchers are well-informed and vigilant.

The MIMIC IV v2.0 datasheet was created based on the template provided by Gebru et al. (2021), with necessary modifications to accommodate the complexities of the database structure in clinical research databases (CRDs). This adaptation was crucial as the original template in Gebru et al. (2021) was designed for datasets rather than databases. To ensure comprehensive coverage of CRD-related information, several changes were made. For example, the question *"Can/How the dataset be/are created from the MIMIC database?"* was included in the datasheet (indicated by '+'), and the question *"What is the composition of the dataset?"* was replaced with *"What is the composition of the database?"*. The complete MIMIC IV v2.0 datasheet is provided below.

A.1 Motivation

For what purpose was the database created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The creation of the MIMIC-IV CRD aimed to improve patient care through knowledge discovery and algorithm development using a historically collected medical dataset. It was developed with an approach that allows permissive access, enabling extensive utilization of the MIMIC-IV database. Consequently, the database has been widely utilized in various healthcare applications, including assessing treatment effectiveness in specific patient groups and predicting critical outcomes such as mortality, readmission and length of stay Johnson et al. (2023).

Who created the database (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The MIMIC-IV database Johnson et al. (2023), developed by Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark from the Massachusetts Institute of Technology at the MIT Laboratory for Computational Physiology, is a collaborative effort involving various research groups.

Who funded the creation of the database?

The work was supported by grants from the National Institute of Biomedical Imaging and Bioengineering (NIBIB) of the National Institutes of Health (NIH) under award numbers R01-EB001659 (2003-2013) and R01-EB017205 (2014-2018)[1].

Any other comments?

MIMIC is a large and freely available database that contains deidentified health-related data from patients admitted to the critical care units of the Beth Israel Deaconess Medical Center. There are multiple versions of MIMIC that have been released:

1. **MIMIC-IV** encompasses data collected from 2008 to 2019, obtained from Metavision bedside monitors Johnson et al. (2023).
2. **MIMIC-III** comprises data collected from 2001 to 2012, obtained from both Metavision and CareVue bedside monitors Johnson et al. (2023).
3. **MIMIC-II** includes data collected from 2001 to 2008, obtained exclusively from CareVue bedside monitors. While MIMIC-II is no longer publicly available, its data can still be obtained from MIMIC-III by selectively including the data from the CareVue monitors Johnson et al. (2023).

Throughout the surveyed timeline, MIMIC III and MIMIC IV have been extensively utilized for healthcare machine learning (HML) prediction models. The datasheet provided is specifically for MIMIC IV v2.0, the latest available version of the database.

License - The licensing for the MIMIC files can be found in the *PhysioNet Credentialed Health Data License 1.5.0 (MIT-LCP)*[1].

A.2 Composition

What is the composition of the database?

MIMIC IV database is grouped into four modules: MIMIC IV (hosp, and icu) , MIMIC IV-ED (ed), MIMIC IV-Note (note) and MIMIC-CXR (cxr)[1]

1. **MIMIC IV**Johnson et al. (2023)
   (a) The ***Hosp*** module grants access to diverse data extracted from the hospital's electronic health record system, including patient and admission details, laboratory measurements, microbiology information, medication administration records, and billed diagnoses. These data are organized in the form of tables, including patient and admission-related tables (patients, admissions, transfers), laboratory measurement tables (labevents, d_labitems), microbiology culture table (microbiologyevents), provider order tables (poe, poe_detail), medication administration tables (emar, emar_detail), medication prescription tables (prescriptions, pharmacy), and hospital billing information tables (diagnoses_icd, d_icd_diagnoses, procedures_icd, d_icd_procedures, services).
   (b) The ***ICU*** module contains information collected from the clinical information system (BIDMC: MetaVision (iMDSoft)) used within the ICU. Documented data includes intravenous administrations, ventilator settings, and other charted items. Data documented in the icu module includes intravenous and fluid inputs (inputevents), ingredients of the aforementioned inputs (ingredientevents), patient outputs (outputevents), procedures (procedureevents), information documented as a date or time (datetimeevents), and other charted information (chartevents).

[1] `https://mimic.mit.edu/`

2. **MIMIC IV-ED**Johnson et al. (2023) - The *ED* module of MIMIC IV-ED focuses on emergency department patients and encompasses information regarding reasons for admission, triage assessments, vital signs, and medication reconciliation. The subject_id and hadm_id identifiers within MIMIC-IV-ED allow for linkage with other MIMIC-IV modules.

3. **MIMIC IV-Note** Johnson et al. (2023) - The *Note* module contains deidentified free-text clinical notes for hospitalized patients..

4. **MIMIC IV-CXR** Johnson et al. (2023) - The *CXR* module of MIMIC IV-CXR provides lookup tables that establish connections between patient identifiers and MIMIC-CXR study_id and dicom_id, facilitating the analysis of patient chest x-rays in conjunction with clinical data from other MIMIC-IV modules.

+ How is the data arranged within each modules and for what purpose?

The data within each module is structured in the form of tables, as MIMIC is a well-organized relational database. Each table within a module represents a specific type of data. Within each table, the data is organized into rows and columns. Each row corresponds to a particular patient or event, while each column represents a specific variable or attribute corresponding to that row. This organized structure allows researchers to efficiently extract customized datasets tailored to their research inquiries and facilitates the construction of machine learning models.

+ Can the modules be linked together to create specific dataset for specific tasks?

Yes. The tables within a module can be connected to other tables within the same module or across different modules using unique identifiers

+ Explain in detail the tables presented in each module?

MIMIC IV

***Hosp* Module**

Table 1: Description of *Hosp* module Tables with detailed information about the features

Table	Description	Features
omr	The Online Medical Record (OMR) table contains miscellaneous information from the EHR	subject_id, chartdate, seq_num, result_name, result_value
provider	The provider table lists deidentified provider identifiers used in the database	provider_id
admission	Detailed information about hospital stays	subject_id, hadm_id, admittime, dischtime, deathtime, admission_type, admit_provider_id, admission_location, discharge_location, insurance, language, marital_status, race, edregtime, edouttime, hospital_expire_flag
diagnoses_icd	Billed ICD-9/ICD-10 diagnoses for hospitalizations	subject_id, hadm_id, seq_num, icd_code, icd_version
drgcodes	Billed diagnosis-related group (DRG) codes for hospitalizations	subject_id, hadm_id, drg_type, drg_code, description, drg_severity, drg_mortality
emar	The Electronic Medicine Administration Record (eMAR); barcode scanning of medications at the time of administration	subject_id, hadm_id, emar_id, emar_seq, poe_id, pharmacy_id, enter_provider_id, charttime, medication, event_txt, scheduletime, storetime
emar_detail	Supplementary information for electronic administrations recorded in eMAR	subject_id, emar_id, emar_seq, parent_field_ordinal, administration_type, pharmacy_id, barcode_type, reason_for_no_barcode, complete_dose_not_given, dose_due, dose_due_unit, dose_given, dose_given_unit, will_remainder_of_dose_be_given, product_amount_given, product_unit, product_code, product_description, product_description_other, prior_infusi, n_rate, infusion_rate, infusion_rate_adjustment, infusion_rate_adjustment_amount, infusion_rate_unit, route, infusion_complete, completion_interval, new_iv_bag_hung, contin, ed_infusion_in_other_location, restart_interval, side, site, non_formulary_visual_verification
hpcsevents	Billed events occurring during the hospitalization. Includes CPT codes	subject_id, hadm_id, chartdate, hcpcs_cd, seq_num, short_description
labevents	Laboratory measurements sourced from patient-derived specimens	labevent_id, subject_id, hadm_id, specimen_id, itemid, order_provider_id, charttime, storetime, value, valuenum, valueuom, ref_range_lower, ref_range_upper, flag, priority, comments
microbiologyevents	Microbiology cultures	microevent_id,subject_id, hadm_id, micro_specimen_id, order_provider_id, chartdate, charttime, spec_itemid, spec_type_desc, test_seq, storedate, storetime, test_itemid, test_n, me, org_itemid, org_name, isolate_num, quantity, ab_itemid, ab_name, dilution_text, dilution_comparison, dilution_value, interpretation, comments
patients	Patients' gender, age, and date of death if information exists	subject_id, gender, anchor_age, anchor_year, anchor_year_group, dod
pharmacy	Formulary, dosing, and other information for prescribed medications	subject_id, hadm_id, pharmacy_id, poe_id, starttime, stoptime, medication, proc_type, status, entertime, verifiedtime, route, frequency, disp_sched, infusion_type, sliding_scale, lockout_interval, basal_rate, one_hr_max, doses_per_24_hrs, duration, duration_interval, expiration_value, expiration_unit, expirationdate, dispensation, fill_quantity

poe	Orders made by providers relating to patient care	poe_id, poe_seq, subject_id, hadm_id, ordertime, order_type, order_subtype, transaction_type, discontinue_of_poe_id, discontinued_by_poe_id, order_provider_id, order_status
poe_detail	Supplementary information for orders made by providers in the hospital	poe_id, poe_seq, subject_id, field_name, field_value
prescriptions	Prescribed medications	subject_id, hadm_id, pharmacy_id, poe_id, poe_seq, order_provider_id, starttime, stoptime, drug_type, drug, formulary_drug_cd, gsn, ndc, prod_strength, form_rx, dose_val_rx, dose_unit_rx, form_val_disp, form_unit_disp, doses_per_24_hrs, route
procedures_icd	Billed procedures for patients during their hospital stay	subject_id, hadm_id, seq_num, chartdate, icd_code, icd_version
services	The hospital service(s) that cared for the patient during their hospitalization	subject_id, hadm_id, transfertime, prev_service, curr_service
transfers	Detailed information about patients' unit transfers	subject_id, hadm_id, transfer_id, eventtype, careunit, intime, outtime
d_hcpcs	Dimension table for hpcsevents; provides a description of CPT codes	code, category, long_description, short_description
d_icd_diagnoses	Dimension table for diagnoses_icd; provides a description of ICD-9/ICD-10 billed diagnoses	icd_code, icd_version, long_title
d_icd_procedures	Dimension table for procedures_icd; provides a description of ICD-9/ICD-10 billed procedures	icd_code, icd_version, long_title
d_labitems	Dimension table for labevents provides a description of all lab items	itemid, label, fluid, category

ICU module

Table 2: Description of *ICU* module Tables with detailed information about the features

Table	Description	Features
caregiver	The caregiver table lists deidentified provider identifiers used in the ICU module	caregiver_id
d_items	Dimension table describing itemid. Defines concepts recorded in the events table in the ICU module	itemid, label, abbreviation, linksto, category, unitname, param_type, lownormalvalue, highnormalvalue
chartevents	Charted items occurring during the ICU stay. Contains the majority of information documented in the ICU	subject_id, hadm_id, stay_id, caregiver_id, charttime, storetime, itemid, value, valuenum, valueuom, warning
datetimeevents	Documented information which is in a date format (e.g., date of last dialysis)	subject_id, hadm_id, stay_id, caregiver_id, charttime, storetime, itemid, value, valueuom, warning
icustays	Tracking information for ICU stays including admission and discharge times	subject_id, hadm_id, stay_id, first_careunit, last_careunit, intime, outtime, los
Ingredientevents	Ingredients of continuous or intermittent administrations including nutritional and water content	subject_id, hadm_id, stay_id, caregiver_id, starttime, endtime, storetime, itemid, amount, amountuom, rate, rateuom, orderid, linkorderid, statusdescription, originalamount, originalrate
inputevents	Information documented regarding continuous infusions or intermittent administrations	subject_id, hadm_id, stay_id, caregiver_id, starttime, endtime, storetime, itemid, amount, amountuom, rate, rateuom, orderid, linkorderid, ordercategoryname, secondaryordercategoryname, ordercomponenttypedescription, ordercategorydescription, patientweight, totalamount, totalamountuom, isopenbag, statusdescription, originalamount, originalrate

outputevents	Information regarding patient outputs including urine, drainage, and so on	subject_id, hadm_id, stay_id, caregiver_id, charttime, storetime, itemid, value, valueuom
procedureevent	Procedures documented during the ICU stay (e.g., ventilation), though not necessarily conducted within the ICU (e.g., x-ray imaging)	subject_id, hadm_id, stay_id, caregiver_id, starttime, endtime, storetime, itemid, value, valueuom, location, locationcategory, orderid, linkorderid, ordercategoryname, ordercategorydescription, patientweight, isopenbag, continueinnextdept, statusdescription, originalamount, originalrate

MIMIC IV-ED

Table 3: Description of *ED* module Tables with detailed information about the features

Table	Description	Features
diagnosis	The diagnosis table provides billed diagnoses for patients. Diagnoses are determined after discharge from the emergency department	subject_id, stay_id, seq_num, icd_code, icd_version, icd_title
edstays	The edstays table 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	subject_id, hadm_id, stay_id, intime, outtime, gender, race, arrival_transport, disposition
medrecon	On admission to the emergency departments, staff will ask the patient what current medications they are taking. This process is called medicine reconciliation, and the medrecon table stores the findings of the care providers	subject_id, stay_id, charttime, name, gsn, ndc, etc_rn, etccode, etcdescription
pyxis	The pyxis table provides information for medicine dispensations made via the Pyxis system	subject_id, stay_id, charttime, med_rn, name, gsn_rn, gsn
triage	The triage table contains information about the patient when they were first triaged in the emergency department	subject_id, stay_id, temperature, heartrate, resprate, o2sat, sbp, dbp, pain, acuity, chiefcomplaint
vitalsign	Patients admitted to the emergency department have routine vital signs taken every 1-4 hours. These vital signs are stored in the vitalsign table	subject_id, stay_id, charttime, temperature, heartrate, resprate, o2sat, sbp, dbp, rhythm, pain

MIMIC IV-CXR

Table 4: Description of *CXR* module Tables with detailed information about the features

Table	Description	Features
cxr_record_list	Lists all records in the MIMIC-CXR database	subject_id, study_id, dicom_id

MIMIC IV-Note

Table 5: Description of *Note* module Tables with detailed information about the features

Table	Description	Features
discharge	Discharge summaries for hospitalizations	note_id, subject_id, hadm_id, note_type, note_seq, charttime, storetime, text
discharge_detail	Auxiliary information for discharge summaries	note_id, subject_id, field_name, field_value, field_ordinal
radiology	Radiology report	note_id, subject_id, hadm_id, note_type, note_seq, charttime, storetime, text
radiology_detail	Auxiliary information for radiology notes	note_id, subject_id, field_name, field_value, field_ordinal
cxr_record_list	Lists all records in the MIMIC-CXR database	subject_id, study_id, dicom_id

+ Can/How the dataset be/are created from the MIMIC database?

The MIMIC database is a comprehensive clinical research database that encompasses various types of data, such as patient admissions, ICU records, triage information, bedside health records, X-rays, and clinician medical notes. It offers researchers the flexibility to create custom datasets tailored to their specific research tasks.

For example, if the objective is to predict **heart failure**, relevant cohorts related to heart failure can be extracted from tables like admission, patient, diagnoses_icd, and d_icd_diagnoses in the hosp module. Additional features associated with heart failure can be obtained by linking tables from the ICU module and ED module. Once the cohort and their corresponding heart-related features are extracted, they undergo pre-processing and cleaning before being represented in either a time series or non-time series format, depending on the prediction task. This allows for the creation of suitable datasets for predictive modeling. Similarly, researchers can curate a wide range of task-specific datasets based on their specific needs.

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?

A dataset derived from the MIMIC contains patient health data. The data can be patient demography (Age, gender, ethnicity, language etc.), ICU details, X-ray images or even Clinician notes. It differs depending on the the intended prediction task.

How many instances are there in total (of each type, if appropriate)?

Dataset is extracted from the MIMIC database based on the intended task and count of instances depend on the dataset extracted.

For instance, If we intend to create a *MIMIC IV ED dataset* by linking ED, hosp and ICU modules then the dataset will have *425087* instances. Similarly several complex datasets can be created and the instance of the dataset vary depending of the prediction task/requirements.

Does the dataset/database contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The MIMIC-IV database is a subset of deidentified electronic health records (EHRs) obtained from patients admitted to BIDMC between 2008 and 2019. It is a curated collection that has undergone validation and quality assurance by a team of interdisciplinary experts. The database includes a diverse range of patients and diagnoses, making it suitable for various research purposes. However, it is important to acknowledge that the dataset is not comprehensive, as it represents a subset of the overall patient population. Researchers should be mindful of potential biases inherent in the dataset and employ appropriate methods to address them when conducting analyses or studies[1].

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

MIMIC IV (hosp and ICU module) and MIMIC-IV-ED (ED module) consists of raw unprocessed text, Date time and number data in comma separated format of the patients admitted to the hospital, ICU and ED. Whereas, MIMIC-CXR and MIMIC-Note contains images of the Chest X-Rays and free-text clinical notes for hospitalized patients respectively. Table **??** provides detailed feature information of the data.

Is there a label or target associated with each instance? If so, please provide a description.

The choice of target variable in the MIMIC dataset depends on the specific prediction task at hand. For example, if the goal is to predict the length of stay in the ICU, the *los* attribute in the *icustay* table can serve as the target variable. On the other hand, if the objective is to predict *in-hospital mortality*, the *hospital_expire_flag* in the *admissions* table can be used as the target variable. **The selection of the target variable is contingent upon the specific prediction task being undertaken**.

Are there recommended data splits (e.g., training, development/validation, testing)?

No

Are there any errors, sources of noise, or redundancies in the database? If so, please provide a description.

Our analysis of the MIMIC IV dataset has revealed several biases and inconsistencies that researchers should be aware of,

1. **Inconsistencies in patient details:**Patient language is inconsistently recorded, with only English being specified while other languages are marked as '?' or unknown.
2. **Inconsistencies in in-hospital expiry information:** The admission table contains multiple reports of the same patient's death, leading to inconsistencies.
3. **Vagueness in insurance coverage information:** The dataset lacks definitive information about insurance coverage, limiting researchers' ability to draw conclusions on insurance choices.
4. **Inconsistencies in hospital admit and discharge timestamps:** The admission table exhibits inconsistencies in the recorded timestamps, and there are also *missing values for death time*.
5. **Potential representation bias in the dataset:** The database owners acknowledge the potential for bias, particularly since the data is derived from a single hospital system and may not be***representative of the entire population***.

The data Johnson et al. (2023) in the database is collected during routine clinical practice, reflecting the specific practices of the hospital. It is important to note that there may be implausible values present in the database due to the archival process[1]. Therefore, caution should be exercised when using the data, and researchers should be mindful of the dataset limitations and potential biases.

We strongly recommend that researchers adhere to best practice guidelines Goldberger et al. (2000) when analyzing the data

Does the database contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor– patient confidentiality, data that includes the content of individuals' non-public communications)?

Yes, the MIMIC IV dataset includes medical records of patients, encompassing confidential personal and health-related information. However, the dataset is constructed with patient privacy as a priority, and all data within the database undergoes de-identification processes to comply with Health Insurance Portability and Accountability Act (HIPAA) regulations.

Does the database identify any subpopulations (e.g., by age, gender)?

Yes. Database (specifically admission and patient tables) has patient demographic data such as age, gender, ethnicity, language, insurance and marital status.

MIMIC IV Distribution statistics

Table 6: Admission distribution statistics

Description	Value
Total records	180,733
Male	47%
Female	53%
Min Age	18
Max Age	91
Predominant Ethnicity	White (67.2%)

Table 7: Patient distribution statistics

Description	Value
Total records	299,712
Male	47%
Female	53%
Min Age	18
Max Age	91

Table 8: ED table distribution statistics

Description	Value
Total records	299712
Male	46%
Female	54%
Predominant Ethnicity	White (58%)
Predominant Disposition	Home

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the database? If so, please describe how.

No, all data in the database is de-identified in accordance with HIPAA regulations.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?

Yes, database recorded demographic information like ethnicity, gender, age, marital status, language and insurance

+ *Does researchers have to take any important measures to handle the data with care?*

To ensure patient privacy, researchers are required to comply data usage agreements mandated in Goldberger et al. (2000), Johnson et al. (2023) and obtain the necessary approvals and certifications before accessing the dataset. Researchers working with healthcare-related data have a responsibility to handle the data carefully and ethically, taking measures to prevent any potential harm or dissatisfaction. While the data is de-identified in accordance with HIPAA regulations, it is crucial to treat the data with respect and caution, following best practices. Additionally, the collection of patient information and the creation of the research resource have been approved by the Institutional Review Board of the Beth Israel Deaconess Medical Center.

A.3 Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The data was extracted from the hospital databases of the Beth Israel Deaconess Medical Center (BIDMC) specifically for patients admitted to the intensive care units. A comprehensive patient list was compiled, including all medical record numbers associated with ICU or emergency department admissions from 2008 to 2019. To ensure the reliability of the database, a multidisciplinary team of scientists and clinicians thoroughly evaluated MIMIC-IV during its development, conducting code reviews and documenting any identified issues using a ticket system Johnson et al. (2023).

What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?

MIMIC-IV is derived from two distinct database systems within the hospital setting: a customized *electronic health record (EHR)* used across the entire hospital and a specialized clinical information system called *MetaVision (iMDSoft)* specifically designed for the intensive care units at the Beth Israel Deaconess Medical Center (BIDMC).

To ensure the accuracy and reliability of the MIMIC-IV dataset, a diverse team of scientists and clinicians conducted a comprehensive evaluation during its development, which included code reviews and the systematic documentation of identified issues using a ticket system Johnson et al. (2023).

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

Over a period of *11 years, from 2008 - 2019*

Were any ethical review processes conducted (e.g., by an institutional review board)?

Yes, the collection of patient information and creation of the research resource was reviewed by the *Institutional Review Board at the Beth Israel Deaconess Medical Center*.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

Data is collected from hospital EHR and ICU specific clinical information system at the BIDMC called *CareVue and MetaVision (iMDSoft)* .

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

Unknown, however the MIMIC data is deidentified[1], and patient identifiers were removed according to the Health Insurance Portability and Accountability Act (HIPAA) Safe Harbor provision Johnson et al. (2023).

A.4 Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.

The data within the MIMIC-IV database underwent reorganization to enhance its suitability for retrospective data analysis Johnson et al. (2023). This involved denormalizing tables, eliminating audit trails, and consolidating the data into a smaller number of tables. The primary objective of this process was to simplify the retrospective analysis of the database. Notably, no data cleaning procedures were applied to ensure that the dataset accurately represents real-world clinical data[1].

To protect patient privacy, patient identifiers were removed in compliance with HIPAA regulations. Random ciphers were used to replace patient identifiers, resulting in deidentified integer values for patients, hospitalizations, and ICU stays. Structured data underwent filtering using look-up tables and allow lists. Additionally, dates and times were randomly shifted into the future by a specific number of days. Consequently, the data for each individual patient remains internally consistent Johnson et al. (2023).

Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.

Unknown. However, authors have stated that the application of a free-text deidentification algorithm was used as a measure to remove personally identifiable information (PHI) from the free-text data, if needed.

A.5 Uses

Has the database/dataset been used for any tasks already? If so, please provide a description.

Yes, MIMIC database is one of the most widely used CRD. It has been widely used for below types of works,

1. Prediction tasks like,
   (a) **Readmission** Assaf and Jayousi (2020), Chen et al. (2022), Partovi et al. (2022), Rojas et al. (2018), Thacker (2023) (30, 60, 90, 120 and custom days) - Predict patients at risk of readmission early in the health care process(helps to prioritize care towards such patients preventing mortality and readmission).
   (b) **Mortality** Ahmad et al. (2021), Caicedo-Torres and Gutierrez (2019), Feng et al. (2018), Kong et al. (2020), Lin et al. (2019) - Predict the likelihood of patients dying.
      i. **In-hospital** Chen et al. (2021), Ding et al. (2021), Theis et al. (2021), Yang et al. (2023b) - Predict the likelihood of patient dying in hospital while they are admitted (helpful to identify high risk patients early on to provide medical interventions).
      ii. **Short term** Gao et al. (2021), Hou et al. (2020), Lu et al. (2021), Luo et al. (2022), Zhang et al. (2022a) - Predict short-term mortality (typically within 2-3 days) after ICU admission
      iii. **Long term** Ahmad et al. (2021), Caicedo-Torres and Gutierrez (2019), Kong et al. (2020), Liu et al. (2021a) - Predict long-term mortality (typically within 30 days to 1 year) after hospital discharge.
   (c) **Length of stay (LOS)** Geethamani and Rangaraj, Geethamani and Rangaraj, Liu et al. (2023), Shu et al. (2023), Wang et al. (2022) - Predict the length of stay of each admission(typically predicting > 3 and 7 days stay. Custom days is also being predicted.)
   (d) **Phenotype label and ICD-9/10 code grouping** - Helpful in tasks like disease prediction, outcome analysis, treatment recommendation and customized treatments.
      i. **Phenotype labelling** Dong et al. (2022), Singh et al. (2020), Yang et al. (2023a), Zhang et al. (2022b)classify patients into specific groups based on their diagnoses, procedures, medications, and other clinical variables.
      ii. **Grouping ICD 9/10 codes** Huang et al. (2019), Li and Yu (2020), Li et al. (2018)into different categories based on patient diagnosis to classify the disease.
2. Prediction for specific **health ailments** like,
   (a) Heart failure Ali et al. (2022), Li et al. (2021)
   (b) Chronic Kidney Disease (CKD) Sun et al. (2022), Yue et al. (2022)
   (c) Chronic obstructive pulmonary disease (COPD) Liu et al. (2021b), Rojas et al. (2018)
   (d) Coronary artery disease (CAD) Yang et al. (2023b), Ye et al. (2023)
   (e) Sepsis Böck et al. (2022), Yue et al. (2022)
   (f) Cancer Kurniati et al. (2018), Magna et al. (2020)
   (g) Ventilation failure Geri et al. (2021), Sayed et al. (2021)

Is there a repository that links to any or all papers or systems that use the database/dataset? If so, please provide a link or other access point.

No, however the owners Johnson et al. (2023) have provided the repository `https://github.com/MIT-LCP/mimic-code` where the code and other discussions related to the database are hosted.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks,

financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

The data available in the database reflects the idiosyncrasies of routine clinical practice, as stated by the owners. The archival process may have introduced implausible values and potential bias into the data. Therefore, it is important for researchers to follow best practice guidelines when using the data for analysis or other purposes Johnson et al. (2023).

Are there tasks for which the dataset should not be used? If so, please provide a description.

Unknown, the owners of the database did not provide clear information in the documentation.

A.6 Distribution

+ Is the data publicly available? How and where can it be accessed (e.g., website, GitHub)?

Yes. The MIMIC-IV data is accessible to the public through the PhysioNet[2] . To gain access, individuals need to become a PhysioNet credentialed user and agree to the data use agreement. Once granted access, users can download the complete set of file or select specific subsets that align with their requirements.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

Access to the MIMIC-IV data is granted through a license agreement called the *Data Use Agreement (DUA)*, which outlines the terms and conditions for data usage. To obtain access, users are required to complete an online course on the ethical use of human subjects research data and obtain a certificate of completion. With the certificate, users can then apply for dataset access through the PhysioNet[2]. The application process involves agreeing to the DUA terms and providing details about the intended use of the data.

A.7 Maintenance

Is the database maintained? Who will be supporting/hosting/maintaining the database?

Yes, MIMIC-IV is maintained by the Laboratory for Computational Physiology at the Massachusetts Institute of Technology (MIT) and BIDMC. They provide ongoing support and maintenance for the database

How can the owner/curator/manager of the database be contacted (e.g., email address)?

For private issue, they can be contacted at `mimic-support@physionet.org` and for issues related to patient health information (PHI) `phi-report@physionet.org` is being used[1].

Will the database be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (e.g., mailing list, GitHub)?

Yes, the MIMIC-IV database is regularly updated by the MIT Laboratory for Computational Physiology team. The latest version, v2.2, has been released, which includes updates from the previous version, v1.0. The frequency of future updates is unknown, but any information regarding updates can be found on the official website `https://physionet.org/content/mimiciv/2.2/` and `https://github.com/MIT-LCP/mimic-code`.

Will older versions of the database continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.

Previous versions of the database will continue to be supported and maintained, however it is not explicitly stated whether they might have any further updates by the owners.

[2] `https://physionet.org/content/mimiciv/2.2/`

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description.

Content for the MIMIC website and documentation is hosted publicly on GitHub: `https://github.com/MIT-LCP/mimic-website`. To raise a problem or to suggest an improvement, new issue can be created at: `https://github.com/MIT-LCP/mimic-website/issues`. To take part in discussion `https://github.com/MIT-LCP/mimic-code/discussions` can be used.

+ What is the Data life cycle of MIMIC database?

1. Data Acquisition
2. Data Archive
3. Data Preparation
4. Data loading

constitutes the life cycle if MIMIC CRD.

Data Acquisition

Data is collected from the source which may be internal, external or both.

Table 9: Data Acquisition

Internal (In-hospital) data	**External data source**
ICU/MICU/SICU/CCU/CVICU/NICU data (vitals, trends, anomalies) Chart details (Fluids, medications etc.) Demographics (age, gender, ethnicity, language, marital status, religion, insurance etc.) Lab reports Billing details Physician notes Provider order entries etc.	Social security death Index etc

Data Archive Data collected from the source is Archived before proceeding with data preparation for later use.

Data Preparation

To ensure compliance with HIPAA regulations, measures such as deidentification, date shifts, and format conversions are applied to the archival data. The data is then reorganized into a more suitable format for retrospective analysis, which involves consolidating tables, denormalizing data, and removing audit trails. It's important to note that no data cleaning procedures were performed to maintain the authenticity of the real-world clinical dataset. Feedback from users will be considered for further iterations, and the final version of the data will be loaded into the database.

Data loading to database

Final version of the data is then loaded to the database which is built on a PostgreSQL relational database management system and is hosted on a secure server infrastructure. The data can either be downloaded locally or accessed on the cloud via BigQuery, AWS or GCS

B Risk Prediction Task Analysis

We analyzed prevalent HML prediction models using the MIMIC dataset across all demographics. Access to the MIMIC IV v2.0 data `https://physionet.org/content/mimiciv/2.2/` is restricted to PhysioNet credentialed users only.

B.1 Credential Acess process

Researchers can gain physionet Credential access by following the below,

1. Submit their personal details for physionet's review [3]
2. Complete CITI Data or Specimens Only Research [4] training and/or sign a Data Use Agreement[5]

Once PhysioNet has reviewed the personal details, CITI training completion, and DUA agreement, they will grant access to the MIMIC IV database. You can retrieve all the necessary files from the files section of the MIMIC IV v2.0 [6].

B.2 Prediction task-specific dataset retrieval

For our analysis, we retrieved datasets for various prediction tasks, including in-hospital mortality, heart failure, chronic kidney disease (CKD), 30-day readmission, and length of stay (LOS) for heart failure. We followed an established pipeline Gupta et al. (2022) for most tasks, except for sepsis mortality, where we directly retrieved the dataset without a specific pipeline to evaluate and authenticate the results obtained from both the strategies.

B.3 Additional Analysis Information

Python 3 software was used for the analysis, leveraging its analytical and statistical libraries. We examined the data distribution, including 33 different ethnic records, and categorized patients into Asian, Black, Hispanic/Latino, Other, and White ethnic subgroups following the previous fairness works Meng et al. (2022), Röösli et al. (2021). Our analysis uncovered disparities in patient treatment based on ethnicity and insurance status, impacting patient outcomes. Statistical tests, such as chi-square and ANOVA, confirmed a significant association between ethnicity and prediction outcome, *highlighting the importance of considering model performance across all demographics for fair, generalizable, and data-aware HML models*.

C Example Usage & Reproducibility

The complete code for analyzing the length of ICU stay for heart failure patients over 7 days is available at `https://github.com/Trustworthy-HML-Models/Unmasking-Data-Bias-for-Data-Aware-Modeling` , and other prediction tasks can be reproduced by substituting the respective datasets.

```
df_new = pd.read_csv('Task-specific-dataset.csv')
```

Distribution of the dataset is analysed to understand the dataset better and the ethnic grouping is performed

```

racial_groupings = {
    'White': ['white', 'white - brazilian','white - eastern european',
         'white - other european','white - russian','portuguese'],
    'Black': ['black/african', 'black/african american', 'black/cape
        verdean','black/caribbean island','south american'],
    'Hispanic/Latino': ['hispanic or latino','hispanic/latino -
        central american', 'hispanic/latino - columbian','hispanic/
        latino - cuban', 'hispanic/latino - dominican','hispanic/
        latino - guatemalan', 'hispanic/latino - honduran','hispanic/
        latino - mexican', 'hispanic/latino - puerto rican','hispanic/
        latino - salvadoran'],
```

[3]Personal detail submission - Available at `https://physionet.org/settings/credentialing/`

[4]CITI Data or Specimens Only Research - `https://physionet.org/content/mimiciv/view-required-training/2.2/`

[5]DUA - Available at `https://physionet.org/content/mimiciv/view-dua/2.2/`

[6]MIMIC IV v2.0 - Available at `https://physionet.org/content/mimiciv/2.2/`

```
    'Asian': ['asian', 'asian - asian indian','asian - chinese', '
        asian - korean', 'asian - south east asian'],
    'Other': ['native hawaiian or other pacific islander','other','
        patient declined to answer', 'unable to obtain', 'unknown','
        american indian/alaska native','multiple race/ethnicity']
}

# Replace the original race values with the new groupings

df_new['ethnicity'] = df_new['ethnicity'].str.strip().str.lower()


def get_race_sub_group(race):
    for group, races in racial_groupings.items():
        if race in races:
            return group
    return 'Unknown'

df_new['race_sub_group'] = df_new['ethnicity'].apply(
    get_race_sub_group)
```

Further detailed analysis of patient treatment based on ethic and Insurance demographics is done and they are visualized for better understanding the insights of the results. This analysis uncovered disparities in patient treatment based on ethnicity and insurance status, impacting patient outcomes.

Statistical tests, such as chi-square and ANOVA is performed to identify the association between ethnicity and prediction outcome

```

# Create a contingency table
contingency_table = pd.pivot_table(df, values='count', index=['
    race_sub_group', 'insurance'], columns='label', fill_value=0)

# Perform the chi-square test
chi2, p, _, _ = chi2_contingency(contingency_table)

# Print the contingency table
print("Contingency Table:")
print(contingency_table)

# Print the test statistic and p-value
print("Chi-square test statistic:", chi2)
print("p-value:", p)

# Perform ANOVA
from scipy import stats
result = stats.f_oneway(
    df[df['race_sub_group'] == 'Asian']['count'],
    df[df['race_sub_group'] == 'Black']['count'],
    df[df['race_sub_group'] == 'Hispanic/Latino']['count'],
    df[df['race_sub_group'] == 'Other']['count'],
    df[df['race_sub_group'] == 'White']['count']
)

# Print the ANOVA test result
print("ANOVA test result:")
print("F-value:", result.statistic)
print("p-value:", result.pvalue)
```

Test results confirmed a significant association between ethnicity and prediction outcome, *highlighting the importance of considering model performance across all demographics for fair, generalizable, and data-aware HML models*

References

Fahad Shabbir Ahmad, Liaqat Ali, Hasan Ali Khattak, Tahir Hameed, Iram Wajahat, Seifedine Kadry, and Syed Ahmad Chan Bukhari. A hybrid machine learning framework to predict mortality in paralytic ileus patients using electronic health records (ehrs). *Journal of Ambient Intelligence and Humanized Computing*, 12:3283–3293, 2021.

Aliya Ali, Saleha Yurf Asghar, Ali Danish Khan Yousafzai, Ali Haider Bangash, Rabia Mohsin, Arshiya Fatima, Saiqa Zehra, Ayesha Khalid Khan, Ali Haider Shah, Syed Mohammad Mehmood Abbas, et al. Prediction of in-hospital mortality among heart failure patients: An automated machine learning analysis of mimic-iii database. *American Heart Journal*, 254:261, 2022.

Rasha Assaf and Rashid Jayousi. 30-day hospital readmission prediction using mimic data. In *2020 IEEE 14th International Conference on Application of Information and Communication Technologies (AICT)*, pages 1–6. IEEE, 2020.

Markus Böck, Julien Malle, Daniel Pasterk, Hrvoje Kukina, Ramin Hasani, and Clemens Heitzinger. Superhuman performance on sepsis mimic-iii data by distributional reinforcement learning. *PLoS One*, 17(11):e0275358, 2022.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

William Caicedo-Torres and Jairo Gutierrez. Iseeu: Visually interpretable deep learning for mortality prediction inside the icu. *Journal of biomedical informatics*, 98:103269, 2019.

Han Chen, Shu-Rong Gong, and Rong-Guo Yu. Association between normalized lactate load and mortality in patients with septic shock: an analysis of the mimic-iii database. *BMC anesthesiology*, 21(1):1–8, 2021.

Qifan Chen, Yang Lu, Charmaine Tam, and Simon Poon. Outcome-oriented predictive process monitoring to predict unplanned icu readmission in mimic-iv database. 2022.

Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

Ning Ding, Cuirong Guo, Changluo Li, Yang Zhou, and Xiangping Chai. An artificial neural networks model for early predicting in-hospital mortality in acute pancreatitis in mimic-iii. *BioMed research international*, 2021, 2021.

Hang Dong, Víctor Suárez-Paniagua, Huayu Zhang, Minhong Wang, Arlene Casey, Emma Davidson, Jiaoyan Chen, Beatrice Alex, William Whiteley, and Honghan Wu. Ontology-based and weakly supervised rare disease phenotyping from clinical notes. *arXiv preprint arXiv:2205.05656*, 2022.

Mengling Feng, Jakob I McSparron, Dang Trung Kien, David J Stone, David H Roberts, Richard M Schwartzstein, Antoine Vieillard-Baron, and Leo Anthony Celi. Transthoracic echocardiography and mortality in sepsis: analysis of the mimic-iii database. *Intensive care medicine*, 44:884–892, 2018.

Qiaoyan Gao, Dandan Wang, Pingping Sun, Xiaorong Luan, Wenfeng Wang, et al. Sentiment analysis based on the nursing notes on in-hospital 28-day mortality of sepsis patients utilizing the mimic-iii database. *Computational and Mathematical Methods in Medicine*, 2021, 2021.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

D Geethamani and R Rangaraj. Heterogeneous multi-model ensemble based length of stay prediction on mimic iii.

Guillaume Geri, Loic Ferrer, Nam Tran, Leo A Celi, Matthieu Jamme, Joon Lee, and Antoine Vieillard-Baron. Cardio-pulmonary-renal interactions in icu patients. role of mechanical ventilation, venous congestion and perfusion deficit on worsening of renal function: Insights from the mimic-iii database. *Journal of critical care*, 64:100–107, 2021.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23): e215–e220, 2000.

Mehak Gupta, Brennan Gallamoza, Nicolas Cutrona, Pranjal Dhakal, Raphael Poulain, and Rahmatollah Beheshti. An extensive data processing pipeline for mimic-iv. In *Machine Learning for Health*, pages 311–325. PMLR, 2022.

Nianzong Hou, Mingzhe Li, Lu He, Bing Xie, Lin Wang, Rumin Zhang, Yong Yu, Xiaodong Sun, Zhengsheng Pan, and Kai Wang. Predicting 30-days mortality for mimic-iii patients with sepsis-3: a machine learning approach using xgboost. *Journal of translational medicine*, 18(1):1–14, 2020.

Jinmiao Huang, Cesar Osorio, and Luke Wicent Sy. An empirical evaluation of deep learning for icd-9 code assignment using mimic-iii clinical notes. *Computer methods and programs in biomedicine*, 177:141–153, 2019.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Benjamin Moody, Brian Gow, Li-wei H Lehman, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

Guilan Kong, Ke Lin, and Yonghua Hu. Using machine learning methods to predict in-hospital mortality of sepsis patients in the icu. *BMC medical informatics and decision making*, 20:1–10, 2020.

Angelina Prima Kurniati, Geoff Hall, David Hogg, and Owen Johnson. Process mining in oncology using the mimic-iii dataset. In *Journal of Physics: Conference Series*, volume 971, page 012008. IOP Publishing, 2018.

Fei Li and Hong Yu. Icd coding from clinical text using multi-filter residual convolutional neural network. In *proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8180–8187, 2020.

Fuhai Li, Hui Xin, Jidong Zhang, Mingqiang Fu, Jingmin Zhou, and Zhexun Lian. Prediction model of in-hospital mortality in intensive care unit patients with heart failure: machine learning-based, retrospective analysis of the mimic-iii database. *BMJ open*, 11(7):e044779, 2021.

Min Li, Zhihui Fei, Min Zeng, Fang-Xiang Wu, Yaohang Li, Yi Pan, and Jianxin Wang. Automated icd-9 coding via a deep learning approach. *IEEE/ACM transactions on computational biology and bioinformatics*, 16(4): 1193–1202, 2018.

Ke Lin, Yonghua Hu, and Guilan Kong. Predicting in-hospital mortality of patients with acute kidney injury in the icu using random forest model. *International journal of medical informatics*, 125:55–61, 2019.

Dongliang Liu, Yiyang Tang, and Qian Zhang. Admission hyperglycemia predicts long-term mortality in critically ill patients with subarachnoid hemorrhage: a retrospective analysis of the mimic-iii database. *Frontiers in Neurology*, 12:678998, 2021a.

Tao Liu, Haochen Xuan, Lili Wang, Xiaoqun Li, Zhihao Lu, Zhaoxuan Tian, Junhong Chen, Chaofan Wang, Dongye Li, and Tongda Xu. The association between serum albumin and long length of stay of patients with acute heart failure: A retrospective study based on the mimic-iv database. *Plos one*, 18(2):e0282289, 2023.

Taotao Liu, Qinyu Zhao, and Bin Du. Effects of high-flow oxygen therapy on patients with hypoxemia after extubation and predictors of reintubation: a retrospective study based on the mimic-iv database. *BMC Pulmonary Medicine*, 21(1):1–15, 2021b.

Zongqing Lu, Jin Zhang, Jianchao Hong, Jiatian Wu, Yu Liu, Wenyan Xiao, Tianfeng Hua, and Min Yang. Development of a nomogram to predict 28-day mortality of patients with sepsis-induced coagulopathy: an analysis of the mimic-iii database. *Frontiers in medicine*, 8:661710, 2021.

Mengdi Luo, Yang Chen, Yuan Cheng, Na Li, and He Qing. Association between hematocrit and the 30-day mortality of patients with sepsis: A retrospective analysis based on the large-scale clinical database mimic-iv. *PloS one*, 17(3):e0265758, 2022.

Andrés Alejandro Ramos Magna, Héctor Allende-Cid, Carla Taramasco, Carlos Becerra, and Rosa L Figueroa. Application of machine learning and word embeddings in the classification of cancer diagnosis using patient anamnesis. *Ieee Access*, 8:106198–106213, 2020.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

Chuizheng Meng, Loc Trinh, Nan Xu, James Enouen, and Yan Liu. Interpretability and fairness evaluation of deep learning models on mimic-iv dataset. *Scientific Reports*, 12(1):7166, 2022.

Andi Partovi, Dickson Lukose, and Geoffrey I Webb. Mipy: A framework for benchmarking machine learning prediction of unplanned hospital and icu readmission in the mimic-iv database. 2022.

Alvin Rajkomar, Michaela Hardt, Michael D Howell, Greg Corrado, and Marshall H Chin. Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine*, 169(12):866–872, 2018.

Juan C Rojas, Kyle A Carey, Dana P Edelson, Laura R Venable, Michael D Howell, and Matthew M Churpek. Predicting intensive care unit readmission with machine learning using electronic health record data. *Annals of the American Thoracic Society*, 15(7):846–853, 2018.

Eliane Röösli, Brian Rice, and Tina Hernandez-Boussard. Bias at warp speed: how ai may contribute to the disparities gap in the time of covid-19. *Journal of the American Medical Informatics Association*, 28(1): 190–192, 2021.

Mohammed Sayed, David Riano, and Jesús Villar. Predicting duration of mechanical ventilation in acute respiratory distress syndrome using supervised machine learning. *Journal of Clinical Medicine*, 10(17):3824, 2021.

Tingting Shu, Jian Huang, Jiewen Deng, Huaqiao Chen, Yang Zhang, Minjie Duan, Yanqing Wang, Xiaofei Hu, and Xiaozhu Liu. Development and assessment of scoring model for icu stay and mortality prediction after emergency admissions in ischemic heart disease: a retrospective study of mimic-iv databases. *Internal and Emergency Medicine*, 18(2):487–497, 2023.

AK Singh, Mounika Guntu, Ananth Reddy Bhimireddy, Judy W Gichoya, and Saptarshi Purkayastha. Multi-label natural language processing to identify diagnosis and procedure codes from mimic-iii inpatient notes. *arXiv preprint arXiv:2003.07507*, 2020.

Chun Sun, Deqing Chen, Xin Jin, Guangtao Xu, Chenye Tang, Xiao Guo, Zhiling Tang, Yixin Bao, Fei Wang, and Ruilin Shen. Association between acute kidney injury and prognoses of cardiac surgery patients: Analysis of the mimic-iii database. *Frontiers in Surgery*, 9, 2022.

Jared Thacker. *A Machine Learning Pipeline for Readmission Prediction with MIMIC-III*. PhD thesis, Auburn University, 2023.

Julian Theis, William L Galanter, Andrew D Boyd, and Houshang Darabi. Improving the in-hospital mortality prediction of diabetes icu patients using a process mining/deep learning architecture. *IEEE Journal of Biomedical and Health Informatics*, 26(1):388–399, 2021.

Dongyan Wang, Xiaoyan Guo, Wenwen Xia, Zhijuan Ru, Yihai Shi, and Zhengyu Hu. Effect of admission serum calcium levels and length of stay in patients with acute pancreatitis: Data from the mimic-iii database. *Emergency Medicine International*, 2022, 2022.

Siyue Yang, Paul Varghese, Ellen Stephenson, Karen Tu, and Jessica Gronsbell. Machine learning approaches for electronic health records phenotyping: a methodical review. *Journal of the American Medical Informatics Association*, 30(2):367–381, 2023a.

Wei Yang, Hong Zou, Meng Wang, Qin Zhang, Shadan Li, and Hongyin Liang. Mortality prediction among icu inpatients based on mimic-iii database results from the conditional medical generative adversarial network. *Heliyon*, 9(2), 2023b.

Zixiang Ye, Shuoyan An, Yanxiang Gao, Enmin Xie, Xuecheng Zhao, Ziyu Guo, Yike Li, Nan Shen, Jingyi Ren, and Jingang Zheng. The prediction of in-hospital mortality in chronic kidney disease patients with coronary artery disease using machine learning models. *European Journal of Medical Research*, 28(1):1–13, 2023.

Suru Yue, Shasha Li, Xueying Huang, Jie Liu, Xuefei Hou, Yumei Zhao, Dongdong Niu, Yufeng Wang, Wenkai Tan, and Jiayuan Wu. Machine learning for the prediction of acute kidney injury in patients with sepsis. *Journal of translational medicine*, 20(1):1–12, 2022.

HuanRui Zhang, Wen Tian, and YuJiao Sun. The value of anion gap for predicting the short-term all-cause mortality of critically ill patients with cardiac diseases, based on mimic-iii database. *Heart & Lung*, 55:59–67, 2022a.

Jingqing Zhang, Luis Daniel Bolanos Trujillo, Ashwani Tanwar, Julia Ive, Vibhor Gupta, and Yike Guo. Clinical utility of automatic phenotype annotation in unstructured clinical notes: intensive care unit use. *BMJ Health & Care Informatics*, 29(1):e100519, 2022b.