**MIMIC Dataset Summary Document**

**MIMIC-III (Critical Care Database)**

**Overview:**

MIMIC-III is a large, publicly available database of de-identified health data for critical care patients at Beth Israel Deaconess Medical Center (Boston, MA) between 2001 and 2012. It contains records for over 40,000 patients (including >60,000 ICU admissions) in the intensive care units (ICUs) of the hospital. MIMIC-III was one of the first open-access clinical databases of this size, enabling broad research in critical care outcomes.

**Data Content:**

• Demographics (patient profile such as age, gender, etc.)

• Vital signs (bedside monitor measurements, ~hourly)

• Laboratory test results (blood tests, etc.)

• Procedures and medications (treatments administered, procedure codes)

• Caregiver notes and observations (free-text clinical notes, discharge summaries)

• Imaging reports (e.g. radiology reports)

• Outcomes (e.g. length of stay, in-hospital mortality and post-discharge survival)

All data are de-identified to meet HIPAA Safe Harbor standards (e.g. dates shifted, free-text de-identified) while preserving clinical utility.

**Access:**

MIMIC-III is freely accessible on PhysioNet to credentialed users. Access requires completing an online training in human subjects research and signing a data use agreement. (For example, users must complete the CITI “Data or Specimens Only Research” course and register on PhysioNet.) Once credentialed, users can download the database for research use.

**Key Characteristics & Uses:**

MIMIC-III is a single-center ICU dataset, but it is comprehensive, supporting diverse studies. It has been widely used in developing predictive models (for tasks like mortality risk prediction, length-of-stay forecasting, physiologic event detection, and phenotype classification). It also supports epidemiological analyses of critical illness and has even been used in medical education and benchmarking exercises. MIMIC-III’s open availability has led to thousands of publications, making it a gold standard for data-driven critical care research and model evaluation.

**MIMIC-IV (Contemporary Critical Care/Hospital Database)**

**Overview:**

MIMIC-IV is the next-generation MIMIC database, covering a decade of hospital admissions (2008–2019) at the same institution (BIDMC in Boston). It builds upon MIMIC-III with more contemporary data and an improved structure. MIMIC-IV contains data for 65,000+ ICU admissions and also integrates general hospital ward and emergency department data (via modules) for a total of hundreds of thousands of patient records. It was introduced to continue MIMIC’s mission with updated content and to enable new research applications.

**Data Content:**

• Demographics, vitals, labs, medications, procedures, diagnoses, notes, etc. (similar to MIMIC-III)

• Electronic medication administration records

• More granular diagnoses/procedures coding (ICD-9 and ICD-10)

• Modular organization (e.g. “hospital” module for in-patient data, “ICU” for critical care, “ED” and “note” modules)

This modular design highlights data provenance and facilitates combining disparate data sources. For example, one can link a patient’s ED visit to their ICU stay and to their radiology images if available.

**Access:**

Access to MIMIC-IV is provided through PhysioNet with the same requirements as MIMIC-III: completion of the human research training and agreement to the data use policies. Once authorized, users can download the MIMIC-IV database (which is split into parts, such as core clinical data and separate modules like notes or ED data).

**Key Characteristics & Uses:**

MIMIC-IV was released in 2020–2021 (with updates through 2023) to provide a more recent EHR dataset for research. Its modular structure allows more flexible analysis (and extension — e.g., linking to external data or future modules). MIMIC-IV is intended to support a wide array of research studies and education, similar to its predecessor. Early uses of MIMIC-IV include developing and validating machine learning models on more current ICU data and creating FHIR (Fast Healthcare Interoperability Resources) representations for interoperability research. As a direct continuation of MIMIC-III, it carries on enabling studies in outcome prediction, treatment effect analysis, and other data-driven investigations in critical care, with the benefit of contemporary data and larger scale.

**MIMIC-CXR (Chest X-ray Imaging Database)**

**Overview:**

MIMIC-CXR is a large publicly available dataset of chest radiographs (X-ray images) and corresponding radiology reports, sourced from BIDMC. It covers 377,110 images from 227,835 imaging studies, taken for 65,000+ patients (primarily from the emergency department and ICU) between 2011 and 2016. Each study typically includes one or more chest X-ray images (frontal/view positions) along with the original free-text report dictated by a radiologist. All identifying information has been removed to ensure the data are de-identified.

**Data Content:**

• Chest X-ray images (in DICOM format, high-resolution) – frontal and lateral views of the chest for a wide variety of patients/conditions.

• Free-text radiology reports – the narrative interpretations and findings written by radiologists for each imaging study.

• Additionally, derived labels: The dataset has an alternative version (MIMIC-CXR-JPG) with images in JPEG format and structured labels extracted from the reports (for machine learning purposes).

All images and reports are de-identified (e.g. obscuring any personal info on images, removing patient names in text) to comply with HIPAA Safe Harbor.

**Access:**

MIMIC-CXR is available via PhysioNet and requires the same credentialed access as other MIMIC datasets (completion of the data usage training and agreement) due to the presence of clinical images. Once access is granted, users can download the image files and reports.

**Key Characteristics & Uses:**

MIMIC-CXR is one of the largest open chest X-ray datasets with paired reports. It has become a key resource for medical imaging AI research. Researchers use it to train and evaluate algorithms for automated chest X-ray interpretation, including disease detection (e.g. finding abnormalities like nodules, heart failure signs, etc.), report generation from images, and multi-modal learning (combining image and text). The availability of free-text reports also enables natural language processing (NLP) research on radiology text (for example, to develop label extraction methods or analyze report language). The dataset’s size and openness allow for reproducible benchmarks in computer vision – many models (such as deep convolutional networks and transformer-based models) are evaluated on MIMIC-CXR for tasks like abnormality classification. Overall, MIMIC-CXR supports a wide body of research in medical image understanding, clinical NLP, and decision support systems, helping to improve diagnostic algorithms and potentially clinical workflows.

**MIMIC-IV-ED (Emergency Department Dataset)**

**Overview:**

MIMIC-IV-ED is an emergency department module complementary to MIMIC-IV. It includes data for approximately 448,972 ED visits occurring at BIDMC from 2011 to 2019. These represent all patients who came to the emergency department (including those who were discharged from the ED as well as those subsequently admitted to the hospital/ICU). MIMIC-IV-ED provides detailed ED encounter information, filling the gap for research on urgent and acute care before ICU admission.

**Data Content:**

• ED triage data: Triage vital signs (initial measurements like heart rate, blood pressure, etc.), triage acuity level, chief complaint as recorded on arrival.

• Vital sign measurements: Ongoing vital signs recorded in the ED.

• Medication reconciliation and administration: Medications that the patient reports taking (medication history) and medications given in the ED.

• Procedures and interventions in ED (if documented separately)

• Discharge diagnoses from ED: The coded diagnoses given at the end of the ED visit (often ICD codes for the condition or reason for visit).

The ED data can be linked with the MIMIC-IV hospital/ICU data: each ED stay has identifiers (subject\_id, hadm\_id, etc.) that connect it to the same patient’s records in MIMIC-IV and even to MIMIC-CXR imaging if applicable. This enables following a patient from ER to ICU to outcomes.

**Access:**

Access to MIMIC-IV-ED is managed as part of MIMIC-IV on PhysioNet. Researchers who have access to MIMIC-IV will also have access to the ED module. The same training and data use agreement requirements apply.

**Key Characteristics & Uses:**

MIMIC-IV-ED is a valuable resource for studying emergency care at scale. It captures a broad, heterogeneous patient group (from minor injuries to life-threatening emergencies), allowing research on triage prioritization, early risk prediction, and decision-making in the ED. Potential uses include: building models to predict which ED patients will need ICU admission or hospitalization, studying workflow and throughput in the ED, analyzing treatments given in the ED and their outcomes, and identifying factors in the ED that correlate with inpatient outcomes. Because it links to in-hospital data, one can investigate, for example, how ED vital signs or interventions predict ICU mortality or how delays in the ED affect outcomes in critical care. MIMIC-IV-ED is relatively new, but it is expected to support research in operational emergency medicine, early warning scores, and AI assistance for triage. Additionally, it supports educational use for those learning about emergency medicine data analysis. In summary, MIMIC-IV-ED extends the MIMIC paradigm to the ED setting, enabling end-to-end studies from emergency admission to hospital discharge.

**Note:**

Beyond the above, the MIMIC project also encompasses other modules (for example, MIMIC-IV contains a module of deidentified free-text clinical notes “MIMIC-IV-Note” and earlier MIMIC versions had waveform data from bedside monitors). These additional data types further increase what researchers can do (e.g., analyzing clinician notes or physiological waveforms), though they often require separate handling. All MIMIC datasets are made to work together; e.g., one can link MIMIC-CXR with MIMIC-IV using patient/study identifiers, facilitating multi-modal research.