

MIMIC-IV

Alistair Johnson , Lucas Bulgarelli , Tom Pollard , Steven Horng , Leo Anthony Celi , Roger Mark 

Published: Jan. 6, 2023. Version: 2.2

Guidelines for creating datasets and models from MIMIC (April 24, 2024, 10:12 a.m.)

We recognize that there is value in creating datasets or models that are either derived from MIMIC or which augment MIMIC in some way (for example, by adding annotations). Here are some guidelines on creating these datasets and models:

- **Any derived datasets or models should be treated as containing sensitive information.** If you wish to share these resources, they should be shared on PhysioNet under the same agreement as the source data.
- **If you would like to use the MIMIC acronym in your project name,** please include the letters “Ext” (for example, MIMIC-IV-Ext-YOUR-DATASET”). Ext may either indicate “extracted” (e.g. a derived subset) or “extended” (e.g. annotations), depending on your use case.

When using this resource, please cite: ([show more options](#))

Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L. A., & Mark, R. (2023). MIMIC-IV (version 2.2). *PhysioNet*. <https://doi.org/10.13026/6mm1-ek67>.

Additionally, please cite the original publication:

[Johnson, A.E.W., Bulgarelli, L., Shen, L. et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data* 10, 1 \(2023\). <https://doi.org/10.1038/s41597-022-01899-x>](#)

Please include the standard citation for PhysioNet: ([show more options](#))

Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation* [Online]. 101 (23), pp. e215–e220.

Abstract

Retrospectively collected medical data has the opportunity to improve patient care through knowledge discovery and algorithm development. Broad reuse of medical data is desirable for the greatest public good, but data sharing must be done in a manner which protects patient privacy. The Medical Information Mart for Intensive Care (MIMIC)-III database provided critical care data for over 40,000 patients admitted to intensive care units at the Beth Israel Deaconess Medical Center (BIDMC). Importantly, MIMIC-III was deidentified, and patient identifiers were removed according to the Health Insurance Portability and Accountability Act (HIPAA) Safe Harbor provision. MIMIC-III has been integral in driving large amounts of research in clinical informatics, epidemiology, and machine learning. Here we present MIMIC-IV, an update to MIMIC-III, which incorporates contemporary data and improves on numerous aspects of MIMIC-III. MIMIC-IV adopts a modular approach to data organization, highlighting data provenance and facilitating both individual and combined use of disparate data sources. MIMIC-IV is intended to carry on the success of MIMIC-III and support a broad set of applications within healthcare.

Background

In recent years there has been a concerted move towards the adoption of digital health record systems in hospitals. In the US, nearly 96% of hospitals had a digital electronic health record system (EHR) in 2015 [1]. Retrospectively collected medical data has increasingly been used for epidemiology and predictive modeling. The latter is in part due to the effectiveness of modeling approaches on large datasets [2]. Despite these advances, access to medical data to improve patient care remains a significant challenge. While the reasons for limited sharing of medical data are multifaceted, concerns around patient privacy are highlighted as one of the most significant issues. Although patient studies have shown almost uniform agreement that deidentified medical data should be used to improve medical practice, domain experts continue to debate the optimal mechanisms of doing so. Uniquely, the MIMIC-III database adopted a permissive access scheme which allowed for broad reuse of the data [3]. This mechanism has been successful in the wide use of MIMIC-III in a variety of studies ranging from assessment of treatment efficacy in well defined cohorts to prediction of key patient outcomes such as mortality. MIMIC-IV aims to carry on the success of MIMIC-III, with a number of changes to improve usability of the data and enable more research applications.

1. *Acquisition.* Data for patients who were admitted to the BIDMC emergency department or one of the intensive care units were extracted from the respective hospital databases. A master patient list was created which contained all medical record numbers corresponding to patients admitted to an ICU or the emergency department between 2008 - 2019. All source tables were filtered to only rows related to patients in the master patient list.
2. *Preparation.* The data were reorganized to better facilitate retrospective data analysis. This included the denormalization of tables, removal of audit trails, and reorganization into fewer tables. The aim of this process is to simplify retrospective analysis of the database. Importantly, data cleaning steps were not performed, to ensure the data reflects a real-world clinical dataset.
3. *Deidentify.* Patient identifiers as stipulated by HIPAA were removed. Patient identifiers were replaced using a random cipher, resulting in deidentified integer identifiers for patients, hospitalizations, and ICU stays. Structured data were filtered using look up tables and allow lists. If necessary, a free-text deidentification algorithm was applied to remove PHI from free-text. Finally, date and times were shifted randomly into the future using an offset measured in days. A single date shift was assigned to each *subject_id*. As a result, the data for a single patient are internally consistent. For example, if the time between two measures in the database was 4 hours in the raw data, then the calculated time difference in MIMIC-IV will also be 4 hours. Conversely, distinct patients are not temporally comparable. That is, two patients admitted in 2130 were not necessarily admitted in the same year.

After these three steps were carried out, the database was exported to a character based comma delimited format.

Data Description

MIMIC-IV is grouped into two modules: *hosp*, and *icu*. The aim of these modules is to highlight their provenance.

hosp

The *hosp* module contains data derived from the hospital wide EHR. These measurements are predominantly recorded during the hospital stay, though some tables include data from outside the hospital as well (e.g. outpatient laboratory tests in *labevents*). Patient demographics (*patients*), hospitalizations (*admissions*), and intra-hospital transfers (*transfers*) are recorded in the *hosp* module.

Notably, the *patients* table provides timing information for each patient through the *anchor_year* and *anchor_year_group* columns. The *anchor_year* is a deidentified year occurring sometime between 2100 - 2200, and the *anchor_year_group* is a three year long date ranges between 2008 - 2019. These pieces of information allow researchers to infer the approximate year a patient received care. For example, if a patient's *anchor_year* is 2158, and their *anchor_year_group* is 2011 - 2013, then any hospitalizations for the patient occurring in the year 2158 actually occurred sometime between 2011 - 2013. Finally, the *anchor_age* provides the patient age in the given *anchor_year*. If the patient was over 89 in the *anchor_year*, this *anchor_age* has been set to 91 (i.e. all patients over 89 have been grouped together into a single group with value 91, regardless of what their real age was).

Date of death is available within the *dod* column of the *patients* table. Date of death is derived from hospital records and state records. If both exist, hospital records take precedence. State records were matched using a custom rule based linkage algorithm based on name, date of birth, and social security number. State and hospital records for date of death were collected two years after the last patient discharge in MIMIC-IV, which should limit the impact of reporting delays in date of death.

Dates of death occurring more than one year after hospital discharge are censored as a part of the deidentification process. As a result, the maximum time of follow up for each patient is exactly one year after their last hospital discharge. For example, if a patient's last hospital discharge occurs on 2150-01-01, then the last possible date of death for the patient is 2151-01-01. If the individual died on or before 2151-01-01, and it was captured in either state or hospital death records, then the *dod* column will contain the deidentified date of death. If the individual survived for at least one year after their last hospital discharge, then the *dod* column will have a NULL value.

Other information in the *hosp* module includes laboratory measurements (*labevents*, *d_labitems*), microbiology cultures (*microbiologyevents*, *d_micro*), provider orders (*poe*, *poe_detail*), medication administration (*emar*, *emar_detail*), medication prescription (*prescriptions*, *pharmacy*), hospital billing information (*diagnoses_icd*, *d_icd_diagnoses*, *procedures_icd*, *d_icd_procedures*, *hpcsevents*, *d_hcpcs*, *drgcodes*), online medical record data (*omr*), and service related information (*services*).

Provider information is available in the *provider* table. The *provider_id* column is a deidentified character string which uniquely represents a single care provider. As *provider_id* is used in different contexts across the module, a prefix is usually present in data tables to contextualize how the provider relates to the event. For example, the provider who admits the patient to the hospital is documented in the *admissions* table as *admit_provider_id*. All columns which have a suffix of *provider_id* may be linked to the *provider* table.

(*ingredientevents*), patient outputs (*outputevents*), procedures (*procedureevents*), information documented as a date or time (*datetimeevents*), and other charted information (*chartevents*). All events tables contain a *stay_id* column allowing identification of the associated ICU patient in *icustays*, and an *itemid* column allowing identification of the concept documented in *d_items*. Additionally, the *caregiver* table contains *caregiver_id*, a deidentified integer representing the care provider who documented data into the system. All events tables (*chartevents*, *datetimeevents*, *ingredientevents*, *inputevents*, *outputevents*, *procedureevents*) have a *caregiver_id* column which links to the *caregiver* table.

Usage Notes

The data described here are collected during routine clinical practice and reflect the idiosyncrasies of that practice. Implausible values may be present in the database as an artifact of the archival process. Researchers should follow best practice guidelines when analyzing the data.

Up to date documentation for MIMIC-IV is available on the MIMIC-IV website [4]. We have created an open source repository for the sharing of code and discussion of the database, referred to as the MIMIC Code Repository [5, 6]. The code repository provides a mechanism for shared discussion and analysis of all versions of MIMIC, including MIMIC-IV.

Release Notes

MIMIC-IV v2.2

MIMIC-IV v2.2 was released in January 2023. It added provider identifiers, imputed *hadm_id* for a number of rows in *emar*, and changed the subset of *subject_id* which are held out. Final row counts are available in the validation scripts published with the MIMIC Code Repository [6]. For clarity, after removal of the test set, the row counts are as follows:

- *patients*: 299,712 (was 315,460 in v2.0)
- *admissions*: 431,231 (was 454,324 in v2.0)
- *icustays*: 73,181 (was 76,943 in v2.0)

icu module

- *caregiver*
 - New table in v2.2. Contains one column: *caregiver_id*, a deidentified integer which uniquely represents a single caregiver or provider. These identifiers are sourced from the MetaVision ICU system. When present in a table, it indicates the user who documented the data into MetaVision. For example, the *caregiver_id* associated with a row indicating mechanical ventilation in the *procedureevents* table represents the user who documented the event, and not the provider who performed the procedure.
- *chartevents*, *datetimeevents*, *ingredientevents*, *inputevents*, *outputevents*, *procedureevents*
 - Added the *caregiver_id* column. This column is a deidentified integer representing the care provider who documented the data for the given row.

hosp module

- *provider*
 - New table in v2.2. Contains one column: *provider_id*, a deidentified string which uniquely represents a single caregiver or provider. These identifiers are sourced from the hospital wide EHR system, and used in a variety of contexts across tables in the module.
- *admissions*
 - New column: *admit_provider_id*, a deidentified string representing the provider who admitted the patient.
- *emar*
 - New column: *enter_provider_id*, a deidentified string representing the provider who entered the medication administration information into the database.
 - Fixed a bug where a subset of *emar* rows (713,117, ~2.5%) did not have an *hadm_id* even though they were associated with a given hospitalization. These rows occur outside of the administratively documented admission and discharge times for a hospitalization, but are still considered as administered during that hospitalization in the raw data.
- *labevents*, *microbiologyevents*, *poe*, *prescriptions*
 - New column: *order_provider_id*, a deidentified string representing the provider who ordered the corresponding event (e.g. the lab test in the case of *labevents*, or the medication in the case of *prescriptions*).

- patients - Removed 15,748 `subject_id` from the table
- admissions - Removed 23,093 `hadm_id` from the table.
- icustays - Removed 3,762 `stay_id` from the table.
- Other tables will have rows removed to reflect the removal of the aforementioned `subject_id`, `hadm_id`, and `stay_id`. Final row counts are available in the validation scripts published with the MIMIC Code Repository [6].

MIMIC-IV v2.0

MIMIC-IV v2.0 was released on June 12, 2022. It focused on expanding the data elements available for patients within MIMIC-IV v1.0. Additional data available includes out-of-hospital date of death, information from the online medical record system (which includes height and weight), and more detail for continuous infusions in the ICU.

Major changes

- The `core` module has been removed to simplify the schema. The `admissions`, `patients`, and `transfers` tables are now in the `hosp` module.
- Neonates have been removed from the dataset. Neonatal data will be released in a separate project with data from the neonatal intensive care unit.

icu module

- `icustays`
 - Around 700 stays (~1%) have changed due to the changes in the `patients` table.
- `chartevents`, `d_items`
 - The problem list from MetaVision has been added. All problems are documented with the same `itemid` now present in `d_items`: 220001. There are just over 1,000 unique problems. Most documented problems are related to the care plan for the patient and documented during nurse shift changes (either 7am or 7pm). Less frequently, the ongoing issues are documented here.
- `ingredientevents`
 - This is a new table associated with `inpuvents`. Each intravenous administration tracked in `inpuvents` is associated with a set of ingredients. These ingredients include water content, caloric information, and so on. The goal of the `inpuvents` table is to support nutrition research and to provide a mechanism for estimating fluid input via summing all instances of the water ingredient. These ingredients have been separated from the `inpuvents` table to simplify analysis and reduce the size of `inpuvents`.
- `inpuvents`
 - Removed a single column which contained only null values: `cancelreason`.
- `procedureevents`
 - Removed columns which contained only null values: `totalamount`, `totalamountuom`, `cancelreason`, `comments_editedby`, `comments_canceledby`, `comments_date`, `secondaryordercategoryname`.

hosp module

- `admissions`
 - Fixed an issue where hospitalizations were missing `edregtime` and `edouttime` when the patient was admitted via the ED (reported in [#1247](#), thanks [@MEladawi](#)).
- `patients`
 - `dod` is now populated with out-of-hospital mortality from state death records. For patients admitted to the ICU, this change has increased capture of date of death from 8,223 records to 23,844 (i.e. we now have out-of-hospital mortality for an additional 15,621 ICU patients).
 - The mechanism for determining patients included in MIMIC was changed. For the most part this has resulted in an improvement, particularly regarding the logic for merging patients who had distinct medical record numbers. As a result of this change, most tables have had a change in the data content. Approximately 1% of stays were affected.
- `transfers`
 - Fixed a bug where the `outtime` for ED stays with no associated `hadm_id` (i.e. an ED stay where the individual was not admitted to the hospital) was incorrect. This resulted in all `transfers` rows with a NULL `hadm_id` having an apparent stay of minutes or less. The `outtime` column has now been corrected.
- `labevents`, `d_labitems`
 - The `itemid` for `d_labitems` has been changed for 43 items. These are extremely infrequently documented and each `itemid` has fewer than 100 observations in `labevents`. The exact `itemid` are provided in the changelog file CHANGELOG.txt.
 - Errors were found in the current values of `loinc_code` (reported in [#938](#), thanks [@Mauvila](#)). In order to enable collaborative improvement, the `loinc_code` column has been removed, and will now be collaboratively developed in the [MIMIC Code](#)

- New organisms, tests, specimens, and antibiotics have been added.
- *omr*
 - A new table has been added: *omr*. The source of this data is the Online Medical Record, and it contains miscellaneous information useful for understanding an individual's health. As of v2.0, the *omr* table has the following information: blood pressure, height, weight, body mass index, and Estimated Glomerular Filtration Rate (eGFR). These values are available from both inpatient and outpatient visits, and in many cases a "baseline" value from before a patient's hospitalization is available.
- *prescriptions*
 - The *formulary_drug_cd* table has been added back (was previously in MIMIC-III). This column has the same set of values as the *product_code* column of *emar_detail*.

MIMIC-IV v1.0

MIMIC-IV v1.0 was released March 16, 2021.

core

- *admissions*
 - A number (~1000, <1%) of erroneous *hadm_id* have been removed.
- *patients*
 - *dod* is now populated using the patient's *deathtime* from their latest hospitalization (reported in [#71](#), thanks [@jinjinzhou](#)).
 - At the moment, out-of-hospital mortality is **not** captured by 'dod'.
- *transfers*
 - Removed erroneous transfers included in the previous version.
 - *transfer_id* has been regenerated. *transfer_id* in MIMIC-IV v1.0 are **not compatible** with *transfer_id* from v0.4. We do not intend to change *transfer_id* when updating MIMIC-IV, but had to update it due to an error in its generation.
 - All *hadm_id* in transfers are also present in *admissions* and vice-versa (reported in [#84](#), thanks [@kokoko12305](#)).

icu

- *icustays*
 - ICU stays were inappropriately assigned in the previous version due to an error in the preprocessing code. Previously, non-ICU ward transfers were included in the ICU stays, and certain ward stays were not treated as ICU stays (reported in [#67](#), thanks [@JHLiu7](#) and [@stefanhgm](#)). The assignment of *stay_id* has been regenerated.
 - The mapping between hospital transfers and ICU stays has been updated.
 - *stay_id* in MIMIC-IV v1.0 are **not compatible** with *stay_id* from v0.4. We do not intend to change *stay_id* when updating MIMIC-IV, but had to update it due to the error identified above.
- The change in *icustays* has re-assigned values to new *stay_id*, as a result all tables have had their content changed (due to a change in *stay_id*), but the structure is unchanged.

hosp

- *hpcsevents*
 - Data has been added for a number of previously excluded hospitalizations.
 - The table now has a *chartdate* column, containing the date associated with the code. Every row is associated with a date.
- *drgcodes*
 - Data has been added for a number of previously excluded hospitalizations.
 - Duplicate DRG codes have been removed from the table.
 - Descriptions have been updated using the latest dictionaries made available from mass.gov and HCUP.
- *diagnoses_icd*, *d_icd_diagnoses*
 - Data has been added for a number of previously excluded hospitalizations (reported in [#27](#), thanks [@yugangjia](#)).
 - The *icd_code* column is now trimmed and stored as a VARCHAR, i.e. codes no longer contain trailing whitespaces ('850 ' -> '850').
 - Missing ICD codes have been added to the dictionary. All ICD codes in the *diagnoses_icd* table have an associated reference in *d_icd_diagnoses*.
- *labevents*
 - The *comments* field has been updated, fixing a bug where comments longer than 4096 characters were truncated. Due to the deidentification, it's unlikely users will see much difference, as these comments will appear as ____.
- *procedures_icd*
 - Data has been added to *procedures_icd* for a number of previously excluded hospitalizations.
 - The table now has a *chartdate* column, containing the date associated with each billed procedure.

v0.4

- *d_micro*
 - This table has been removed
- *microbiologyevents*
 - Added the column *spec_type_desc*, *test_name*, *org_name*, and *ab_name* columns
 - These columns contain the textual name of the organism/antibiotic/test/specimen
 - Added the *comments* column: this column contains information about the test, and in some cases (e.g. viral load tests), contains the result

v0.3

- Fixed a bug in the timing between *hosp* and *icu*

v0.2

- Updated demographics in the patient table
 - *anchor_year* -> *anchor_year_group*
 - *anchor_year_shifted* -> *anchor_year*
 - See the patients table in the MIMIC online documentation for detail on these columns
- *transfers*
 - Deleted the *los* column
- *emar*
 - *mar_id* -> *emar_id*
 - *emar_id* is now a composite of *subject_id* and *emar_seq*, and has form "subject_id-emar_seq"
 - *emar_seq* column - a monotonically increasing integer starting with the first eMAR administration
 - Added *poe_id* and *pharmacy_id* columns for linking to those tables
- *emar_detail*
 - *mar_id* -> *emar_id* (changed as above)
 - Deleted the *mar_detail_id* column
- *hpcsevents*
 - *ticket_id_seq* -> *seq_num*
- *labevents*
 - Many previously NULL values are now populated - these were removed originally due to deidentification
 - Added the *comments* column. This contains deidentified free-text comments with labs. PHI is replaced with three underscores (___). If an entire comment is ___, then the entire comment was scrubbed.
- *microbiologyevents*
 - *stay_id* column removed
 - *spec_id* -> *micro_specimen_id*
- Added the *poe* and *poe_detail* tables
 - Documentation of provider orders for various treatments and other aspects of patient management
- Added the *prescriptions* table
 - Documentation of medicine prescriptions via the provider order interface
- Added the *pharmacy* table
 - Detailed information regarding prescriptions provided by the pharmacy including formulary dose, route, frequency, dose, and so on.
- *inpuvents*
 - Fixed an error in the calculation of the *amount* column
- *icustays*
 - Re-derived *stay_id* - the new *stay_id* are distinct from the previous version.

Ethics

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, who granted a waiver of informed consent and approved the data sharing initiative.

Conflicts of Interest

None to declare.

References

1. Henry, J., Pylypchuk, Y., Searcy T. & Patel V. (May 2016). Adoption of Electronic Health Record Systems among U.S. Non-Federal Acute Care Hospitals: 2008-2015. ONC Data Brief, no.35. Office of the National Coordinator for Health Information Technology: Washington DC. +
2. Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. IEEE Intelligent Systems, 24(2), 8-12.
3. Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.H., Feng, M., Ghassemi, M., ... & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. Scientific data, 3(1), 1-9.
4. MIMIC Online Documentation. <https://mimic.mit.edu>
5. Johnson AE, Stone DJ, Celi LA, Pollard TJ. The MIMIC Code Repository: enabling reproducibility in critical care research. Journal of the American Medical Informatics Association. 2018 Jan;25(1):32-9.
6. Alistair Johnson, Tom Pollard, Jim Blundell, Brian Gow, erinhong, Nicolas Paris, et al. MIT-LCP/mimic-code: MIMIC Code v2.1.1. Zenodo; 2021. <https://doi.org/10.5281/zenodo.821871>

Contents

Share



Access

Access Policy:

Only credentialed users who sign the DUA can access the files.

License (for files):

[PhysioNet Credentialed Health Data License 1.5.0](#)

Data Use Agreement:

[PhysioNet Credentialed Health Data Use Agreement 1.5.0](#)

Required training:

[CITI Data or Specimens Only Research](#)

Discovery

DOI (version 2.2):

<https://doi.org/10.13026/6mm1-ek67>

DOI (latest version):

<https://doi.org/10.13026/07hj-2a80>

Topics:

[mimic](#) [critical care](#) [machine learning](#) [intensive care unit](#)

Corresponding Author

Alistair Johnson
Massachusetts Institute of Technology.
alistairewj@gmail.com

Versions

[0.3](#) - Aug. 13, 2020

[0.4](#) - Aug. 13, 2020

[1.0](#) - March 16, 2021

[2.0](#) - June 12, 2022

[2.1](#) - Nov. 16, 2022

[2.2](#) - Jan. 6, 2023

Files

This is a restricted-access resource. To access the files, you must fulfill all of the following requirements:

- be a [credentialed user](#)
- complete required training:
 - [CITI Data or Specimens Only Research](#)
You may submit your training [here](#).
- [sign the data use agreement](#) for the project

PhysioNet is a repository of freely-available medical research data, managed by the MIT Laboratory for Computational Physiology.

Supported by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) under NIH grant number R01EB030362.

For more accessibility options, see the [MIT Accessibility Page](#).

[Back to top](#)