



MIMIC-IV

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

Published: Aug. 13, 2020. Version: 0.3 [<View latest version>](#)

This is **not** the latest version. Click [here](#) for the latest version.



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.
- Please select the relevant "Parent Projects" in the Discovery tab of the submission portal when preparing your project for submission.

When using this resource, please cite:

“Cite”

Copy BibTeX

Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L. A., & Mark, R. (2020). MIMIC-IV (version 0.3). *PhysioNet*. RRID:SCR_007345. <https://doi.org/10.13026/a2mm-bn44>

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. RRID:SCR_007345.

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 (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.

Methods

MIMIC-IV is sourced from two in-hospital database systems: a custom hospital wide EHR and an ICU specific clinical information system. The creation of MIMIC-IV was carried out in three steps:

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. Dates were shifted randomly into the future, but consistently for an individual patient, retaining the temporal dynamics of their stay. 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.

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 three modules: *core*, *hosp*, and *icu*. The aim of these modules is to highlight their intended use and provenance. Up to date documentation for MIMIC-IV is available on the [MIMIC-IV website](#).

core

The *core* module stores patient tracking information necessary for any data analysis using MIMIC-IV. The *core* module contains three tables: *patients*, *admissions*, and *transfers*. These tables provide demographics for the patient, a record for each hospitalization, and a record for each ward stay within a hospitalization.

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).

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*). Information 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*, *hcpcsevents*, *d_hcpcs*, *drgcodes*), and service related information (*services*).

icu

The *icu* module contains data sourced from the clinical information system at the BIDMC: MetaVision (iMDSoft). MetaVision tables were denormalized to create a star schema where the *icustays* and *d_items* tables link to a set of data tables all suffixed with "events". Data documented in the *icu* module includes intravenous and fluid inputs (*inpuvents*), 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*.

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.

We have created an open source repository for the sharing of code and discussion of the database, referred to as the [MIMIC-IV Code Repository](#). The code repository provides a mechanism for shared discussion and analysis of MIMIC-IV.

Release Notes

Current version

The current version of MIMIC-IV is v0.3. As the database is still in development, we may change the schema in future versions. Our aim is to eventually release MIMIC-IV v1.0, at which point schema changes will respect the semantic versioning.

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](#) 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
 - *hcpcsevents*
 - *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.
 - *inputevents*
 - 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.
-

Acknowledgements

We would like to thank the Beth Israel Deaconess Medical Center for their continued support of the MIMIC project. In particular we would like to thank Carolyn Conti, Alvin Gayles, Larry Markson, Ayad Shammout, Lu Shen, and Manu Tandon for their assistance. This work was supported by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) under NIH grant number R01EB030362.

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.

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 0.3):
<https://doi.org/10.13026/a2mm-bn44>

DOI (latest version):
<https://doi.org/10.13026/07hj-2a80>

Topics:

critical care

intensive care unit

mimic

machine learning

Project Website:
<https://mimic-iv.mit.edu>

Project Views

0

Current Version

51348

All Versions

Project Views by Unique Registered Users

View Details

Corresponding Author

You must be logged in to view the contact information.

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
3.0	July 23, 2024
3.1	Oct. 11, 2024

Files

The files for this version of the project (0.3) are no longer available. The latest version of this project is [3.1](#).



Maintained by the MIT Laboratory for Computational Physiology

Supported by the National Institute of Biomedical Imaging and Bioengineering (NIBIB), National Heart Lung and Blood Institute (NHLBI), and NIH Office of the Director under NIH grant numbers U24EB037545 and R01EB030362

Navigation

- [Discover Data](#)
- [Share Data](#)
- [About](#)
- [News](#)

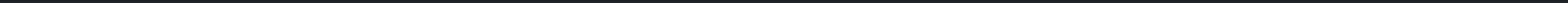
Explore

[Data](#)

[Software](#)

[Challenges](#)

[Tutorials](#)



[Accessibility](#)