MIMIC-IV-ECG: Diagnostic Electrocardiogram Matched Subset

Brian Gow , Tom Pollard , Larry A Nathanson , Alistair Johnson , Benjamin Moody , Chrystinne Fernandes , Nathaniel Greenbaum , Jonathan W Waks , Parastou Eslami , Tanner Carbonati , Ashish Chaudhari , Elizabeth Herbst , Dana Moukheiber , Seth Berkowitz , Roger Mark , Steven Horng

Published: Sept. 15, 2023. Version: 1.0

MIMIC-IV-ECG module released *(Sept. 15, 2023, 4:14 p.m.)*

The MIMIC-IV-ECG module is now available. This module contains approximately 800,000 diagnostic electrocardiograms across nearly 160,000 unique patients. The vast majority of ECGs for patients who appear in the MIMIC-IV Clinical Database are included. The patients in MIMIC-IV-ECG have been matched against the MIMIC-IV Clinical Database, making it possible to link to information across the MIMIC-IV modules. When a cardiologist report is available for a given ECG, we provide information for linking to it.

When using this resource, please cite: (show more options)
Gow, B., Pollard, T., Nathanson, L. A., Johnson, A., Moody, B., Fernandes, C., Greenbaum, N., Waks, J. W., Eslami, P., Carbonati, T., Chaudhari, A., Herbst, E., Moukheiber, D., Berkowitz, S., Mark, R., & Horng, S. (2023). MIMIC-IV-ECG: Diagnostic Electrocardiogram Matched Subset (version 1.0). *PhysioNet*. https://doi.org/10.13026/4nqg-sb35.

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.

Contents

Parent Projects

MIMIC-IV-ECG: Diagnostic Electrocardiogram Matched Subset was derived from:

- MIMIC-IV v2.2

Please cite them when using this project.

Share

Access

Discovery

DOI (version 1.0):
https://doi.org/10.13026/4nqg-sb35

DOI (latest version):
https://doi.org/10.13026/b95v-ff39

Corresponding Author

You must be logged in to view the contact information.

Versions

0.1 - Dec. 23, 2022

0.2 - Feb. 8, 2023

0.3 - July 21, 2023

1.0 - Sept. 15, 2023

Abstract

The MIMIC-IV-ECG module contains approximately 800,000 diagnostic electrocardiograms across nearly 160,000 unique patients. These diagnostic ECGs use 12 leads and are 10 seconds in length. They are sampled at 500 Hz. This subset contains all of the ECGs for patients who appear in the MIMIC-IV Clinical Database. When a cardiologist report is available for a given ECG, we provide the needed information to link the waveform to the report. The patients in MIMIC-IV-ECG have been matched against the MIMIC-IV Clinical Database, making it possible to link to information across the MIMIC-IV modules.

Background

An Electrocardiogram or ECG / EKG measures the electrical activity associated with the heart [1]. Diagnostic ECGs are a standard part of a patients care [2]. The standard ECG leads are denoted as lead I, II, III, aVF, aVR, aVL, V1, V2, V3, V4, V5, V6. They are routinely obtained when admitted to the Emergency Department or to a hospital floor. ECGs will typically be repeated for patients who exhibit cardiac symptoms such as chest pain or abnormal rhythms. Daily ECGs may be obtained following acute cardiovascular events such as myocardial infarction. Patients in the Intensive Care Unit (ICU) are continuously monitored to detect rhythm abnormalities, but full ECGs are needed to evaluate evidence of cardiac ischemia or infarction. However, diagnostic ECGs typically only comprise a small part of understanding the overall condition of a subject at the hospital. To fully understand how to best treat a given patient, a broader set of data is collected which may include: patient demographics, diagnosis, medications, lab tests, and additional information. This broader set of clinical information is shared as part of the MIMIC-IV Clinical Database [3]. The MIMIC-IV-ECG Matched Subset contains the vast majority of diagnostic ECGs collected between 2008 - 2019 for subjects in MIMIC-IV.

Methods

As part of routine care, diagnostic ECGs are collected across Beth Israel Deaconess Medical Center (BIDMC). Three types of information associated with an ECG are presented here. The electrocardiogram waveforms themselves, the machine measurements (ex: average RR interval as calculated by the machine), and the cardiologist reports. Identifiers connected to the ECGs allow this information to be connected back to the patients overall electronic health record. All of the information is de-identified to satisfy the US Health Insurance Portability and Accountability Act of 1996 (HIPAA) Safe Harbor requirements.

Electronic Health Record

Patients from the MIMIC-IV Clinical Database who had ECGs collected between 2008 - 2019 are included as part of MIMIC-IV-ECG. The diagnostic ECGs are collected on machines from various manufacturers including Burdick/Spacelabs, Philips, and General Electric. When the ECG is collected, the machine is populated with the patient's demographics and their medical record number (MRN).

As part of de-identification the raw identifiers are shifted. The patient's MRN was used to match a given 12-lead ECG record to the corresponding subject ID in the MIMIC-IV Clinical Database. As another part of the de-identification, the date-time information was shifted to obscure the actual date and time. Relative date-time information for a given subject is preserved though. The shifted date-times were matched against date-times in the subject's MIMIC-IV Clinical Database records. A unique `study_id` was generated for each record.

Electrocardiogram Waveforms

If a patient appears in the MIMIC-IV Clinical Database, all of their available ECG waveforms were pulled. This includes ECGs from the BIDMC emergency department, hospital (including the ICU), and outpatient care centers. We converted the ECGs from the manufacturers format to the open WFDB format 16 [4] with each WFDB record comprised of a header (.hea) file and a signal (.dat) file. The files were then transferred from BIDMC to MIT for additional processing.

We scrubbed the WFDB header files for PHI such that only the signal information, subject ID, and shifted date-time are provided. Timestamps for events in the MIMIC-IV Clinical Database, such as drug administration, are aligned with the timestamps in MIMIC-IV-ECG. However, some of the diagnostic ECGs provided here were collected outside of ED or ICU visits at the hospital. Since the MIMIC-IV Clinical Database is comprised solely of ED and ICU data, the ECG timestamp can occur before or after a visit from the clinical database.

Machine Measurements

The ECG machine generates summary reports and summary measures (ex: RR interval, QRS onset and end, etc.) for each diagnostic ECG. We collectively refer to these as machine measurements. The machine output is parsed and any PHI is removed. In particular, the MRN is shifted to `subject_id`, the de-identified `study_id` is assigned in a manner consistent with the ECG waveform files, and the raw Cart ID is randomly shifted to create a de-identified `cart_id`. There was no PHI in the report lines.

The global machine measures are provided in this release. These global measures are calculated across all 12 leads. Machine measurements for individual leads may be released in a future version of this project.

Cardiologist Reports

Most ECG waveforms get read by a cardiologist and an associated report is generated from the reading. We provide information for linking a waveform with its associated report where available.

The de-identified free-text notes from these ECG reports will be made available as part of the MIMIC-IV-Note module [5] at a later time. These ECG reports are de-identified using a rule-based approach [6, 7, 8], similar to that used for other MIMIC reports.

Data Description

Electrocardiogram Waveforms

Approximately 800,000 ten-second-long 12 lead diagnostic ECGs across nearly 160,000 unique subjects are provided in the MIMIC-IV-ECG module. Around 5% of the available diagnostic ECGs were withheld from this release so they can be used as a hidden test set in workshops and challenges. The ECGs are sampled at 500 Hz. The patients in this module have been matched with the MIMIC-IV Clinical Database. Many of the provided diagnostic ECGs overlap with a MIMIC-IV hospital or emergency department stay but a number of them do not overlap. Approximately 55% of the ECGs overlap with a hospital admission and 25% overlap with an emergency department visit.

The ECGs are grouped into subdirectories based on `subject_id`. Each DICOM record path follows the pattern: `files/pNNNN/pXXXXXXXX/sZZZZZZZZ/ZZZZZZZZ`, where:

- `NNNN` is the first four characters of the `subject_id`,
- `XXXXXXXX` is the `subject_id`,
- `ZZZZZZZZ` is the `study_id`

An example of the file structure is as follows:

```
files
├── p1000
|   └── p10001725
|       └── s41420867
|           ├── 41420867.dat
|           └── 41420867.hea
└── p1002
    └── p10023771
        ├── s42745010
        |   ├── 42745010.dat
        |   └── 42745010.hea
        ├── s46989724
        |   ├── 46989724.dat
        |   └── 46989724.hea
        └── s42460255
            ├── 42460255.dat
            └── 42460255.hea
```

Above we find two subjects `p10001725` (under the `p1000` group level directory) and `p10023771` (under the `p1002` group level directory). For subject `p10001725` we find one study: `s41420867`. For `p10023771` we find three studies: `s42745010`, `s46989724`, `s42460255`. The study identifiers are completely random, and their order has no implications for the chronological order of the actual studies. Each study has a like named .hea and .dat file, comprising the WFDB record.

The `record_list.csv` file contains the file name and path for each WFDB record. It also provides the corresponding subject ID and study ID. The subject ID can be used to link a subject from MIMIC-IV-ECG to the other modules in the MIMIC-IV Clinical Database.

Machine Measurements

Machine measurements for each ECG waveform are provided in the `machine_measurements.csv` file. A data dictionary provides a description for each of the columns in `machine_measurements_data_dictionary.csv`. The machine measurements table provides the machine generated reports in columns `report_0..report_17`. The report lines are provided as generated by the machine. In some cases there will be a column with no text in between columns with text (ex: report_0: <text_a>, report_1: empty, report_2: <text_b>). In addition to the summary measurements (`rr_interval, qrs_onset, qrs_end`, etc.) columns for the machine's `bandwidth` and filter settings (`filtering`) are provided. A `cart_id` is provided which can be used to track which machine was used for a given ECG. Finally, the `subject_id`, `study_id`, and `ecg_time` are provided, consistent with the ECG waveform files themselves.

Cardiologist Reports

A little more than 600,000 cardiologist reports are available for the ~800,000 diagnostic ECGs. Not all diagnostic ECGs get read by a cardiologist. This is the primary reason that there are fewer reports than waveforms.

The `waveform_note_links.csv` table provides a `note_id` for the associated ECG waveform. This `note_id` can be used to link between a waveform and the free-text note in the MIMIC-IV-Note module. Each `note_id` is composed of the subject ID, the abbreviation for the domain (EK) that the report comes from, and a sequential integer. The sequential integer is also listed in its own column, `note_seq`, and can be used to decipher the order in which ECGs were collected for a given subject across all of their visits. This table also contains the subject ID, study ID, and waveform path.

BigQuery

The information from the `record_list.csv`, `machine_measurements.csv`, and `waveform_note_links.csv` tables are available on BigQuery [9].

Usage Notes

This module provides MIMIC-IV users an additional, potentially important piece of information for their research using MIMIC.

There are some limitations with this dataset. The date and time for each ECG were recorded by the machine's internal clock, which in most cases was not synchronized with any external time source. As a result, the ECG time stamps could be significantly out of sync with the corresponding time stamps in the MIMIC-IV Clinical Database, MIMIC-IV Waveform Database, or other modules in MIMIC-IV. An additional limitation, as noted above, is that some of the ECGs provided here were collected outside of the ED and ICU. This means that the timestamps for those ECGs won't overlap with data from the MIMIC-IV Clinical Database.

The signals can be viewed in Lightwave by clicking the Visualize waveforms links in the Files section below. Additionally, the signals can be read by using the WFDB toolboxes provided on PhysioNet: WFDB (in C) [10], WFDB-Matlab [11], and WFDB-Python [12]. Here is a basic script for reading a downloaded record from this project and plotting it by using the WFDB-Python toolbox:

```
import wfdb
rec_path = '/files/p1000/p10001725/s41420867/41420867'
rd_record = wfdb.rdrecord(rec_path)
wfdb.plot_wfdb(record=rd_record, figsize=(24,18), title='Study 41420867
example', ecg_grids='all')
```

where `rec_path` is the path to the name of the .hea and .dat files for the record you'd like to plot.

Here we provide an example of how subject `p10023771` from MIMIC-IV-ECG can be linked to their admission information in the MIMIC-IV Clinical Database. Executing this from BigQuery:

```
SELECT * FROM `physionet-data.mimiciv_hosp.admissions` WHERE subject_id=10023771
```

we see that the patient only has one admission to the hospital with an `admittime = 2113-08-25T07:15:00` and a `dischtime = 2113-08-30T14:15:00`. We also need to check to see if they were seen in the emergency department and not admitted to the hospital:

```
SELECT * FROM `physionet-data.mimiciv_ed.edstays` WHERE subject_id = 10023771
```

We observe that they did not have a stay in the emergency department.

Next, we get the timestamps from the diagnostic ECGs by checking the `base_date` and `base_time` variables. These are the variables used in the WFDB format for storing date and time. They correspond with the timestamps for the diagnostic ECGs that are provided in the summary tables. We then save the result to a csv file:

```
from pathlib import Path
import pandas as pd

import wfdb

# get the path to all the study .hea files for p10023771
paths = list(Path("p10023771/.").rglob("*.hea"))

# get date and time for each study
date_times = {'study':[],'date':[],'time':[]} # use a dictionary to store the
date and time for each study
for file in paths:
    study = file.stem
    metadata = wfdb.rdheader(f'{file.parent}/{file.stem}')
    date_times['study'].append(study)
    date_times['date'].append(metadata.base_date)
    date_times['time'].append(metadata.base_time)

df_date_times = pd.DataFrame(data=date_times)
df_date_times.to_csv('p10023771_date_times.csv', index=False)
```

We observe the following for the 3 diagnostic ECGs for `p10023771`:

study	datetime
42745010	2110-07-23T08:43
46989724	2113-08-19T07:18
42460255	2113-08-25T13:58

where the date is given before the T as YYYY-MM-DD and the time is given after the T as HH:MM. Comparing this to the subjects admission in the MIMIC-IV Clinical Database:

admittime	dischtime
2113-08-25T07:15	2113-08-30T14:15

we observe that `s42745010` and `s46989724` occurred prior to their only hospital admission while `s42460255` occurred during their hospital admission.

We can also check the available cardiologist reports for this subject by running this command in BigQuery:

```
SELECT * FROM `lcp-consortium.mimic_ecg.reports` WHERE subject_id = 10023771
```

We find that there are cardiologist reports available for `s46989724` and `s42460255` but not `s42745010`. Please note that only members who are part of our consortium can access the cardiologist reports / notes from `lcp-consortium` on BigQuery.

Release Notes

MIMIC-IV-ECG v1.0

This release removes the sensitive information (i.e. free-text note) from the cardiologist reports. We now simply provide information for linking between the waveforms in this module and their associated free-text note in MIMIC-IV-Note module. Since that sensitive information has been removed, the project access has been changed to open instead of requiring credentialling.

Ethics

The project was approved by the Institutional Review Boards of Beth Israel Deaconess Medical Center (Boston, MA) and the Massachusetts Institute of Technology (Cambridge, MA). Requirement for individual patient consent was waived because the project did not impact clinical care and all protected health information was deidentified.

Acknowledgements

SH, RM, BG, DM, and TP are funded by the Massachusetts Life Sciences Center, Nov. 30, 2020. NG is supported by National Institutes of Health National Library of Medicine Biomedical Informatics and Data Science Research Training Program under grant number T15LM007092-30. BG, TP, AJ, BM, CF, DM, and RM are supported by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) under NIH grant number R01EB030362.

Conflicts of Interest

The author(s) have no conflicts of interest to declare.

References

1. Geselowitz DB. On the theory of the electrocardiogram. Proceedings of the IEEE. 1989 Jun;77(6):857-76.
2. Harris PR. The Normal electrocardiogram: resting 12-Lead and electrocardiogram monitoring in the hospital. Critical Care Nursing Clinics. 2016 Sep 1;28(3):281-96.
3. Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L. A., & Mark, R. (2021). MIMIC-IV (version 1.0). PhysioNet. https://doi.org/10.13026/s6n6-xd98.
4. Documentation for the Waveform Database (WFDB) file format. https://wfdb.io/ [Accessed 21 June 2022]
5. Johnson, A., Pollard, T., Horng, S., Celi, L. A., & Mark, R. (2023). MIMIC-IV-Note: Deidentified free-text clinical notes (version 2.2). PhysioNet. https://doi.org/10.13026/1n74-ne17.
6. Margaret Douglass, Computer-assisted de-identification of free-text nursing notes. Master's Thesis, 2005. MIT.
7. Neamatullah, I., Douglass, M.M., Lehman, L.H., Reisner, A., Villarroel, M., Long, W.J., Szolovits, P., Moody, G.B., Mark, R.G., Clifford, G.D. (2007). De-Identification Software Package (version 1.1). PhysioNet. doi:10.13026/C20M3F
8. Neamatullah I, Douglass MM, Lehman LH, Reisner A, Villarroel M, Long WJ, Szolovits P, Moody GB, Mark RG, Clifford GD. Automated de-identification of free-text medical records. BMC medical informatics and decision making. 2008 Dec;8(1):1-7. doi:10.1186/1472-6947-8-32
9. Documentation about using the Medical Information Mart for Intensive Care (MIMIC) Database with Google BigQuery. https://mimic.mit.edu/docs/gettingstarted/cloud/ [Accessed 21 June 2022]
10. Documentation for the Waveform Database (WFDB) toolbox in C. https://physionet.org/content/wfdb/10.7.0/ [Accessed 21 June 2022]
11. Documentation for the Waveform Database (WFDB) toolbox for Matlab. https://physionet.org/content/wfdb-matlab/0.10.0/ [Accessed 21 June 2022]
12. Documentation for the Waveform Database (WFDB) toolbox for Python. https://physionet.org/content/wfdb-python/3.4.1/ [Accessed 21 June 2022]

Files

Total uncompressed size: 90.4 GB.

Access the files

- Download the ZIP file (33.8 GB)
- Download the files using your terminal: `wget -r -N -c -np https://physionet.org/files/mimic-iv-ecg/1.0/`

Visualize waveforms

Folder Navigation: <base>

Name	Size	Modified
files		
LICENSE.txt	25.2 KB	2023-09-15
RECORDS	12.7 KB	2023-08-31
SHA256SUMS.txt	167.9 MB	2023-09-15
machine_measurements.csv	174.2 MB	2023-08-31
machine_measurements_data_dictionary.csv	1.0 KB	2023-08-31
record_list.csv	67.1 MB	2023-08-31
waveform_note_links.csv	56.2 MB	2023-08-31