scientific data

OPEN

DATA DESCRIPTOR

MC-MED, multimodal clinical monitoring in the emergency department

Aman Kansal[1], Emma Chen[2,3], Boyang Tom Jin[4], Pranav Rajpurkar[2] & David A. Kim[4] ✉

Emergency Department (ED) patients often present with undiagnosed complaints, and can exhibit rapidly evolving physiology. Therefore, data from continuous physiologic monitoring, in addition to the electronic health record, is essential to understand the acute course of illness and responses to interventions. The complexity of ED care and the large amount of unstructured multimodal data it produces have limited the accessibility of detailed ED data for research. We release Multimodal Clinical Monitoring in the Emergency Department (MC-MED), a comprehensive, multimodal, and de-identified clinical and physiological dataset. MC-MED includes 118,385 adult ED visits to an academic medical center from 2020 to 2022. Data include continuously monitored vital signs, physiologic waveforms (electrocardiogram, photoplethysmogram, respiration), patient demographics, medical histories, orders, medication administrations, laboratory and imaging results, and visit outcomes. MC-MED is the first dataset to combine detailed physiologic monitoring with clinical events and outcomes for a large, diverse ED population.

Background & Summary

Emergency Departments (EDs) play a critical role in evaluating and treating patients with a wide range of medical conditions, and ED care has major implications for patient outcomes, healthcare costs, and downstream inpatient and ambulatory care[1,2]. ED patients often present with undifferentiated complaints, and the nature and severity of their condition may become clear only over the course of the visit, which can include rapidly changing physiology. Thus, ED patients undergo continuous monitoring while receiving a large variety of diagnostic and therapeutic interventions, resulting in a large volume of heterogeneous, time-varying data. These data include triage reports, free-text notes, continuously monitored vital signs and physiologic waveforms such as electrocardiogram (ECG) and photoplethysmography (PPG), medication administration logs, laboratory and imaging results, diagnoses, disposition decisions, and subsequent encounters. While structured data elements are typically stored in the electronic health record (EHR), other types of data, such as the high-resolution time series produced by bedside monitoring, are seldom integrated with clinical data, and are often discarded due to their size[3]. Moreover, noise, missingness, and the ubiquity of protected health information can make these data challenging to navigate, and limit their availability for research[4].

Few comprehensive ED datasets exist for general use. Currently, the only publicly available ED-specific dataset is MIMIC-IV-ED[5], a module of MIMIC-IV[6]. MIMIC-IV contains data from the ED and intensive care unit (ICU) of Beth Israel Deaconess Medical Center, from 2008 to 2019, including 73 K ICU admissions and 358 K ED visits. While ICU data include continuous vital signs and physiologic waveforms, ED visits include only infrequent vital sign measurements. Other ICU-only datasets include MIMIC-III[7], eICU[8,9], HiRID[10,11], and AmsterdamUMCdb[12,13]. These contain structured EHR data, and vital signs recorded at various frequencies, with 1-minute intervals being most common. None contain physiologic waveforms (Table 1).

We present Multimodal Clinical Monitoring in the Emergency Department (MC-MED)[14,15], a first-of-its-kind dataset containing multimodal clinical and physiological data from 118,385 adult ED visits to monitored beds of the Stanford Health Care ED between September 2020 and September 2022. The dataset includes: patient demographics, medical histories, and home medications; continuously monitored vital

[1]Department of Computer Science, Stanford University, Stanford, USA. [2]Department of Biomedical Informatics, Harvard Medical School, Boston, USA. [3]Harvard John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, USA. [4]Department of Emergency Medicine, Stanford University, Stanford, USA. ✉e-mail: davidak@stanford.edu

Dataset	Source	Time range	Patients	Visits	ED data	ICU data	Structured EHR data	Free text data	Wave-forms	Typical vital sign frequency
MIMIC-III[7]	Beth Israel Deaconess Medical Center ICU	2001–2012	38,597 (adults)	53,423 (adults)		✓	✓	✓		1 hour (variable)
MIMIC-IV (v2.2)[5,6]	Beth Israel Deaconess Medical Center ICU	2008–2019	299,712	73,181 ICU stays		✓	✓	✓	ECG, PPG, BP	1 hour (variable), 1 minute or less where waveforms available
	Beth Israel Deaconess Medical Center ED			358,050 ED visits	✓		✓	✓		1–4 hours
eICU (v2.0)[8,9]	335 ICUs at 208 US hospitals	2014–2015	139,367	200,859		✓	✓	◯ (Semi-structured notes)		1–2 hours (charted), 5 minutes (monitor)
HiRID[10,11]	Bern University Hospital ICU	2008–2016	33,905	55,602		✓	✓			1 minute or less
AmsterdamUMCdb (v1.0.2)[12,13]	Amsterdam University Medical Center ICU	2003–2016	20,109	23,106		✓	✓	✓		1–30 minutes
EHRSHOT[20]	Stanford Medicine	1990–2023	6,739	921,499 (all encounters, not restricted to ED or ICU)	✓	✓	✓			1–4 hours
MC-MED[14,15]	Stanford Health Care Emergency Department	2020–2022	70,545	118,385	✓		✓	✓	ECG, PPG, respiration	1 minute (monitored) or as recorded (charted)

Table 1. Characteristics of ED and ICU datasets.

signs and ECG, PPG, and respiratory waveforms; orders placed and medications administered during the visit; laboratory and imaging results; diagnoses, visit disposition, and length of stay. Figure 1 presents an overview of the time-varying data modalities included throughout an ED visit. MC-MED differs from existing datasets in its focus on a diverse ED population and its inclusion of continuously recorded vital signs and physiologic waveforms. Moreover, it is the first dataset to exclusively cover ED patients during and after the peak of the COVID-19 pandemic. Thus, MC-MED represents a valuable resource for researchers exploring many aspects of modern emergency care, with a focus on granular physiological measurements. The dataset has been used in the Multimodal Clinical Benchmark for Emergency Care (MC-BEC)[16], which focuses on clinically relevant prediction tasks across different timescales, including patient decompensation, disposition, and ED revisit.

2.

Methods

Acquisition & transformation.

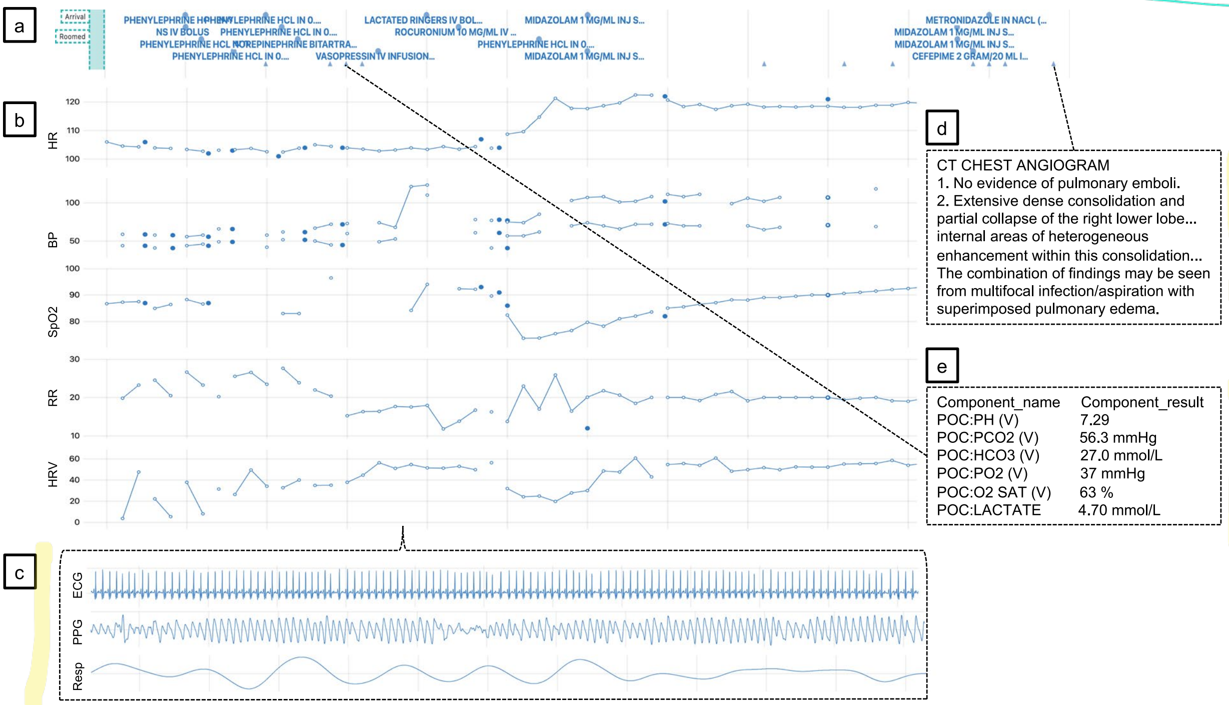
MC-MED includes 118,385 adult ED visits to monitored beds of the Stanford Health Care Emergency Department between 2020 and 2022, from 70,545 unique patients aged 18 or older at the time of visit. Clinical EHR data were derived from the STAnford medicine Research data Repository (STARR), a clinical data warehouse containing data from the Epic EHR at Stanford Health Care, and from auxiliary hospital applications such as the radiology Picture Archiving and Communications System. Continuously monitored vital signs and physiologic waveforms were captured with Philips IntelliVue bedside monitors, stored in a separate data warehouse, and extracted using Philips PIC iX DWC Toolkit (C.03.31). The data acquisition, transformation, and de-identification processes are documented below. The creation and release of this de-identified dataset were approved by the Stanford University Institutional Review Board (58581) and Secondary Data Use Committee with waiver of consent for retrospective research on de-identified, routinely collected data.

MC-MED data is organized into four categories: (1) structured EHR data (visit data, prior diagnoses, home medications, laboratory results, orders), (2) free-text radiology reports, (3) continuously monitored vital signs, and (4) ECG, PPG, and respiratory waveforms. Data categories 1–3 are consolidated in tables and stored as CSV files. Physiologic waveforms are stored as WaveForm DataBase (WFDB) files, in folders nested by visit identifier (CSN).

Deidentification.

MC-MED underwent a comprehensive deidentification process to remove all patient identifiers specified in the HIPAA Privacy Rule[17]. Patient (MRN) and visit (CSN) identifiers were mapped to random integers, and ages and date-times were randomly shifted at the patient level. Free-text radiology impressions were processed to remove any protected health information or information about specific providers, and manually verified by human reviewers. Figure 2 illustrates the de-identification process. Specific data elements were de-identified as follows:

- **Medical Record Number (MRN)** is a unique patient-level identifier. We generated a new unique integer to replace each of n original MRNs, randomly generating n values between large number $L = 99999999$ and L-20n without replacement, and ensuring no overlap with original MRNs.



CT CHEST ANGIOGRAM
1. No evidence of pulmonary emboli.
2. Extensive dense consolidation and partial collapse of the right lower lobe... internal areas of heterogeneous enhancement within this consolidation... The combination of findings may be seen from multifocal infection/aspiration with superimposed pulmonary edema.

Component_name	Component_result
POC:PH (V)	7.29
POC:PCO2 (V)	56.3 mmHg
POC:HCO3 (V)	27.0 mmol/L
POC:PO2 (V)	37 mmHg
POC:O2 SAT (V)	63 %
POC:LACTATE	4.70 mmol/L

Fig. 1 Time-varying data modalities recorded throughout a patient's ED visit (CSN 99797372). (**a**) shows the orders placed by the ED physician. (**b**) shows selected numeric monitoring data, including heart rate (HR), blood pressure (BP), peripheral oxygen saturation (SpO2), respiratory rate (RR), and heart rate variability (HRV). (**c**) shows electrocardiogram (ECG), photoplethysmogram (PPG), and respiration (Resp) waveforms at a specific point in the visit. (**d**) shows a free-text radiology report. (**e**) depicts one of numerous time-stamped laboratory results.

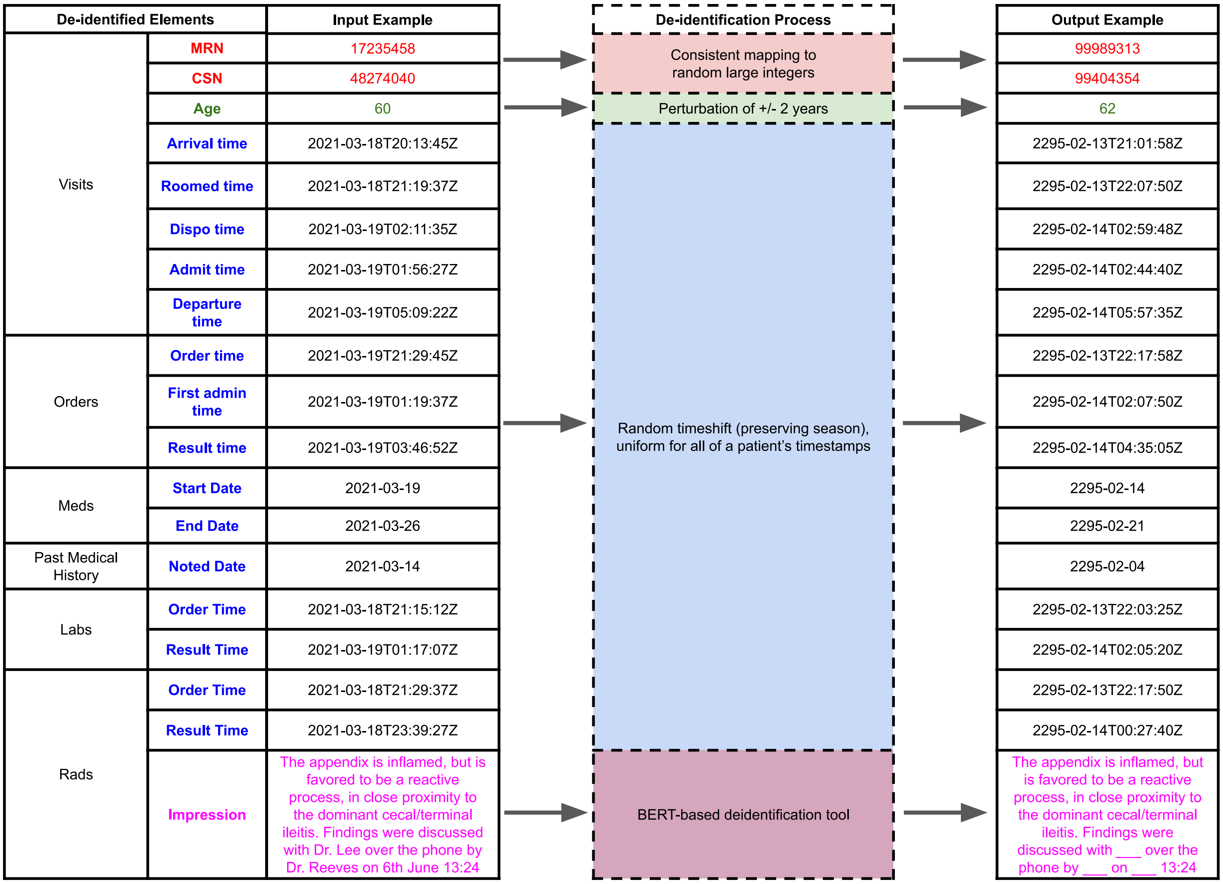
- **Contact Serial Number (CSN)** is a unique visit identifier, allowing linkage of the various data elements of MC-MED. We randomly generated a unique integer to replace each CSN, with the same procedure used for MRNs.
- **Age** at time of visit was altered by adding or subtracting a uniformly random number of years ranging from 0 to 2. Adjusted ages below 18 (all reflecting actual ages of at least 18) were set to 18, and ages exceeding 90 were set to 90. This ensures that actual ages are obscured, while providing researchers with an accurate age range for analysis.
- **Datetime fields** are shifted by a random interval for each patient (MRN). We generated a time-shift value for each patient, measured in seconds, that is applied consistently to that patient's datetime fields in all data elements. All datetime values are anchored to new dates between 2150 and 2350, preserving the approximate seasonality of each patient's first visit (January-March, April-June, July-September, and October-December), as well as of most subsequent visits (which are shifted by the same amount but may change "seasons" depending on where that visit fell in the quarter). For example, a first visit occurring in February might be shifted to January or March, but not to July. Because time-shifts are applied at the patient level, time intervals between events in a patient's trajectory are preserved. Shifted datetime fields include: Arrival_time, Roomed_time, Admit_time, Dispo_time, and Departure_time (visits table); Order_time, Result_time, and First_admin_time (orders table); Entry_date, Start_date and End_date (home medications table); Noted_date (past medical history table); Order_time and Result_time (labs and rads tables).
- **Laboratory results** with qualitative free-text interpretations unique to a specific patient were removed.
- **Radiology reports** were deidentified with the Stanford MIDRC Penn Deidentifier[18], an automated deidentification model designed to remove PHI from free-text radiology reports. Any PHI was replaced with three underscores ("___").

De-identification of all data elements was manually inspected for PHI by author and non-author human reviewers.

Waveform and vital sign preprocessing.

In the study hospital's monitoring database, waveform and numeric data recorded by the bedside monitors are associated with ED beds rather than patient visits. We therefore used data on patient rooming locations, and rooming and departure times to segment continuously recorded vital signs and waveforms by patient visit. We used Python's WFDB library to process and export this waveform data.

Because visits are associated with variable periods of monitoring of different modalities (for instance, ECG leads and PPG probes may be detached during patient movement and transport, then reattached upon return



De-identified Elements		Input Example	De-identification Process	Output Example
Visits	MRN	17235458	Consistent mapping to random large integers	99989313
	CSN	48274040		99404354
	Age	60	Perturbation of +/- 2 years	62
	Arrival time	2021-03-18T20:13:45Z	Random timeshift (preserving season), uniform for all of a patient's timestamps	2295-02-13T21:01:58Z
	Roomed time	2021-03-18T21:19:37Z		2295-02-13T22:07:50Z
	Dispo time	2021-03-19T02:11:35Z		2295-02-14T02:59:48Z
	Admit time	2021-03-19T01:56:27Z		2295-02-14T02:44:40Z
	Departure time	2021-03-19T05:09:22Z		2295-02-14T05:57:35Z
Orders	Order time	2021-03-19T21:29:45Z		2295-02-13T22:17:58Z
	First admin time	2021-03-19T01:19:37Z		2295-02-14T02:07:50Z
	Result time	2021-03-19T03:46:52Z		2295-02-14T04:35:05Z
Meds	Start Date	2021-03-19		2295-02-14
	End Date	2021-03-26		2295-02-21
Past Medical History	Noted Date	2021-03-14		2295-02-04
Labs	Order Time	2021-03-18T21:15:12Z		2295-02-13T22:03:25Z
	Result Time	2021-03-19T01:17:07Z		2295-02-14T02:05:20Z
Rads	Order Time	2021-03-18T21:29:37Z		2295-02-13T22:17:50Z
	Result Time	2021-03-18T23:39:27Z		2295-02-14T00:27:40Z
	Impression	The appendix is inflamed, but is favored to be a reactive process, in close proximity to the dominant cecal/terminal ileitis. Findings were discussed with Dr. Lee over the phone by Dr. Reeves on 6th June 13:24	BERT-based deidentification tool	The appendix is inflamed, but is favored to be a reactive process, in close proximity to the dominant cecal/terminal ileitis. Findings were discussed with ___ over the phone by ___ on ___ 13:24

Fig. 2 De-identification process. Patient (MRN) and visit (CSN) identifiers are mapped to random integers. Patient age is randomly perturbed by 0–2 years. All timestamps are shifted by a patient-level random time interval, maintaining seasonality. PHI is stripped from free-text radiology impressions using a BERT-based de-identification tool.

to the room), we present waveform data for each visit in multiple segments for each modality (ECG, PPG, respiration), and exclude recordings without physiologically meaningful signals (for instance, from detached leads). Specifically, waveform segments with constant values for 10 seconds or longer were removed. We computed derivative waveforms, w', where w'[i + 1] = w[i + 1]-w[i]. We then removed waveform segments for which w' was 0 for 10 seconds or greater. This processing ensures efficient representation and reliability of waveform data linked to complete ED visits.

Train-validation-test splits. Though researchers may segment MC-MED in the manner most appropriate for their research question, we release two training/validation/test splits for general use. For both splits, the training set contains 80% of visits, and validation and test sets each contain 10% of visits.

Random patient-level split. CSNs (visits) corresponding to the same MRN (patient) are present in the same set: split_random_train.csv, split_random_val.csv and split_random_test.csv.

Chronological split. All visits in the validation set occur after the final visit in the training set, and all visits in the test set occur after the final visit in the validation set. To prevent patient data leakage between sets, each patient (MRN) is again restricted to only one of the training, validation, or test sets. This results in 13,007 visits being removed from these sets, and exact splits of 78%, 11%, and 11% for split_chrono_train.csv, split_chrono_val.csv, and split_chrono_test.csv.

3,

Data Records

MC-MED is accessible through PhysioNet[15,19] (where data can be downloaded by eligible researchers after signing a data use agreement) and Nightingale Open Science[14] (where data can be analyzed on a secure cloud computing environment after signing a data use agreement). The dataset contains seven primary tables: visits.csv, orders.csv, meds.csv, pmh.csv, rads.csv, labs.csv, and numerics.csv. Waveform data availability is described in waveform_summary.csv, and the waveform data themselves are contained in folders named by the last 3 digits of de-identified CSNs, where each folder contains waveform data as WFDB records, including Electrocardiogram

(ECG), Photoplethysmogram (PPG), and Respiration (Resp). Additional description of data elements and patient/visit characteristics is available at https://physionet.org/content/mc-med.

Visits. The visits table describes high-level characteristics of each visit. Data available at the time of patient arrival include: patient demographics ("Age", "Gender", "Race", "Ethnicity"), means of arrival to the ED ("Means_of_arrival"), triage vital signs ("Triage_Temp", "Triage_HR", "Triage_RR", "Triage_SpO2", "Triage_SBP", "Triage_DBP"), triage acuity by Emergency Severity Index (ESI) ("Triage_acuity"), and chief complaint ("CC"). Data summarizing the visit itself include ED disposition ("ED_dispo"), ED length of stay in hours ("ED_LOS"), class of primary visit payor ("Payor_class"), and primary diagnosis, by ICD9 (International Classification of Diseases, Ninth Revision) and ICD10 (International Classification of Diseases, Tenth Revision) codes, accompanied by free-text descriptions ("Dx_name"). For patients admitted to the hospital, the table includes admitting service ("Admit_service"), hospital length of stay in days ("Hosp_LOS"), and disposition on hospital discharge ("DC_dispo"). Shifted timestamps include: "Arrival_time" (arrival in ED), "Roomed_time" (first rooming), "Dispo_time" (time of disposition decision), "Admit_time" (time of admission), and "Departure_time" (time of departure from ED). Finally, the visits table includes the number of visits from a given patient in the dataset ("Visits"), the sequence of a given visit ("Visit_no"), the hours from ED departure until a patient's next ED visit ("Hours_to_next_visit"), and the disposition of the next ED visit ("Dispo_class_next_visit"). The visits table can be linked to other tables by CSN (orders, labs, rads, numerics) or MRN (meds, PMH).

Orders. The orders table contains all orders placed by the ED physician during the visit, and is linked to other tables by CSN. "Order_type" categorizes orders, e.g., lab tests, imaging, medications, consults, nursing orders. "Procedure_name" describes the order, and "Procedure_ID" gives an accompanying procedure code. The following timestamps are shifted at the MRN level: "Order_time" describes when an order is placed, "First_admin_time" when a medication order is administered to a patient, and "Result_time" when a laboratory or imaging order produces a reported result.

Meds. The meds table contains patient home medications, organized by patient (MRN). "Med_ID" gives a unique medication code, and "NDC" the National Drug Code, where available. "Name" and "Generic_name" describe the medication. "Med_class" gives a high-level classification of the medication, and "Med_subclass" a more detailed classification. "Active" indicates whether a patient was thought to be using the medication at the time of the visit. "Start_date" and "End_date" give shifted dates of medication initiation and termination, where applicable. These dates can be used to identify home medications at the time of a given visit.

PMH. The PMH (Past Medical History) table contains prior diagnoses, organized by patient (MRN). "Noted_date" gives the shifted date when the diagnosis was recorded, and can be used to identify known diagnoses at the time of a given visit. "CodeType" specifies whether the "Code" should be interpreted as an ICD9 or ICD10 code. "Desc10" gives a text description of the ICD code. "CCS" gives the Clinical Classification Software category of the diagnosis, and "DescCCS" a text description of the CCS category.

Labs. The labs table gives results for lab tests ordered during the ED visit (CSN). "Display_name" describes the test or panel of tests (e.g. comprehensive metabolic panel), while "Component_name" describes the specific measurement (e.g. serum sodium). "Abnormal" indicates whether any result in the test falls outside the normal range, and "Component_abnormal" whether a specific measurement is abnormal. "Component_result" gives the specific result, which may be numeric or categorical, while "Component_value" assigns a numeric value to all results. "Component_units" gives the units in which "Component_value" is measured, and "Component_nml_low" and "Component_nml_high" describe the normal range, where applicable. "Order_time" is the shifted time the test was ordered (which may not exactly match the timestamps in the orders table, since the laboratory may modify or further specify orders), and "Result_time" gives the time the result was reported.

Rads. The rads table contains results of imaging studies ordered during each visit (CSN). "Study" is the type of imaging test (e.g., "XR HAND 3 VIEWS LEFT"), and "Impression" gives the de-identified free text summary of the resulting radiology report. As for the labs table, "Order_time" is the shifted time the study was ordered, and "Result_time" reflects when the attending radiologist's impression was posted.

Numerics. The numerics table contains the non-waveform monitoring data, by visit (CSN). "Measure" indicates one of 12 measurements: heart rate (HR), respiratory rate (RR), oxygen saturation by pulse oximetry (SpO2), systolic blood pressure (SBP), diastolic blood pressure (DBP), mean arterial pressure (MAP), temperature in degrees Fahrenheit (Temp), mean last-minute perfusion index derived from the PPG waveform (Perf), pain rating on 0–10 scale (Pain), liters per minute of supplemental oxygen (LPM_O2), and heart rate variability over the last 1 minute (1min_HRV) or 5 minutes (5min_HRV), calculated as the standard deviation of the beat-to-beat RR interval of the ECG waveform over this period. "Value" gives the accompanying numeric value of the observation. Where underlying observations are made more frequently than once per minute, they are aggregated to the mean value over the 60 seconds preceding the timestamp given in "Time". "Source" indicates whether a value was recorded by nursing (Chart) or derived directly from the monitoring database (Monitor).

Waveforms. The waveforms folder contains ECG (electrocardiogram), PPG (photoplethysmogram), and Resp (respiration) waveform data as WFDB records. The waveform data is organized into folders by visit. Parent folders are named by the last 3 digits of the CSN. These folders contain CSN-level folders, which contain subfolders for each waveform data type (ECG, PPG, respiration), containing multi-segment WFDB-compatible records of the waveforms.

Technical Validation

4,

We confirmed the completeness and consistency of de-identification, ensuring that all HIPAA identifiers were removed and that all perturbations were consistent across data elements. Perturbed values include patient identifiers (MRN), visit identifiers (CSN), ages, timestamps, and free-text radiology report impressions. Validation included the following checks:

1. Original and the de-identified MRNs are disjoint sets.
2. Original and the de-identified CSNs are disjoint sets.
3. All shifted timestamps range between years 2150 and 2350.
4. Since ages were randomly perturbed by adding a number from {-2,-1,0,1,2} to the original age, we ensured that the distribution of differences between true and perturbed ages was uniform across this range.
5. All event timestamps in a visit fall between arrival and departure times, with certain appropriate exceptions (e.g. results reported after ED departure).
6. All timestamps for a given patient were shifted by the same amount, with rare exceptions: some diagnoses had missing Noted_date in the pmh table (typically marked as 1900-01-01), for which the de-identified timestamp was set to 2150-01-01.
7. De-identification of the radiology report impressions in the rads table was confirmed by independent human reviewers.

Code availability

We demonstrate how to read, link, and visualize the data in this GitHub repository: https://github.com/dkimlab/MCMED.

Received: 5 March 2025; Accepted: 17 June 2025;
Published online: 01 July 2025

References

1. Burke, L. G., Burke, R. C., Epstein, S. K., Orav, E. J. & Jha, A. K. Trends in Costs of Care for Medicare Beneficiaries Treated in the Emergency Department From 2011 to 2016. *JAMA Netw Open* **3**, e208229, https://doi.org/10.1001/jamanetworkopen.2020.8229 (2020).
2. Sun, B. C. *et al*. Effect of emergency department crowding on outcomes of admitted patients. *Ann. Emerg. Med.* **61**, 605–611.e6, https://doi.org/10.1016/j.annemergmed.2012.10.026 (2013).
3. De Georgia, M. A., Kaffashi, F., Jacono, F. J. & Loparo, K. A. Information technology in critical care: review of monitoring and data acquisition systems for patient care and research. *Scientific World Journal* **2015**, 727694, https://doi.org/10.1155/2015/727694 (2015).
4. Seh, A. H. *et al*. Healthcare Data Breaches: Insights and Implications. *Healthcare (Basel)* **8**, https://doi.org/10.3390/healthcare8020133 (2020).
5. Johnson, A. *et al*. MIMIC-IV-ED. *PhysioNet*. https://doi.org/10.13026/5NTK-KM72 (2023).
6. Johnson, A. *et al*. MIMIC-IV, a freely accessible electronic health record dataset. *Sci. Data* **10**, 1, https://doi.org/10.1038/s41597-022-01899-x (2023).
7. Johnson, A. *et al*. MIMIC-III, a freely accessible critical care database. *Sci. Data* **3**, 160035, https://doi.org/10.1038/sdata.2016.35 (2016).
8. Pollard, T. J. *et al*. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Sci. Data* **5**, 180178, https://doi.org/10.1038/sdata.2018.178 (2018).
9. Pollard, T. J. *et al*. eICU Collaborative Research Database. *PhysioNet* https://doi.org/10.13026/C2WM1R (2019).
10. Yèche, H. *et al*. HiRID-ICU-Benchmark–A Comprehensive Machine Learning Benchmark on High-resolution ICU Data. https://arxiv.org/abs/2111.08536 (2021).
11. Faltys, M. *et al*. HiRID, a high time-resolution ICU dataset. *PhysioNet* https://doi.org/10.13026/nkwc-js72 (2021).
12. Thoral, P. J. *et al*. Sharing ICU Patient Data Responsibly Under the Society of Critical Care Medicine/European Society of Intensive Care Medicine Joint Data Science Collaboration: The Amsterdam University Medical Centers Database (AmsterdamUMCdb) Example. *Crit. Care Med.* **49**, e563–e577, https://doi.org/10.1097/CCM.0000000000004916 (2021).
13. Elbers, P. W. G. AmsterdamUMCdb v1.0.2. *DANS Data Station Life Sciences* https://doi.org/10.17026/dans-22u-f8vd (2019).
14. Kansal, A., Chen, E., Jin, B. T., Rajpurkar, P. & Kim, D. A. Multimodal Clinical Monitoring in the Emergency Department (MC-MED). *Nightingale Open Science* https://doi.org/10.48815/N57P4G (2025).
15. Kansal, A., Chen, E., Jin, T., Rajpurkar, P. & Kim, D. Multimodal Clinical Monitoring in the Emergency Department (MC-MED). *PhysioNet*. https://doi.org/10.13026/jz99-4j81 (2025).
16. Chen, E. *et al*. Multimodal Clinical Benchmark for Emergency Care (MC-BEC): A Comprehensive Benchmark for Evaluating Foundation Models in Emergency Medicine. *Advances in Neural Information Processing Systems* **36**, 45794–45811, https://arxiv.org/abs/2311.04937 (2023).
17. Department of Health and Human Services. *The HIPAA Privacy Rule*. https://www.hhs.gov/hipaa/for-professionals/privacy/index.html (2024).
18. Chambon, P. J. *et al*. Automated deidentification of radiology reports combining transformer and 'hide in plain sight' rule-based methods. *J. Am. Med. Inform. Assoc.* **30**, 318–328, https://doi.org/10.1093/jamia/ocac219 (2023).
19. Goldberger, A. L. *et al*. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* **101**, E215–20, https://doi.org/10.1161/01.cir.101.23.e215 (2000).
20. Wornow, M., Thapa, R., Steinberg, E., Fries, J. A. & Shah, N. H. EHRSHOT: An EHR Benchmark for Few-Shot Evaluation of Foundation Models. *Advances in Neural Information Processing Systems* **36**, https://arxiv.org/abs/2307.02028 (2023).

Acknowledgements

We thank Stephanie Bogdan and Xiaoli Yang for their independent verification of de-identification. We thank Eric Helfenbein for his guidance on processing waveform data. We thank the Gordon and Betty Moore Foundation (AWD102625) for funding this work.

Author contributions

D.A.K. and P.R. conceived of the de-identified dataset. D.A.K. acquired and linked the data elements. A.K., E.C. and B.T.J. performed de-identification and data transformations. All authors contributed to drafting and revising the manuscript, and to verification of de-identification. D.A.K. and P.R. supervised the project.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to D.A.K.