scientific data

Check for updates

OPEN

DATA DESCRIPTOR

BOLD: Blood-gas and Oximetry Linked Dataset

João Matos[1,2,3,4] ✉, Tristan Struja[5], Jack Gallifant[1,6], Luis Nakayama[1,7], Marie-Laure Charpignon[8], Xiaoli Liu[9], Nicoleta Economou-Zavlanos[10], Jaime S. Cardoso[2,3], Kimberly S. Johnson[11], Nrupen Bhavsar[12,13], Judy Gichoya[14], Leo Anthony Celi[1,15,16] & An-Kwok Ian Wong[4,17] ✉

Pulse oximeters measure peripheral arterial oxygen saturation (SpO_2) noninvasively, while the gold standard (SaO_2) involves arterial blood gas measurement. There are known racial and ethnic disparities in their performance. BOLD is a dataset that aims to underscore the importance of addressing biases in pulse oximetry accuracy, which disproportionately affect darker-skinned patients. The dataset was created by harmonizing three Electronic Health Record databases (MIMIC-III, MIMIC-IV, eICU-CRD) comprising Intensive Care Unit stays of US patients. Paired SpO_2 and SaO_2 measurements were time-aligned and combined with various other sociodemographic and parameters to provide a detailed representation of each patient. BOLD includes 49,099 paired measurements, within a 5-minute window and with oxygen saturation levels between 70–100%. Minority racial and ethnic groups account for ~25% of the data – a proportion seldom achieved in previous studies. The codebase is publicly available. Given the prevalent use of pulse oximeters in the hospital and at home, we hope that BOLD will be leveraged to develop debiasing algorithms that can result in more equitable healthcare solutions.

Background & Summary

The measurement and management of arterial blood gas (ABG) and pulse oximetry in the Intensive Care Unit (ICU) have long been the subject of clinical interest but are often under-studied. Pulse oximeters and ABG are tools for evaluating systemic oxygen saturation and providing guidance for clinical decision-making. Standardization in pairing arterial blood gas samples with pulse oximeter readings, a critical component for effective patient monitoring and management, is particularly scarce. This is due in part to challenges in coordinating large electronic health record (EHR) datasets and synchronizing clinical protocols across multiple medical centers.

Recent research by Sjoding *et al.*, Wong *et al.*, Valbuena *et al.*, and Gottlieb *et al.*, has added another layer of complexity by uncovering racial disparities in pulse oximeter reading accuracy, which have critical implications for patient care and outcomes[1–4]. Such disparities further emphasize the urgent need for robust and inclusive datasets allowing the conduct of thorough comparative analysis across subpopulations. Given these pervasive challenges and recent findings, the retrospective investigation of real-world data can offer invaluable insights.

[1]Laboratory for Computational Physiology, Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA, USA. [2]Faculty of Engineering, University of Porto (FEUP), Porto, Portugal. [3]Institute for Systems and Computer Engineering, Technology and Science (INESCTEC), Porto, Portugal. [4]Duke University, Department of Medicine, Division of Pulmonary, Allergy, and Critical Care Medicine, Durham, NC, USA. [5]Medical University Clinic, Kantonsspital Aarau, Aarau, Switzerland. [6]Department of Critical Care, Guy's and St Thomas' NHS Trust, London, United Kingdom. [7]Department of Ophthalmology, São Paulo Federal University, São Paulo, SP, Brazil. [8]Institute for Data Systems and Society, Massachusetts Institute of Technology, Cambridge, MA, US. [9]Center for Artificial Intelligence in Medicine, The General Hospital of PLA, Beijing, China. [10]Duke University, AI Health, Durham, NC, USA. [11]Duke University, Department of Medicine, Division of Geriatrics, Durham, NC, USA. [12]Duke University, Department of Biostatistics and Bioinformatics, Division of Translational Biomedical Informatics, Durham, NC, USA. [13]Duke University, Department of Surgery, Division of Surgical Sciences, Durham, NC, USA. [14]Emory University, Department of Radiology, Atlanta, GA, USA. [15]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [16]Department of Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA. [17]Duke University, Department of Biostatistics and Biomedical Informatics, Division of Translational Biomedical Informatics, Durham, NC, USA. ✉e-mail: jcmatos@mit.edu; med@aiwong.com

All of the above-listed studies have used EHR data, which was stored in multiple formats that may not be easy to use and may not be available to external researchers, representing a significant barrier to entry.

Existing large-scale EHR datasets, even when available in open access or under a formal research protocol, are often not in a form that can help readily answer nuanced yet urgent clinical questions such as the need of applying potential corrections by skin tone to the measurements output by FDA-approved devices[5–7]. The understanding of health systems, data schemas, and critical care physiology necessary to make individual ICU-EHR datasets usable in practice is nontrivial; thus, our effort to preprocess raw time series to create a unified dataset removes a barrier to entry.

This paper aims to present a comprehensive approach to the extraction, processing, and analysis of arterial blood gas samples and pulse oximeter readings from electronic health records (EHR). We demonstrate the application of this principled approach to the issue of racial and ethnic disparities in device measurement and hope it can guide future studies.

We propose a clinically-grounded and reproducible methodology to convert unprocessed database queries into a clinically useful dataset. Our multidisciplinary team, which includes clinicians (i.e., pulmonologists and intensivists) and data scientists, has developed rules based on clinical and physiological standards for pairing each arterial blood gas sample with a corresponding pulse oximeter reading as well as with clinical scores, vital signs, and laboratory test values.

Our primary objective is to facilitate extensive analysis of pulse oximetry by merging data from three major, publicly available, ICU-EHR databases – MIMIC-III, MIMIC-IV, and eICU-CRD. A combined dataset not only offers a solution to the paucity of large and diverse datasets but also a unique platform for identifying disparities in pulse oximetry readings and designing approaches to remediate such inequities. By making this robust dataset publicly available, we provide researchers with the means to develop models that address known racial and ethnic disparities and those yet to be discovered, with the potential to improve fairness in healthcare delivery. Furthermore, by making the platform and code available, this platform can serve as an example for conducting similar studies that would benefit from linked databases.

Our work stands as a necessary and fundamental milestone to any machine learning (ML)[8] or advanced data analysis of oxygen readings that can be performed to produce actionable insights. Data curation is a critical step, especially considering the volume of data in our base datasets; e.g., MIMIC-III v.1.4 alone contains over 58,000 hospital admissions from approximately 38,600 adults, resulting in 6.2GB of data. A traditional manual chart review (e.g., to identify patients at risk of hypoxemia) would be impractical given this volume of data, emphasizing the need for an automated, yet clinically-validated approach. As an example, we anticipate value in a machine learning model that, based on a patient's oxygen saturation trajectory since ICU admission, could predict the likelihood of hypoxemia in the next hours.

The dataset we present not only addresses the critical issue of pulse oximetry disparities but also offers a versatile tool for the broader medical research community. In the future, we plan to extend this dataset to other EHR databases and to include waveform data. By detailing our methodologies and sharing our modular scripts, we provide avenues for other researchers to build upon this work, potentially extending it to other biometric readings (e.g., body temperature, blood pressure) and clinical contexts (e.g., home-based care, primary care, emergency room).

Overall, our dataset aims to serve as a pivotal resource for the clinical and research communities alike, informing respiratory parameter management in the ICU with a particular focus on addressing racial and ethnic disparities in pulse oximetry accuracy. We operationalize the evaluation of racial and ethnic disparities in pulse oximetry by quantifying differences in the occurrence of hidden hypoxemia, defined as SaO2 > 88% but $SpO_2 \geq 88\%0$[2]. Finally, we provide the complete codebases for data curation and validation assays to encourage ongoing, collaborative research in this critical area.

As observational data collected in hospital settings are often used retrospectively to inform the development, manufacturing, and quality control of pulse oximeters, our effort should prompt other parties (e.g., pulse oximetry equipment manufacturers) with access to such paired measurements but in different settings (e.g., randomized trials) to also share the underlying datasets with the public.

Methods

Data sources.

Three EHR databases were used: MIMIC-III, MIMIC-IV, and eICU-CRD.

MIMIC-III. MIMIC-III (Medical Information Mart for Intensive Care III) is a comprehensive and publicly accessible database that contains de-identified health data associated with over 40,000 patients who stayed in critical care units of the Boston-based Beth Israel Deaconess Medical Center (BIDMC) between 2001 and 2012. It is maintained by the Laboratory for Computational Physiology (LCP) at MIT and is shared through the PhysioNet platform. The database includes information such as demographics, vital sign measurements, laboratory test results, medications, and more. Since its release in 2016, it has served as a valuable resource for a wide range of research studies in healthcare, including those focused on critical care and machine learning applications in medicine[6].

MIMIC-IV. MIMIC-IV builds on the foundation laid by MIMIC-III, extending the dataset to include patients admitted to the ICU from 2008 to 2019. Unique features of MIMIC-IV include clinical progress notes and physiological data collected from bedside monitors. Approximately 70,000 de-identified medical records are archived in the MIMIC-IV database[7].

As there is potential overlap of patients between 2008–2012 across both MIMIC versions, where the same patient may have distinct but not linkable identifiers, users of our dataset may consider dropping MIMIC-III encounters entirely or restricting their analysis to those corresponding to 2013–2019.

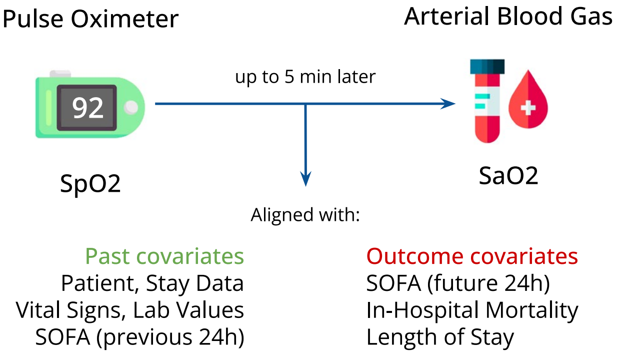

Fig. 1 Rationale and variables included in the dataset.

eICU-CRD. The eICU-CRD database is a publicly available multi-center database sourced from the Philips Healthcare eICU Telehealth Program. It contains information about over 200,000 admissions from 208 hospitals or ICUs monitored by eICU programs across the United States, between 2014 and 2015. The eICU-CRD patients are distinct from MIMIC-III and MIMIC-IV subjects, alleviating any concerns regarding the potential overlap of their underlying populations[5].

Software.

For data extraction, BigQuery through Google Colaboratory (Python 3.10) was used.

Publicly-available derived views.

The derived tables available on the MIMIC-Code (*icustay_detail*, *vitalsign*, *complete_blood_count*, *coagulation*, *chemistry*, *bg*, *enzyme*, and *sofa*) and eICU-Code (*pivoted_vital* and *pivoted_lab*) repositories, made available on the MIT-LCP GitHub repository, were used[9].

Inclusion and exclusion criteria.

All patients admitted to a hospital or ICU captured by one of the aforementioned databases who had valid ABG and pulse oximetry data were included. Two versions of the dataset were obtained: an extended dataset created primarily for validation purposes and a preprocessed, validated dataset shared in the present study:

1. Extended dataset (mainly for technical validation purposes):
   - (SaO_2, SpO_2) pairs are captured within 90 minutes
   - No range for the oxygen saturation is set
   - All pairs per hospital admission are considered
2. Preprocessed dataset (the shared version):
   - (SaO2, SpO_2) pairs are captured within 5 minutes[10,11]
   - A range of 70–100% is set
   - Only the first pair per hospital admission is considered

$SaO_2 – SpO_2$ matching.

We require each pulse oximetry reading (SpO_2) to precede the ABG measurement (SaO_2). Missing ABG data is not allowed. For the extended dataset, the window is [−90, 0] minutes; for the preprocessed dataset, [−5, 0] minutes. Figure 1 depicts the rationale of the final dataset.

Time alignment and curation across different databases.

To facilitate modifications, ensure that definitions remain consistent across databases, and promote subsequent code reuse by other teams, we followed these steps to align each (SaO_2, SpO_2) pair with time-varying covariates:

1. Create pivoted views of the lab measurements, vital signs, and hourly SOFA scores (either publicly-available on BigQuery, or generated by our team);
2. List the variables to be pulled, each with the following fields:
   a. Variable type (for the prefix)
   b. Original name (for the pull)
   c. New name (harmonized across databases)
   d. Time window for the value to be considered (variable-specific)
   e. Source table (the *id* of the pivoted view)
   f. Used foreign key (that links the (SaO_2, SpO_2) pair with the source table)
   g. Name of the timestamp variable (specific to the source table)
3. Parse the new table (saved as an editable *Google Spreadsheet*) through *Pandas*, on Google Colab

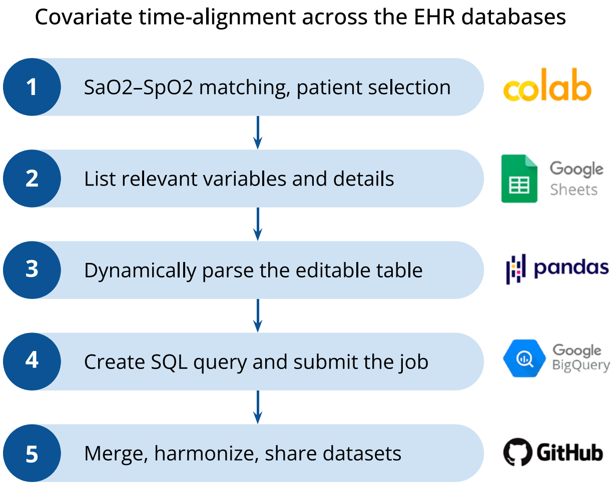
Covariate time-alignment across the EHR databases
1
SaO2–SpO2 matching, patient selection
colab
2
List relevant variables and details
Google Sheets
3
Dynamically parse the editable table
pandas
4
Create SQL query and submit the job
Google BigQuery
5
Merge, harmonize, share datasets
GitHub

Fig. 2 Pipeline created to curate and merge the datasets.

4. Create the complete SQL query automatically, feeding it with all listed variables, previously aligned with the (SaO_2, SpO_2) pairs through separate subqueries
5. Query the databases through the *Python BigQuery API*

The above-described stepwise process is summarized in Fig. 2. By setting relevant time windows for each variable, we ensure to extract relevant data only. For example, a temperature reading will only be aligned with a (SaO_2, SpO_2) pair if registered up to 8 hours before the SaO_2value was measured. Missingness is kept as is to give users the flexibility to adopt their own imputation strategy, if needed.

Harmonization of concepts among databases. Data from different tables were harmonized into the same format across all databases. To minimize missingness, only variables that are available in all three databases were included. Patient-level variables (*identifiers*; *demographics; admission characteristics* and *patient outcomes*) were unified as shown in the Supplemental Table 1. Time-varying variables (*vital signs*; *laboratory test values*; *hourly SOFA scores*) were pulled and harmonized as shown in Supplemental Table 2. Each variable maps to an *itemid* (respectively, *label*) in the MIMIC (respectively, eICU-CRD) databases.

Since eICU-CRD is designed around time offsets (e.g., minutes from ICU admission, minutes from hospital admission, etc.), all dates were converted from offsets by a reference date of January 1st, 2014 (since eICU-CRD encompasses 2014-2015 data).

For race and ethnicity, the NIH Policy on Reporting Race and Ethnicity was used as the reference[12]. The eight unified categories were: *“American Indian / Alaska Native”, “Asian”, “Black or African American”, “Hispanic OR Latino”, “More Than One Race”, “Native Hawaiian / Pacific Islander”, “White”, and “Unknown”*. We mapped the original race and ethnicity labels present in the databases we studied to these categories. The exact mappings are further depicted in Supplemental Tables 3a,b, and c. Based upon how data was captured, none of the databases are able to distinguish between race (e.g., Asian, Black, White) and ethnicity (e.g., “Hispanic OR Latino” vs “Not Hispanic or Latino”). We denote “Hispanic OR Latino” as “Hispanic”, a coded value. As such, a patient could be addressed as either any race (but non-Hispanic ethnicity) or Hispanic ethnicity (of any race). This limitation remains present in BOLD.

Data types. Several additional variables can help augment analyses and characterize patients receiving a temporally proximate (SaO_2, SpO_2) pair. These are further described below.

All adjunctive (e.g., vital signs, laboratory test values, etc.) data are referenced from the time of the ABG. For the purpose of this manuscript, a time delta (*delta_* prefix) refers to the time difference between the time of the most recently recorded covariate of interest and that of the ABG measurement. Each time-varying covariate is accompanied by a time delta. The ABG measurement time is set as the reference; any covariate measurement or reading must occur before this reference, unless otherwise noted.

Identifiers. Each encounter has three identifiers, at different levels: patient, hospital, and ICU admission. The original identifiers are kept to allow linking the data with the original databases and eventually pull other variables of interest. However, to avoid overlap among the databases, we created new, unique identifiers for our preprocessed dataset; they reflect each of these three identifiers. In addition, each encounter has an identifier to reflect the source database.

Among the three considered databases, only eICU-CRD has hospital identifiers, since the MIMIC databases come from one single hospital, i.e., BIDMC. As a result, MIMIC data was assigned a hospital index of 9999, which is outside the range of eICU-CRD hospital indices. Other hospital-related variables (number of beds, US region, and teaching status) were harmonized accordingly.

Demographics. Demographics, such as age at admission, sex, race and ethnicity, were extracted from the demographics tables of each database. Age at admission age was unified, with values between 18–89 kept intact and values of 90 and above taken equal to 90.

Admission characteristics and patient outcomes. Comorbidities are calculated by van Walraven Elixhauser score[13] (MIMIC-III) and Charlson Comorbidity Index[14] (MIMIC-IV, eICU-CRD). BMI was computed with the weight and height on admission, for each database.

Hospital-level (e.g., hospital size) and patient-specific admission characteristics (e.g., admission time) as well as patient-specific outcomes (e.g., in-hospital mortality) were recorded for each encounter.

Vital signs. Vital sign data were merged in accordance with Supplemental Table 2. Temperature, blood pressure (both non-invasive and invasive), heart rate, respiratory rate, and SpO_2 were extracted. These data were obtained from the *chartevents* and *nursecharting* tables of the original MIMIC and eICU databases, respectively. The prefix “*vitals_*” is used for each variable of this type, except for SpO_2.

Laboratory test values. Common laboratory test values were merged within variable-specific time windows as noted. Measurements of the following categories were pulled: ABG (no prefix), complete blood count (“*cbc_*” prefix); coagulation (“*coag_*” prefix); basic metabolic panel (“*bmp_*” prefix); hepatic function panel (“*hfp_*” prefix); and other enzymes (“*other_*” prefix). In the MIMIC databases, all laboratory test data were collected from the original *labevents* table; in eICU-CRD, data were collected from the *labs* table.

Hourly SOFA scores. To characterize organ dysfunction and severity of illness, sequential organ failure assessment (SOFA) scores were used[15]. SOFA scores for each dataset were calculated hourly. SOFA scores were extracted in the hour prior to the ABG to ensure that the latter has no impact on characterizing underlying organ dysfunction and thus avoid reverse causation (“*sofa_past_*” prefix). To quantify the impact of hypoxemia on organ dysfunction, subsequent SOFA scores were also extracted 24 hours after the ABG (“*sofa_future_*” prefix).

In the MIMIC databases, the publicly available derived table with hourly SOFA scores were used. In the eICU-CRD, we used an auxiliary query created by our team.

Data storage. The preprocessed dataset, meeting the defined inclusion / exclusion criteria, is stored on PhysioNet as a single comma separated value (CSV) file.

Descriptive analytics and technical validation. We now present the methodology followed to support the criteria we set to select patients and clean the data.

Flow diagrams depicting the application of inclusion and exclusion criteria to select our cohort were created and analyzed. At each exclusion step, we analyzed the composition of the patients who are dropped in terms of demographics.

We created descriptive tables highlighting patients' characteristics across source databases; race and ethnicity; and hidden hypoxemia – when $SpO_{2\geq}$88% but $SaO_2 < 88\%$, as defined by Wong *et al.*[2] The *tableone* package was used[16].

We employed Modified Bland-Altman plots, based on the methodology proposed by Wong *et al.*[2], to evaluate the agreement between $SaO_2$and SpO_2 measurements. We assessed the calibration performance across two different time window sizes — 5 and 30 minutes — to justify our final selection of a 5-minute window. Moreover, we conducted separate analyses across racial and ethnic groups to highlight disparities in calibration accuracy.

Oxyhemoglobin dissociation curves[17] are also reported as a referential integrity of our ABG data. To verify the existence of left and right shifts, we plotted, in different colors, the pairs with pH in the 90th and 10th percentiles, respectively.

The root mean squared error (RMSE) of each (SaO_2, SpO_2) pair was computed across different window limits, from 0 to 90, and stratified by race and ethnicity (for simplicity, considered groups were White, Black, Hispanic OR Latino, and Asian). RMSE was computed using Eq. (1) for each pair, aggregated with a mean, and then 95% confidence intervals were computed assuming normal distributions.

$$RMSE = \sqrt{\frac{1}{N}\sum\nolimits_{i=1}^{N}\left(x_i - \hat{x}_i\right)^2} \tag{1}$$

Finally, the *missingno* package[18] was used to assess the completeness of the data, reported as a bar plot.

Data Records

BOLD is available on PhysioNet as a credentialled database[19].

Description of fields

Demographics. ***subject_id*:** Describes a unique subject. This is unique per component dataset and mapped directly to the equivalent term in each. A subject may have multiple admissions, denoted by hospital_admission_id. Same as in the original database.

***hospital_admission_id*:** Describes a unique hospital admission. This is unique per component dataset and mapped directly to the equivalent term in each. Same as in the original database.

***icustay_id*:** Describes a unique ICU admission. This is unique per component dataset and mapped directly to the equivalent term in each. Each hospital admission may have multiple ICU stays. Same as in the original database.

unique_subject_id: Describes a unique subject. Guarantees that no subjects coming from different databases have the same identifier.

unique_hospital_admission_id: Describes a unique hospital admission. Guarantees that no subjects coming from different databases have the same identifier.

unique_icustay_id: Describes a unique ICU admission. Guarantees that no subjects coming from different databases have the same identifier.

source_db: Labeled as *mimic_iii*, *mimic_iv*, *eicu* to distinguish between component datasets.

race_ethnicity: Harmonized race and ethnicity.

Hospital characteristics.

hospitalid: Unique hospital ID. Beth Israel Deaconess (MIMIC-III, -IV) was denoted as 9999.

numbedscategory: Hospital size, in numbers of beds. This was recorded as “*<100*”; “*100–250*”; “*250–500*”; “*≥500*” beds for eICU-CRD. For MIMIC-III and -IV (BIDMC), we fixed the value at “*≥500” beds.*

teachingstatus: This field marks whether a hospital was identified as a teaching hospital, where a value of 1 implies that it was a teaching hospital. BIDMC has teaching status = 1.

region: Maps to the US census region distribution per eICU. It is either *Midwest, Northeast*, *South*, or *West*. MIMIC (BIDMC) was set as *Northeast*.

admission_age: Harmonized age on admission. eICU admission age mapped to admission_age. MIMIC-III admission age mapped to admission age. MIMIC-IV admission_age mapped to admission_age.

sex_female: Assigned a value of 1 if the patient is of female sex.

weight_admission: Weight on admission (in kilograms).

height_admission: Height on admission (in centimeters).

BMI_admission: Calculated BMI on admission, based on weight_admission and height_admission. If either weight_admission or height_admission is missing, BMI_admission is missing.

Admission characteristics.

datetime_hospital_admit, datetime_hospital_discharge, datetime_icu_admit, datetime_icu_discharge: Date and time of hospital and ICU admission (_admit) and discharge (_discharge).

los_hospital, los_ICU: Length of stay for hospital (_hospital) and ICU (_ICU) in days.

in_hospital_mortality: This variable is true if the patient died during the hospital admission, regardless if the patient died during the ICU admission or not.

comorbidity_score_name, comorbidity_score_value: Comorbidity score (either Elixhauser or Charlson), along with score.

ABG data.

SaO2_timestamp: Date and time of ABG test.

pH: pH value.

pCO2: Partial pressure of CO2 (mmHg).

pO2: Partial pressure of O2 (mmHg).

SaO2: Arterial oxygen saturation (%).

Carboxyhemoglobin: Percentage of hemoglobin bound to CO (%).

Methemoglobin: Percentage of methemoglobin (%).

Vitals data.

SpO2: Pulse oximetry saturation (%).

SpO2_timestamp: Date and time of SpO$_2$ recording.

vitals_heart_rate: Heart rate.

vitals_resp_rate: Respiratory rate.

vitals_mbp_ni, vitals_sbp_ni, vitals_dbp_ni: Mean arterial pressure (MAP), Systolic pressure (SBP), Diastolic pressure (SBP) calculated from noninvasive (cuff) blood pressure.

vitals_mbp_i, vitals_sbp_i, vitals_dbp_i: Mean arterial pressure (MAP), Systolic pressure (SBP), and Diastolic pressure (DBP) calculated from invasive (arterial line) blood pressure.

vitals_tempc: Temperature, from any body measuring site, in Celsius (°C).

Labs.

Complete Blood Count (CBC). ***cbc_wbc***: White blood cell. (10^9/L)

cbc_hemoglobin: Measured hemoglobin (g/L).

cbc_hematocrit: Measured hematocrit. (%)

cbc_platelet: Measured platelet count.(10^9/L)

cbc_mch: Measured mean corpuscular hemoglobin. (pg)

cbc_mchc: Measured mean corpuscular hemoglobin concentration. (g/L)

cbc_mcv: Measured mean corpuscular volume.(fL)

cbc_rbc: Measured red blood cells (RBC). (10^{12}/L)

cbc_rdw: Measured RBC distribution width. (%)

Coagulation labs. ***coag_fibrinogen***: Measured fibrinogen.

coag_pt: Measured prothrombin time. (s)

coag_inr: Measured international normalized ratio.

coag_ptt: Measured partial thromboplastin time. (s)

Basic Metabolic Panel (BMP). ***bmp_sodium***: Measured sodium levels. (mmol/L)

bmp_potassium: Measured potassium levels. (mmol/L)

bmp_chloride: Measured chloride levels. (mmol/L)

Race and Ethnicity	Average No. Pairs (Standard Deviation) ↓	N
White	4.47 (7.74)	43,925
Black	4.65 (8.23)	5,467
American Indian / Alaska Native	4.71 (8.49)	397
Hispanic OR Latino	5.09 (8.65)	2,436
Asian	5.12 (10.88)	1,061
Unknown	4.96 (9.47)	5,030

Table 1. Average number of pairs per race and ethnicity, extended dataset.

bmp_bicarbonate: Measured bicarbonate levels. (mg/dL)
bmp_bun: Measured blood urea nitrogen levels. (mg/dL)
bmp_creatinine: Measured creatinine levels. (mg/dL)
bmp_glucose: Measured glucose levels. (mg/dL)
bmp_aniongap: Measured anion gap. (mmol/L)
bmp_calcium: Measured calcium levels. (mg/dL)
bmp_lactate: Measured lactate levels. (mmol/L)

Hepatic function panel (HFP). ***hfp_alt***: Measured alanine aminotransferase (ALT) levels. (U/L)
hfp_alp: Measured alkaline phosphatase (ALP) levels. (U/L)
hfp_ast: Measured aspartate aminotransferase (AST) levels. (U/L)
hfp_bilirubin_total: Measured total bilirubin levels. (mg/dL)
hfp_bilirubin_direct: Measured direct bilirubin levels. (mg/dL)
hfp_albumin: Measured albumin levels. (g/dL)

Other labs (enzyme). ***others_ck_cpk***: Measured creatine kinase (CK) levels, also known as creatine phosphokinase (CPK). (U/L)
others_ck_mb: Measured creatine kinase MB (CK-MB) levels. (U/L)
others_ld_ldh: Measured lactate dehydrogenase (LDH) levels. (U/L)

SOFA scores. ***sofa_past_overall_24hr***: SOFA score, calculated from component values below, measured in the hour window prior to the ABG.

sofa_past_coagulation_24hr, sofa_past_liver_24hr, sofa_past_cardiovascular_24hr, sofa_past_cns_24hr, sofa_past_renal_24hr: SOFA score components, with highest value for each component in the past 24 hours. The hour window just prior to the hour window containing the ABG is recorded here to characterize baseline patient status.

sofa_future_overall_24hr: SOFA score, calculated from component values below, measured 24 hours after the ABG window.

sofa_future_coagulation_24hr, sofa_future_liver_24hr, sofa_future_cardiovascular_24hr, sofa_future_cns_24hr, sofa_future_renal_24hr: SOFA score components, with highest value for each component in the 24 hours after the ABG to characterize the 24 hour impact of discrepancies.

Technical Validation

In the extended dataset, at the loose cut-off of 90 minutes, we obtained on average 5 pairs for *Asian* and *Hispanic OR Latino* patients, 4.7 pairs for *Black* and *American Indian / Alaska Native* patients, and 4.5 pairs for *White* patients (see Table 1 with the average number of pairs per race and ethnicity).

To ensure that the distributions of SaO_2 and PaO_2 values obtained in the three considered EHR were concordant with the literature, we plotted an oxyhemoglobin dissociation curve (Fig. 3). We did not observe substantial deviation from the known dissociation curve. Specifically, the pH-associated left and right shifts were verified, ensuring that the data curation process in the original databases was not flawed.

We examined the pairs of the extended dataset to study the agreement of SpO_2 measurements by pulse oximeter with the SaO_2 measurements by ABG. We assessed various lengths of the eligible time window (*delta_SpO2*) to pair readings. The modified Bland-Altman plots presented in Fig. 4 revealed no significant differences between the two readings over time windows of length 5 and 30 minutes, respectively. The patterns remained the same irrespective of the database, patient race, and patient ethnicity.

However, increasing *delta_SpO2* tolerances yielded a sharp increase in the RMSE for Asian, Black, and Hispanic patients (Fig. 5); this change was most pronounced for a time delta of 60 minutes or more, most likely due to lower sample sizes among patients from minority groups than among White patients. Although increasing the time delta mechanistically resulted in an incremental increase in the number of eligible paired samples, we selected 5 minutes as the optimal cut-off. Indeed, it coincided with a marked increase in paired samples, while still yielding a relatively small RMSE.

The final dataset consisted of 49,099 first (SaO_2, SpO_2) pairs. Most pairs emanated from eICU-CRD (43,438), followed by MIMIC-IV (4,921), and then by MIMIC-III (740) (see Fig. 6a–c, with the flow diagram per database; Fig. 7 with the overall flow diagram). The distribution of eligible pairs by race and ethnicity varied across the three databases. Notably, the application of our exclusion criteria in the two MIMIC databases resulted in an

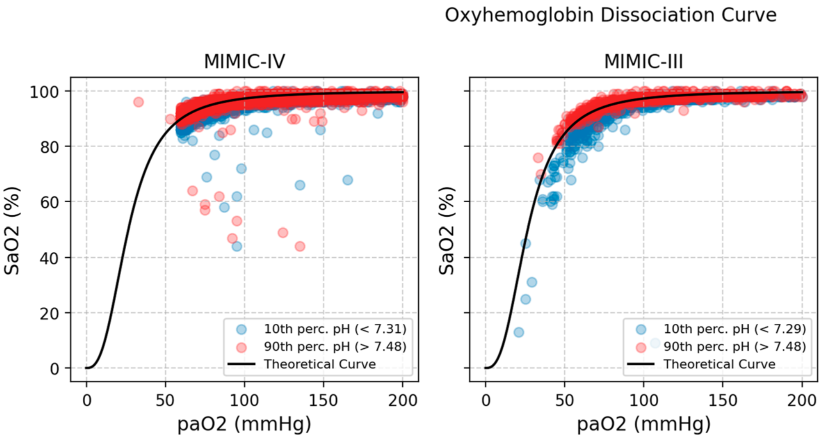

Fig. 3 Oxyhemoglobin dissociation curve, per database, with the pH shift highlighted, on the extended dataset.

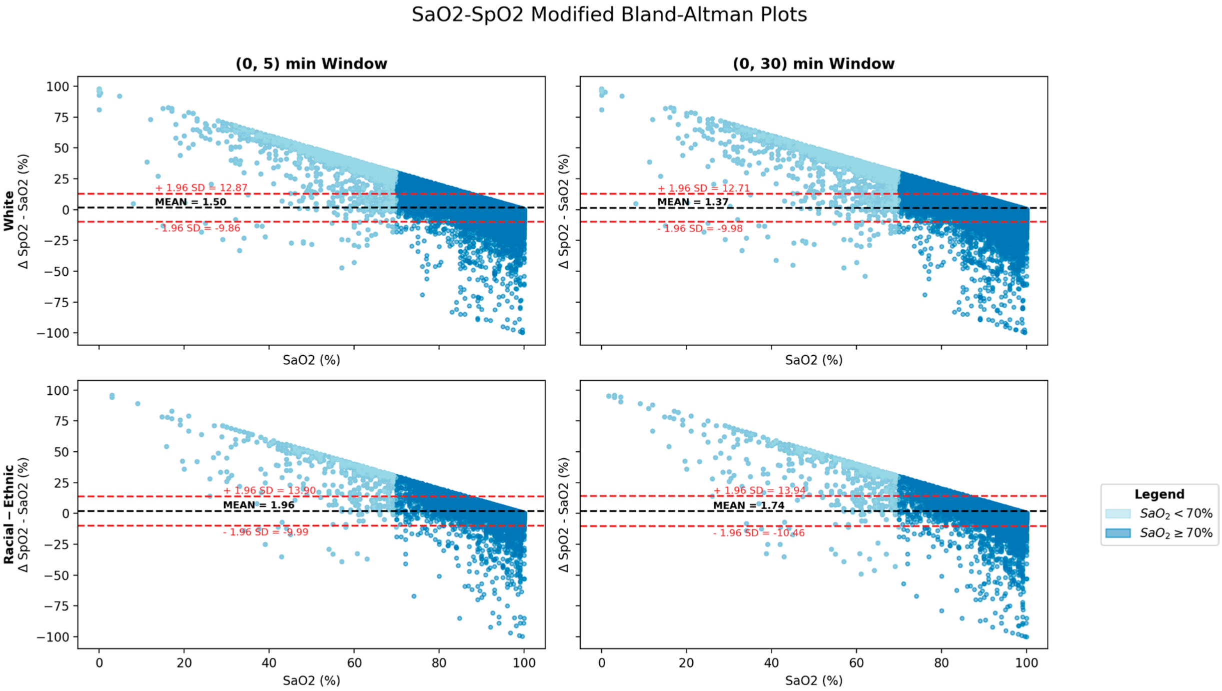

Fig. 4 Modified Bland-Altman plots, across race and ethnicity (White compared with racial and ethnic group), and across 2 time windows, on the extended dataset.

overrepresentation of White and male patients, while patients from racial and ethnic groups were dropped at a higher rate. This disproportionate exclusion rate may owe to the lower likelihood of ABG draws among patients from minority racial and ethnic groups[2].

In sum, across all three databases, White patients formed the most prevalent racial and ethnic subgroup, accounting for ~75% of cases in the resulting dataset after application of our inclusion and exclusion criteria. In addition, the majority of patients in this dataset were male (55.3% in eICU-CRD; 61.6% in MIMIC-III; 65.0% in MIMIC-IV).

As noted in Table 2, the distribution of patients among different regions of the US varied by database, with the Midwest (%), South (%), and West (%) being the primary regions represented overall. The median admission age was approximately 66.0 years in eICU-CRD and 68.0 years in the two MIMIC databases, respectively. The median admission weight, height, and BMI were consistent across the three databases. The median Charlson comorbidity scores were consistent among MIMIC-IV and eICU-CRD; in MIMIC-III, this score was not available. Length of stay (LoS) measures were more variable across databases; in particular, all forms of LoS were consistently shorter in eICU-CRD than in MIMIC-III and MIMIC-IV ($p < 0.001$, as determined by a Kruskal-Wallis test). In-hospital mortality rates were found to be 17.8% in eICU-CRD, 17.4% in MIMIC-III, and 15.5% in MIMIC-IV ($p < 0.001$, as determined by a Chi-squared test). Patient characteristics by race and ethnicity, and

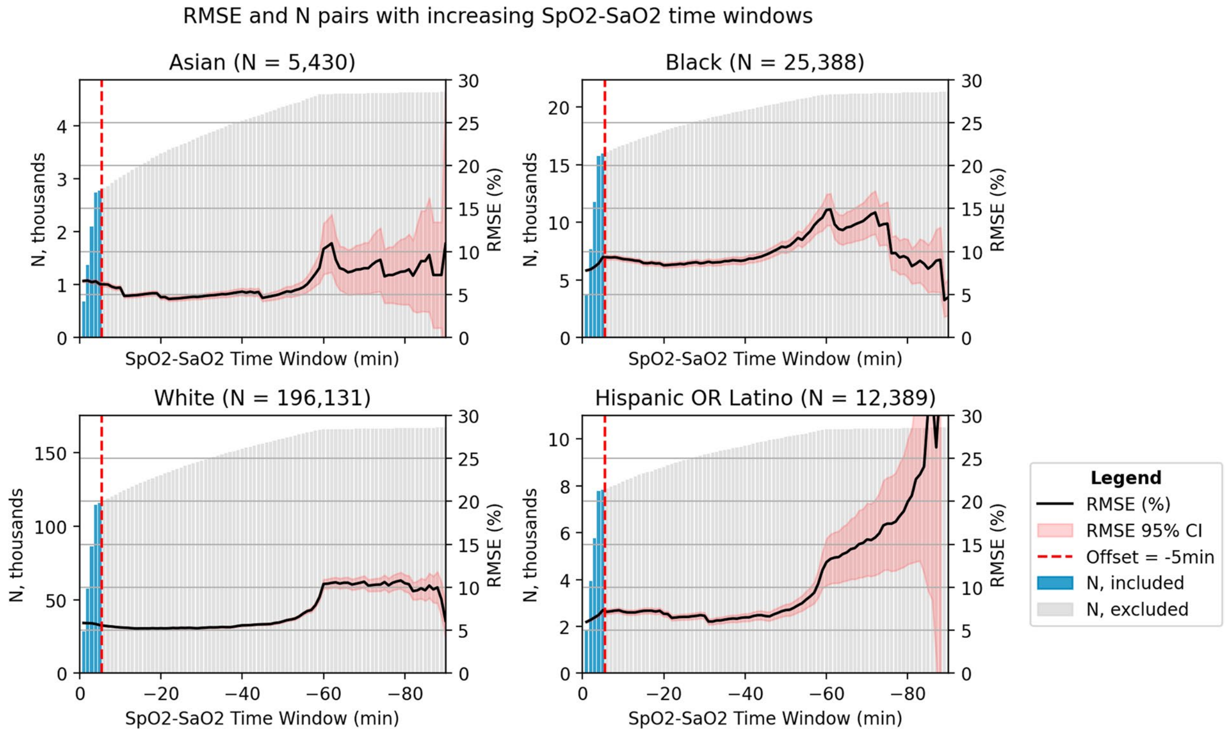

Fig. 5 RMSE and number of pairs with varying window between SaO_2 and SpO_2, per race and ethnicity, on the extended dataset.

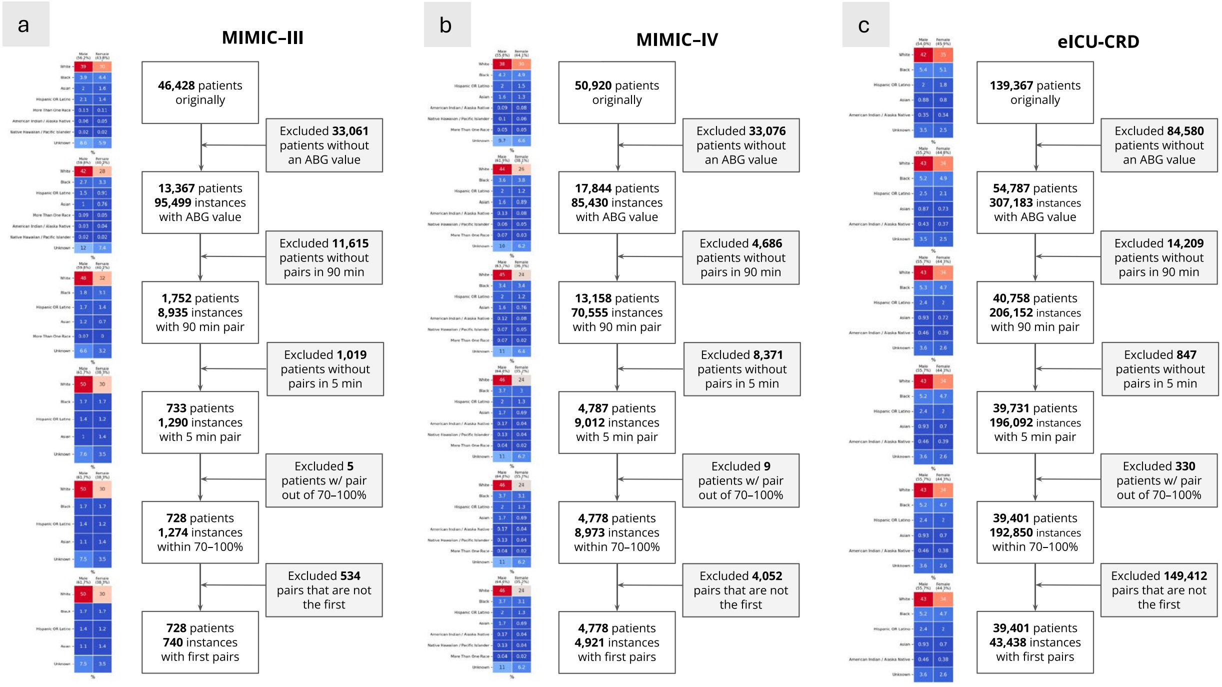
a
MIMIC–III
46,428 patients originally
Excluded 33,061 patients without an ABG value
13,367 patients 95,499 instances with ABG value
Excluded 11,615 patients without pairs in 90 min
1,752 patients 8,935 instances with 90 min pair
Excluded 1,019 patients without pairs in 5 min
733 patients 1,290 instances with 5 min pair
Excluded 5 patients w/ pair out of 70–100%
728 patients 1,274 instances within 70–100%
Excluded 534 pairs that are not the first
728 patients 740 instances with first pairs
b
MIMIC–IV
50,920 patients originally
Excluded 33,076 patients without an ABG value
17,844 patients 85,430 instances with ABG value
Excluded 4,686 patients without pairs in 90 min
13,158 patients 70,555 instances with 90 min pair
Excluded 8,371 patients without pairs in 5 min
4,787 patients 9,012 instances with 5 min pair
Excluded 9 patients w/ pair out of 70–100%
4,778 patients 8,973 instances within 70–100%
Excluded 4,052 pairs that are not the first
4,778 patients 4,921 instances with first pairs
c
eICU-CRD
139,367 patients originally
Excluded 84,580 patients without an ABG value
54,787 patients 307,183 instances with ABG value
Excluded 14,209 patients without pairs in 90 min
40,758 patients 206,152 instances with 90 min pair
Excluded 847 patients without pairs in 5 min
39,731 patients 196,092 instances with 5 min pair
Excluded 330 patients w/ pair out of 70–100%
39,401 patients 192,850 instances within 70–100%
Excluded 149,412 pairs that are not the first
39,401 patients 43,438 instances with first pairs

Fig. 6 **a**. Flow diagram for MIMIC-III depicting cohort selection. **b**. Flow diagram for MIMIC-IV depicting cohort selection. **c**. Flow diagram for eICU-CRD depicting cohort selection.

presence of hidden hypoxemia are present in Supplemental Tables 4 and 5, respectively. The noted differences reflect significant differences across healthcare systems and should be handled carefully when using BOLD.

Static variables, such as the patient's biological sex, or outcomes, like in-hospital mortality, had very few missing values (see Fig. 8 for covariates' completeness). However, time-varying variables, such as laboratory test values, were more sporadic. For lab tests with significant temporal volatility (e.g., lactate), data up to a maximum

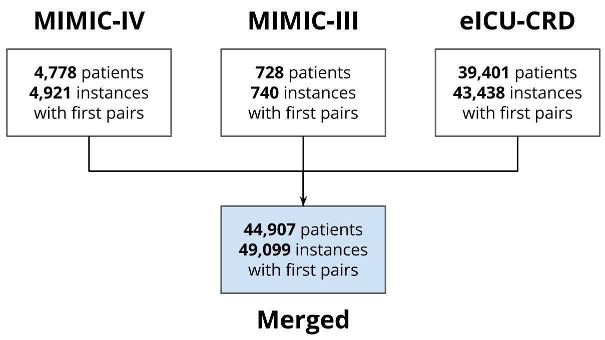

Fig. 7 Flow diagram for the merged dataset.

		eICU-CRD	MIMIC-III	MIMIC-IV
N	**Class**	43,438	740	4,921
Covariates				
Age (admission), median [Q1,Q3]		66.0 [55.0,76.0]	68.0 [58.0,77.0]	68.0 [59.0,77.0]
Race and Ethnicity, N (%)	American Indian / Alaska Native	371 (0.9)	0 (0)	9 (0.2)
	Asian	723 (1.7)	16 (2.2)	119 (2.4)
	Black	4,405 (10.1)	42 (5.7)	336 (6.8)
	Hispanic OR Latino	1,934 (4.5)	20 (2.7)	162 (3.3)
	More Than One Race	0	0	3 (0.1)
	Native Hawaiian / Pacific Islander	0	0	9 (0.2)
	Unknown	2,648 (6.1)	72 (9.7)	846 (17.2)
	White	33,357 (76.8)	590 (79.7)	3,437 (69.8)
Sex N (%)	Female	19,431 (44.7)	284 (38.4)	1,720 (35.0)
BMI (Admission), median [Q1,Q3]		28.0 [23.7,33.7]	28.3 [24.7,33.5]	28.3 [24.7,32.7]
Charlson Comorbidity Index, median [Q1,Q3]		4.0 [2.0,6.0]	N/A	5.0 [3.0,7.0]
Elixhauser Comorbidity Index, median [Q1,Q3]		N/A	9.0 [3.0,16.0]	N/A
Hospital Region, N (%)	Midwest	13,979 (34.5)	N/A	N/A
	Northeast	3397 (8.4)	740 (100.0)	4921 (100.0)
	South	14,018 (34.6)	N/A	N/A
	West	9,069 (22.4)	N/A	N/A
Outcomes				
ICU LoS if dead, median [Q1,Q3], days		3.2 [1.4,6.8]	9.0 [3.0,17]	8.9 [4.1,15]
ICU LoS if survived, median [Q1,Q3], days		2.9 [1.7,5.7]	4.0 [2.0,10]	3.9 [2.0,8.7]
In-Hospital Mortality, N (%)		7,651 (17.8)	129 (17.4)	763 (15.5)

Table 2. Descriptive patient characteristics by individual dataset. LoS, length of stay.

of 4 hours before the SaO_2 measurement were considered. For lab tests often drawn on a daily basis (e.g., basic metabolic panel), this window was extended to 24 hours before baseline. For labs drawn less frequently (e.g., hepatic function panel/complete metabolic panel), data up to 7 days before baseline were included. There were no relevant differences when assessing covariates' completeness across racial and ethnic groups, as depicted in the Supplemental Figs. 1a,b, c, and d.

Researchers should carefully choose the most appropriate time window length for their study, based on our data. If they choose to include repeated measurements of the (SaO_2, SpO_2) pair for the same patient in their analysis, we strongly recommend replicating our data validation steps to mitigate the risk of introducing systematic errors and limit selection bias. To ensure the highest level of data fidelity, we suggest that a cross-disciplinary team of data scientists and domain experts be involved in the data analysis process.

Limitations.

There are several noteworthy limitations associated with the preprocessed dataset presented in this paper that researchers and clinicians should consider.

First, imbalances in the sampling rate of arterial blood gas (ABG) across patient sociodemographics, including by race and ethnicity, limit the potential for downstream model development[2]. Indeed, low ABG sampling rates make it challenging to merge pulse oximeter readings with gold-standard ABG data effectively to characterize outcome heterogeneity in the population and in sufficient quantity to train correction models[2]. For example, the absence of a uniform rate of ABG sampling across sociodemographics may result in the poor estimation of the differential prevalence of hidden hypoxemia in subpopulations, thereby hindering the evaluation

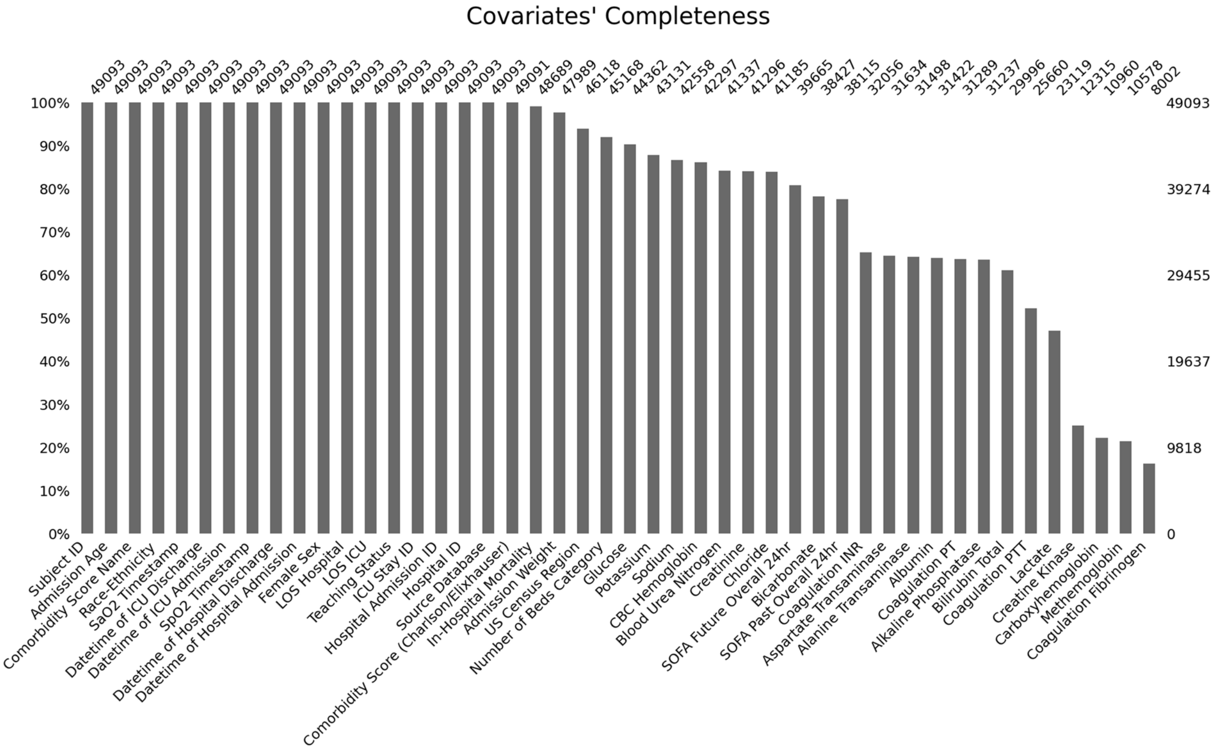

Fig. 8 Completeness of the aligned covariates in the preprocessed dataset.

of disparities and the downstream implementation of subpopulation-specific corrections. This issue adds to the reality of limited patient sample sizes for certain racial and ethnic subgroups, posing a further challenge for recalibration efforts aimed at addressing documented disparities and those yet to be identified.

Second, a more general limitation of EHR data that affects our preprocessed dataset is the lack of objective information on skin tone, a factor known to bias pulse oximeter readings. In fact, some initial studies have suggested that skin tone (on top of self-reported race and ethnicity, when measured using administered visual scales, reflectance colorimetry, or reflectance spectrophotometry) is associated with pulse oximetry discrepancies. However, it is still not clear which method is more suitable in the context of pulse oximetry[20]. As additional studies are needed to determine the contributions of skin tone and other potential confounders on pulse oximetry discrepancies, we advocate for efforts to thoughtfully collect such variables, using a diverse array of methodologies, upon hospital entry in future EHR systems, in order to better address disparities. Finally, as in any EHR-based dataset, we also acknowledge potential errors in data entry.

Strengths. Despite its limitations, our dataset offers several strengths that make it a robust foundation for future research. Notably, it provides a unique platform for quantifying the extent racial and ethnic disparities in intensive care, thereby laying the groundwork for innovative, data-driven solutions to enhance the outcomes of critically-ill patients.

We curated almost fifty thousand rigorously paired (SaO_2, SpO_2) measurements, obtained under strict, clinically relevant criteria. Our curated dataset eliminates a key barrier to entry in the field of data science for critical care. Its creation required the involvement of specialized and multidisciplinary teams of data scientists and clinicians who can navigate complex EHR databases from different health systems. With its public release, our hope is that it will serve as a test bed for future generations of trainees.

The data formats we present are harmonized, user-friendly, and well documented. Original identifiers are kept in the curated dataset to allow the inclusion of further information from the original MIMIC and eICU databases by future users, upon its release in open access.

Finally, to facilitate broader use and encourage careful data engineering, we present what we believe to be a set of best practices in the field of data curation for health equity research. Our methodology for curating EHR data — specifically arterial blood gas and pulse oximetry readings — is fully accompanied by open-source code, which can be easily modified by interested users to accommodate new needs.

Usage Notes

The data of this paper employs three publicly available datasets MIMIC-III, MIMIC-IV, and eICU-CRD, all available on PhysioNet as CSV files, or on Google Cloud BigQuery. The BOLD dataset serves as an authorized extension originating from MIMIC-III, MIMIC-IV, and eICU-CRD, published on PhysioNet. It is duly acknowledged that MIT-LCP and PhysioNet hold the legitimate rights for the redistribution of this extension data set.

BOLD is available on PhysioNet as a credentialled database[19]. To access BOLD, users must be registered on PhysioNet, have proper ethics training, and sign a data use agreement outlining the data usage and security standards, prohibiting any effort to identify the patients of the dataset.

(https://physionet.org/content/blood-gas-oximetry/1.0/)

Please note that the three source datasets are subject to restrictions on re-distribution. This prevents re-publication within secondary datasets as per the license terms (Physionet Data Usage Agreement, version 1.5). To manage this, approval was sought and granted from the original data owner, host institution, and repository to confirm this was possible in this case. This dataset therefore serves as an authorized extension originating from MIMIC-III, MIMIC-IV, and eICU-CRD, published on PhysioNet. It is duly acknowledged that MIT-LCP and PhysioNet hold the legitimate rights for the redistribution of this extension data set. This derived dataset is published under the same license as those source outputs (Physionet Data Usage Agreement, version 1.5)

We also share on GitHub all the code to recreate the dataset curation process.

(https://github.com/joamats/pulse-ox-dataset)

The 1_dataset.ipynb notebook contains all the necessary queries optimized to be used on Google's BigQuery (SQL standard) to generate the final CSV file. We did softcode the important inclusion criteria of lower SaO_2, upper SaO_2, and lower and upper time windows to facilitate any changes to these key parameters. Analysts need to make sure they set up a BigQuery project according to the instructions in our notebook. We also share the notebooks 2_consort_diagrams.ipynb, 3_tableones.ipynb, and 4_technical_validation.ipynb to recreate all the analyses provided in this paper.

(https://docs.google.com/spreadsheets/d/1W4PS3__-jF3m8OemERsv2r_b9sfACWIr-JQcPxW2A7)

This Google's Spreadsheet file contains the details for all time-varying variables that are encoded, as well as the necessary field for them to be pulled from the databases. These details can be changed either globally, or separately for each database.

(https://docs.google.com/spreadsheets/d/1Hv_sOd0–6TPYiB3Crjdn_JrIhIazXXJc05mL4GefOU)

Finally, this Google's Spreadsheet file contains the unified mappings for the static variables (for reference), as well as the race and ethnicity mappings (which are then fed to the created scripts).

Code availability

All code used for data extraction, processing, visualization, and technical validation is available as SQL queries (Google's Bigquery syntax) and Jupyter notebooks in the corresponding PhysioNet page and on GitHub.

https://github.com/joamats/pulse-ox-dataset

1. The publicly-available scripts are structured as follows: 1. The folders MIMIC-III, MIMIC-IV, and eICU-CRD contain the SQL queries to fetch the data, alongside auxiliary tables that need to be created first in a user's BigQuery environment. 2. The source folder contains the Jupyter notebook (1_dataset.ipynb) to create the dataset, which is calling the main SQL scripts needed to create the final CSV file. It also contains the notebooks *2_CONSORT_diagram.ipynb*, *3_tableones.ipynb*, *4_technical_validation.ipynb, and 5_missingness.ipynb* to recreate all the analyses.
2. The source folder contains the Jupyter notebook (1_dataset.ipynb) to create the dataset, which is calling the main SQL scripts needed to create the final CSV file. It also contains the notebooks *2_CONSORT_diagram.ipynb*, *3_tableones.ipynb*, *4_technical_validation.ipynb, and 5_missingness.ipynb* to recreate all the analyses.

Received: 11 October 2023; Accepted: 4 April 2024;
Published online: 24 May 2024

References

1. Sjoding, M. W., Dickson, R. P., Iwashyna, T. J., Gay, S. E. & Valley, T. S. Racial Bias in Pulse Oximetry Measurement. *N. Engl. J. Med.* **383**, 2477–2478 (2020).
2. Wong, A. I. *et al.* Analysis of discrepancies between pulse oximetry and arterial oxygen saturation measurements by Race/Ethnicity and association with organ dysfunction and mortality. *JAMA Network Open*, https://doi.org/10.1001/jamanetworkopen.2021.31674 (2021).
3. Valbuena, V. S. M. *et al.* Racial bias and reproducibility in pulse oximetry among medical and surgical inpatients in general care in the Veterans Health Administration 2013-19: multicenter, retrospective cohort study. *BMJ* **378**, e069775 (2022).
4. Gottlieb, E. R., Ziegler, J., Morley, K., Rush, B. & Celi, L. A. Assessment of Racial and Ethnic Differences in Oxygen Supplementation Among Patients in the Intensive Care Unit. *JAMA Internal Medicine* vol. 182 849 https://doi.org/10.1001/jamainternmed.2022.2587 (2022).
5. Pollard, T. J., Johnson, A. E. W., Raffa, J. & Badawi, O. The eICU Collaborative Research Database. *physionet.org* https://doi.org/10.13026/C2WM1R (2017).
6. Johnson, A. E. W. *et al.* MIMIC-III, a freely accessible critical care database. *Sci Data* **3**, 160035 (2016).
7. Johnson, A. *et al.* MIMIC-IV. *PhysioNet*, https://doi.org/10.13026/S6N6-XD98 (2021).
8. Matos, J. *et al.* Shining Light on Dark Skin: Pulse Oximetry Correction Models. in *2023 IEEE 7th Portuguese Meeting on Bioengineering (ENBENG)* 211–214 (2023).

9. Johnson, A. E., Stone, D. J., Celi, L. A. & Pollard, T. J. The MIMIC Code Repository: enabling reproducibility in critical care research. *J. Am. Med. Inform. Assoc.* **25**, 32–39 (2018).
10. Gruber, P., Kwiatkowski, T., Silverman, R., Flaster, E. & Auerbach, C. Time to equilibration of oxygen saturation using pulse oximetry. *Acad. Emerg. Med.* **2**, 810–815 (1995).
11. Cakar, N. *et al.* Time required for partial pressure of arterial oxygen equilibration during mechanical ventilation after a step change in fractional inspired oxygen concentration. *Intensive Care Med.* **27**, 655–659 (2001).
12. Flanagin, A., Frey, T. & Christiansen, S. L., AMA Manual of Style Committee. Updated Guidance on the Reporting of Race and Ethnicity in Medical and Science Journals. *JAMA* **326**, 621–627 (2021).
13. van Walraven, C., Austin, P. C., Jennings, A., Quan, H. & Forster, A. J. A modification of the Elixhauser comorbidity measures into a point system for hospital death using administrative data. *Med. Care* **47**, 626–633 (2009).
14. Charlson, M. E., Pompei, P., Ales, K. L. & MacKenzie, C. R. A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. *J. Chronic Dis.* **40**, 373–383 (1987).
15. Vincent, J. L. *et al.* The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine. *Intensive Care Med.* **22**, 707–710 (1996).
16. Pollard, T. J., Johnson, A. E. W., Raffa, J. D. & Mark, R. G. tableone: An open source Python package for producing summary statistics for research papers. *JAMIA Open* **1**, 26–31 (2018).
17. Collins, J.-A., Rudenski, A., Gibson, J., Howard, L. & O'Driscoll, R. Relating oxygen partial pressure, saturation and content: the haemoglobin-oxygen dissociation curve. *Breathe (Sheff)* **11**, 194–201 (2015).
18. Bilogur, A. Missingno: a missing data visualization suite. *J. Open Source Softw.* **3**, 547 (2018).
19. Matos, J. *et al.* BOLD, a blood-gas and oximetry linked dataset. *PhysioNet* https://doi.org/10.13026/phvt-3277 (2023).
20. Hao, S. *et al.* Utility of skin tone on pulse oximetry in critically ill patients: a prospective cohort study. *medRxiv*, https://doi.org/10.1101/2024.02.24.24303291 (2024).

Acknowledgements

We would like to thank Dr. Tom Pollard for providing thoughtful and constructive suggestions for improving the dataset. J.M. was supported by a Fulbright / FLAD Grant, Portugal, AY 2022/2023. T.S. is supported by the Swiss National Science Foundation, P400PM_194497 / 1). M.L.C. is supported by a doctoral fellowship from the Eric and Wendy Schmidt Center of the MIT-Harvard Broad Institute. J.G. and L.A.C. are supported by the NIBIB, under R01 EB001659. A.I.W. is supported by the Duke CTSI by the National Center for Advancing Translational Sciences of the NIH under UL1TR002553 and the National Institute on Minority Health and Health Disparities REACH Equity Award under 5U54MD012530.

Author contributions

All authors contributed to writing the manuscript. J.M., T.S., and A.I.W. collaborated on the data extraction, visualization, and analysis. J.M., T.S., J.G., L.N., M.L.C., X.L., J.S.C., N.E.Z., K.S.J., N.B., and J.G. interpreted, validated results, design of the work and supervised data extraction. L.A.C. and A.I.W. reviewed the paper and supervised the work.

Competing interests

A.I.W. holds equity and management roles in Ataia Medical. All other authors report no conflicts of interest.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-024-03225-z.

Correspondence and requests for materials should be addressed to J.M. or A.-K.I.W.