apm ANESTHESIA & PAIN MEDICINE

Anesth Pain Med 2023;18:213-219
https://doi.org/10.17085/apm.23076
pISSN 1975-5171 • eISSN 2383-7977

Review

Open datasets in perioperative medicine: a narrative review

Leerang Lim and Hyung-Chul Lee

Department of Anesthesiology and Pain Medicine, Seoul National University College of Medicine, Seoul National University Hospital, Seoul, Korea

Received June 26, 2023
Revised July 9, 2023
Accepted July 10, 2023

Corresponding author
Hyung-Chul Lee, M.D., Ph.D.
Department of Anesthesiology and Pain Medicine, Seoul National University College of Medicine, Seoul National University Hospital, 101 Daehak-ro, Jongno-gu, Seoul 03080, Korea
Tel: 82-2-2072-0723
Fax: 82-2-747-8363
E-mail: vital@snu.ac.kr

With the growing interest of researchers in machine learning and artificial intelligence (AI) based on large data, their roles in medical research have become increasingly prominent. Despite the proliferation of predictive models in perioperative medicine, external validation is lacking. Open datasets, defined as publicly available datasets for research, play a crucial role by providing high-quality data, facilitating collaboration, and allowing an objective evaluation of the developed models. Among the available datasets for surgical patients, VitalDB has been the most widely used, with the Medical Informatics Operating Room Vitals and Events Repository recently launched and the Informative Surgical Patient dataset for Innovative Research Environment expected to be released soon. For critically ill patients, the available resources include the Medical Information Mart for Intensive Care, the eICU Collaborative Research Database, the Amsterdam University Medical Centers Database, and the High time Resolution ICU Dataset, with the anticipated release of the Intensive Care Network with Million Patients' information for the AI Clinical decision support system Technology dataset. This review presents a detailed comparison of each to enrich our understanding of these open datasets for data science and AI research in perioperative medicine.

Keywords: Artificial intelligence; Big data; Critical care; Data science; Machine learning; Perioperative medicine.

INTRODUCTION

There has been a significant increase in the development of machine learning and artificial intelligence (AI) models in medicine. Despite the proliferation of AI-based predictive models, there is a significant lack of external validation [1]. Before these models can be implemented in clinical practice, it is crucial to perform reproducible validation. Open datasets, which are publicly available datasets for research, play a critical role by providing researchers with high-quality medical data, facilitating collaboration, and enabling the objective evaluation of performance metrics for the developed models.

Previous efforts in outcome research have resulted in various registries such as the National Surgical Quality Program, American Society of Anesthesiologists Closed Claims Project, Multicenter Perioperative Outcomes Group, National Anesthesia Clinical Outcomes Registry in the United States (US), Critical Care Minimum Data Set, and National Perioperative Data Standard Program by the National Health Service of the United Kingdom [2]. However, most registries have limited access to data within the participating institutions or societies involved in data collection. On the other hand, open datasets can be accessed by any credentialed researcher who agrees to a data use agreement.

Due to the nature of medical data, the risk of re-identifica-

tion may remain in open datasets; thus, data use agreements typically include a prohibition on unauthorized use for purposes other than research and restrictions on re-identification and redistribution. According to a US government website [3], over 10 million medical records are breached annually. However, reports on re-identification of open datasets are extremely rare. Therefore, when considering the risks and benefits of releasing open datasets, potential problems such as the development of biased models and the proliferation of unvalidated models can outweigh the risk of re-identification.

There is a great example of how open datasets can be used to change medical practices. In 2020, researchers from the University of Michigan used their database, along with the eICU Collaborative Research Database (eICU-CRD), to analyze data from 10,789 patients. They reported that oxygen saturation levels measured using pulse oximetry were higher than the actual ABGA results in Black patients than in White patients, resulting in a higher incidence of occult hypoxemia [4]. Subsequently, another group conducted a causal analysis using the Medical Information Mart for Intensive Care (MIMIC)-IV dataset and found that Black patients were less likely to receive oxygen supplementation and mechanical ventilation, which was attributed to the overestimation of oxygen saturation levels by pulse oximetry [5]. Although this issue has been raised for decades [6], recent discussions on using large open datasets during the Coronavirus disease (COVID-19) pandemic have attracted attention. Consequently, the US Food and Drug Administration reviewed the regulation of pulse oximeters.

In this review, we presented a detailed comparison of each open dataset to enrich our understanding of these datasets for data science and AI research in perioperative medicine.

SURGICAL DATASETS

Although relatively few open datasets for surgical patients are available, the VitalDB has been widely used in numerous studies. Additionally, there is a recently released dataset or an upcoming dataset that has recently been scheduled to be released.

VitalDB

VitalDB is a single-center perioperative open dataset that includes 6,388 noncardiac surgeries performed under anesthesia at the Seoul National University Hospital in South Korea between August 2016 and June 2017 [7]. The dataset primarily consists of physiological and high-resolution vital sign data along with clinical information and laboratory results (Fig. 1).

One of the key features of VitalDB is its high temporal resolution, which distinguishes it from other datasets. The authors used a free research software called Vital Recorder [8] to achieve a significantly higher time resolution of the dataset. The waveform data had a high resolution (62.5-500 Hz), while the numerical data were recorded at 1–7 s intervals. The physiological signal data in VitalDB include data collected in real time from patient monitors and anesthesia machines, as well as data extracted from various medical devices, including electroencephalogram monitors, regional cerebral oxygen saturation monitors, specialized hemodynamic monitors, and drug infusion devices. VitalDB provides 12 types of waveform data, including photoplethysmography, invasive arterial pressure waveforms, and processed electroencephalogram waveforms, as well as 184 types of numerical data from heart rate, blood pressure, and pulse oximetry. Additionally, the dataset includes demographic information, such as age, sex, height, and weight, as well as clinical information, such as surgical procedures, preoperative comorbidities, and 34 types of blood tests performed from 3 months before surgery to 3 months after surgery. These clinical and laboratory data are provided in a comma-separated value (CSV) format.

The dataset can be accessed and downloaded from a website (https://vitaldb.net) after signing up and signing a data usage agreement without any user credentials. To facilitate easy utilization of the waveform and numerical data in VitalDB, researchers can utilize the open-source VitalDB Python library and PyVital, another open-source biosignal analysis library, to calculate secondary variables such as pulse pressure variation or heart rate variability.

INSPIRE dataset

The Informative Surgical Patient dataset for Innovative Research Environment (INSPIRE) dataset is a single-center perioperative open dataset that includes approximately 260,000 patients who underwent anesthesia for surgery at the Seoul National University Hospital in Korea over 10 years, from January 2011 to December 2020. Data were extracted from the clinical data warehouse of the hospital in six tables linked by a patient identifier (Fig. 1). Each patient has a unique subject identifier and more than one hospital

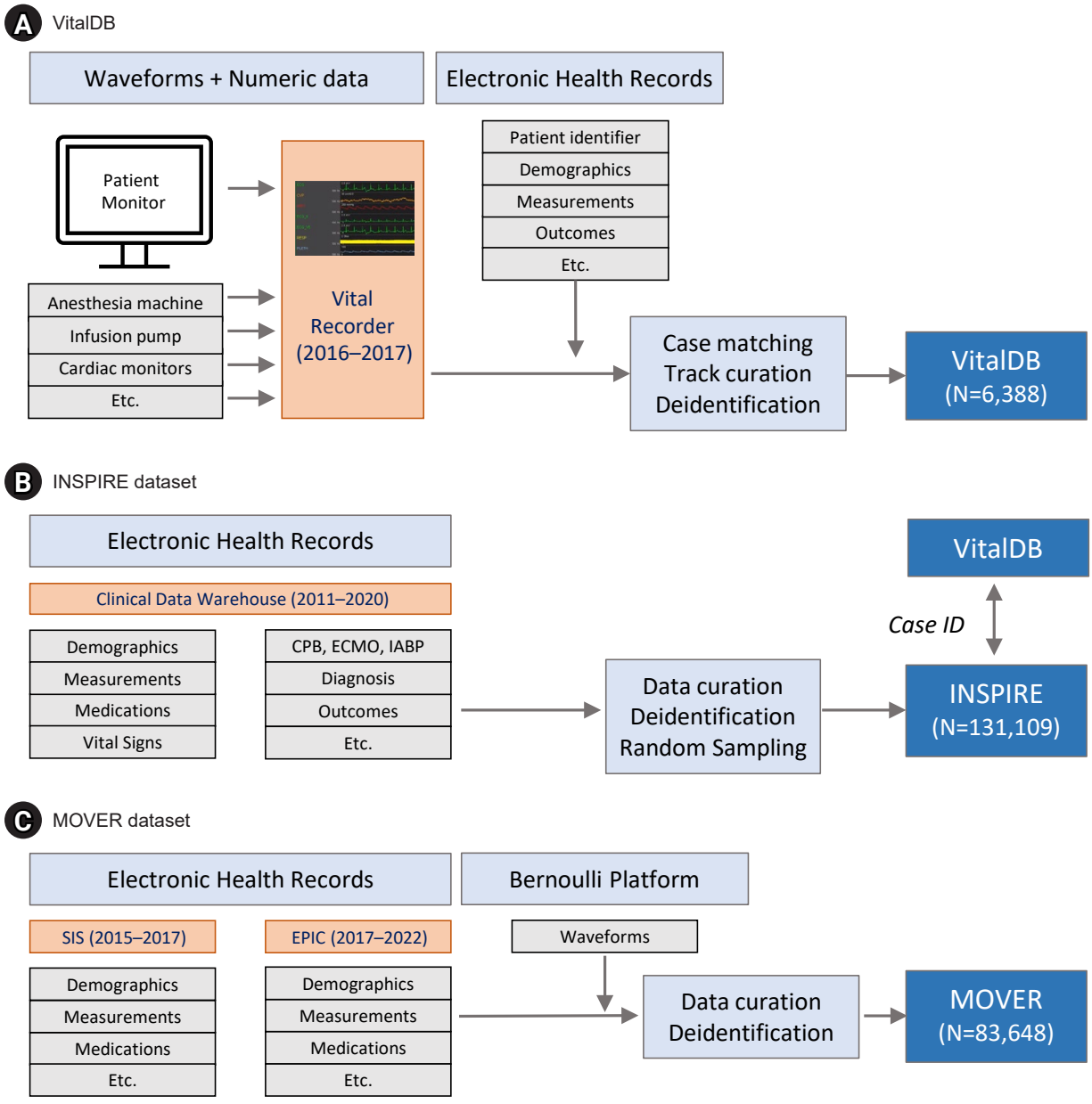

Fig. 1. Schematic representation of data extraction and creation of surgical datasets: (A) VitalDB, (B) INSPIRE dataset, and (C) MOVER dataset. INSPIRE: the informative surgical patient dataset for innovative research environment, CPB: cardiopulmonary bypass, ECMO: extracorporeal membrane oxygenation, IABP: intra-aortic balloon pump, SIS: surgical information systems, EPIC: Epic Systems, MOVER: Medical Informatics Operating Room Vitals and Events Repository.

admission identifier, and each operation has an operation identifier. The dataset includes various patient characteristics such as age, sex, American Society of Anesthesiologists classification, diagnosis, and surgical information such as surgical procedure, surgical department, and anesthesia duration. It also encompasses vital signs measured during surgery by patient monitoring and using an anesthetic machine, and is recorded from admission to discharge in regular wards or intensive care units (ICU). Additionally, the dataset provides laboratory results from six months before the first admission to six months after the last discharge and medication records during the hospitalization period. Complication information includes ICU admissions, total hospital and ICU lengths of stay, and in-hospital deaths.

The vital signs in the dataset have a maximum temporal resolution of 5 min. The laboratory results and medication

records include all predefined parameters conducted during that time. All time-related variables were transformed into relative time (min), with the first admission of each subject set to 0 to protect personal information. Diagnoses and surgical procedures are partially provided in the form of the International Classification of Diseases 10th revision (ICD-10) codes. Certain diagnoses that required special protection, such as mental and behavioral disorders and sexually transmitted infections, are also excluded.

Because the INSPIRE dataset has an overlapping patient population with VitalDB [7], it provides a linker that can be used for matching with VitalDB. Using this information, researchers can utilize pre- and postoperative vital signs, medication history, and the use of special devices in the ICU (such as continuous renal replacement therapy and extracorporeal membrane oxygenation) that are not included in the existing VitalDB. Moreover, the dataset includes rare occurrences such as patient mortality and the use of special devices, making it suitable for the external validation of various prediction models for surgical patients. However, the dataset consists of single institutional data from a single ethnicity. Therefore, caution should be exercised when validating predictive models developed for different populations.

The INSPIRE dataset will be released via the website https://inspire.or.kr, or PhysioNet [9] by 2023. It will include approximately 130,000 surgical cases randomly selected from 260,000 patients. User credentialing will be performed through PhysioNet and made available for research purposes only to researchers who sign the data use agreements.

MOVER Dataset

The Medical Informatics Operating Room Vitals and Events Repository Dataset is a single-center perioperative open dataset that includes surgeries conducted on approximately 59,000 patients, accounting for approximately 83,000 surgical cases at the University of California, Irvine Medical Center from 2015 to 2022 [10]. The dataset is divided into two parts: the Surgical Information Systems (SIS, Surgical Information Systems, USA) and the EPIC (Epic Systems, USA) datasets (Fig. 1). The SIS dataset consists of nine tables, whereas the EPIC dataset consists of ten tables. The collected data is primarily composed of waveform and alphanumeric data. Waveform data include electrocardiograms (ECG), photoplethysmography, and invasive arterial pressure measurements. The alphanumeric data includes variables such as age, sex, underlying conditions, surgical clinical information, vital signs, medication history during surgery, input/output information, intubation, various administrations, and drainage tube placement.

The SIS and EPIC datasets differ slightly in terms of included data. While demographics, medications, and vital signs are included in both datasets, the SIS dataset includes anesthesia machine data, such as ventilator settings, whereas the EPIC dataset includes information on drainage tubes, arterial and venous lines, intubation, complications, and billing codes. Furthermore, the SIS dataset has only a surgical identifier without a patient identifier, whereas the EPIC dataset includes both patient and hospital visit identifiers, allowing the tracking of patients who underwent multiple surgeries. The maximum resolution of the numerical data is 1 min.

The MOVER dataset is the largest perioperative open dataset currently available. As mentioned previously, it encompasses a wide range of data, including vital signs, laboratory results, medications, and information on drainage tubes. This dataset provides extensive research opportunities in various domains. Specifically, it includes a comprehensive categorization of postoperative complications into airway, respiratory, and cardiovascular complications, including various complications such as laryngospasm and cardiac arrest. This makes it suitable for the development of predictive models of postoperative complications. However, some data are presented in a free-text format, requiring individual researchers to perform data preprocessing as part of their analysis.

The MOVER dataset can be accessed through its website (https://mover.ics.edu). By completing the data usage agreement on the website, individuals receive an ID, password, and address to download the data via a registered e-mail.

INTENSIVE CARE DATASETS

In a recent systematic review [11], four publicly available datasets were identified for adult ICU patients: MIMIC, eICU-CRD, Amsterdam University Medical Centers Database (AmsterdamUMCdb), and the High time Resolution ICU Dataset (HiRID). Soon, we expect the release of the Intensive Care Network with Million Patients' information for the AI-CDSS Technology (IMPACT) dataset, a multicenter dataset from Korea that is anticipated to facilitate a wide range of research studies.

MIMIC

The Laboratory for Computational Physiology (LCP) at the Massachusetts Institute of Technology (MIT) has actively released numerous open datasets since the 1980s, starting with the MIT-BIH arrhythmia dataset [12]. In 1996, researchers expanded their dataset by collecting multichannel data beyond ECG in the ICU. They named it Multi-Parameter Intelligent Monitoring for Intensive Care (MIMIC) and later changed it to Medical Information Mart for Intensive Care. This was the first publicly available multi-parameter dataset for critically ill patients, containing 20 h of data from 90 patients, including electrocardiogram, arterial blood pressure, pulmonary artery pressure, and photoplethysmography.

MIMIC-II, released between 2001 and 2008, provided ICU data matched with clinical records and social security death data. In 2013, the MIMIC-III was introduced, featuring a large-scale dataset of over 40,000 patients spanning more than 10 years [13]. The MIMIC-III also included a waveform dataset called the MIMIC-III Waveform Database, which contained a matched subset of 22,317 waveform records and 22,247 numerical records of 10,282 distinct ICU patients.

The MIMIC-IV, released in March 2021, consists of 299,712 patients including those in the emergency department [14]. Additionally, it provides interpretive reports of chest radiography (MIMIC-CXR) [15], 12-Lead ECG (MIMIC-ECG) [14], and echocardiography (MIMIC-Echo) data. The MIMIC dataset has become the most widely cited open dataset for critical care, turning the Beth Israel Deaconess Hospital into one of the most intensively studied critical care cohorts worldwide.

The MIMIC dataset is available through PhysioNet after user credentialing and signing of the data-use agreement. A training program called the Collaborative Institutional Training Initiative (CITI) is required for user credentialing.

IMPACT dataset

In 2021, the Korean government initiated the Korean Medical Information Market for Intensive Care (K-MIMIC) Project to establish a multi-institutional open dataset for intensive care. The IMPACT consortium, responsible for the K-MIMIC project, plans to collect data sourced from 19 hospitals nationwide from over 500,000 individuals by 2023 and expand it to over 1 million individuals by 2025. This dataset will be made available to researchers, both domestically and internationally. It also aims to ensure compatibility with the structure and format of the MIMIC dataset from the US, allowing for sharing of the source code. Given the limited availability of ICU datasets in Asian countries, the release of such datasets has the potential to address data imbalance issues and serve as a valuable resource for research and development in critical care medicine.

eICU-CRD

The eICU-CRD dataset, released in 2018, is a multi-institutional dataset that includes data from 335 ICUs across 208 hospitals in the US as part of the Philips eICU program [16]. The main advantage of this dataset is that it is a multi-institutional dataset covering the entire US, allowing for an assessment of model generalizability. This dataset includes data from 200,859 critically ill patients hospitalized between 2014 and 2015. It comprised admission and discharge records, medication administration records, laboratory results, diagnoses, nursing notes, and treatment records. The laboratory results encompass approximately 160 standard laboratory tests, whereas the vital sign data includes 16 variables such as blood pressure, heart rate, oxygen saturation, and body temperature recorded at 5-min intervals. However, there may be variations in the recording frequency across different institutions. Unlike the MIMIC dataset, the eICU-CRD dataset did not include patients who are also included in the MIMIC dataset, allowing cross-validation between the two datasets.

Similar to the MIMIC dataset, the eICU-CRD dataset is also available through PhysioNet after user credentialing and signing a data use agreement. For user credentialing, the CITI must be completed.

AmsterdamUMCdb

The AmsterdamUMCdb is a single-center open dataset for critical care based on data collected from 44 surgical and high-dependency intensive care beds at the Amsterdam University Medical Center in the Netherlands from 2003 to 2017 [17]. After its initial release in 2019, the dataset was further updated, and the final version, 1.0.2, which included 23,106 cases of intensive care admissions, was released in March 2020. The AmsterdamUMCdb consists of seven alphanumeric data tables provided in CSV format. It includes demographic information as well as data related to medication history, various test results, vital signs, procedures, and special interventions during the ICU stay.

The AmsterdamUMCdb is the first open dataset for critical care collected and released outside the US, contributing to increased diversity in terms of race and population groups [17]. The final dataset was de-identified to comply with the more stringent requirements of the general data protection regulation, surpassing the health insurance portability and accountability act standards. Prior to its release, the dataset was thoroughly reviewed and approved by external experts in various fields, including privacy and ethics. In particular, it includes an assessment of the risk of re-identification under different re-identification attack scenarios, demonstrating significantly lower re-identification risks than de-identification based on the HIPAA Privacy Rule.

AmsterdamUMCdb dataset can be requested through the Amsterdam Medical Data Science website (https://amsterdammedicaldatascience.nl/) and is available after signing a data use agreement and user credentialing, including completion of the CITI course.

HiRID

The HiRID dataset is the second open dataset for intensive care in Europe, following the release of the AmsterdamUMCdb in 2021 [18]. It is a single-center ICU dataset from Bern University Hospital in Switzerland, which includes records of approximately 34,000 patients admitted to the ICU between 2008 and 2016. The dataset was developed through the collaboration between the Swiss Federal Institute of Technology (ETH) Zürich and Bern University Hospital's ICUs. The dataset consists of more than 600 variables, including basic demographic data, vital signs, laboratory results, medication records, and fluid and nutritional data. It is also divided into two main categories: Raw and preprocessed data. The Raw data corresponds to the original dataset consisting of three tables. In contrast, Preprocessed data is derived from the preprocessing phase described in the original article study [18]. This involves merging multiple variables representing a single clinical concept into a metavariable, resulting in a set of 18 metavariables.

The HiRID dataset has a higher temporal resolution than other datasets, particularly in terms of vital sign data, which are recorded at 2-min interval. This followed the anonymization strategies employed for the MIMIC-III and AmsterdamUMCdb datasets. Time information has been modified by setting the initial admission time to an arbitrary date between 2,100 and 2,200, while maintaining seasonality, date, and time consistency.

Compared to other open datasets, the HiRID dataset was not originally developed for publication but as a step toward building a prediction model using machine learning [18]. Therefore, preprocessed data are provided, enabling researchers to quickly assess the feasibility of constructing an AI model at the initial stage of individual research.

The dataset was downloaded from PhysioNet. Users must complete a data-use agreement and the CITI training course, followed by user credentialing to access data. Compared to other open datasets, users are also required to submit their research plans with a dataset request. The dataset and corresponding code for predictive modeling can be found at https://github.com/HIRID/HiRID_v1.

CONCLUSION

The advancement of scientific knowledge has been achieved through the accumulation of reproducible knowledge. Research that cannot be reproduced is not considered scientific, because it cannot be refuted. However, many medical studies are difficult to reproduce and validate owing to the inability to share data publicly [19], often due to privacy concerns. Nevertheless, similar to all other sciences, medical research needs to be transparent and reproducible, from data collection to final performance evaluation, so that future generations can build on and improve it. Thus, we can accelerate the advancement of medical knowledge and ensure the development of reliable and well-validated predictive models for use in clinical practice. Open perioperative datasets play a crucial role in knowledge advancement. Therefore, we hope that more open datasets will be released and that research using them will become more active. Standing on the shoulders of giants, we can observe this further.

FUNDING

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean Government (MSIT) (No. NRF-2020R1C1C1014905).

CONFLICTS OF INTEREST

No potential conflict of interest relevant to this article was reported.

DATA AVAILABILITY STATEMENT

All datasets cited in this study are publicly available and can be downloaded from the address provided in the manuscript.

AUTHOR CONTRIBUTIONS

Conceptualization: Leerang Lim, Hyung-Chul Lee. Formal analysis: Leerang Lim, Hyung-Chul Lee. Writing – original draft: Leerang Lim. Writing – review & editing: Hyung-Chul Lee. Supervision: Hyung-Chul Lee.

ORCID

Leerang Lim, https://orcid.org/0000-0002-4015-8123
Hyung-Chul Lee, https://orcid.org/0000-0003-0048-7958

REFERENCES

1. Yoon HK, Yang HL, Jung CW, Lee HC. Artificial intelligence in perioperative medicine: a narrative review. Korean J Anesthesiol 2022; 75: 202-15.
2. Sessler DI. Big Data–and its contributions to perioperative medicine. Anaesthesia 2014; 69: 100-5.
3. Notice to the Secretary of HHS Breach of Unsecured Protected Health Information. U.S. Department of Health & Human Services [Internet]. [cited 2023 Jun 25]. Available from https://ocrportal.hhs.gov/ocr/breach/breach_report.jsf.
4. Sjoding MW, Dickson RP, Iwashyna TJ, Gay SE, Valley TS. Racial bias in pulse oximetry measurement. N Engl J Med 2020; 383: 2477-8.
5. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. Nat Med 2018; 24: 1716-20.
6. Jiang LY, Liu XC, Nejatian NP, Nasir-Moin M, Wang D, Abidin A, et al. Health system-scale language models are all-purpose prediction engines. Nature. 2023. doi: 10.1038/s41586-023-06160-y. [Epub ahead of print].
7. Lee HC, Park Y, Yoon SB, Yang SM, Park D, Jung CW. VitalDB, a high-fidelity multi-parameter vital signs database in surgical patients. Sci Data 2022; 9: 279.
8. Lee HC, Jung CW. Vital Recorder-a free research tool for automatic recording of high-resolution time-synchronized physiological data from multiple anaesthesia devices. Sci Rep 2018; 8: 1527.
9. Moody GB, Mark RG, Goldberger AL. PhysioNet: a research resource for studies of complex physiologic and biomedical signals. Comput Cardiol 2000; 27: 179-82.
10. Samad M, Rinehart J, Angel M, Kanomata Y, Baldi P, Cannesson M. MOVER: Medical Informatics Operating Room Vitals and Events Repository. medRxiv 2023; 2023.03.03.23286777.
11. Sauer CM, Dam TA, Celi LA, Faltys M, De La Hoz MAA, Adhikari L, et al. Systematic review and comparison of publicly available ICU aata sets-a decision guide for clinicians and data scientists. Crit Care Med 2022; 50: e581-8.
12. Moody GB, Mark RG. The impact of the MIT-BIH arrhythmia database. IEEE Eng Med Biol Mag 2001; 20: 45-50.
13. Johnson AEW, Pollard TJ, Shen L, Lehman LWH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016; 3: 160035.
14. Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data 2023; 10: 1.
15. Johnson AEW, Pollard TJ, Berkowitz SJ, Greenbaum NR, Lungren MP, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci Data 2019; 6: 317.
16. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multicenter database for critical care research. Sci Data 2018; 5: 180178.
17. Thoral PJ, Peppink JM, Driessen RH, Sijbrands EJG, Kompanje EJO, Kaplan L, 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 2021; 49: e563-77.
18. Hyland SL, Faltys M, Hüser M, Lyu X, Gumbsch T, Esteban C, et al. Early prediction of circulatory failure in the intensive care unit using machine learning. Nat Med 2020; 26: 364-73.
19. Stupple A, Singerman D, Celi LA. The reproducibility crisis in the age of digital medicine. NPJ Digit Med 2019; 2: 2.