www.nature.com/scientificdata

SCIENTIFIC DATA

OPEN

DATA DESCRIPTOR

PIC, a paediatric-specific intensive care database

Xian Zeng[1,2,3], Gang Yu[1,3], Yang Lu[2,3], Linhua Tan[1], Xiujing Wu[1], Shanshan Shi[1], Huilong Duan[2], Qiang Shu[1*] & Haomin Li[1*]

PIC (Paediatric Intensive Care) is a large paediatric-specific, single-centre, bilingual database comprising information relating to children admitted to critical care units at a large children's hospital in China. The database is deidentified and includes vital sign measurements, medications, laboratory measurements, fluid balance, diagnostic codes, length of hospital stays, survival data, and more. The data are publicly available after registration, which includes completion of a training course on research with human subjects and signing of a data use agreement mandating responsible handling of the data and adherence to the principle of collaborative research. Although the PIC can be considered an extension of the widely used MIMIC (Medical Information Mart for Intensive Care) database in the field of paediatric critical care, it has many unique characteristics and can support database-based academic and industrial applications such as machine learning algorithms, clinical decision support tools, quality improvement initiatives, and international data sharing.

Background & Summary

Over the past ten years, electronic health records (EHRs) have rapidly been adopted worldwide. In the US, the percentage of non-federal acute care hospitals that have adopted basic EHRs increased from 9.4% in 2008 to 83.8% in 2015[1]. In China, the healthcare system has been undergoing the New Medical Reform that has stimulated several nation-wide initiatives to encourage and assess adoption and implementation of EHRs since 2009[2]. The analysis of data contained in EHRs, which contain a large volume of structured and unstructured information regarding patient care, is a promising avenue of research for clinicians and data scientists[3–5]. However, there are several important barriers to the widespread adoption of big data in health care, including no strong incentives for its use within groups of clinicians, considerable privacy concerns, and limited interoperability[6,7]. Moreover, barriers to data sharing in healthcare have limited the usage of clinical data and largely prevented the reproducibility of published studies[8,9].

Intensive care units (ICUs) provide care for severely ill patients who require invasive life-saving treatment. Large amounts of data are routinely collected for these critical care patients. The widely used MIMIC (Medical Information Mart for Intensive Care) critical care database has been developed for more than a decade contains comprehensive clinical data of patients admitted to the Beth Israel Deaconess Medical Centre in Boston, Massachusetts[10,11]. For many years, it has been the only freely accessible database in critical care medicine, and it supports a broad range of research areas, as evidenced by many publications[12–15]. Building upon the success of MIMIC-III, Philips Healthcare developed a multi-centre intensive care unit database called the eICU Collaborative Research Database[16]. However, all these databases have an obvious age gap in their patients. Although MIMIC-III contains data on neonates in the NICU, comprehensive paediatric patient (age from 0 to 18 years old) data are not available in it. Children are not just 'small adults' and often have different diseases, developmental issues, and possibly differential responses to therapies and the recovery of function when compared with adult patients in critical care[17–20]. There is vast variability in age-related differences that are all linked to the developmental stage and ability of children. Therefore, all clinical evidence should also be validated in children before being applied in paediatrics. Although there are some large-scale paediatric critical care research databases based on case registries, such as the Virtual PICU Systems (VPS) database in the US[21] and the Paediatric Intensive Care Audit Network (PICANet) in the UK[22], there is a lack of a freely accessible database that contains the fidelity and completeness of raw paediatric intensive care clinical data for researchers. A freely accessible paediatric-specific

[1]The Children's Hospital of Zhejiang University School of Medicine and National Clinical Research Center for Child Health, Hangzhou, China. [2]The College of Biomedical Engineering and Instrument Science, Zhejiang University, Hangzhou, China. [3]These authors contributed equally: Xian Zeng, Gang Yu and Yang Lu. *email: shuqiang@zju.edu.cn; hmli@zju.edu.cn

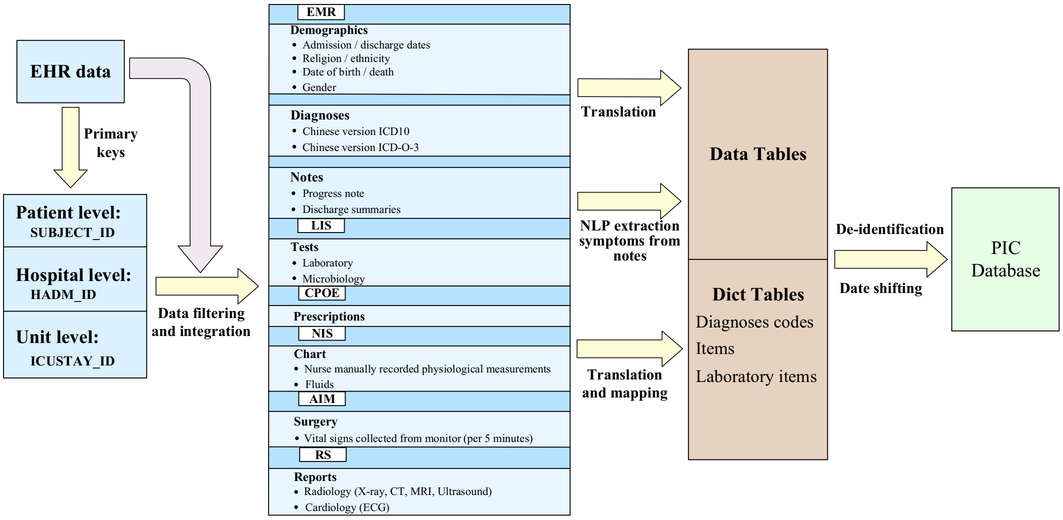

Fig. 1 Overview of the PIC database.

critical care database will stimulate more researchers to pay attention to children's intensive care and improve the quality of critical care for children.

Here, we report the release of the PIC (Paediatric Intensive Care) database, a freely accessible paediatric-specific critical care database (http://pic.nbscn.org). Like the widely used MIMIC database, the PIC database integrates deidentified, comprehensive clinical data of paediatric patients admitted to the Children's Hospital of Zhejiang University School of Medicine and makes it internationally accessible to researchers under a data use agreement (Fig. 1).

The PIC database is notable for the following reasons:

- It is publicly and freely available to researchers worldwide as a paediatric critical care supplement to the MIMIC-III
- It encompasses a very large population of paediatric patients and spans almost a decade;
- It contains high temporal resolution and high-fidelity clinical data;
- It is the first freely accessible English-Chinese bilingual clinical database.

Methods

The PIC database was populated with data that had been acquired during routine hospital care The Children's Hospital, Zhejiang University School of Medicine. This children's hospital, with more than 1900 beds, is the largest comprehensive paediatric medical centre in Zhejiang Province and the Chinese National Clinical Research Centre of Child Health. There are more than 3 million outpatient and emergency visits per year in this children's only medical center. It also accepts critical paediatric patients who referred from lower level hospital across 11 cities of Zhejiang province. It has 119 critical care beds in 5 intensive care units: general ICU, paediatric ICU (PICU), surgical ICU (SICU), cardiac ICU (CICU) and neonatal ICU (NICU). The clinical data of patients admitted to any of the ICUs between 2010 and 2018 was used to construct the PIC database. This project was approved by the Institutional Review Board of the Children's Hospital, Zhejiang University School of Medicine (Hangzhou, China). The requirement for individual patient consent was waived because the project did not impact clinical care, and all protected health information was deidentified.

Database Development

Data were extracted and downloaded from several information systems in the hospital, including the following:

- Hospital electronic medical record system
- Laboratory information system
- Computerized physician order entry system
- Nursing information system
- Anaesthesia information management system
- Reporting system of different examination departments (radiology, ultrasound, ECG, pathology, *et al*.).

The core database schema followed the classic MIMIC-III database with some adjustments to adapt it for local Chinese data. This will help researchers who have experience with MIMIC to easily understand and utilize the PIC and to compare it with the MIMIC database.

All patients who transferred to any of the ICUs and had records in the hospital medical record system between 2010 and 2018 were included in this project. Based on these data, three core tables were created to describe the patients (PATIENTS), admissions (ADMISSIONS) and ICU stays (ICUSTAYS). The primary key of the three core tables was used to index all other clinical data. SUBJECT_ID in the PATIENTS table is a unique identifier

Table name	Description	Rows
ADMISSIONS	Every unique hospitalization for each patient in the database (defines HADM_ID)	13,449
CHARTEVENTS	All charted observations for the patients from the hospital database, not including bedside monitor trends and lab events.	2,278,978
D_ICD_DIAGNOSES	Dictionary of International Statistical Classification of Diseases and Related Health Problems (ICD-10) codes and International Classification of Diseases for Oncology (ICD-O-3) relating to diagnoses.	25,378
D_ITEMS	Dictionary of local codes ('ITEMIDs') appearing in the PIC database, except those that relate to laboratory tests.	466
D_LABITEMS	Dictionary of local codes ('ITEMIDs') appearing in the PIC database that relate to laboratory tests.	832
DIAGNOSES_ICD	Hospital assigned diagnoses, coded using the ICD-10 Chinese code and the ICD-O-3 Chinese code.	13,365
EMR_SYMPTOMS	Structured symptoms extracted from notes, including nursing and physician notes, discharge summaries and so on.	402,142
ICUSTAYS	Every unique ICU stay in the database (defines ICUSTAY_ID).	13,941
INPUTEVENTS	Calculated fluid input data every morning for each patient in ICU.	26,884
LABEVENTS	Laboratory measurements for patients from the hospital database.	10,094,117
MICROBIOLOGYEVENTS	Microbiology culture results and antibiotic sensitivities from the hospital database.	183,869
OR_EXAM_REPORTS	Contains all exams performed during the patient's stay.	183,809
OUTPUTEVENTS	Output information for patients from the hospital database.	39,891
PATIENTS	Every unique patient in the database (defines SUBJECT_ID).	12,881
PRESCRIPTIONS	Medications ordered for a given patient	1,256,591
SURGERY_VITAL_SIGNS	Vital signs recorded every 5 minutes during the surgery	1,216,011

Table 1. An overview of the data tables in the PIC database.

that specifies an individual patient, HADM_ID in the ADMISSIONS table is the encounter number that uniquely identifies a particular hospitalization for patients who might have been admitted multiple times, and ICUSTAY_ID in the ICUSTAYS table refers to a unique admission to an intensive care unit. Each SUBJECT_ID has one or more related HADM_IDs, and each HADM_ID can have one or more related ICUSTAY_ID.

After collecting the original data, we performed post-processing and integrated medical records to ensure that each patient had relatively complete data. Since the time spans of the data in each table from the different systems are inconsistent, we used SUBJECT_ID and HADM_ID, which appear in the ADMISSIONS, PATIENTS, and ICUSTAYS table for data integration and filtering. In parallel, we carefully checked for impossible data entries (for example, the year of the visit date was 1900), erroneous characters (for example, Chinese characters appearing as numbers), and extreme outliers in the collected data and removed or updated these data.

Following the MIMIC-III database schema, structured clinical data, which included patient demographics, medications, fluid balances, comprehensive laboratory results, and microbiological information (tests performed and sensitivities) from the patients' entire hospital stay, not only periods in the ICU, were collected from different systems. Another specially structured type of data is time-series data, such as vital signs directly collected from monitoring devices, as there are no critical care information systems to store the high-frequency automatically collected vital signs. The PIC only used the physiological measurements manually recorded by nurses. However, the PIC integrated the vital signs collected during surgery every 5 minutes from the anaesthesia information management system. To help research to distinguish these vital data from MIMIC, we created a new table named SURGERY_VITAL_SIGNS.

To make this database widely used worldwide, the largest challenge is the language barrier. All the lab test items, medication names, examinations and diagnoses were recorded using Chinese in the original information systems. To address this problem, the PIC provided English and Chinese bilingual dictionary tables to make the data can be interpretable by most researchers. For terms with widely used standard codes, such as the ICD-10 codes for diagnostics, the English terms were based on the corresponding English version of the standard codes. However, the local ICD-10 version is an extended version with 7 characters compared to the standard 5-character ICD-10 code, and therefore, the Chinese diagnosis is more specific than the English diagnosis in the PIC. Except some of lab test items with LOINC code, other clinical terms without a standard code were translated to English and reviewed manually by authors (X.Z., Y.L. and H.L.). We keep the original Chinese terms in all the tables with corresponding English terms and will help researchers identify language issues if there are any. Such a bilingual Chinese-English clinical dictionary table also creates a data environment that can serve other potential international data sharing projects.

The free-text clinical documents and reports were also recorded in Chinese. Translating a large volume of narrative clinical documents and reports is not feasible, and automatically de-identification tools and algorithms for Chinese clinical documents is also not available. We could not release the original clinical documents at this stage. However, much important clinical information is embedded in the narrative of clinical documents, so common symptoms were extracted from clinical progress notes and discharge summaries using Natural Language Processing (NLP) technology[23]. These symptom terms and their negation status (present symptom with "+" and absence symptom with "−") data are combined with the document recording time and published as the EMR_SYMPTOMS table in the first release of the PIC. To initially evaluate the accuracy of the NLP extraction. We randomly selected 100 notes including admission note, discharge note, progress note, transferred note, and so on from the original notes, extracted symptoms from them using NLP technology, and finally resulted in 3410 symptoms. Two authors independently checked the extracted results. The average accuracy is 91.9% (91.3%, 92.4%). Several NLP typical errors should be mentioned. First, wrong segmentation of long Chinese phrases which do not have natural word

Critical care unit	CICU	GICU	NICU	PICU	SICU	Total
Patients, no. (%)	2583 (20.1%)	2642 (20.5%)	3137 (24.4%)	1953 (15.2%)	2566 (19.9%)	12881 (100%)
Admissions, no. (%)	2638 (19.6%)	2725 (20.3%)	3205 (23.8%)	2084 (15.5%)	2797 (20.8%)	13449 (100%)
ICU stay, no. (%)	2803 (20.1%)	2788 (20.0%)	3282 (23.5%)	2166 (15.5%)	2902 (20.8%)	13941 (100%)
Age, years, mean (Q1–Q3)	2.5 (0.5–3.1)	3.8 (0.4–6.3)	14.5 days (0.0–18.0)	3.6 (0.5–5.7)	3.2 (0.4–4.9)	2.5 (0.1–3.3)
Gender, male, % of unit stays	1391 (49.6%)	1712 (61.4%)	1984 (60.5%)	1250 (57.7%)	1680 (57.9%)	8017 (57.5%)
ICU LOS, mean days (Q1–Q3)	3.9 (0.9–4.0)	7.3 (0.9–8.9)	21.6 (2.5–32.8)	9.7 (2.0–11.1)	2.3 (0.8–1.6)	9.3 (0.9–9.2)
HLOS, mean days (Q1–Q3)	16.6 (9.0–19.2)	12.8 (3.8–16.2)	27.1 (9.6–37.8)	14.5 (4.6–16.7)	14.7 (7.0–18.7)	17.6 (7.0–21.0)
ICU mortality, %	48 (1.7%)	414 (14.8%)	236 (7.2%)	200 (9.2%)	57 (2.0%)	955 (6.9%)
Hospital mortality, %	53 (2.0%)	417 (15.3%)	239 (7.5%)	205 (9.8%)	57 (2.0%)	971 (7.2%)

Table 2. Details of the PIC patient population by the first critical care unit on hospital admission. CICU: cardiac intensive care unit; GICU: general intensive care unit; NICU: neonatal intensive care unit; PICU: paediatric intensive care unit; SICU: surgical intensive care unit; LOS: length of stay.

	MIMIC-III	PIC
Language	English	English-Chinese bilingual
Data source	673-bed general hospital with OB/GYN	1900-bed children's hospital
Category of critical care	CCU, CSRU, MICU, SICU, TSICU, NICU Total 77 critical care beds	CICU, SICU, PICU, NICU, GICU Total 119 paediatric critical care beds
Number of records	50,000+	10,000+
Patient age, median (Q1–Q3)	60.6 years (42.1–74.1)	0.8 years (0.1–3.5)
Vital signs	Bedside monitor generated (1 hour)	Daily nurse recorded & surgery monitor generated (5 minute)
Laboratory/clinical data	Yes	Yes
Clinician notes	Yes	Extracted symptoms from notes
Diagnoses codes	ICD-9 Codes	ICD-10 Codes and ICD-O-3 Codes
Mortality	Death recorded in the hospital and social security database	Death recorded in the hospital

Table 3. Differences of the PIC and the MIMIC-III.

space such as English. Second, falsely negative detection both in positive and negative. Third, the complicated sentence that talk about the family symptoms or the risk of some condition were also extracted as symptoms of patient. Taking into account the accuracy of NLP, we remind researchers to use these data with caution.

For many examinations and their reports, the PIC only records an examination event without the content of the reports in this release. We will try to standardize most examination results in the future and publish it in the release of the next iteration of the database.

Deidentification

Before data were integrated into the PIC database, it was fully anonymized by removing all HIPAA (Health Insurance Portability and Accountability Act) protected health information identifiers such as patient name, address, telephone numbers, and any other characteristic that could uniquely identify the individual in structured data sources. For free-text fields, protected health information was not extracted in accordance with HIPAA standards. When creating the dataset, patient IDs including SUBJECT_ID, HADM_ID and ICUSTAY_ID were randomly assigned a unique number, and the original identifier was not retained. As a result, the identifiers in the PIC cannot be linked back to the original, identifiable data. In addition, dates such as birth date and death date were shifted to the future in a consistent manner by the random offset (50–100 years) of each patient to maintain the interval, resulting in stays that occur sometimes between 2060 and 2120. Time of day, day of the week, and approximate seasonality (we divided 12 months into four seasons: spring: March, April, May; summer: June, July, August; autumn: September, October, November; winter: December, January, February) were conserved during the date transformation. Shift the time in such a wide range may limit the utilization of PIC in many time-sensitive studies. At the same time there are also algorithm to obscures date information to any desired granularity[24]. While, the proposed approach definitely can reduce the risk of re-identification of death children. The ethnicity of patients from ethnic minorities with a small population was not specified in the PIC. As all patients allow to be accepted by this Children's Hospital must under the age of 18, we do not need to handle the birthday of patients who were older than 90 years old to hide their real age as MIMIC.

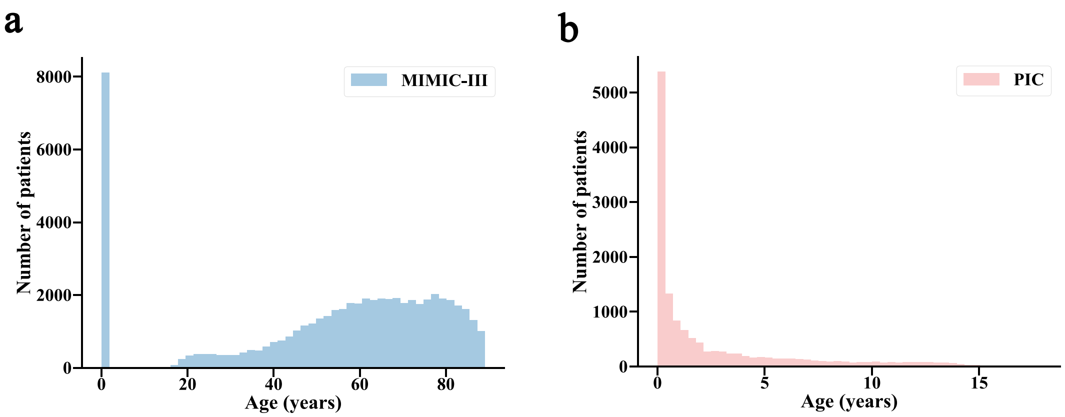

Fig. 2 Age distributions of the MIMIC-III and the PIC databases. (**a**) MIMIC-III database. (**b**) PIC database.

Table Name	Coverage Rate	Coverage Group
CHARTEVENTS	79.7%	high
DIAGNOSES_ICD	99.4%	excellent
EMR_SYMPTOMS	6.7%	low
INPUTEVENTS	41.1%	medium
LABEVENTS	93.9%	excellent
MICROBIOLOGYEVENTS	89.7%	excellent
OR_EXAM_REPORTS	91.9%	excellent
OUTPUTEVENTS	11.7%	low
PRESCRIPTIONS	51.5%	medium
SURGERY_VITAL_SIGNS	47.7%	medium

Table 4. Amount of complete data grouped by table. Coverage rate was calculated at the admission level. Low (0–20%), medium (20–60%), high (60–80%), and excellent (80–100%).

Data Records

Data available in the PIC database includes laboratory measurements, charted observations during a patient's stay, structured symptoms extracted from notes and vital signs recorded while a patient was present in the operating room. The PIC consists of 16 tables that are linked by unique identifiers[25]. Tables with a 'D' prefix are dictionary tables and provide a definition of the identifiers. For example, each row of LABEVENTS is associated with an ITEMID that represents the concept of the measurement, but it does not contain the actual name of the measurement. By joining LABEVENTS and D_LABITEMS on the ITEMID, the concept can be identified represented by a given ITEMID.

Table 1 provides summary descriptions of the data tables, and more detail data schema and technique document about the data tables is available on the PIC website[26]. Briefly, there are three tables used to define and track patient hospitalization: ADMISSIONS; PATIENTS; and ICUSTAYS. The three identifiers described earlier (SUBJECT_ID, HADM_ID, ICUSTAY_ID) are present in all three of the above tables. The other three tables are dictionaries for defining disease codes and items appearing in the PIC database. The remaining tables contain data related to patient care during the hospital stay, such as demographics, laboratory test results, medications, symptoms, vital sign measurements, and mortality. All the tables are distributed as a collection of comma separated value (CSV) files that can be loaded into many relational database systems.

Technical Validation

Paediatric-specific data features.

The PIC encompasses 13,499 distinct hospital admissions from 12,881 distinct paediatric patients (aged 0–18 years) admitted to critical care unit between 2010 and 2018. Of those patients, 468 (3.6%) had multiple hospital admissions. The mean age of the patients was 2.5 years (Q1–Q3: 0.1–3.3), 57.5% of the patients were male, and the in-hospital mortality was 7.1%. The mean hospital stay was 17.6 days (Q1–Q3: 7.0–21.0), and the mean ICU stay was 9.3 days (Q1–Q3: 0.9–9.2). The mean ICU stay was longest in the neonatal ICU, 21.6 days (Q1–Q3: 2.5–32.8), and the shortest stay was in the surgical ICU, 2.3 days (Q1–Q3: 0.8–1.6). The summary statistics and detailed patient demographics of the paediatric population in the PIC stratified by care unit are listed in Table 2.

The primary diagnosis of patients in the PIC shows many unique features of paediatric populations, as shown in Online-only Table 1. The top three disease categories of the discharge codes were Q00–Q99, congenital malformations, deformations and chromosomal abnormalities (25.4%); P00–P96, certain conditions originating in the perinatal period (14.1%); and J00–J99, diseases of the respiratory system (10.3%). This is completely different from what is found in adult critical care patients and also confirms the uniqueness and necessity of the PIC.

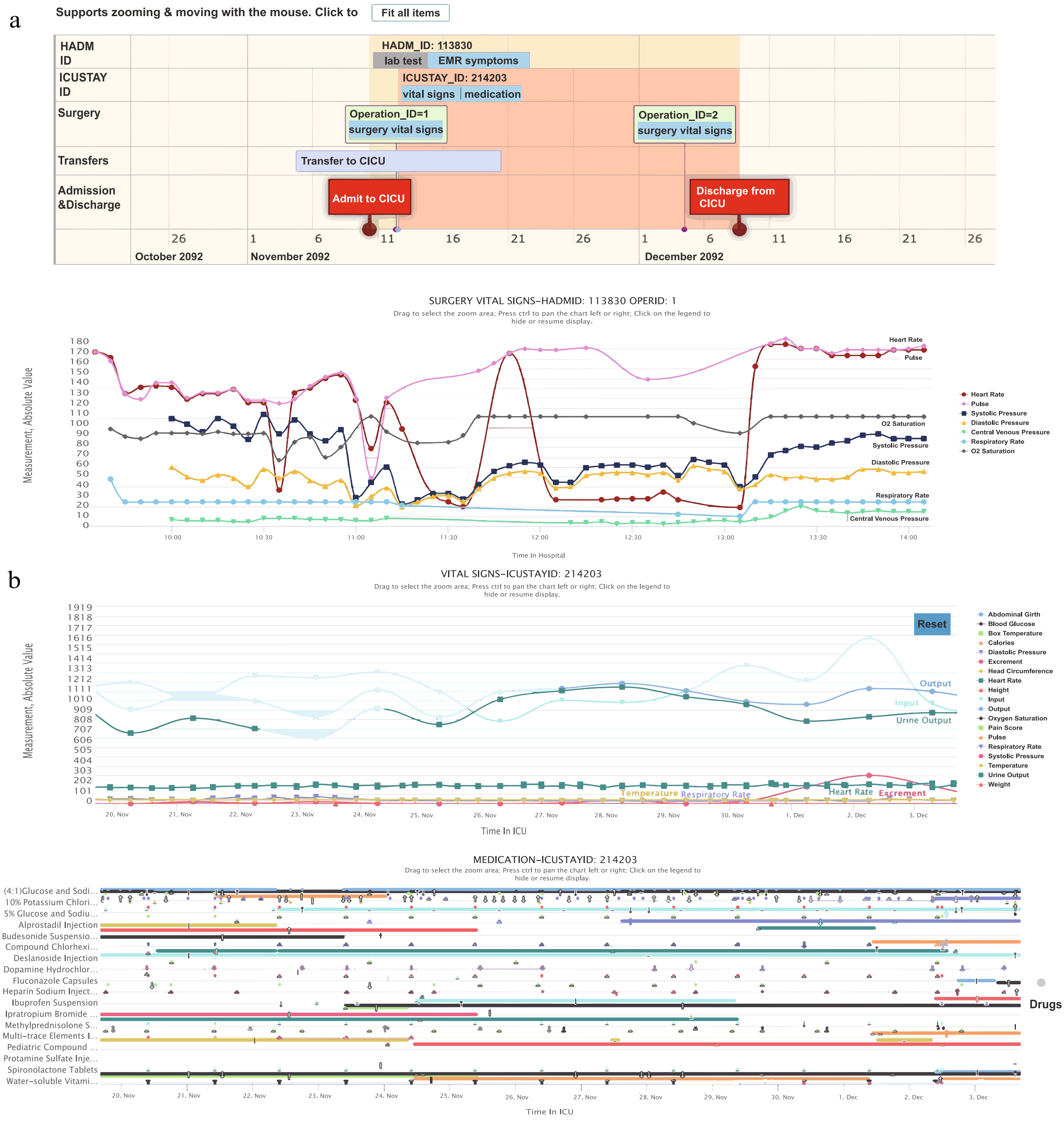

Fig. 3 Visualization of the PIC data. (**a**) The timeline of a patient during the hospital stay. (**b**) Synchronized timelines of different data types.

Comparing the PIC with the MIMIC-III. Table 3 provides a comparison of the PIC with the MIMIC-III database, and Fig. 2 shows the difference in age distributions between patients in the PIC and MIMIC-III. The salient database features that are compared in Table 3 include the data source, number of records and record completeness (availability of different clinical data, physiological waveforms and outcomes). Although there are many limitations in the current release of the PIC, such as that the PIC is smaller than the MIMIC-III based on the number of ICU patients, there are no high-frequency bedside monitor data, and clinical notes are described in Chinese and are not suitable for direct release. The PIC is still notable as a supplement to the paediatric intensive care unit data lacking in the MIMIC-III and will definitely encourage more researchers to focus on paediatric critical care. Furthermore, the PIC database is under continuous development with a better critical care environment, and we hope to supplement the PIC with more and more additional data in the future, including vital signs and physiology waveforms collected at bedsides, as much structured information extracted from the clinical notes as possible, and images corresponding to any medical imaging performed during the patient's stay.

Data quality and completeness. To maintain data fidelity, the original Chinese data were kept, and most of the translated English terms were manually reviewed except the ICD-10 English terms that were considered high-quality. All data except the symptoms that were extracted from narrative documents by NLP technology had very little post-processing performed.

As different clinical information systems were implemented at different times which means the data from some information systems was not available before it implemented, the completeness of the different data tables is different. Furthermore, the vendors of nursing systems and EMR systems have changed during the time of data collected which make the data more complicated to integrated. Table 4 shows the completion of the different tables in the PIC. As a long-term development project, we are improving and upgrading various clinical information systems so that the PIC will have progressively better data quality and completeness in future versions. The source code in python to calculate the completion of different tables was released on the Github.

Best practice for scientific computing. Based on best practice for scientific computing[27], the code used to build PIC was version controlled and developed collaboratively within the laboratory. The code for community to utilize and visualize the PIC dataset was hosted on GitHub. Issue tracking is used to ensure that any issues about the code and dataset can be managed and maintained. This approach encouraged and facilitated sharing of readable code and document, as well as feedback from community.

Usage Notes

Data access. The PIC is provided as a collection of comma separated value (CSV) files, along with example scripts to help with importing the data into the database system MySQL and PostgreSQL. As the database contains detailed information regarding the clinical care of patients, it must be treated with appropriate care and respect. Following the MIMIC protocol, researchers are required to formally request access via a process documented on the PIC website[26] and the PhysioNet[25]. There are two key steps that must be completed before access is granted:

- the researcher must complete a recognized course in protecting human research participants that includes HIPAA requirements or obtain a GCP certification from a local institute in China.
- the researcher must sign a data use agreement, which outlines appropriate data usage and security standards, and forbids efforts to identify individual patients.

Any researchers that have been approved to access the MIMIC-III will be familiar with this process and can easily get started with the PIC. Approval requires approximately three working days. Once an application has been approved, the researcher will receive emails containing instructions for downloading the database.

Collaborative documentation and visualization. The core aims for publicly developing the PIC database are to allow as much free access as possible to interested researchers and to promote secondary analysis of electronic health records. Detailed documentation is available on the PIC website and includes information regarding data access, table contents, and a schematic of the relationships between the tables in the data. The PIC is a large relational database with many different data types. This method of data storage does not visually indicate the patient's hospitalization, nor does it quickly and easily retrieve information specific to an individual patient. It is a time-consuming process for new researchers to perform joint multiple table queries to fully understand which data are useful and available. For researchers who prefer to explore the data, we provide a website that can easily be used to explore the PIC and visualize clinical data at the patient level. As shown in Fig. 3a, admission to the hospital, surgery, transfer to the ICU and discharge from the ICU and hospital is clearly shown on a timeline. There are several data buttons for users to choose clinical data of interest. Clicking on the button will show corresponding data on a timeline. It is possible to zoom in and out on the timeline to help users better explore the details and overview of the data. We also developed a mechanism to synchronize different timelines on the webpage to help users identify relationships among different data. For example, the medicine used at a time point may explain the change in vital signs at the same time as shown in Fig. 3b. The source code for this visualization is also openly available.

To help user utilize the PIC dataset, a Jupyter Notebook was released to demonstrate how to use PIC to create a machine learning project.

Code availability

The code that was used to create the PIC database, calculate statistics of this paper, demonstrate a machine learning task and source code which underpins the PIC website and documentation is openly available, and contributions from the research community are encouraged: https://github.com/Healthink/PIC.

Received: 24 September 2019; Accepted: 20 December 2019;
Published online: 13 January 2020

References

1. Henry, J., Pylypchuk, Y., Searcy, T. & Patel, V. *Adoption of Electronic Health Record Systems among U.S. Non-Federal Acute Care Hospitals: 2008–2015*. ONC Data Brief, no. 35. (The Office of the National Coordinator for Health Information Technology, 2016).
2. Shu, T. *et al*. EHR adoption across China's tertiary hospitals: A cross-sectional observational study. *Int. J. Med. Inform.* **83**, 113–21 (2014).
3. MIT Critical Data. *Secondary Analysis of Electronic Health Records*. *Secondary Analysis of Electronic Health Records* (Springer International Publishing, 2016).
4. Sun, H. *et al*. Semantic processing of EHR data for clinical research. *J. Biomed. Inform.* **58**, 247–259 (2015).

5. Finlayson, S. G., LePendu, P. & Shah, N. H. Building the graph of medicine from millions of clinical narratives. *Sci. Data* **1**, 140032 (2014).
6. Murdoch, T. B. & Detsky, A. S. The inevitable application of big data to health care. *JAMA - J. Am. Med. Assoc.* **309**, 1351–1352 (2013).
7. Vayena, E., Salathé, M., Madoff, L. C. & Brownstein, J. S. Ethical Challenges of Big Data in Public Health. *PLoS Comput. Biol.* **11**, e1003904 (2015).
8. Collins, F. S. & Tabak, L. A. NIH plans to enhance reproducibility. *Nature* **505**, 612–613 (2014).
9. Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature* **533**, 452–454 (2016).
10. Saeed, M. *et al*. Multiparameter intelligent monitoring in intensive care II: A public-access intensive care unit database. *Crit. Care Med.* **39**, 952–960 (2011).
11. Johnson, A. E. W. *et al*. MIMIC-III, a freely accessible critical care database. *Sci. data* **3**, 160035 (2016).
12. Nemati, S. *et al*. An Interpretable Machine Learning Model for Accurate Prediction of Sepsis in the ICU. *Crit. Care Med.* **46**, 547–553 (2018).
13. Feng, M. *et al*. Transthoracic echocardiography and mortality in sepsis: analysis of the MIMIC-III database. *Intensive Care Med.* **44**, 884–892 (2018).
14. Deliberato, R. O. *et al*. An Evaluation of the Influence of Body Mass Index on Severity Scoring. *Crit. Care Med.* **47**, 247–253 (2019).
15. Tyler, P. D. *et al*. Assessment of Intensive Care Unit Laboratory Values That Differ From Reference Ranges and Association With Patient Mortality and Length of Stay. *JAMA Netw. Open* **1**, e184521 (2018).
16. Pollard, T. J. *et al*. The eICU collaborative research database, a freely available multi-center database for critical care research. *Sci. Data* **5**, 180178 (2018).
17. Czaja, A. S. Children Are Not Just Little Adults. *Pediatr. Crit. Care Med.* **17**, 178–180 (2016).
18. Giza, C. C., Mink, R. B. & Madikians, A. Pediatric traumatic brain injury: Not just little adults. *Curr. Opin. Crit. Care* **13**, 143–152 (2007).
19. Goel, R., Cushing, M. M. & Tobian, A. A. R. Pediatric Patient Blood Management Programs: Not Just Transfusing Little Adults. *Transfus. Med. Rev.* **30**, 235–241 (2016).
20. Soleimani, T. *et al*. Pediatric pyoderma gangrenosum: is it just big wounds on little adults? *J. Surg. Res.* **206**, 113–117 (2016).
21. Wetzel, R. Pediatric Intensive Care Databases for Quality Improvement. *J. Pediatr. Intensive Care* **5**, 81–88 (2016).
22. Ramnarayan, P. *et al*. Interhospital Transport of Critically Ill Children to PICUs in the United Kingdom and Republic of Ireland: Analysis of an International Dataset. *Pediatr. Crit. Care Med.* **19**, e300–3311 (2018).
23. Li, H.-M., Li, Y., Duan, H.-L. & Lv, X.-D. Term extraction and negation detection method in Chinese clinical document. *Chinese J. Biomed. Eng.* **27**, 716–721 (2008).
24. Hripcsak, G., Mirhaji, P., Low, A. F. H. & Malin, B. A. Preserving temporal relations in clinical data while maintaining privacy. *J. Am. Med. Informatics Assoc.* **23**, 1040–1045 (2016).
25. Li, H., Zeng, X. & Yu, G. Paediatric Intensive Care Database. *PhysioNet*. https://doi.org/10.13026/bcbd-4t11 (2019).
26. PIC database: documentation and website, http://pic.nbscn.org (2019).
27. Wilson, G. *et al*. Best Practices for Scientific Computing. *PLoS Biol.* **12**, e1001745 (2014).

Acknowledgements

This work was supported by the National Natural Science Foundation of China [81871456], the Chinese State's Key Project of R&D Plan [2016YFC0901905,2016YFC0901703]; the Zhejiang Provincial Program for the Cultivation of High-level Innovative Health Talents [2016-6]; the Fundamental Research Funds for the Central Universities [2019XZZX003-16]. We also want to thank the MIMIC project and the PhysioNet especially thank Dr. Alistair E. W. Johnson and Dr. Tom J. Pollard; many of the protocols in this project were based on it.

Author contributions

X.Z., G.Y., L.Y., Q.S. and H.L. built the PIC database. G.Y., L.T., J.W., S.S., Q.S. and H.L. collected the original clinical data. L.Y., X.Z., H.D. and H.L. developed the PIC website. All authors gave input during the database development process and contributed to writing the paper.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Q.S. or H.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.