scientific data

OPEN

DATA DESCRIPTOR

Chest computed tomography for patients with sepsis in the emergency intensive care unit

Senjun Jin[1,6], Wenwei Cai[1,6], Qiang Shen[2,6], Lingfan Yang[3], Hu Sheng'an[1,6], Jin Fu[1] & Zhongheng Zhang[4,5] ✉

Sepsis is a systemic inflammatory response syndrome (SIRS) caused by infection, which may lead to multiple organ dysfunction in susceptible patient. The most frequently involved organs/systems include the lung, kidney and circulation system. It is well established that sepsis is a risk factor for acute lung injury. While overt pulmonary infiltrates can be well captured by human operators, subtle structural changes of the lung might be ignored. Since the advantage of chest computed tomography (CT) is its capability of providing fine structural changes in high spatial resolution, the study of chest CT by means of computer science may provide further insights into the underlying pathophysiology. The integration of chest CT into the study of sepsis is limited partly due to the lack of well-curated database. The study aims to establish a database comprising detailed clinical tabular data, as well as the raw chest CT images. The database is intended to support a wide array of research studies involving radiomics in sepsis patients, helping to reduce barriers to the reproducibility of clinical research.

Background & Summary

Sepsis is a systemic inflammatory response syndrome (SIRS) caused by infection, which is also a leading cause of mortality and morbidity for hospitalized patients[1]. SIRS can lead to multiple organ dysfunction in susceptible patients, and the most frequently involved organs/systems include lung, kidney and circulation[2]. While the inflammatory response and immune profile of sepsis have been extensively investigated[3,4], the integrated analysis of the structural changes of lung parenchyma and clinical features is rarely reported.

Computed tomography (CT) provides high spatial resolution of the structural changes of lung parenchyma in response to the SIRS[5]. Sepsis-induced acute lung injury is a well-established form of lung involvement during sepsis, which is presented as bilateral infiltrates on Chest CT. However, more subtle changes not visible to human eyes might be ignored. With the development of computer vision and deep learning technologies, complicated features can be well represented and extracted from CT images[6]. These features have been shown to provide valuable insights into disease prognosis, subtyping and medical decision-making[7–9]. However, due to lack of well curated publicly available CT datasets for sepsis patients, studies exploring the lung CT radiomics are scarce. Thus, current study aims to establish a publicly available lung CT datasets, together with high granularity clinical tabular data. This dataset will arouse enthusiasm for studies on sepsis by integrating medical images and clinical tabular data.

[1]Emergency and Critical Care Center, Department of Emergency Medicine, Zhejiang Provincial People's Hospital, Affiliated People's Hospital, Hangzhou Medical College, Hangzhou, Zhejiang, China. [2]Center for Rehabilitation Medicine, Department of Radiology, Zhejiang Provincial People's Hospital, Affiliated People's Hospital, Hangzhou Medical College, Hangzhou, Zhejiang, China. [3]The information center, Zhejiang Provincial People's Hospital, Affiliated People's Hospital, Hangzhou Medical College, Hangzhou, Zhejiang, China. [4]Department of Emergency Medicine, Provincial Key Laboratory of Precise Diagnosis and Treatment of Abdominal Infection,Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, 310016, China. [5]School of Medicine, Shaoxing University, Shaoxing, Zhejiang, 312000, P.R. China. [6]These authors contributed equally: Senjun Jin, Wenwei Cai, Qiang Shen, Hu Sheng'an. ✉e-mail: zh_zhang1984@zju.edu.cn

Methods

Study setting and population. The structure of tabular data was formatted as described in previous work[10], and the descriptions of all tables were repeated in the following sections for the ease of reading. However, the population included in the study was different from our previous work and the current work is focusing more on the computed tomography.

The study was conducted in Zhejiang Provincial People's Hospital, Zhejiang, China from January 2019 to December 2022. All sepsis patients admitted to the Emergency ICU of the hospital were eligible. Sepsis was defined in accordance with the sepsis-3.0 criteria, which included suspected or documented infection plus an acute rise of sequential organ failure assessment (SOFA) score greater than or equal to 2 points[11]. The study was approved by the ethics committee of Zhejiang Provincial People's Hospital (approval number: 2023-397), and the approval included the publication of the data records. The study was conducted in accordance with the Declaration of Helsinki.

Informed consent was waived as determined by the institutional review board, due to the retrospective design of the study.

Database structure and development. The database comprises two types of data. One is the clinical tabular data, which is distributed as comma-separated value (CSV) files that can be managed by any relational database language such as SQL. The other is the CT image data, which is distributed as *NIfTI* files with nii.gz suffix. The CT image files can be linked to the clinical data by the *CT2hospitalID* table. Each individual patient can be identified by a series number (patient_SN) with the combination of digits and letters such as "5810787d01cf52e6973eef9819b7d2ac". The patient_SN is deidentified. Each unique hospital stay is denoted by a *Hospital_ID* with examples such as "337016968172517". The unique ICU stay can be identified by the *HospitalTransfer* table, which contains intrahospital transfer events. All tables are linked by *Hospital_ID* to identify sequential medical events during an individual hospital stay.

Deidentification. The Health Insurance Portability and Accountability Act (HIPAA) is employed as the standard to conduct de-identification. All protected information such as addresses, date of birth, date of hospital admission, date of medical order, personal numbers (e.g. name, phone, social security, and hospital number), date of discharge, exact age on admission (age is discretized into bins) are removed. When creating the dataset, patients were randomly assigned a unique identifier (patient_SN and hospital_ID) and the original hospital identifiers were not retained. As a result, the identifiers in the tables cannot be linked back to the original, identifiable data. All identifiers related to doctors, nurses, and pharmacists have been removed to protect the privacy of contributing providers. Date-time variables/columns are de-identified by showing only days in reference to hospital admission.

Data Records

The data reported in this paper have been deposited in the OMIX, China National Center for Bioinformation/Beijing Institute of Genomics, Chinese Academy of Sciences[12], as well as the PhysioNet repository[13]. The database comprises 728 hospital visits (i.e. including outpatient visits) for 337 unique admissions from January 2012 to December 2022. Table 1 shows the baseline demographics of hospital admissions (outpatient visits are excluded). There are 103 female and 234 male patients in the dataset. The length of hospital days was 22 days (Q1 to Q3: 12 to 36). Male patients showed slightly longer hospital stay.

Individual chest CT files are distributed as the *NIfTI* format, since the format is a popular file format for storing medical imaging data and is widely used in medical research and related fields. These CT files are converted from the original DICOM file by using the SimpleITK package (v2.2.0). if there are multiple series in a CT volume, the one containing chest CT is extracted. Patients can have multiple CT scans during hospital stay and all scans are curated in the database. There are 836 CT scans for 327 hospital admissions. There are many packages to handle such file type. For instance, the *RNifti* package can be utilized to manipulate and visualize the CT images. Sample chest CT slices can be found in Fig. 1.

Classes of data. The data are organized into two categories which are clinical tabular data and *NIfTI* files. The structure of clinical data is quite like the ones reported previously. To keep the content of this data descriptor intact, we describe these tables again in supplemental digital contents, with more focuses on their associations with lung CT and sepsis (Supplementary Table S1 to S10). There are a total of 14 tables comprising patient demographic data, serial ID of Chest CT image, medical order, laboratory findings, image studies, microbiology and hospital transfer events (Tables 2 and 3). The unstructured data, including natural language reports from chest CT scans and other examinations, have been incorporated into the current version of our dataset. Specifically, these data are contained within the "ExamReport.csv" table, which forms part of our submission.

The first column gives the file names for each table. The *MD5_hashes* column gives the MD5 hashes. The MD5 (message-digest algorithm) is a cryptographic protocol used for authenticating messages as well as content verification and digital signatures. MD5 is based on a hash function that verifies that a file you sent matches the file received by the person you sent it to. The *NumObs* column describes the number of rows in each table.

The CT2hospitalID table. The *CT2hospitalID* table contains information corresponding the CT file names to the hospital ID (Table 4). The *serialID* column gives the CT serial ID, which is also the file name in the *CTImage* folder. *CTexame_DateTime* gives the days offset by the hospital admission date time. Some numbers are negative in this column because some CT scans are performed in the emergency department before hospital admission. The *patient_SN* and *Hospital_ID* give the unique identifier for each patient and hospital admission. The

Variables	Total (n = 337)	Female (n = 103)	Male (n = 234)	P*
Age_cut, n (%)				0.673
(0,18]	2 (1)	0 (0)	2 (1)	
(18,30]	8 (2)	2 (2)	6 (3)	
(30,40]	10 (3)	2 (2)	8 (3)	
(40,50]	23 (7)	4 (4)	19 (8)	
(50,60]	41 (12)	11 (11)	30 (13)	
(60,70]	87 (26)	26 (25)	61 (26)	
(70,80]	95 (28)	30 (29)	65 (28)	
(80,90]	52 (15)	21 (20)	31 (13)	
(90,150]	19 (6)	7 (7)	12 (5)	
DaysHospitalStay, Median (Q1,Q3)	22 (12, 36)	20 (12, 30)	22 (12.75, 38)	0.224
StatusOnDischarge, n (%)				0.427
Cured	185 (56)	55 (53)	130 (57)	
Not cured	121 (36)	37 (36)	84 (37)	
Unknown	26 (8)	11 (11)	15 (7)	

Table 1. Demographics and discharge status of the 337 hospital admissions in the database. *The P value represents the probability of observing the current or more extreme differences in continuous or categorical variables between female and male patients, assuming there is no true difference between the groups (i.e., the null hypothesis). A p-value less than the predetermined significance level, often 0.05, suggests that the differences observed are statistically significant, indicating a potential biological or clinical relevance in gender comparisons. The t-test is applicable when data are normally distributed and assumes equal variances between groups, calculating the t-statistic to determine the p-value from the t-distribution. In cases where normality is questionable, the non-parametric Mann-Whitney U test is used, which ranks data across both groups and calculates the p-value based on the sum of ranks. For categorical data, the chi-square test of independence is often used to compare the distribution of categorical variables between genders, with the p-value derived from the chi-square distribution. Fisher's Exact Test is an alternative when sample sizes are small or expected frequencies are too low, providing an exact p-value based on the hypergeometric distribution.

STUDYRESULT column gives the description of the CT finding in text. The *DIAGRESULT* gives the impression of diagnosis reported by radiologists.

Technical Validation

Data were verified for integrity during the data transfer process from the hospital information system to the database platform using MD5 checksums (Table 2). The MD5_hashes presented in Table 2 can also be used to check the integrity of the downloaded datasets. Since the tables are extracted from different information systems, the identifiers for linking the tables may not consistent. Thus, we check the consistency to ensure that all *Hospital_ID* from each table can be matched to the *Hospital_ID* in the *PtAdmiTable* table.

All text information extracted from our medical information system are in Chinese. To ease the usage of the dataset by international users, some meta-data and short texts are translated to English. The translation was first performed by using the paid BaiDu academic translation service (service number: MPE2022102608424528825) and then checked by two authors (Zhongheng Zhang and Senjun Jin) of the project. However, in order to maintain data fidelity, very little post-processing has been performed for other long text fields such as present history, progress notes, and text reports of image studies, because any translations may alter the results of natural language processing or text mining[14,15]. Academic language translation services (including API) can be employed for enormous translation work.

We have conducted a thorough assessment of our data's FAIRness and have taken the following steps to enhance its compliance with these principles[16]:

1. **Findable:** We have ensured that our dataset is registered with a globally unique identifier (OMIX005655) and is properly indexed in relevant data repositories (https://ngdc.cncb.ac.cn/omix), allowing for easy discovery through search engines and databases.
2. **Accessible:** We have provided clear documentation and metadata within the document, describing the data's content, structure, and any conditions for access. This information is available to all potential users to facilitate understanding and use of the data.
3. **Interoperable:** We have formatted our data using standard structures and formats (e.g., CSV for tabular data) and have included metadata in standardized schemas to ensure compatibility with other datasets and tools.
4. **Reusable:** We have documented the data's provenance, including information on data collection, processing steps, and any transformations applied. This documentation enables users to assess the data's suitability for their intended purposes and to give appropriate credit.

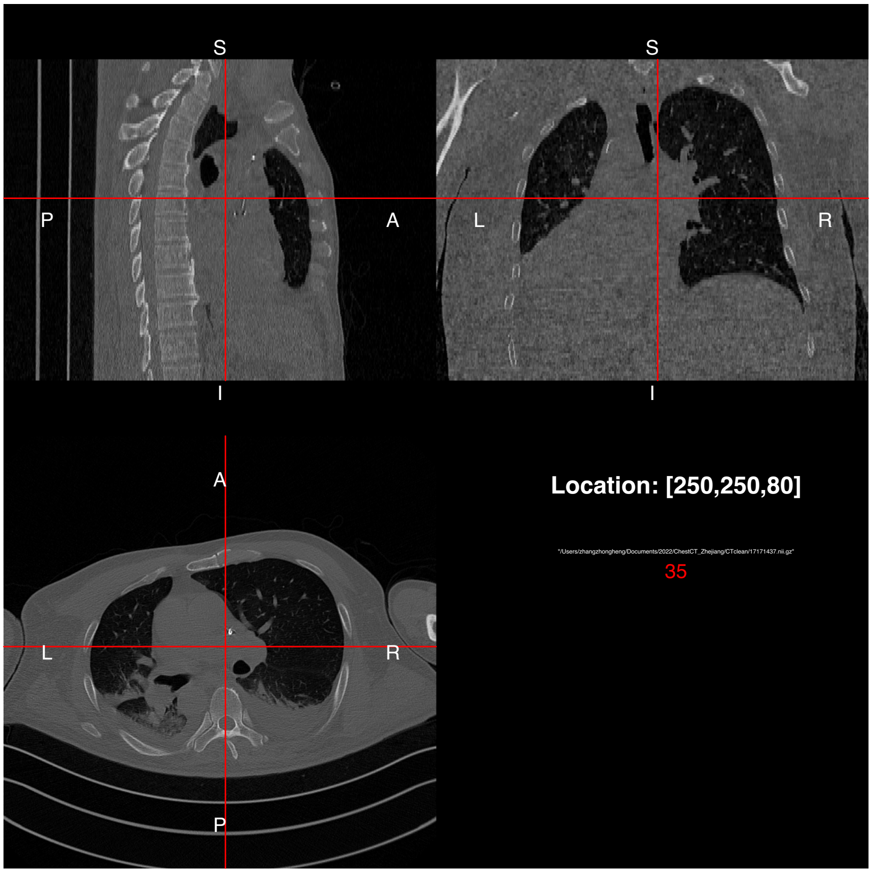
S
P
A
I
S
L
R
I
A
L
R
P
Location: [250,250,80]
"/Users/zhangzhongheng/Documents/2022/ChestCT_Zhejiang/CTclean/17171437.nii.gz"
35

Fig. 1 A sample chest CT scan showing the sagittal, coronary and axial views. The spatial coordinates are indicated by [250, 250, 80]. A = anterior; P = posterior; S = superior; I = inferior; R = right; L = left.

Tables	MD5_hashes	NumObs	Description
CT2hospitalID.csv	77c2d45f22daa1586d57db12fa12db99	837	Map CT file name to the hospital ID
Diagnosis.csv	86e97703829b8d0d67e7c93566ba3cce	8459	Diagnosis
DrugSens.csv	13983f8544031b340a19550e349c8ec2	55029	Sensitivity of pathogen to antibiotics for cultured bacteria
ExamReport.csv	6ba91fca92f0f9b5a5f2aa2e6a41c94e	6037	Examination report including CT, ultrasound and MRI
HospitalTransfer.csv	1d58eab7d1cf4f0a2c869aa5a3206bba	277	intrahospital transfer events
Lab_dictionary.csv	9faf7e9021d1310e9eab4dcaa335d071	246	Dictionary for laboratory events
Lab.csv	29c64f065a907e9c8bd3dfee2654ded2	789010	Laboratory findings

Table 2. A general description of the tables in the database.

Usage notes

Data access. The data reported in this paper have been deposited in the OMIX, China National Center for Bioinformation/Beijing Institute of Genomics, Chinese Academy of Sciences (https://ngdc.cncb.ac.cn/omix/release/OMIX005655)[17], as well as in the Physionet repository[13]. Data access also requires the users to sign a data use agreement, which stipulates that the user will not try to re-identify any subjects, will release code associated with any publication using the data, and will not share the data. After approval of the accessibility, the plain CSV tables and CT image files can be downloaded from the repository.

Use cases. The dataset presented in this study offers a plethora of utilities and use cases that hold significant value for the medical and scientific community (Table 5). It has been meticulously curated to advance the field of medical imaging, particularly in the analysis of chest computed tomography scans for patients with sepsis. The comprehensive nature of the dataset makes it an invaluable tool for researchers aiming to develop and refine machine learning algorithms that can accurately predict and diagnose septic conditions. Furthermore, it serves as a robust foundation for enhancing clinical decision support systems, enabling healthcare professionals to make more informed and timely treatment decisions. The dataset also extends its utility to the educational realm, providing a rich source of information for medical students and practitioners to hone their diagnostic skills. Lastly, it contributes to public health by offering data that can inform healthcare policy and strategic planning around

Tables	MD5_hashes	NumObs	Description
Medication.csv	f888056e674134b66209de647bb15f65	109246	Medication events
MedOrder.csv	e8944f8f62cc523549919acb6fc0ba25	117751	Medical order
MicrobiologyCulture.csv	99e08c5362e32def924c08aba30043ac	20661	Microbiology cuture
NursingChart_IO.csv	bc4ab64f975606ee025db9e75d16b3d9	98231	Fluid Input and output
NursingChart_VitalSign.csv	653b56fbef7d500e679d7d46e93e62cb	2251285	Vital Sign from Nursing chart
PtAdmiTable.csv	e96f4f717884c0d94ad96db06d02ecb3	728	Patient admission table
VitalSign.csv	d63b966aabb0d75bde66aeff3c1135d2	505151	Vital signs

Table 3. A general description of the tables in the database.

Variables	Explanation
serialID	CT serial ID corresponding to the file name in the *CTImage* folder
CTexame_DateTime	The time of CT examination in relative to the hospital admission time
patient_SN	Patient series number: unique to each individual subject
Hospital_ID	unique to each hospital admission
STUDYRESULT	Description of the CT finding in text
DIAGRESULT	Diagnosis for the CT finding in text

Table 4. Explanations for the variables in the *CT2hospitalID* table.

Use case	Description
Medical Imaging Research	Analysis of chest computed tomography scans to identify sepsis-related features and patterns.
Machine Learning Model Development	Utilization of dataset to train and validate algorithms for sepsis diagnosis and prognosis.
Clinical Decision Support	Integration into clinical tools to support healthcare providers in making informed treatment decisions for sepsis patients. For example, fluid resuscitation can be informed by analyzing the chest CT.
Medical Education	Application in educational programs to improve analytic skills of researchers.
Research Advancement	Driving innovative research to improve diagnostic methods, treatment strategies, and patient outcomes in sepsis care.

Table 5. Use cases of the dataset in various aspects of sepsis management.

sepsis management. Collectively, the dataset's multifaceted applications have the potential to significantly improve clinical outcomes, streamline medical education, and drive innovative research in sepsis care.

Code availability

Code are available at: https://github.com/zh-zhang1984/ZhejiangProvinceICU/blob/main/ZhejiangProvinceICU.md.

Received: 5 January 2024; Accepted: 13 November 2024;
Published online: 20 November 2024

References

1. Rudd, K. E. *et al.* Global, regional, and national sepsis incidence and mortality, 1990-2017: analysis for the Global Burden of Disease Study. *Lancet* **395**, 200–211 (2020).
2. Li, W. *et al.* Classic Signaling Pathways in Alveolar Injury and Repair Involved in Sepsis-Induced ALI/ARDS: New Research Progress and Prospect. *Dis Markers* **2022**, 6362344 (2022).
3. Michels, E. H. A. *et al.* Association between age and the host response in critically ill patients with sepsis. *Crit Care* **26**, 385 (2022).
4. Zhang, Z. *et al.* Gene signature for the prediction of the trajectories of sepsis-induced acute kidney injury. *Crit Care* **26**, 398 (2022).
5. Vliegenthart, R., Fouras, A., Jacobs, C. & Papanikolaou, N. Innovations in thoracic imaging: CT, radiomics, AI and x-ray velocimetry. *Respirology* **27**, 818–833 (2022).
6. Suri, J. S. *et al.* A narrative review on characterization of acute respiratory distress syndrome in COVID-19-infected lungs using artificial intelligence. *Comput Biol Med* **130**, 104210 (2021).
7. Bouchareb, Y. *et al.* Artificial intelligence-driven assessment of radiological images for COVID-19. *Comput Biol Med* **136**, 104665 (2021).
8. Ter Maat, L. S. *et al.* Imaging to predict checkpoint inhibitor outcomes in cancer. A systematic review. *Eur J Cancer* **175**, 60–76 (2022).
9. Röhrich, S. *et al.* Prospects and Challenges of Radiomics by Using Nononcologic Routine Chest CT. *Radiology: Cardiothoracic Imaging* **2**, e190190 (2020).
10. Jin, S. *et al.* Establishment of a Chinese critical care database from electronic healthcare records in a tertiary care medical center. *Sci Data* **10**, 49 (2023).
11. Singer, M. *et al.* The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA* **315**, 801–810 (2016).
12. CNCB-NGDC Members and Partners Database Resources of the National Genomics Data Center, China National Center for Bioinformation in 2024. *Nucleic Acids Res* **52**, D18–D32 (2024).
13. Jin, S. & Zhang, Z. Chest Computed Tomography for patients with sepsis in the Emergency Department (version 1.0.0). *PhysioNet* https://doi.org/10.13026/zne5-qh18 (2024).

14. Li, S. *et al*. Deep Phenotyping of Chinese Electronic Health Records by Recognizing Linguistic Patterns of Phenotypic Narratives With a Sequence Motif Discovery Tool: Algorithm Development and Validation. *J Med Internet Res* **24**, e37213 (2022).
15. Gong, L., Zhang, Z. & Chen, S. Clinical Named Entity Recognition from Chinese Electronic Medical Records Based on Deep Learning Pretraining. *J Healthc Eng* **2020**, 8829219 (2020).
16. Wilkinson, M. D. *et al*. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016).
17. Jin, S. *et al*. Chest computed tomography for patients with sepsis in the emergency intensive care unit. OMIX https://ngdc.cncb.ac.cn/omix/release/OMIX005655 (2024).

Acknowledgements

Z.Z. received funding from China National Key Research and Development Program (2023YFC3603104), the Huadong Medicine Joint Funds of the Zhejiang Provincial Natural Science Foundation of China under Grant No. LHDMD24H150001, China National Key Research and Development Program (2022YFC2504500), National natural science foundation of China (82472243, 82272180), A collaborative scientific project co-established by the Science and Technology Department of the National Administration of Traditional Chinese Medicine and the Zhejiang Provincial Administration of Traditional Chinese Medicine (GZY-ZJ-KJ-24082), General Health Science and Technology Program of Zhejiang Province (2024KY1099), and Project of Zhejiang University Longquan Innovation Center (ZJDXLQCXZCJBGS2024016). S.J. received funding from Youth Talents Project of Health Commission of Zhejiang Province (Project number: 2019RC103) and Health Science and Technology Plan of Zhejiang Province (2023KY051).

Author contributions

S.J. and Z.Z. conceived the idea; W.C., Q.S., L.Y. and H.S. curated data; J.F., Q.H. and N.L. checked the accuracy of the data. Z.Z. drafted the manuscript. All authors reviewed and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-024-04132-z.

Correspondence and requests for materials should be addressed to Z.Z.