Using freely accessible databases for laboratory medicine research: experience with MIMIC database

Yuan-Lan Huang[1], Tony Badrick[2,3], Zhi-De Hu[4]

[1]Department of Laboratory Medicine, No. 455 Hospital of the Chinese People's Liberation Army, Shanghai 200030, China; [2]Faculty of Health Sciences and Medicine, Bond University, Robina, Queensland, Australia; [3]Sullivan Nicolaides Pathology, Taringa, Queensland, Australia; [4]Department of Laboratory Medicine, General Hospital of Ji'nan Military Region, Ji'nan 250031, China
Correspondence to: Zhi-De Hu, PhD, MD. Department of Laboratory Medicine, General Hospital of Ji'nan Military Region, Ji'nan 250031, China. Email: hzdlj81@163.com.

Abstract: As staffs in clinical laboratories are not routinely involved in the management of patients, it is difficult for them to perform researches designed to explore the clinical utility of laboratory tests. Using publicly accessible clinical databases for clinical research represents an alternative way of performing studies investigating the clinical aspects of laboratory tests. Here, we introduced a freely accessible critical care database named Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II). This database contains the results of many laboratory tests and thus can be used to perform laboratory medicine research. The steps to access this database were indicated, and the structure of this database was introduced. Some tips in using this database were shared.

Keywords: Laboratory medicine; database; critical care; clinical research

Laboratory medicine research usually focuses on the analytical and clinical aspects of laboratory tests. The analytical aspect of laboratory medicine research includes the following areas: (I) establishment of novel, inexpensive, easily quantified, rapid and reliable assays for laboratory tests; (II) assessment of the performance of an analytical method (e.g., precision, limit of detection, linearity, accuracy, quality control); (III) analytical and pre-analytical errors impacting on the interpretation of a test. The clinical aspect of laboratory medicine research is usually focused on the clinical significance of laboratory tests, including their utility in disease diagnosis, prognosis or disease severity/activity estimation, risk stratification, and treatment monitoring. A critical step in the clinical aspect of laboratory medicine research is data collection, either prospectively or retrospectively. However, staffs in clinical laboratories are not routinely involved in the management of patients, and this situation makes it difficult for them to perform research designed to explore the clinical utility of laboratory tests. During past years, the authors have performed some clinical research (1-3) in laboratory medicine using the Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II), a freely accessible critical care database. The aim of this paper is to describe this database, and share our experiences using it in laboratory medicine research.

Brief introduction about MIMIC II

This database includes more than 20,000 patients admitted to various intensive care units (ICUs) (e.g., medical, surgical, coronary care, and neonatal) of Beth Israel Deaconess Medical Center (BIDMC, Boston, MA, USA) between 2001 and 2008 (4). The Institutional Review Boards (IRB) of the Massachusetts Institute of Technology (MIT, Cambridge, MA, USA) approved the establishment of this database. All patients in this database are de-identified, and key demographics (e.g., admitted time, birthday) are shifted to protect their privacy.

The clinical data of the patients were obtained from

bedside workstations and hospital archives. These data include demographics, laboratory tests, medications, fluid balance, physiological scores [e.g., simplified acute physiology score I (SAPS I), sepsis-related organ failure assessment (SOFA)], More importantly, the long-term (1 year all-cause mortality) and short-term (hospital mortality) outcomes of patients are recorded. One year all-cause mortality of the patients is obtained from the social security database.

Recently, MIMIC II has been updated to MIMIC III, with a renamed full title of "Medical Information Mart for Intensive Care" (5). MIMIC-III comprises over 58,000 hospital admissions for 38,645 adults and 7,875 neonates, with data spanning the period from June 2001 to October 2012. Data are collected by two different information systems namely Philips CareVue Clinical Information System (models M2331A and M1215A; Philips Health-care, Andover, MA) and iMDsoft MetaVision ICU (iMDsoft, Needham, MA). For patients recorded by CareVue system, their out-of-hospital mortality was obtained from the social security database, with a minimum follow-up time of 4 years. While for Metavision patients, the minimum follow-up time is only 90 days.

Accessing MIMIC II

The process of applying for access to MIMIC II, initially requires completion of the CITI "Data or Specimens Only Research" course (https://www.citiprogram.org/index.cfm?pageID=154&icat=0&ac=0) followed by the creation of an account on PhysioNet (https://physionet.org/pnw/login). After successful completion of the course, and obtaining certification, the applicant is permitted to download database files from PhysioNet to their personal computer. The details for data downloading and database installment are well documented by Dr. Zhang (6).

The structure of MIMIC database

The MIMIC II database contains 38 tables, while MIMIC III contains 40. These tables record the clinical details of the patients, including demographics, laboratory tests and medications. Tables pre-fixed with "D_" represent dictionaries. For example, all laboratory tests are defined as ITEMID in a table named D_LABITEMS. This table can be joined with a table named LABEVENTS to obtain laboratory tests of a patient.

All tables can be linked by identifiers which usually have

the suffix "ID". Three IDs are used to specify the patient: SUBJECT_ID is a unique identification for a patient; HADM_ID and ICUSTAY_ID refer to a unique hospital stay and ICU stay, respectively.

Data of patients can be extracted from MIMIC using structure query language (SQL), an open source administration and development platform for PostgreSQL. Some of the example for data querying can be found at following link: https://mimic.physionet.org/tutorials/intro-to-mimic-iii/.

Using MIMIC database for laboratory research

The hospital and long-term mortality are recorded in MIMIC database and can be used to investigate the prognostic value of laboratory tests for a certain disease. For example, we have investigated the prognostic value of admission red blood cell distribution width (RDW) and prognosis of patients with acute myocardial infarction (1), acute pancreatitis (2) and subarachnoid haemorrhage (3). The protocols of these studies are similar:

❖ Tables named DIAGNOSIS_ICD in MIMIC III and ICD9 in MIMIC II were used to define the patients with a specified disease. Please note that a column titled SEQUENCE in MIMIC II and SEQ_NUM in MIMIC III provides the order of the diagnosis for a patient. The ICD diagnoses are ordered by priority;

❖ D_LABITEMS and LABEVENTS are used to extract the data on laboratory tests defined by a column titled CHARTTIME;

❖ A table named ICUDETAILS in MIMIC II is used to extract the short and long term outcomes of patients. In MIMIC III, the short and long term outcomes of patients can be calculated in a table named PATIENTS. The patients' demographic characteristics are also included in these tables.

The medical interventions are recorded in a table named INPUTEVENTS and the vital signs can be found in a table named CHARTEVENT.

The most difficult part of the MIMIC based studies is the data extraction. The data in MIMIC database can be extracted by SQL query. This can be a challenge for researchers with little or no background of database management.

Building models that predict the prognosis of patients has attracted much attention. For example, Zhou et al has developed an easy-to-use prognostic model in cirrhotic patients for named quick CLIF-SOFA (qCLIF-SOFA)

using MIMIC III (7). As well as performing studies investigating the prognostic value of laboratory tests, researchers can also investigate the diagnostic value of laboratory tests for a certain disease or disorder (e.g., sepsis, heart failure). However, to date it seems that this type of study is rare.

Limitations of research based on MIMIC database

The MIMIC database has some advantages, including the extremely large sample size and data available for each patient during hospitalization, as well as long-term follow up time. However, it has the following weaknesses:

❖ Although the data in MIMIC database are prospectively collected, all MIMIC based studies are of a retrospective design. This may affect the representativeness of the subjects and the reliability of results;

❖ Only routine laboratory tests are recorded in MIMIC database, thus it is impossible for the researcher to investigate the clinical value of novel biomarkers such as circulating microRNA;

❖ The database only records the all-cause mortality and hospital mortality. It is impossible to investigate the association between laboratory tests and disease specific mortality, such as major adverse cardiovascular events (MACE) in cardiovascular diseases.

Nevertheless, the MIMIC database represents a new opportunity for laboratory clinicians to conduct research. In our opinion, in the era of big data (8-10) and data sharing, publicly accessed databases with full pictures of patients will be more and more widely used. These studies, although presenting many challenges at the current stage of these databases development, will greatly shape the profiles of clinical research in future. For laboratory clinicians, data may be not a problem in future.

Acknowledgments

Footnote

Provenance and Peer Review: This article was commissioned by the editorial office, *Journal of Laboratory and Precision Medicine* for the series "Clinical Database in Laboratory Medicine Research Column". The article has undergone external peer review.

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at http://dx.doi.org/10.21037/jlpm.2017.06.06). The series "Clinical Database in Laboratory Medicine Research Column" was commissioned by the editorial office without any funding or sponsorship. Tony Badrick served as Guest Editor of the series and serves as an unpaid editorial board member of *Journal of Laboratory and Precision Medicine* from December 2016 to November 2018. Zhi-De Hu serves as an unpaid Executive Editor of *Journal of Laboratory and Precision Medicine* from November 2016 to October 2021. The authors have no other conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Open Access Statement: This is an Open Access article distributed in accordance with the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International License (CC BY-NC-ND 4.0), which permits the non-commercial replication and distribution of the article with the strict proviso that no changes or edits are made and the original work is properly cited (including links to both the formal publication through the relevant DOI and the license). See: https://creativecommons.org/licenses/by-nc-nd/4.0/.

References

1. Huang YL, Hu ZD. Lower mean corpuscular hemoglobin concentration is associated with poorer outcomes in intensive care unit admitted patients with acute myocardial infarction. Ann Transl Med 2016;4:190.

2. Hu ZD, Wei TT, Tang QQ, et al. Prognostic value of red blood cell distribution width in acute pancreatitis patients admitted to intensive care units: an analysis of a publicly accessible clinical database MIMIC II. Clin Chem Lab Med 2016;54:e195-7.

3. Huang YL, Han ZJ, Hu ZD. Red blood cell distribution width and neutrophil to lymphocyte ratio are associated with outcomes of adult subarachnoid haemorrhage patients admitted to intensive care unit. Ann Clin Biochem 2017:4563216686623.

4. Saeed M, Villarroel M, Reisner AT, et al. Multiparameter Intelligent Monitoring in Intensive Care II: a public-access intensive care unit database. Crit Care Med 2011;39:952-60.

5. Johnson AE, Pollard TJ, Shen L, et al. MIMIC-III, a freely

accessible critical care database. Sci Data 2016;3:160035.

6. Zhang Z. Accessing critical care big data: a step by step approach. J Thorac Dis 2015;7:238-42.

7. Zhou XD, Zhang JY, Liu WY, et al. Quick chronic liver failure-sequential organ failure assessment: an easy-to-use scoring model for predicting mortality risk in critically ill cirrhosis patients. Eur J Gastroenterol Hepatol 2017;29:698-705.

8. Wang SD, Shen Y. Big-data Clinical Trial (BCT): the third talk. J Thorac Dis 2015;7:E243-4.

9. Wang SD, Shen Y. Redefining big-data clinical trial (BCT). Ann Transl Med 2014;2:96.

10. Wang SD. Opportunities and challenges of clinical research in the big-data era: from RCT to BCT. J Thorac Dis 2013;5:721-3.