Chapter 2
Review of Clinical Databases

Jeff Marshall, Abdullah Chahin and Barret Rush

Take Home Messages

- There are several open access health datasets that promote effective retrospective comparative effectiveness research.
- These datasets hold a varying amount of data with representative variables that are conducive to specific types of research and populations. Understanding these characteristics of the particular dataset will be crucial in appropriately drawing research conclusions.

2.1 Introduction

Since the appearance of the first EHR in the 1960s, patient driven data accumulated for decades with no clear structure to make it meaningful and usable. With time, institutions began to establish databases that archived and organized data into central repositories. Hospitals were able to combine data from large ancillary services, including pharmacies, laboratories, and radiology studies, with various clinical care components (such as nursing plans, medication administration records, and physician orders). Here we present the reader with several large databases that are publicly available or readily accessible with little difficulty. As the frontier of healthcare research utilizing large datasets moves ahead, it is likely that other sources of data will become accessible in an open source environment.

2.2 Background

Initially, EHRs were designed for archiving and organizing patients' records. They then became coopted for billing and quality improvement purposes. With time, EHR driven databases became more comprehensive, dynamic, and interconnected.

MIT Critical Data, *Secondary Analysis of Electronic Health Records*,
DOI 10.1007/978-3-319-43742-2_2

However, the medical industry has lagged behind other industries in the utilization of big data. Research using these large datasets has been drastically hindered by the poor quality of the gathered data and poorly organised datasets. Contemporary medical data evolved to more than medical records allowing the opportunity for them to be analyzed in greater detail. Traditionally, medical research has relied on disease registries or chronic disease management systems (CDMS). These repositories are a priori collections of data, often specific to one disease. They are unable to translate data or conclusions to other diseases and frequently contain data on a cohort of patients in one geographic area, thereby limiting their generalizability.

In contrast to disease registries, EHR data usually contain a significantly larger number of variables enabling high resolution of data, ideal for studying complex clinical interactions and decisions. This new wealth of knowledge integrates several datasets that are now fully computerized and accessible. Unfortunately, the vast majority of large healthcare databases collected around the world restrict access to data. Some possible explanations for these restrictions include privacy concerns, aspirations to monetize the data, as well as a reluctance to have outside researchers direct access to information pertaining to the quality of care delivered at a specific institution. Increasingly, there has been a push to make these repositories freely open and accessible to researchers.

2.3 The Medical Information Mart for Intensive Care (MIMIC) Database

The MIMIC database (http://mimic.physionet.org) was established in October 2003 as a Bioengineering Research Partnership between MIT, Philips Medical Systems, and Beth Israel Deaconess Medical Center. The project is funded by the National Institute of Biomedical Imaging and Bioengineering [1].

This database was derived from medical and surgical patients admitted to all Intensive Care Units (ICU) at Beth Israel Deaconess Medical Center (BIDMC), an academic, urban tertiary-care hospital. The third major release of the database, MIMIC-III, currently contains more than 40 thousand patients with thousands of variables. The database is de-identified, annotated and is made openly accessible to the research community. In addition to patient information driven from the hospital, the MIMIC-III database contains detailed physiological and clinical data [2]. In addition to big data research in critical care, this project aims to develop and evaluate advanced ICU patient monitoring and decision support systems that will improve the efficiency, accuracy, and timeliness of clinical decision-making in critical care.

Through data mining, such a database allows for extensive epidemiological studies that link patient data to clinical practice and outcomes. The extremely high granularity of the data allows for complicated analysis of complex clinical problems.

2.3.1 Included Variables

There are essentially two basic types of data in the MIMIC-III database; clinical data driven from the EHR such as patients' demographics, diagnoses, laboratory values, imaging reports, vital signs, etc (Fig. 2.1). This data is stored in a relational database of approximately 50 tables. The second primary type of data is the bedside monitor waveforms with associated parameters and events stored in flat binary files (with ASCII header descriptors). This unique library includes high-resolution data driven from tracings recorded from patients' electroencephalograms (EEGs), electrocardiograms (EKGs or ECGs), and real-time, second to second tracings of vital signs of patients in the intensive care unit. IRB determined the requirement for individual patient consent was waived, as all public data were de-identified.

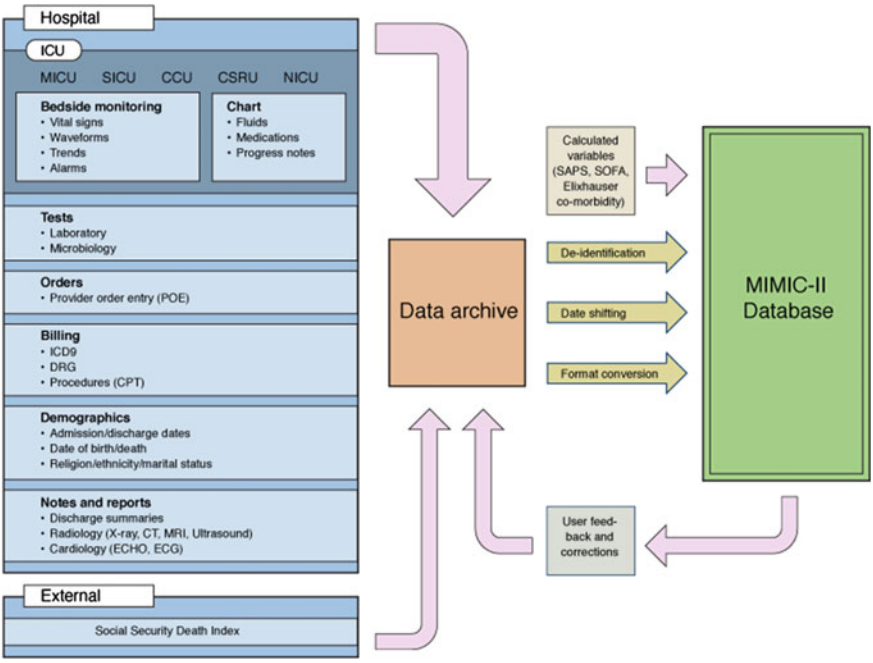

Fig. 2.1 Basic overview of the MIMIC database

2.3.2 Access and Interface

MIMIC-III is an open access database available to any researchers around the globe who are appropriately trained to handle sensitive patient information. The database is maintained by PhysioNet (http://physionet.org), a diverse group of computer scientists, physicists, mathematicians, biomedical researchers, clinicians, and educators around the world. The third release was published in 2015 and is anticipated to continually be updated with additional patients as time progresses.

2.4 PCORnet

PCORnet, the National Patient-Centered Clinical Research Network, is an initiative of the Patient-Centered Outcomes Research Institute (PCORI). PCORI involves patients as well as those who care for them in a substantive way in the governance of the network and in determining what questions will be studied. This PCORnet initiative was started in 2013, hoping to integrate data from multiple Clinical Data Research Networks (CDRNs) and Patient-Powered Research Networks (PPRNs) [3]. Its coordinating center bonds 9 partners: Harvard Pilgrim Health Care Institute, Duke Clinical Research Institute, AcademyHealth, Brookings Institution, Center for Medical Technology Policy, Center for Democracy & Technology, Group Health Research Institute, Johns Hopkins Berman Institute of Bioethics, and America's Health Insurance Plans. PCORnet includes 29 individual networks that together will enable access to large amounts of clinical and healthcare data. The goal of PCORnet is to improve the capacity to conduct comparative effectiveness research efficiently.

2.4.1 Included Variables

The variables in PCORnet database are driven from the various EHRs used in the nine centers forming this network. It captures clinical data and health information that are created every day during routine patient visits. In addition, PCORNet is using data shared by individuals through personal health records or community networks with other patients as they manage their conditions in their daily lives. This initiative will facilitate research on various medical conditions, engage a wide range of patients from all types of healthcare settings and systems, and provide an excellent opportunity to conduct multicenter studies.

2.4.2 Access and Interface

PCORnet is envisioned as a national research resource that will enable teams of health researchers and patients to work together on questions of shared interest. These teams will be able to submit research queries and receive to data conduct studies. Current PCORnet participants (CDRNs, PPRNs and PCORI) are developing the governance structures during the 18-month building and expansion phase [4].

2.5 Open NHS

The National Health Services (NHS England) is an executive non-departmental public body of the Department of Health, a governmental entity. The NHS retains one of the largest repositories of data on people's health in the world. It is also one of only a handful of health systems able to offer a full account of health across care sectors and throughout lives for an entire population.

Open NHS is one branch that was established in October of 2011. The NHS in England has actively moved to open the vast repositories of information used across its many agencies and departments. The main objective of the switch to an open access dataset was to increase transparency and trace the outcomes and efficiency of the British healthcare sector [5]. High quality information is hoped to empower the health and social care sector in identifying priorities to meet the needs of local populations. The NHS hopes that by allowing patients, clinicians, and commissioners to compare the quality and delivery of care in different regions of the country using the data, they can more effectively and promptly identify where the delivery of care is less than ideal.

2.5.1 Included Variables

Open NHS is an open source database that contains publicly released information, often from the government or other public bodies.

2.5.2 Access and Interface

Prior to the creation of Open NHS platform, SUS (Secondary Uses Service) was set up as part of the National Programme for IT in the NHS to provide data for planning, commissioning, management, research and auditing. Open NHS has now replaced SUS as a platform for accessing the national database in the UK.

The National Institute of Health Research (NIHR) Clinical Research Network (CRN) has produced and implemented an online tool known as the Open Data Platform.

In addition to the retrospective research that is routinely conducted using such databases, another form of research is already under way to compare the data quality derived from electronic records with that collected by research nurses. Clinical Research Network staff can access the Open Data Platform and determine the number of patients recruited into research studies in a given hospital as well as the research being done at that hospital. They then determine which hospitals are most successful at recruiting patients, the speed with which they recruit, and in what specialty fields.

2.6 Other Ongoing Research

The following are other datasets that are still under development or have more restrictive access limitations:

2.6.1 eICU—Philips

As part of its collaboration with MIT, Philips will be granting access to data from hundreds of thousands of patients that have been collected and anonymized through the Philips Hospital to Home eICU telehealth program. The data will be available to researchers via PhysioNet, similar to the MIMIC database.

2.6.2 VistA

The **Veterans Health Information Systems and Technology Architecture (VistA)** is an enterprise-wide information system built around the Electronic Health Record (EHR), used throughout the United States Department of Veterans Affairs (VA) medical system. The VA health care system operates over 125 hospitals, 800 ambulatory clinics and 135 nursing homes. All of these healthcare facilities utilize the VistA interface that has been in place since 1997. The VistA system amalgamates hospital, ambulatory, pharmacy and ancillary services for over 8 million US veterans. While the health network has inherent research limitations and biases due to its large percentage of male patients, the staggering volume of high fidelity records available outweighs this limitation. The VA database has been used by numerous medical researchers in the past 25 years to conduct landmark research in many areas [6, 7].

The VA database has a long history of involvement with medical research and collaboration with investigators who are part of the VA system. Traditionally the

dataset access has been limited to those who hold VA appointments. However, with the recent trend towards open access of large databases, there are ongoing discussions to make the database available to more researchers. The vast repository of information contained in the database would allow a wide range of researchers to improve clinical care in many domains. Strengths of the data include the ability to track patients across the United States as well as from the inpatient to outpatient settings. As all prescription drugs are covered by the VA system, the linking of this data enables large pharmacoepidemiological studies to be done with relative ease.

2.6.3 NSQUIP

The National Surgical Quality Improvement Project is an international effort spearheaded by the American College of Surgeons (ACS) with a goal of improving the delivery of surgical care worldwide [8]. The ACS works with institutions to implement widespread interventions to improve the quality of surgical delivery in the hospital. A by-product of the system is the gathering of large amounts of data relating to surgical procedures, outcomes and adverse events. All information is gathered from the EHR at the specific member institutions.

The NSQUIP database is freely available to members of affiliated institutions, of which there are over 653 participating centers in the world. This database contains large amounts of information regarding surgical procedures, complications, and baseline demographic and hospital information. While it does not contain the granularity of the MIMIC dataset, it contains data from many hospitals across the world and thus is more generalizable to real-world surgical practice. It is a particularly powerful database for surgical care delivery and quality of care, specifically with regard to details surrounding complications and adverse events from surgery.

References

1. Lee J, Scott DJ, Villarroel M, Clifford GD, Saeed M, Mark RG (2011) Open-access MIMIC-II database for intensive care research. In: Annual international conference of the IEEE engineering in medicine and biology society, pp 8315–8318
2. Scott DJ, Lee J, Silva I et al (2013) Accessing the public MIMIC-II intensive care relational database for clinical research. BMC Med Inform Decis Mak 13:9
3. Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS (2014) Launching PCORnet, a national patient-centered clinical research network. J Am Med Inform Assoc JAMIA 21 (4):578–582
4. Califf RM (2014) The patient-centered outcomes research network: a national infrastructure for comparative effectiveness research. N C Med J 75(3):204–210
5. Open data at the NHS [Internet]. Available from: http://www.england.nhs.uk/ourwork/tsd/data-info/open-data/
6. Maynard C, Chapko MK (2004) Data resources in the department of veterans affairs. Diab Care 27(Suppl 2):B22–B26
7. Smith BM, Evans CT, Ullrich P et al (2010) Using VA data for research in persons with spinal cord injuries and disorders: lessons from SCI QUERI. J Rehabil Res Dev 47(8):679–688
8. NSQUIP at the American College of Surgeons [Internet]. Available from: https://www.facs.org/quality-programs/acs-nsqip