Part I
Setting the Stage: Rationale Behind and Challenges to Health Data Analysis

Introduction

While wonderful new medical discoveries and innovations are in the news every day, healthcare providers continue to struggle with using information. Uncertainties and unanswered clinical questions are a daily reality for the decision makers who provide care. Perhaps the biggest limitation in making the best possible decisions for patients is that the information available is usually not focused on the specific individual or situation at hand.

For example, there are general clinical guidelines that outline the ideal target blood pressure for a patient with a severe infection. However, the truly best blood pressure levels likely differ from patient to patient, and perhaps even change for an individual patient over the course of treatment. The ongoing computerization of health records presents an opportunity to overcome this limitation. By analyzing electronic data from many providers' experiences with many patients, we can move ever closer to answering the age-old question: What is truly best for each patient?

Secondary analysis of routinely collected data—contrasted with the primary analysis conducted in the process of caring for the individual patient—offers an opportunity to extract more knowledge that will lead us towards the goal of optimal care. Today, a report from the National Academy of Medicine tells us, most doctors base most of their everyday decisions on guidelines from (sometimes biased) expert opinions or small clinical trials. It would be better if they were from multi-center, large, randomized controlled studies, with tightly controlled conditions ensuring the results are as reliable as possible. However, those are expensive and difficult to perform, and even then often exclude a number of important patient groups on the basis of age, disease and sociological factors.

Part of the problem is that health records are traditionally kept on paper, making them hard to analyze en masse. As a result, most of what medical professionals might have learned from experiences is lost, or is inaccessible at least. The ideal digital system would collect and store as much clinical data as possible from as many patients as possible. It could then use information from the past—such as blood pressure, blood sugar levels, heart rate, and other measurements of patients'

body functions—to guide future providers to the best diagnosis and treatment of similar patients.

But "big data" in healthcare has been coated in "Silicon Valley Disruptionese", the language with which Silicon Valley spins hype into startup gold and fills it with grandiose promises to lure investors and early users. The buzz phrase "precision medicine" looms large in the public consciousness with little mention of the failures of "personalized medicine", its predecessor, behind the façade.

This part sets the stage for secondary analysis of electronic health records (EHR). Chapter 1 opens with the rationale behind this type of research. Chapter 2 provides a list of existing clinical databases already in use for research. Chapter 3 dives into the opportunities, and more importantly, the challenges to retrospective analysis of EHR. Chapter 4 presents ideas on how data could be systematically and more effectively employed in a purposefully engineered healthcare system. Professor Roger Mark, the visionary who created the Medical Information Mart for Intensive Care or MIMIC database that is used in this textbook, narrates the story behind the project in Chap. 5. Chapter 6 steps into the future and describes integration of EHR with non-clinical data for a richer representation of health and disease. Chapter 7 focuses on the role of EHR in two important areas of research—outcome and health services. Finally, Chap. 8 tackles the bane of observational studies using EHR: residual confounding.

We emphasize the importance of bringing together front-line clinicians such as nurses, pharmacists and doctors with data scientists to collaboratively identify questions and to conduct appropriate analyses. Further, we believe this research partnership of practitioner and researcher gives caregivers and patients the best individualized diagnostic and treatment options in the absence of a randomized controlled trial. By becoming more comfortable with the data available to us in the hospitals of today, we can reduce the uncertainties that have hindered healthcare for far too long.

Chapter 1
Objectives of the Secondary Analysis of Electronic Health Record Data

Sharukh Lokhandwala and Barret Rush

Take Home Messages

- Clinical medicine relies on a strong research foundation in order to build the necessary evidence base to inform best practices and improve clinical care, however, large-scale randomized controlled trials (RCTs) are expensive and sometimes unfeasible. Fortunately, there exists expansive data in the form of electronic health records (EHR).
- Data can be overwhelmingly complex or incomplete for any individual, therefore we urge multidisciplinary research teams consisting of clinicians along with data scientists to unpack the clinical semantics necessary to appropriately analyze the data.

1.1 Introduction

The healthcare industry has rapidly become computerized and digital. Most healthcare delivered in America today relies on or utilizes technology. Modern healthcare informatics generates and stores immense amounts of detailed patient and clinical process data. Very little real-world patient data have been used to further advance the field of health care. One large barrier to the utilization of these data is inaccessibility to researchers. Making these databases easier to access as well as integrating the data would allow more researchers to answer fundamental questions of clinical care.

1.2 Current Research Climate

Many treatments lack proof in their efficacy, and may, in fact, cause harm [1]. Various medical societies disseminate guidelines to assist clinician decision-making and to standardize practice; however, the evidence used to formulate these guidelines is inadequate. These guidelines are also commonly derived from RCTs with

MIT Critical Data, *Secondary Analysis of Electronic Health Records*,
DOI 10.1007/978-3-319-43742-2_1

limited patient cohorts and with extensive inclusion and exclusion criteria resulting in reduced generalizability. RCTs, the gold standard in clinical research, support only 10–20 % of medical decisions [2] and most clinical decisions have never been supported by RCTs [3]. Furthermore, it would be impossible to perform randomized trials for each of the extraordinarily large number of decisions clinicians face on a daily basis in caring for patients for numerous reasons, including constrained financial and human resources. For this reason, clinicians and investigators must learn to find clinical evidence from the droves of data that already exists: the EHR.

1.3 Power of the Electronic Health Record

Much of the work utilizing large databases in the past 25 years have relied on hospital discharge records and registry databases. Hospital discharge databases were initially created for billing purposes and lack the patient level granularity of clinically useful, accurate, and complete data to address complex research questions. Registry databases are generally mission-limited and require extensive extracurricular data collection. The future of clinical research lies in utilizing big data to improve the delivery of care to patients.

Although several commercial and non-commercial databases have been created using clinical and EHR data, their primary function has been to analyze differences in severity of illness, outcomes, and treatment costs among participating centers. Disease specific trial registries have been formulated for acute kidney injury [4], acute respiratory distress syndrome [5] and septic shock [6]. Additionally, databases such as the Dartmouth Atlas utilize Medicare claims data to track discrepancies in costs and patient outcomes across the United States [7]. While these coordinated databases contain a large number of patients, they often have a narrow scope (i.e. for severity of illness, cost, or disease specific outcomes) and lack other significant clinical data that is required to answer a wide range of research questions, thus obscuring many likely confounding variables.

For example, the APACHE Outcomes database was created by merging APACHE (Acute Physiology and Chronic Health Evaluation) [8] with Project IMPACT [9] and includes data from approximately 150,000 intensive care unit (ICU) stays since 2010 [1]. While the APACHE Outcomes database is large and has contributed significantly to the medical literature, it has incomplete physiologic and laboratory measurements, and does not include provider notes or waveform data. The Phillips eICU [10], a telemedicine intensive care support provider, contains a database of over 2 million ICU stays. While it includes provider documentation entered into the software, it lacks clinical notes and waveform data. Furthermore, databases with different primary objectives (i.e., costs, quality improvement, or research) focus on different variables and outcomes, so caution must be taken when interpreting analyses from these databases.

Since 2003, the Laboratory for Computational Physiology at the Massachusetts Institute of Technology partnered in a joint venture with Beth Israel Deaconess Medical Center and Philips Healthcare, with support from the National Institute of Biomedical Imaging and Bioinformatics (NIBIB), to develop and maintain the Medical Information Mart for Intensive Care (MIMIC) database [11]. MIMIC is a public-access database that contains comprehensive clinical data from over 60,000 inpatient ICU admissions at Beth Israel Deaconess Medical Center. The de-identified data are freely shared, and nearly 2000 investigators from 32 countries have utilized it to date. MIMIC contains physiologic and laboratory data, as well as waveform data, nurse verified numerical data, and clinician documentation. This high resolution, widely accessible, database has served to support research in critical care and assist in the development of novel decision support algorithms, and will be the prototype example for the majority of this textbook.

1.4 Pitfalls and Challenges

Clinicians and data scientists must apply the same level of academic rigor when analyzing research from clinical databases as they do with more traditional methods of clinical research. To ensure internal and external validity, researchers must determine whether the data are accurate, adjusted properly, analyzed correctly, and presented cogently [12]. With regard to quality improvement projects, which frequently utilize hospital databases, one must ensure that investigators are applying rigorous standards to the performance and reporting of their studies [13].

Despite the tremendous value that the EHR contains, many clinical investigators are hesitant to use it to its full capacity partly due to its sheer complexity and the inability to use traditional data processing methods with large datasets. As a solution to the increased complexity associated with this type of research, we suggest that investigators work in collaboration with multidisciplinary teams including data scientists, clinicians and biostatisticians. This may require a shift in financial and academic incentives so that individual research groups do not compete for funding or publication; the incentives should promote joint funding and authorship. This would allow investigators to focus on the fidelity of their work and be more willing to share their data for discovery, rather than withhold access to a dataset in an attempt to be "first" to a solution.

Some have argued that the use of large datasets may increase the frequency of so-called "p-hacking," wherein investigators search for significant results, rather than seek answers to clinically relevant questions. While it appears that p-hacking is widespread, the mean effect size attributed to p-hacking does not generally undermine the scientific consequences from large studies and meta-analyses. The use of large datasets may, in fact, reduce the likelihood of p-hacking by ensuring that researchers have suitable power to answer questions with even small effect

sizes, making the need for selective interpretation and analysis of the data to obtain significant results unnecessary. If significant discoveries are made utilizing big databases, this work can be used as a foundation for more rigorous clinical trials to confirm these findings. In the future, once comprehensive databases become more accessible to researchers, it is hoped that these resources can be used as hypothesis generating and testing ground for questions that will ultimately undergo RCT. If there is not a strong signal observed in a large preliminary retrospective study, proceeding to a resource-intensive and time-consuming RCT may not be advisable.

1.5 Conclusion

With advances in data collection and technology, investigators have access to more patient data than at any time in history. Currently, much of these data are inaccessible and underused. The ability to harness the EHR would allow for continuous learning systems, wherein patient specific data are able to feed into a population-based database and provide real-time decision support for individual patients based on data from similar patients in similar scenarios. Clinicians and patients would be able to make better decisions with those resources in place and the results would feed back into the population database [14].

The vast amount of data available to clinicians and scientists poses daunting challenges as well as a tremendous opportunity. The National Academy of Medicine has called for clinicians and researchers to create systems that "foster continuous learning, as the lessons from research and each care experience are systematically captured, assessed and translated into reliable care" [2]. To capture, assess, and translate these data, we must harness the power of the EHR to create data repositories, while also providing clinicians as well as patients with data-driven decision support tools to better treat patients at the bedside.

References

1. Celi LA, Mark RG, Stone DJ, Montgomery RA (2013) "Big data" in the intensive care unit. Closing the data loop. Am J Respir Crit Care Med 187:1157–1160
2. Smith M, Saunders R, Stuckhardt L, McGinnis JM (2013) Best care at lower cost: the path to continuously learning health care in America. National Academies Press
3. Mills EJ, Thorlund K, Ioannidis JP (2013) Demystifying trial networks and network meta-analysis. BMJ 346:f2914
4. Mehta RL, Kellum JA, Shah SV, Molitoris BA, Ronco C, Warnock DG, Levin A, Acute Kidney Injury N (2007) Acute Kidney Injury Network: report of an initiative to improve outcomes in acute kidney injury. Crit Care 11:R31
5. The Acute Respiratory Distress Syndrome Network (2000) Ventilation with lower tidal volumes as compared with traditional tidal volumes for acute lung injury and the acute respiratory distress syndrome. N Engl J Med 342:1301–1308
6. Dellinger RP, Levy MM, Rhodes A, Annane D, Gerlach H, Opal SM, Sevransky JE, Sprung CL, Douglas IS, Jaeschke R, Osborn TM, Nunnally ME, Townsend SR, Reinhart K, Kleinpell RM, Angus DC, Deutschman CS, Machado FR, Rubenfeld GD, Webb SA, Beale RJ, Vincent JL, Moreno R, Surviving Sepsis Campaign Guidelines Committee including the Pediatric S (2013) Surviving sepsis campaign: international guidelines for management of severe sepsis and septic shock: 2012. Crit Care Med 41:580–637
7. The Dartmouth Atlas of Health Care. Lebanon, NH. The Trustees of Dartmouth College 2015. Accessed 10 July 2015. Available from http://www.dartmouthatlas.org/
8. Zimmerman JE, Kramer AA, McNair DS, Malila FM, Shaffer VL (2006) Intensive care unit length of stay: Benchmarking based on Acute Physiology and Chronic Health Evaluation (APACHE) IV. Crit Care Med 34:2517–2529
9. Cook SF, Visscher WA, Hobbs CL, Williams RL, Project ICIC (2002) Project IMPACT: results from a pilot validity study of a new observational database. Crit Care Med 30:2765–2770
10. eICU Program Solution. Koninklijke Philips Electronics N.V, Baltimore, MD (2012)
11. Saeed M, Villarroel M, Reisner AT, Clifford G, Lehman L-W, Moody G, Heldt T, Kyaw TH, Moody B, Mark RG (2011) Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): a public-access intensive care unit database. Crit Care Med 39:952
12. Meurer S (2008) Data quality in healthcare comparative databases. MIT Information Quality Industry symposium
13. Davidoff F, Batalden P, Stevens D, Ogrinc G, Mooney SE, group Sd (2009) Publication guidelines for quality improvement studies in health care: evolution of the SQUIRE project. BMJ 338:a3152
14. Celi LA, Zimolzak AJ, Stone DJ (2014) Dynamic clinical data mining: search engine-based decision support. JMIR Med Informatics 2:e13

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.

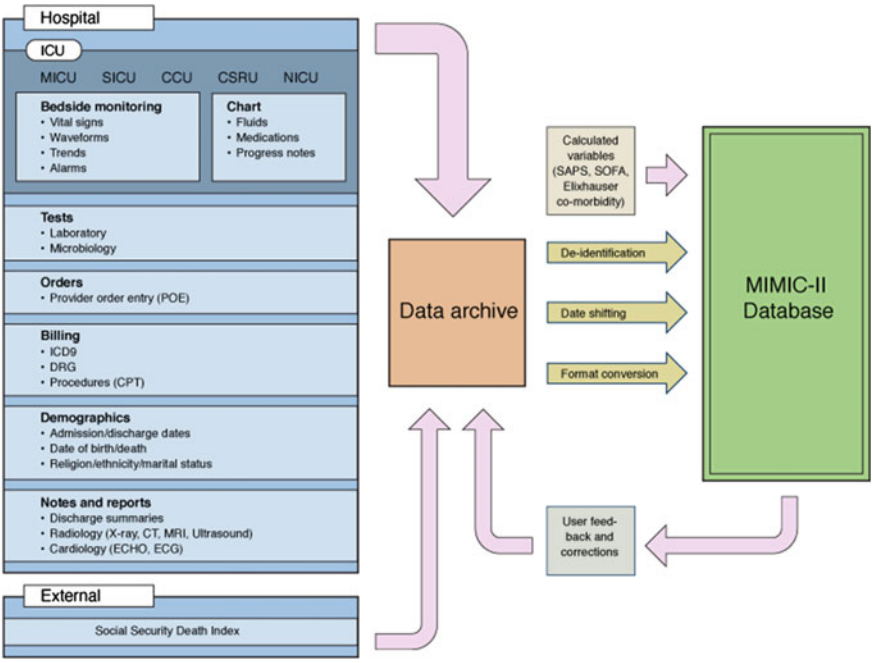

Fig. 2.1 Basic overview of the MIMIC database

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

Chapter 3
Challenges and Opportunities in Secondary Analyses of Electronic Health Record Data

Sunil Nair, Douglas Hsu and Leo Anthony Celi

Take Home Messages

- Electronic health records (EHR) are increasingly useful for conducting secondary observational studies with power that rivals randomized controlled trials.
- Secondary analysis of EHR data can inform large-scale health systems choices (e.g., pharmacovigilance) or point-of-care clinical decisions (e.g., medication selection).
- Clinicians, researchers and data scientists will need to navigate numerous challenges facing big data analytics—including systems interoperability, data sharing, and data security—in order to utilize the full potential of EHR and big data-based studies.

3.1 Introduction

The increased adoption of EHR has created novel opportunities for researchers, including clinicians and data scientists, to access large, enriched patient databases. With these data, investigators are in a position to approach research with statistical power previously unheard of. In this chapter, we present and discuss challenges in the secondary use of EHR data, as well as explore the unique opportunities provided by these data.

3.2 Challenges in Secondary Analysis of Electronic Health Records Data

Tremendous strides have been made in making pooled health records available to data scientists and clinicians for health research activities, yet still more must be done to harness the full capacity of big data in health care. In all health related

MIT Critical Data, *Secondary Analysis of Electronic Health Records*,
DOI 10.1007/978-3-319-43742-2_3

fields, the data-holders—i.e., pharmaceutical firms, medical device companies, health systems, and now burgeoning electronic health record vendors—are simultaneously facing pressures to protect their intellectual capital and proprietary platforms, ensure data security, and adhere to privacy guidelines, without hindering research which depends on access to these same databases. Big data success stories are becoming more common, as highlighted below, but the challenges are no less daunting than they were in the past, and perhaps have become even more demanding as the field of data analytics in healthcare takes off.

Data scientists and their clinician partners have to contend with a research culture that is highly competitive—both within academic circles, and among clinical and industrial partners. While little is written about the nature of data secrecy within academic circles, it is a reality that tightening budgets and greater concerns about data security have pushed researchers to use such data as they have on-hand, rather than seek integration of separate databases. Sharing data in a safe and scalable manner is extremely difficult and costly or impossible even within the same institution. With access to more pertinent data restricted or impeded, statistical power and the ability for longitudinal analysis are reduced or lost. None of this is to say researchers have hostile intentions—in fact, many would appreciate the opportunity for greater collaboration in their projects. However, the time, funding, and infrastructure for these efforts are simply deficient. Data is also often segregated into various locales and not consistently stored in similar formats across clinical or research databases. For example, most clinical data is kept in a variety of unstructured formats, making it difficult to query directly via digital algorithms [1]. Within many hospitals, emergency department or outpatient clinical data may exist separately from the hospital and the Intensive Care Unit (ICU) electronic health records, so that access to one does not guarantee access to the other. Images from Radiology and Pathology are typically stored separately in yet other different systems and therefore are not easily linked to outcomes data. The Medical Information Mart for Intensive Care (MIMIC) database described later in this chapter, which contains ICU EHR data from the Beth Israel Deaconess Medical Center (BIDMC), addresses and resolves these artificial divisions, but requires extensive engineering and support staff not afforded to all institutions.

After years of concern about data secrecy, the pharmaceutical industry has recently turned a corner, making detailed trial data available to researchers outside their organizations. GlaxoSmithKline was among the first in 2012 [2], followed by a larger initiative—the Clinical Trial Data Request—to which other large pharmaceutical firms have signed-on [3]. Researchers can apply for access to large-scale information, and integrate datasets for meta-analysis and other systematic reviews. The next frontier will be the release of medical records held at the health system level. The 2009 Health Information Technology for Economic and Clinical Health (HITECH) Act was a boon to the HIT sector [4], but standards for interoperability between record systems continue to lag [5]. The gap has begun to be resolved by government sponsored health information exchanges, as well as the creation of novel research networks [6, 7], but most experts, data scientists, and working clinicians continue to struggle with incomplete data.

Many of the commercial and technical roadblocks alluded to above have their roots in the privacy concerns held by vendors, providers and their patients. Such concerns are not without merit—data breaches of large health systems are becoming distressingly common [8]. Employees of Partners Healthcare in Boston were recently targeted in a "phishing" scheme, unwittingly providing personal information that allowed hackers unauthorized access to patient information [9]; patients of Seton Healthcare in Texas suffered a similar breach just a few months prior [10]. Data breaches aren't limited to healthcare providers—80 million Anthem enrollees may have suffered loss of their personal information to a cyberattack, the largest of its kind to-date [11]. Not surprisingly in the context of these breaches, healthcare companies have some of the lowest scores of all industries in email security and privacy practices [12]. Such reports highlight the need for prudence amidst exuberance when utilizing pooled electronic health records for big data analytics—such use comes with an ethical responsibility to protect population- and personal-level data from criminal activity and other nefarious ends. For this purpose, federal agencies have convened working groups and public hearings to address gaps in health information security, such as the de-identification of data outside HIPAA-covered entities, and consensus guidelines on what constitutes "harm" from a data breach [13].

Even when issues of data access, integrity, interoperability, security and privacy have been successfully addressed, substantial infrastructure and human capital costs will remain. Though the marginal cost of each additional big data query is small, the upfront cost to host a data center and employ dedicated data scientists can be significant. No figures exist for the creation of a healthcare big data center, and these figures would be variable anyway, depending on the scale and type of data. However, it should not be surprising that commonly cited examples of pooled EHRs with overlaid analytic capabilities—MIMIC (BIDMC), STRIDE (Stanford), the MemorialCare data mart (Memorial Health System, California, $2.2 Billion annual revenue), and the High Value Healthcare Collaborative (hosted by Dartmouth, with 16 other members and funding from the Center for Medicare and Medicaid Services) [14]—come from large, high revenue healthcare systems with regional big-data expertise.

In addition to the above issues, the reliability of studies published using big data methods is of significant concern to experts and physicians. The specific issue is whether these studies are simply amplifications of low-level signals that do not have clinical importance, or are generalizable beyond the database from which they are derived. These are genuine concerns in a medical and academic atmosphere already saturated with innumerable studies of variable quality. Skeptics are concerned that big data analytics will only, "add to the noise," diverting attention and resources from other venues of scientific inquiry, such as the traditional randomized controlled clinical trial (RCT). While the limitations of RCTs, and the favorable comparison of large observational study results to RCT findings are discussed below, these sentiments nevertheless have merit and must be taken seriously as

secondary analysis of EHR data continues to grow. Thought leaders have suggested expounding on the big data principles described above to create open, collaborative learning environments, whereby de-identified data can be shared between researchers—in this manner, data sets can be pooled for greater power, or similar inquiries run on different data sets to see if similar conclusions are reached [15]. The costs for such transparency could be borne by a single institution—much of the cost of creating MIMIC has already been invested, for instance, so the incremental cost of making the data open to other researchers is minimal—or housed within a dedicated collaborative—such as the High Value Healthcare Collaborative funded by its members [16] or PCORnet, funded by the federal government [7]. These collaborative ventures would have transparent governance structures and standards for data access, permitting study validation and continuous peer review of published and unpublished works [15], and mitigating the effects of selection bias and confounding in any single study [17].

As pooled electronic health records achieve even greater scale, data scientists, researchers and other interested parties expect that the costs of hosting, sorting, formatting and analyzing these records are spread among a greater number of stakeholders, reducing the costs of pooled EHR analysis for all involved. New standards for data sharing may have to come into effect for institutions to be truly comfortable with records-sharing, but within institutions and existing research collaboratives, safe practices for data security can be implemented, and greater collaboration encouraged through standardization of data entry and storage. Clear lines of accountability for data access should be drawn, and stores of data made commonly accessible to clarify the extent of information available to any institutional researcher or research group. The era of big data has arrived in healthcare, and only through continuous adaptation and improvement can its full potential be achieved.

3.3 Opportunities in Secondary Analysis of Electronic Health Records Data

The rising adoption of electronic health records in the U.S. health system has created vast opportunities for clinician scientists, informaticians and other health researchers to conduct queries on large databases of amalgamated clinical information to answer questions both large and small. With troves of data to explore, physicians and scientists are in a position to evaluate questions of clinical efficacy and cost-effectiveness—matters of prime concern in 21st century American health care—with a qualitative and statistical power rarely before realized in medical research. The commercial APACHE Outcomes database, for instance, contains physiologic and laboratory measurements from over 1 million patient records across 105 ICUs since 2010 [18]. The Beth Israel Deaconess Medical Center—a tertiary

care hospital with 649 licensed beds including 77 critical care beds—provides an open-access single-center database (MIMIC) encompassing data from over 60,000 ICU stays [19].

Single- and multi-center databases such as those above permit large-scale inquiries without the sometimes untenable expense and difficulty of a randomized clinical trial (RCT), thus answering questions previously untestable in RCTs or prospective cohort studies. This can also be done with increased precision in the evaluation of diagnostics or therapeutics for select sub-populations, and for the detection of adverse events from medications or other interventions with greater expediency, among other advantages [20]. In this chapter, we offer further insight into the utility of secondary analysis of EHR data to investigate relevant clinical questions and provide useful decision support to physicians, allied health providers and patients.

3.4 Secondary EHR Analyses as Alternatives to Randomized Controlled Clinical Trials

The relative limitations of RCTs to inform real-world clinical decision-making include the following: many treatment comparisons of interest to clinicians have not been addressed by RCTs; when RCTs have been performed and appraised, half of systemic reviews of RCTs report insufficient evidence to support a given medical intervention; and, there are realistic cost and project limitations that prevent RCTs from exploring specific clinical scenarios. The latter include rare conditions, clinically uncommon or disparate events, and a growing list of combinations of recognized patient sub-groups, concurrent conditions (genetic, chronic, acute and healthcare-acquired), and diagnostic and treatment options [20, 21].

Queries on EHR databases to address clinical questions are essentially large, nonrandomized observational studies. Compared to RCTs, they are relatively more efficient and less expensive to perform [22], the majority of the costs having been absorbed by initial system installation and maintenance, and the remainder consisting primarily of research personnel salaries, server or cloud space costs. There is literature to suggest a high degree of correlation between treatment effects reported in nonrandomized studies and randomized clinical trials. Ioannidis et al. [23] found significant correlation (Spearman coefficient of 0.75, $p < 0.001$) between the treatment effects reported in randomized trials versus nonrandomized studies across 45 diverse topics in general internal medicine, ranging from anticoagulation in myocardial infarction to low-level laser therapy for osteoarthritis. Of particular interest, significant variability in reported treatment outcome "was seen as frequently among the randomized trials as between the randomized and nonrandomized studies," and they observed that variability was common among *both* randomized trials and nonrandomized studies [23]. It is worth pointing out that larger treatment effects were more frequently reported in nonrandomized studies than randomized trials (exact $p = 0.009$) [23]; however, this need not be evidence

of publication bias, as relative study size and conservative trial protocol could also cause this finding. Ioannidis et al.'s [24] results are echoed by a more recent Cochrane meta-analysis, which found no significant difference in effect estimates between RCTs and observational studies regardless of the observational study design or heterogeneity.

To further reduce confounding in observational studies, researchers have employed propensity scoring [25], which allows balancing of numerous covariates between treatment groups as well as stratification of samples by propensity score for more nuanced analysis [26]. Kitsios and colleagues matched 18 unique propensity score studies in the ICU setting with at least one RCT evaluating the same clinical question and found a high degree of agreement between their estimates of relative risk and effect size. There was substantial difference in the magnitude of effect sizes in a third of comparisons, reaching statistically significance in one case [27]. Though the RCT remains atop the hierarchy of evidence-based medicine, it is hard to ignore the power of large observational studies that include adequate adjusting for covariates, such as carefully performed studies derived from review of EHRs. The scope of pooled EHR data—whether sixty thousand or one million records—affords insight into small treatment effects that may be under-reported or even missed in underpowered RCTs. Because costs are small compared to RCTs, it is also possible to investigate questions where realistically no study-sponsor will be found. Finally, in the case of databased observational studies, it becomes much more feasible to improve and repeat, or simply repeat, studies as deemed necessary to investigate accuracy, heterogeneity of effects, and new clinical insights.

3.5 Demonstrating the Power of Secondary EHR Analysis: Examples in Pharmacovigilance and Clinical Care

The safety of pharmaceuticals is of high concern to both patients and clinicians. However, methods for ensuring detection of adverse events post-release are less robust than might be desirable. Pharmaceuticals are often prescribed to a large, diverse patient population that may have not been adequately represented in pre-release clinical trials. In fact, RCT cohorts may deliberately be relatively homogeneous in order to capture the intended effect(s) of a medication without "noise" from co-morbidities that could modulate treatment effects [28]. Humphreys and colleagues (2013) reported that in highly-cited clinical trials, 40 % of identified patients with the condition under consideration were not enrolled, mainly due to restrictive eligibility criteria [29]. Variation in trial design (comparators, endpoints, duration of follow-up) as well as trial size limit their ability to detect low-frequency or long-term side-effects and adverse events [28]. Post-market surveillance reports are imperfectly collected, are not regularly amalgamated, and may not be publically accessible to support clinical-decision making by physicians or inform decision-making by patients.

Queries on pooled EHRs—essentially performing secondary observational studies on large study populations—could compensate for these gaps in pharmacovigilance. Single-center approaches for this and similar questions regarding medication safety in clinical environments are promising. For instance, the highly publicized findings of the Kaiser Study on Vioxx® substantiated prior suspicions of an association between celecoxib and increased risk of serious coronary heart disease [30]. These results were made public in April 2004 after presentation at an international conference; Vioxx® was subsequently voluntarily recalled from the market in September of the same year. Graham and colleagues were able to draw on *2,302,029* person-years of follow-up from the Kaiser Permanente database, to find 8143 cases of coronary heart disease across all NSAIDs under consideration, and subsequently drill-down to the appropriate odds ratios [31].

Using the MIMIC database mentioned above, researchers at the Beth Israel Deaconess Medical Center were able to describe for the first time an increased mortality risk for ICU patients who had been on selective serotonin reuptake inhibitors prior to admission [32]. A more granular analysis revealed that mortality varied by specific SSRI, with higher mortality among patients taking higher-affinity SSRIs (i.e., those with greater serotonin inhibition); on the other hand, mortality could not be explained by common SSRI adverse effects, such as impact on hemodynamic variables [32].

The utility of secondary analysis of EHR data is not limited to the discovery of treatment effects. Lacking published studies to guide their decision to potentially anticoagulate a pediatric lupus patient with multiple risk factors for thrombosis, physicians at Stanford turned to their own EHR-querying platform (the Stanford Translational Research Integrated Database Environment—STRIDE) to create an electronic cohort of pediatric lupus patients to study complications from this illness [33]. In four hours' time, a single clinician determined that patients with similar lupus complications had a high relative risk of thrombosis, and the decision was made to administer anticoagulation [33].

3.6 A New Paradigm for Supporting Evidence-Based Practice and Ethical Considerations

Institutional experiences such as those above, combined with evidence supporting the efficacy of observational trials to adequately inform clinical practice, validate the concept of pooled EHRs as large study populations possessing copious amounts of information waiting to be tapped for clinical decision support and patient safety. One can imagine a future clinician requesting a large or small query such as those described above. Such queries might relate to the efficacy of an intervention across a subpopulation, or for a single complicated patient whose circumstances are not satisfactorily captured in any published trial. Perhaps this is sufficient for the clinician to recommend a new clinical practice; or maybe they will design a

pragmatic observational study for more nuance—evaluating dose-responsiveness, or adverse effect profiles across subpopulations. As clinical decisions are made and the patient's course of care shaped, this intervention and outcomes information is entered into the electronic health record, effectively creating a feedback loop for future inquiries [34].

Of course, the advantages of secondary analysis of electronic health records must always be balanced with ethical considerations. Unlike traditional RCTs, there is no explicit consent process for the use of demographic, clinical and other potentially sensitive data captured in the EHR. Sufficiently specific queries could yield very narrow results—theoretically specific enough to re-identify an individual patient. For instance, an inquiry on patients with a rare disease, within a certain age bracket, and admitted within a limited timeframe, could include someone who may be known to the wider community. Such an extreme example highlights the need for compliance with federal privacy laws as well as ensuring high institutional standards of data security such as secured servers, limited access, firewalls from the internet, and other data safety methods.

Going further, data scientists should consider additional measures intentionally designed to protect patient anonymity, e.g. date shifting as implemented in the MIMIC database (see Sect. 5.1, Chap. 5). In situations where queries might potentially re-identify patients, such as in the investigation of rare diseases, or in the course of a contagious outbreak, researchers and institutional research boards should seek accommodation with this relatively small subset of potentially affected patients and their advocacy groups, to ensure their comfort with secondary analyses. Disclosure of research intent and methods by those seeking data access might be required, and a patient option to embargo one's own data should be offered.

It is incumbent on researchers and data scientists to explain the benefits of participation in a secondary analysis to patients and patient groups. Such sharing allows the medical system to create a clinical database of sufficient magnitude and quality to benefit individual- and groups of patients, in real-time or in the future. Also, passive clinical data collection allows the patient to contribute, at relatively very low risk and no personal cost, to the ongoing and future care of others. We believe that people are fundamentally sufficiently altruistic to consider contributions their data to research, provided the potential risks of data usage are small and well-described.

Ultimately, secondary analysis of EHR will only succeed if patients, regulators, and other interested parties are assured and reassured that their health data will be kept safe, and processes for its use are made transparent to ensure beneficence for all.

References

1. Riskin D (2012) Big data: opportunity and challenge. HealthcareITNews, 12 June 2012. URL: http://www.healthcareitnews.com/news/big-data-opportunity-and-challenge
2. Harrison C (2012) GlaxoSmithKline opens the door on clinical data sharing. Nat Rev Drug Discov 11(12):891–892. doi:10.1038/nrd3907 [Medline: 23197021]
3. Clinical Trial Data Request. URL: https://clinicalstudydatarequest.com/. Accessed 11 Aug 2015. [WebCite Cache ID 6TFyjeT7t]
4. Adler-Milstein J, Jha AK (2012) Sharing clinical data electronically: a critical challenge for fixing the health care system. JAMA 307(16):1695–1696
5. Verdon DR (2014) ONC's plan to solve the EHR interoperability puzzle: an exclusive interview with National Coordinator for Health IT Karen B. DeSalvo. Med Econ. URL: http://medicaleconomics.modernmedicine.com/medical-economics/news/onc-s-plan-solve-ehr-interoperability-puzzle?page=full
6. Green M (2015) 10 things to know about health information exchanges. Becker's Health IT CIO Rev. URL: http://www.beckershospitalreview.com/healthcare-information-technology/10-things-to-know-about-health-information-exchanges.html
7. PCORnet. URL: http://www.pcornet.org/. Accessed 11 Aug 2015
8. Dvorak K (2015) Big data's biggest healthcare challenge: making sense of it all. FierceHealthIT, 4 May 2015. URL: http://www.fiercehealthit.com/story/big-datas-biggest-healthcare-challenge-making-sense-it-all/2015-05-04
9. Bartlett J (2015) Partners healthcare reports data breach. Boston Bus J. URL: http://www.bizjournals.com/boston/blog/health-care/2015/04/partners-healthcare-reports-potential-data-breach.html
10. Dvorak K (2015) Phishing attack compromises info of 39 K at Seton healthcare family. FierceHealthIT, 28 April 2015. URL: http://www.fiercehealthit.com/story/phishing-attack-compromises-info-39k-seton-healthcare-family/2015-04-28
11. Bowman D (2015) Anthem hack compromises info for 80 million customers. FierceHealthPayer, 5 February 2015. URL: http://www.fiercehealthpayer.com/story/anthem-hack-compromises-info-80-million-customers/2015-02-05
12. Dvorak K (2015) Healthcare industry 'behind by a country mile' in email security. FierceHealthIT, 20 February 2015. URL: http://www.fiercehealthit.com/story/healthcare-industry-behind-country-mile-email-security/2015-02-20
13. White house seeks to leverage health big data, safeguard privacy. HealthData Manage. URL: http://www.healthdatamanagement.com/news/White-House-Seeks-to-Leverage-Health-Big-Data-Safeguard-Privacy-50829-1.html
14. How big data impacts healthcare. Harv Bus Rev. URL: https://hbr.org/resources/pdfs/comm/sap/18826_HBR_SAP_Healthcare_Aug_2014.pdf. Accessed 11 Aug 2015
15. Moseley ET, Hsu DJ, Stone DJ, Celi LA (2014) Beyond open big data: addressing unreliable research. J Med Internet Res 16(11):e259

16. High value healthcare collaborative. URL: http://highvaluehealthcare.org/. Accessed 14 Aug 2015
17. Badawi O, Brennan T, Celi LA et al (2014) Making big data useful for health care: a summary of the inaugural mit critical data conference. JMIR Med Inform 2(2):e22
18. APACHE Outcomes. Available at: https://www.cerner.com/Solutions/Hospitals_and_Health_Systems/Critical_Care/APACHE_Outcomes/. Accessed Nov 2014
19. Saeed M, Villarroel M, Reisner AT et al (2011) Multiparameter intelligent monitoring in intensive care II (MIMIC-II): a public-access intensive care unit database. Crit Care Med 39:952
20. Ghassemi M, Celi LA, Stone DJ (2015) State of the art review: the data revolution in critical care. Crit Care 19:118
21. Mills EJ, Thorlund K, Ioannidis J (2013) Demystifying trial networks and network meta-analysis. BMJ 346:f2914
22. Angus DC (2007) Caring for the critically ill patient: challenges and opportunities. JAMA 298:456–458
23. Ioannidis JPA, Haidich A-B, Pappa M et al (2001) Comparision of evidence of treatment effects in randomized and nonrandomized studies. JAMA 286:7
24. Anglemyer A, Horvath HT, Bero L (2014) Healthcare outcomes assess with observational study designs compared with those assessed in randomized trials. Cochrane Database Syst Rev 29:4
25. Gayat E, Pirracchio R, Resche-Rigon M et al (2010) Propensity scores in intensive care and anaesthesiology literature: a systematic review. Intensive Care Med 36:1993–2003
26. Glynn RJ, Schneeweiss S, Stürmer T (2006) Indications for propensity scores and review of their use in pharmacoepidemiology. Basic Clin Pharmacol Toxicol 98:253–259
27. Kitsios GD, Dahabreh IJ, Callahan S et al (2015) Can we trust observational studies using propensity scores in the critical care literature? A systematic comparison with randomized clinical trials. Crit Care Med (Epub ahead of print)
28. Celi LA, Moseley E, Moses C et al (2014) from pharmacovigilance to clinical care optimization. Big Data 2(3):134–141
29. Humphreys K, Maisel NC, Blodgett JC et al (2013) Extent and reporting of patient nonenrollment in influential randomized clinical trials, 2001 to 2010. JAMA Intern Med 173:1029–1031
30. Vioxx and Drug Safety. Statement of Sandra Kweder M.D. (Deputy Director, Office of New Drugs, US FDA) before the Senate Committee on Finance. Available at: http://www.fda.gov/NewsEvents/Testimony/ucm113235.htm. Accessed July 2015
31. Graham DJ, Campen D, Hui R et al (2005) Risk of acute myocardial infarction and sudden cardiac death in patients treated with cyclo-oxygenase 2 selective and non-selective non-steroidal anti-inflammatory drugs: nested case-control study. Lancet 365(9458):475–481
32. Ghassemi M, Marshall J, Singh N et al (2014) Leveraging a critical care database: selective serotonin reuptake inhibition use prior to ICU admission is associated with increased hospital mortality. Chest 145(4):1–8
33. Frankovich J, Longhurst CA, Sutherland SM (2011) Evidence-based medicine in the EMR era. New Engl J Med 365:19
34. Celi LA, Zimolzak AJ, Stone DJ (2014) Dynamic clinical data mining: search engine-based decision support. JMIR Med Inform 2(1):e13

Chapter 4
Pulling It All Together: Envisioning a Data-Driven, Ideal Care System

David Stone, Justin Rousseau and Yuan Lai

Take Home Messages

- An Ideal Care System should incorporate fundamental elements of control engineering, such as effective and data-driven sensing, computation, actuation, and feedback.
- These systems must be carefully and intentionally designed to support clinical decision-making, rather than being allowed to evolve based on market pressures and user convenience.

This chapter presents ideas on how data could be systematically more effectively employed in a purposefully engineered healthcare system. We have previously written on potential components of such a system—e.g. dynamic clinical data mining, closing the loop on ICU data, optimizing the data system itself, crowdsourcing, etc., and will attempt to 'pull it all together' in this chapter, which we hope will inspire and encourage others to think about and move to create such a system [1–10]. Such a system, in theory, would support clinical workflow by [1] leveraging data to provide both accurate personalized, or 'precision,' care for individuals while ensuring optimal care at a population level; [2] providing coordination and communication among the users of the system; and [3] defining, tracking, and enhancing safety and quality. While health care is intrinsically heterogeneous at the level of individual patients, encounters, specialties, and clinical settings, we also propose some general systems-based solutions derived from contextually defined use cases. This chapter describes the fundamental infrastructure of an Ideal Care System (ICS) achieved through identifying, organizing, capturing, analyzing, utilizing and appropriately sharing the data.

MIT Critical Data, *Secondary Analysis of Electronic Health Records*,
DOI 10.1007/978-3-319-43742-2_4

4.1 Use Case Examples Based on Unavoidable Medical Heterogeneity

The intrinsic heterogeneities inherent in health care at the level of individual patients, encounters, specialties, and clinical settings has rendered the possibility of a single simple systems solution impossible. We anticipate requirements in an ICS

Table 4.1 Clinical use cases with pertinent clinical and data objectives

Clinical use case	Clinical objective(s)	Data objectives
Outpatient in state of good health	Provide necessary preventive care; address mild intermittent acute illnesses	Health maintenance documentation: vaccination records, cancer screening records, documentation of allergies; data on smoking and obesity
Outpatient with complex chronic medical problems	Connect and coordinate care among diverse systems and caregivers	Ensure accurate and synchronized information across care domains without need for oversight by patient and/or family; targeted monitors to prevent admission, readmission
Inpatient—elective surgery	Provide a safe operative and perioperative process	Track processes relevant to safety and quality; track outcomes, complication rates, including safety related outcomes
Inpatient (emergency department, inpatient wards, intensive care units)	Identify and predict ED patients who require ICU care; ICU safety and quality; Identify and predict adverse events	Track outcomes of ED patients including ICU transfers and mortality; Track adverse events; Track usual and innovative ICU metrics
Nursing home patient	Connect and coordinate care among diverse locations and caregivers for a patient who may not be able to actively participate in the process	Ensure accurate and synchronized information across care domains without need for oversight by patient and/or family
Recent discharge from hospital	Prevent re-admission	Data mining for predictors associated with re-admission and consequent interventions based on these determinations; Track functional and clinical outcomes
Labor and delivery	Decision and timing for caesarian section; Lower rates of intervention and complications	Data mining for predictors associated with c-section or other interventions; track complication rates and outcomes
Palliative care/end of life	Decision and timing for palliative care; Ensure comfort and integrity	Data mining to determine characteristics that indicate implementation of palliative care

of identifying common core elements that apply to the medical care of all patients (e.g. safety principles, preventive care, effective end of life care, accurate and up-to-date problem list and medication list management), and subsequently formulating pathways based on specific context. One should note that an individual patient can cross over multiple categories. Any complex outpatient will also have the baseline requirements of meeting objectives of an outpatient in good health and may at some point have an inpatient encounter. Table 4.1 identifies a variety of use cases including abbreviated forms of the pertinent clinical and data issues associated with them.

4.2 Clinical Workflow, Documentation, and Decisions

The digitalization of medicine has been proceeding with the wide adoption of electronic health records, thanks in part to meaningful use as part of the Health Information Technology for Economic and Clinical Health (HITECH) Act [11], but has received varying responses by clinicians. An extensive degree of digitalization is a fundamental element for creating an ICS. Defined at the highest level, a system is a collection of parts and functions (a.k.a. components and protocols) that accepts inputs and produces outputs [3]. In healthcare, the inputs are the patients in various states of health and disease, and the outputs are the outcomes of these patients. Figure 4.1 provides a simple control loop describing the configuration of a data driven health system.

The practice of medicine has a long history of being data driven, with diagnostic medicine dating back to ancient times [12]. Doctors collect and assemble data from histories, physical exams, and a large variety of tests to formulate diagnoses, prognoses, and subsequent treatments. However, this process has not been optimal in the sense that these decisions, and the subsequent actuations based on these decisions, have been made in relative isolation. The decisions depend on the prior experience and current knowledge state of the involved clinician(s), which may or may not be based appropriately on supporting evidence. In addition, these decisions have, for the most part, not been tracked and measured to determine their impact on safety and quality. We have thereby lost much of what has been done that was good and failed to detect much of what was bad [1]. The digitization of medicine provides an opportunity to remedy these issues. In spite of the suboptimal usability of traditional paper documentation, the entries in physicians' notes in natural language constitute the core data required to fuel an ideal care system. While data items such as lab values and raw physiological vital signs may be reasonably reliable and quantitative, they generally do not represent the decision-making and the diagnoses that are established or being considered, which are derived from the analysis and

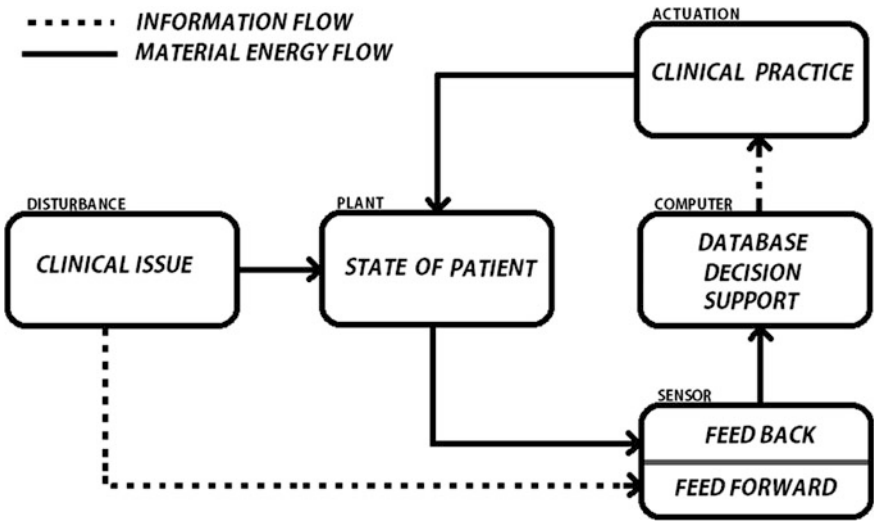

Fig. 4.1 Control loop depicting a data-driven care system. A clinical issue such as an infection or vascular occlusion affects the state of the patient. Subsequently, the system sensor detects this change and submits the relevant data to the computer for storage and analysis. This may or may not result in actuation of a clinical practice intervention that further affects the state of the patient, which feeds back into the system for further analysis. Feed-forward control involves the transmission of disturbances directly to the sensor without first affecting the state of the patient. The detection of a risk factor for venous thromboembolism that triggers prophylaxis in a protocol-based manner represents a clinical example of feed-forward control [3]

synthesis of the available data (the assessment with differential diagnosis) as well as the data to be acquired in the diagnostic workup (the plan).

The digitalization of medicine has encountered two key issues: [1] How does one develop a digitally based workflow that supports rapid, accurate documentation so that the clinician feels enlightened rather than burdened by the process? [2] How can the documentation process of data entry support and enhance the medical decision-making process? The first iteration of electronic health records (EHRs) has simply attempted to replicate the traditional paper documentation in a digital format. In order to address the first issue, smarter support of the documentation process will require innovative redesigns to improve the EHR as it evolves. Rather than requiring the clinician to sit at a keyboard facing away from a patient, the process needs to capture real-time input from the patient encounter in such potential modes as voice and visual recognition. This must be done so that the important details are captured without unduly interfering with personal interactions or without erroneous entries due to delayed recall. The receiving system must ‘consider’ the patient’s prior information in interpreting new inputs in order to accurately recognize and

assimilate the essential information from the current encounter. Furthermore, the data that is collected should not be functionally lost as the patient advances through time and moves between geographic locales. A critical issue is one that has been perpetuated in the current practice of medicine from one encounter to another—the physician and patient should not need to 'reinvent the informational wheel' with every encounter. While each physician should provide a fresh approach to the patient, this should not require refreshing the patient's entire medical story with each single encounter, wasting time and effort. Furthermore, what is documented should be transparent to the patient in contrast to the physician beneficence model that has been practiced for most of the history of medicine where it was considered beneficial to restrict patients' access to their own records. Steps are being taken toward this goal of transparency with the patient with the OpenNotes movement that began in 2010. The effects of this movement are being recognized nationally with significant potential benefits in many areas relating to patient safety and quality of care [13].

Regarding the second issue, we have written of how quality data entry can support medical decision-making [14]. Future iterations of an innovatively redesigned EHR in an ideal care system should assist in the smart assembly and presentation of the data as well as presentation of decision support in the form of evidence and education. The decision-maker is then able to approach each encounter with the advantage of prior knowledge and supporting evidence longitudinally for the individual patient as well as comparisons of their states of health with patients with similar data and diagnoses (Fig. 4.2). Patterns and trends in the data can be recognized, particularly in the context of that patient's prior medical history and evolving current state (Fig. 4.3).

Population data should be leveraged to optimize decisions for individuals, with information from individual encounters captured, stored and utilized to support the care of others as we have described as 'dynamic clinical data mining [2].' This also is similar to what has been described as a 'learning healthcare system' or by a 'green button' for consulting such population data for decision support [15, 16].

In summary, an ICS must have tools (e.g. enhanced versions of current EHRs) to capture and utilize the data in ways that make documentation and decision-making effective and efficient rather than isolated and burdensome. While we realize that individual clinicians function brilliantly in spite of the technical and systems-level obstacles and inefficiencies with which they are faced, we have reached a point of necessity, one recognized by the Institute of Medicine threatening the quality and safety of healthcare, requiring the development of digital tools that facilitate necessary data input and decisions as well as tools that can interact with and incorporate other features of an integrated digitally-based ICS [17]. This will require close interactions and collaborations among health care workers, engineers including software and hardware experts, as well as patients, regulators, policymakers, vendors and hospital business and technical administrators [5].

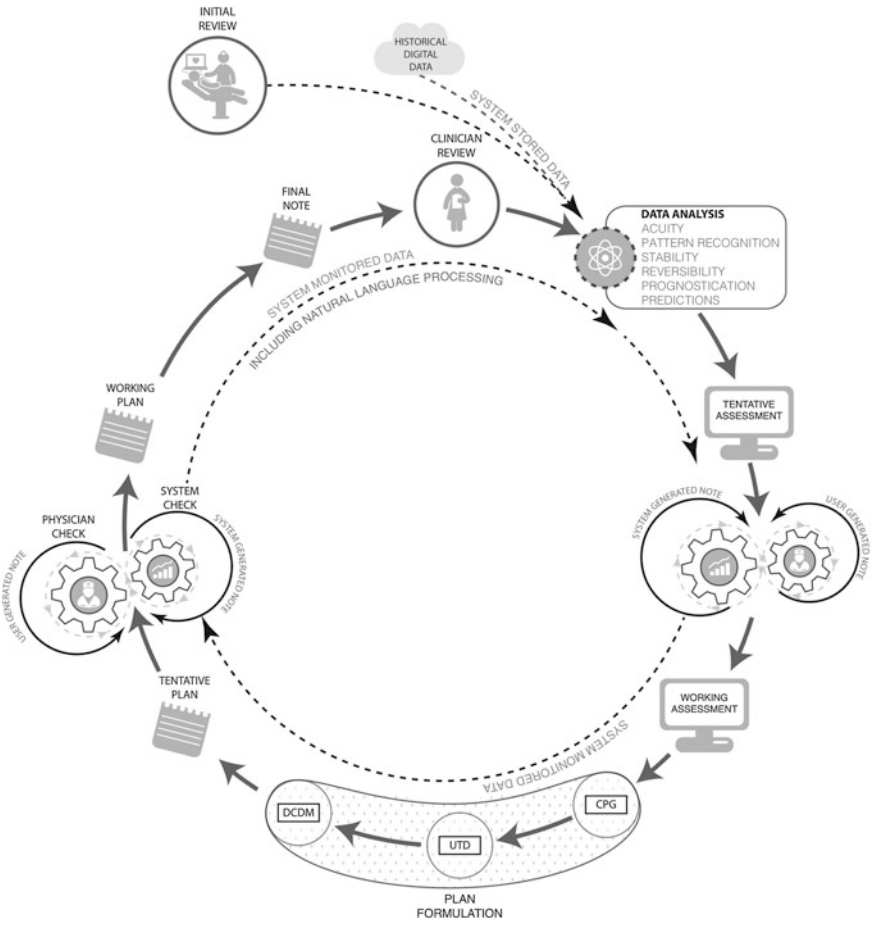

Fig. 4.2 Clinician documentation with fully integrated data systems support. Prior notes and data are input for future notes and decisions. The digital system analyzes input and displays suggested diagnoses and problem list, and then diagnostic test and treatment recommendations hierarchically based on various levels of evidence: CPG—clinical practice guidelines, UTD—Up to Date®, DCDM—Dynamic clinical data mining [14]

4.3 Levels of Precision and Personalization

Many of the tools available to clinicians have become fantastically sophisticated, including technical devices and molecular biological and biochemical knowledge. However, other elements, including those used intensively on a daily basis, are more primitive and would be familiar to clinicians of the distant past. These elements include clinical data such as the heart rates and blood pressures recorded in a

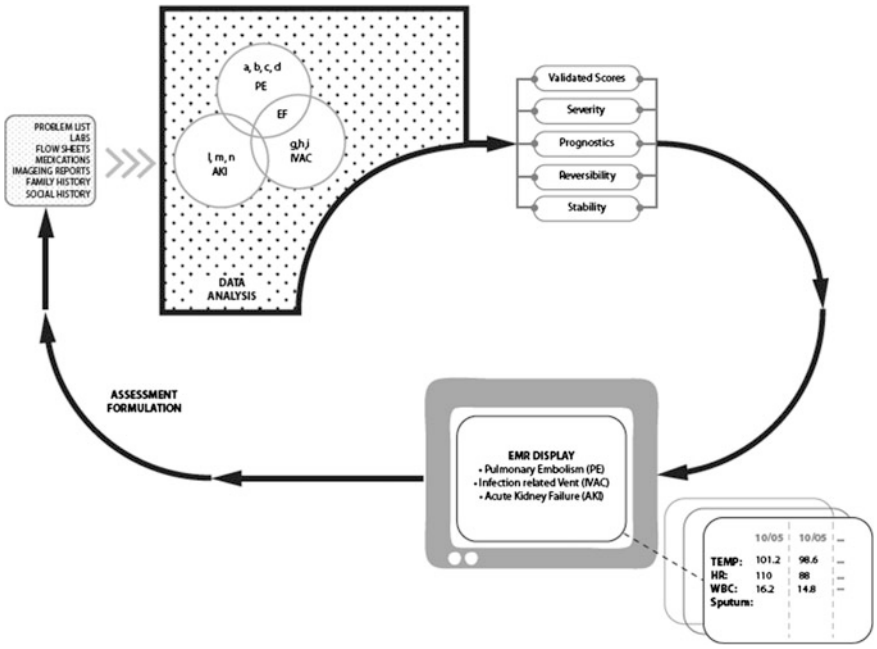

Fig. 4.3 Mock screenshot for the Assessment screen with examples of background data analytics. Based on these analytics that are constantly being performed by the system and are updated as the user begins to enter a note, a series of problems are identified and suggested to the user by EMR display. After consideration of these suggestions in addition to their own analysis, the user can select or edit the problems that are suggested or input entirely new problems. The final selection of problems is considered with ongoing analytics for future assessments [14]

nursing flowsheet. Patient monitoring is not generally employed on a data driven basis, particularly decisions regarding who gets monitored with what particular signals, the duration of monitoring, and whether the data are stored, analyzed, and utilized beyond the current time. Furthermore, it is questionable whether the precedent of setting common numeric thresholds for abnormally high or low values extracts maximal clinical information from those signals. This recognition of abnormal values has become a significant problem of excessive false alarms and alarm fatigue [18]. Data analysis should provide clinicians with personalized and contextualized characterizations of individual vital signs (e.g. heart and respiratory rate variability patterns, subtle ECG waveform shapes, etc.) so that truly important changes can be recognized quickly and effectively while not overwhelming the cognitive load of the clinician. This would constitute 'personalized data driven monitoring' in which the raw data on the monitor screen is analyzed in real time to provide more information regarding the state of the patient. This will become more important and pressing as monitoring becomes more ubiquitous both in the hospital

and in outpatient settings, which is not far from a reality with the exponential development of mobile health monitors and applications. A potential approach to this issue would be to treat monitors as specialized component of the EHR rather than standalone devices that display the heart rate and beep frequently, at times even when there is no good reason. In fact, this has occurred to some functional extent as monitors have become networked and in many cases can import data into the EHR. The loop will be closed when information flows bi-directionally so that the EHR (and other elements such as infusion pumps) can assist in providing clinical contexts and personalized information to enhance the performance potential of the monitors [14]. Whereas the user interface of the monitor is currently solely one of adjusting the monitored channels and the alarm settings, the user interface will also be increasingly rich so that the user could, for instance with the proper credentials, access, edit and annotate the EHR from a bedside or central monitor, or add information directly to the monitor to calibrate the monitoring process.

The data from monitors is beginning to be used for prospective analytic purposes in terms of predicting neonatal sepsis and post cardiac surgery problems [19, 20]. The HeRO neonatal alert focuses on diminution in heart rate variability and increase in decelerations to identify potential sepsis, whereas the Etiometry alert employs a sophisticated statistical analysis of those monitored elements reflecting cardiac function to detect and define problems earlier than humans could ordinarily do. The HeRO team is now working to develop predictive analytics for respiratory deterioration, significant hemorrhage, and sepsis in adults [21]. The essential point is that monitors employing such predictive analytics, as well as streaming and retrospective analytics, can leverage large amounts of personal data to improve the monitoring process as well as the healthcare encounter experience, particularly in areas of quality and safety. However, it is essential that such individual applications, exponentially growing in complexity and sophistication, not be introduced as unrelated bits into an already data-overburdened and under-engineered health care system. In the current state of the healthcare system, there is already plenty of data. However, it is not being systematically handled, utilized and leveraged. It is essential that such new applications be embedded thoughtfully into workflows. They must also be systematically interfaced and interoperable with the core care system, represented by the next generation of EHRs, so that the information can be used in a coordinated fashion, audited in terms of its impact on workflows, and tracked in terms of its impact on patient outcomes, quality, and safety. The addition of further system elements should be planned, monitored, and evaluated in a data-driven fashion. New elements should contribute to the system that uses data in a targeted, well-managed fashion rather than simply collecting it. The introduction of elements outside the core EHR requires communication and coordination among all system elements, just as effectively using the EHR alone requires communication and coordination among caregivers and patients.

4.4 Coordination, Communication, and Guidance Through the Clinical Labyrinth

Coordination and communication would be fundamental properties of an ICS contrasted with the enormous individual efforts required to achieve these goals in the current state. Patients and caregivers should be able to assume that the system captures, stores, and shares their information where and when it is needed. When the patient leaves her nursing home to be seen in a local emergency room or by her neurologist, the clinicians should have all previously available information necessary to treat her. This should also be the case when she returns to the nursing home with the system updating her record with events from her previous encounter as well as implementing new orders reflecting that encounter. This seamless communication and coordination is especially important for the kinds of patients who cannot provide this support themselves: people who are elderly, cognitively impaired, acutely ill, etc. Unfortunately, the current system was developed as a tool to aid in billing and reimbursement of interventions and the challenge that we face with transforming and continuing to develop it into an ICS is to transition its focus to patient care. Currently, patients and their advocates must battle with unrelenting challenges of opacity and obstruction facing immense frustration and threats to patient safety and quality of care where such risks would not be tolerated in any other industry.

Data and the efficient transmission of information where and when it is needed are at the core of an ICS. Information networks that permeate all the relevant locales must be created employing all the interoperability, privacy, and security features necessary. The system must maintain its focus on the patient and must instantly (or sufficiently quickly to meet clinical needs) update, synchronize, and transmit the information to all those who need to know, including qualified and permitted family members and the patients themselves relevant to the care of the patient. Many clinicians may be misinterpreted as being unresponsive, or even uncaring, in response to their continuing frustration with the difficulty of obtaining timely and accurate information. The current state of siloed healthcare systems makes obtaining information from other locales prohibitively challenging with no particular reward for continuing to struggle to obtain pertinent information for the continued care of patients, evoking reactions from caregivers including rudeness, neglect, hostility, or burnout. This challenge to obtain information from outside sources also leads to repeat diagnostic testing exposing patients to unnecessary risks and exposures such as is seen when a patient is transferred from one institution to another but the imaging obtained at the first institution is not able to be transferred appropriately [22]. Unfortunately, the Health Insurance Portability and Accountability Act of 1996 (HIPAA), the very legislation designed to enable the portability of information relevant to patient care, has further hindered this transmission of information. An efficient system of communication and coordination would benefit the caregiver experience in addition to the patients by providing them with the tools and information that they need to carry out their jobs.

The scope of those affected by the challenges inherent in the current healthcare system is broad. Not only does it affect those that are cognitively impaired, but also those with limited education or resources. It affects those that have complicated medical histories as well as those without previous histories. Even when patients are capable of contributing to the management of their own clinical data, there is potential to be overwhelmed and incapacitated through the complexities of the system when affected by illness, no matter the acuity, severity, or complexity. Interoperable EHRs focused on patients rather than locations or brands would provide the necessary and updated information as a patient moves from office A to hospital system B to home and back to emergency room C. When people are sick, they and their caregivers should be supported by the system rather than forced to battle it.

The sharing of data among patients and caregivers in a safe and efficient manner is not primarily a technical problem at this time, although there are many technical challenges to achieving such seamless interoperability. It is also a business as well as a political problem. This complex interaction can be seen in efforts toward healthcare architecture and standards supporting interoperability described in the JASON report, "A Robust Health Data Infrastructure" with responses from industry and EHR vendors in the development and adoption of HL7 Fast Healthcare Interoperability Resources (FHIR) standards [23, 24]. In an ICS, all parties must cooperate to interconnect EHRs among caregivers and locals so that the accurate and reliable data essential for healthcare can be coordinated, synchronized, and communicated across practice domains but within each patient's domain. As we have seen on individual patient levels, an overabundance of data is not useful if it is not processed, analyzed, placed into the appropriate context, and available to the right people at the right places and times.

4.5 Safety and Quality in an ICS

There are many examples in healthcare, such as with bloodletting with leeches, where what was thought to be best practice, based on knowledge or evidence at the time, was later found to be harmful to patients. Our knowledge and its application must be in a continual state of assessment and re-assessment so that unreliable elements can be identified and action taken before, or at least minimal, harm is done [4]. There is currently no agreement on standard metrics for safety and quality in healthcare and we are not going to attempt to establish standard definitions in this chapter [25]. However, in order to discuss these issues, it is important to establish a common understanding of the terminologies and their meaning.

At a conceptual level, we conceive clinical **safety** as a strategic optimization problem in which the maximum level of permissible actuation must be considered and implemented in the simultaneous context of allowing the minimal degree of care-related harm. The objective is to design and implement a care system that minimizes safety risks to approach a goal of zero. The digitization of medicine

affords a realistic chance of attaining this goal in an efficient and effective manner. The application of systems engineering principles also provides tools to design these kinds of systems.

The overall **quality** of healthcare is a summation of the experience of individuals, and for these individuals, there may be varying degrees of quality for different periods of their experience. Similar to safety, we also think of quality as a strategic optimization problem in which outcomes and benefits are maximized or optimized, while the costs and risks involved in the processes required to achieve them, are minimized. The provision of quality via optimized outcomes in clinical care is, to a large extent, a problem in engineering information reliability and flow, providing the best evidence at the right times to assist in making the best decisions [3]. The concepts of the 'best evidence' and 'best decisions' themselves depend on input sources that range from randomized control trials to informed expert opinion to local best practices. To provide actual actuation, information flows must be supplemented by chemical (medications), mechanical (surgery, physical therapy, injections, human touch) and electromagnetic (imaging, ultrasound, radiation therapy, human speech) modalities, which can institute the processes indicated by those information flows.

Furthermore, quality may also be defined with respect to the degree of success in treatment of the disease state. Diseases addressed in modern medicine are, to a surprisingly large and increasingly recognized extent, those of control problems in bioengineering [10]. These diseases may stem from control problems affecting inflammation, metabolism, physiological homeostasis, or the genome. However, these all represent failure in an element or elements of a normally well-controlled biological system. The quality of the clinical response to these failures is best improved by understanding them sufficiently and thoroughly enough so that targeted and tolerable treatments can be developed that control and/or eliminate the systems dysfunction represented by clinical disease. This should be accomplished in a way that minimizes undue costs in physical, mental, or even spiritual suffering. Ultimately, medical quality is based primarily on outcomes, but the nature of the processes leading to those outcomes must be considered. Optimal outcomes are desirable, but not at any cost, in the broad definition of the term. For example, prolonging life indefinitely is not an optimal outcome in some circumstances that are contextually defined by individual, family, and cultural preferences.

Having defined safety and quality in our context, the next step is to develop systems that capture, track and manage these concepts in retrospective, real-time, and predictive manners. It is only when we know precisely what static and dynamic elements of safety and quality we wish to ensure that we can design the systems to support these endeavors. These systems will involve the integration of hardware and software systems such as physiologic monitors with the EHR (including Computerized Provider Order Entry, Picture Archiving and Communication System, etc.), and will require a variety of specialized, domain-specific data analytics as well as technical innovations such as wireless body sensor networks to capture patient status in real time. The system will connect and communicate pertinent information among caregivers by populating standardized, essential access

and alert nodes with timely and accurate information. It is also necessary that information flows bi-directionality (from the records of individuals to the population record, and from the population record to individuals) so that both can benefit from the data [2, 14]. Clearly, this will require an overall monitoring and information system that is interoperable, interactive both with its own components and its users, and actively but selectively informative. Future generations of clinicians will receive their education in an environment in which these systems are ubiquitous, selectively modifiable based on inputs such as crowdsourcing, and intrinsic to the tasks at hand, in contrast to the siloed and apparently arbitrarily imposed applications current clinicians may resist and resent [5, 8].

We noted the importance of control problems in disease, and control will also represent a fundamental component in the design of future safety and quality systems. The detection and prevention of adverse events is a significant challenge when depending on self-reporting methods or chart review and this issue is of high importance in the US [26, 27]. Predictive analytics can be developed as elements of the system to prospectively inform users of threats to safety and quality [19–21]. Carefully designed feed-forward components will inform participants in real time that an high risk activity is occurring so that it can be rectified without requiring retroactive analysis (Fig. 4.4—safety control loop below). Retrospective data analytics will track the factors affecting quality and safety so that practice,

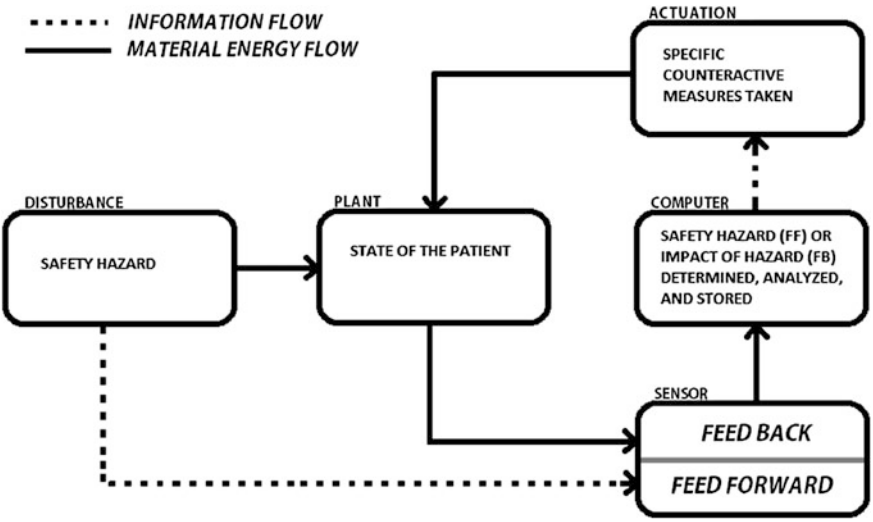

Fig. 4.4 Control loop depicting a data-driven safety system. A clinical safety issue affects the state of the patient. Subsequently, the system sensor detects this change and submits the relevant data to the computer for storage and analysis. This may or may not result in actuation of a counteractive intervention that further affects the state of the patient, which feeds back into the system for further analysis. Feed-forward control involves the transmission of disturbances directly to the sensor without first affecting the state of the patient. An example of such a feed-forward control includes a faulty device or a biohazard

workflow, and technological systems can be accordingly modified. Such an ICS will be capable of monitoring medical errors, adverse events, regulatory and safety agency concerns and metrics, and compliance with best practice as well as meaningful use in parallel with costs and outcomes.

4.6 Conclusion

The basic systems solutions to the health care data problem rest on fully and inclusively addressing the axes of patient, care giver and care system considerations, which at times are apparently independent, but are ultimately interactive and interdependent. The required systems design will also greatly benefit from basic incorporation of the fundamental elements of control engineering such as effective and data-driven sensing, computation, actuation, and feedback. An Ideal Care System must be carefully and intentionally designed rather than allowed to evolve based on market pressures and user convenience.

The patient's data should be accurate, complete, and up-to-date. As patients progress in time, their records must be properly and timely updated with new data while concurrently, old data are modified and/or deleted as the latter become irrelevant or no longer accurate. New entry pipelines such as patient-generated and remotely generated data, as well as genomic data, must be taken into consideration and planned for. These data should be securely, reliably, and easily accessible to the designated appropriate users including the patient. The caregiver should have access to these data via a well-designed application that positively supports the clinical documentation process and includes reasonable and necessary decision support modalities reflecting best evidence, historical data of similar cases in the population, as well as the patient's own longitudinal data. All should have access to the data so far as it is utilized to construct the current and historical patterns of safety and quality. In addition to the data of individuals, access to the data of populations is required for the above purposes as well as to provide effective interventions in emergency situations such as epidemics. The creation of this kind of multimodal systems solution (Fig. 4.5—Ideal Care System Architecture below) will require the input of a great variety of experts including those from the EHR, monitoring devices, data storage, and data analytic industries along with leaders in healthcare legislation, policy makers, regulation, and administration.

Many important engineering, economic, and political questions remain that are not addressed in this chapter. What and who will provide the infrastructure and who will pay for it? Will this kind of system continue to work with current hardware and software or require fundamental upgrades to function at the required level of reliability and security? How and where will the controls be embedded in the system?

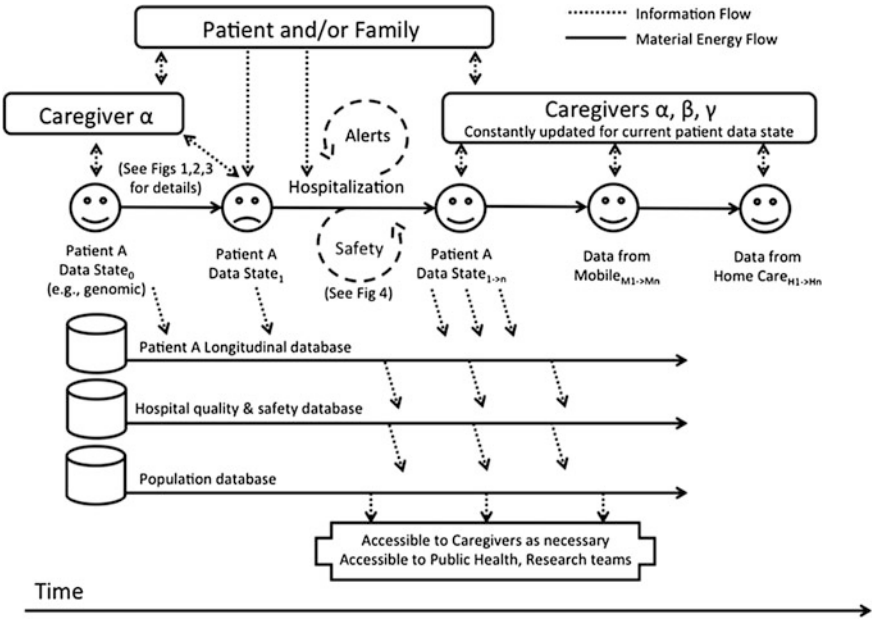

Fig. 4.5 Information Architecture of an Ideal Care System. This diagram integrates the concepts described in this chapter depicting data driven care systems, safety systems, along with connection and coordination of patient data across multiple modalities to achieve an Ideal Care System. Patients move through time and interact with the ICS in different contexts. Parallel databases are integrated with the patient data states in time including an individual patient's longitudinal database, hospital quality and safety database, and a population database. Data from the patient, mobile technologies and from the home care entities keep caregivers informed of the most current patient data state

For example, will they be at the individual smart monitoring level or at a statewide public health level? How will the metadata obtained be handled for the good of individuals and populations? It is critical that the addition of new modalities and devices be fully integrated into the system rather than adding standalone components that may contribute more complexity and confusion than benefit. These goals will require cooperation previously unseen among real and potential competitors and those who have previously been able to work in relative isolation.

References

1. Celi LA, Mark RG, Stone DJ, Montgomery R (2013) "Big data" in the ICU: closing the data loop. Am J Respir Crit Care Med 187(11):1157–1160
2. Celi LA, Zimolzak AJ, Stone DJ (2014) Dynamic clinical data mining: search engine-based clinical decision support. J Med Internet Res Med Inform 2(1):e13. doi:10.2196/medinform.3110
3. Celi LA, Csete M, Stone D (2014) Optimal data systems: the future of clinical predictions and decision support. Curr Opin Crit Care 20:573–580
4. Moseley ET, Hsu D, Stone DJ, Celi LA (2014) Beyond data liberation: addressing the problem of unreliable research. J Med Internet Res 16(11):e259
5. Celi LA, Ippolito A, Montgomery R, Moses C, Stone DJ (2014) Crowdsourcing knowledge discovery and innovations in medicine. J Med Internet Res 16(9):e216. doi:10.2196/jmir.3761
6. Celi LA, Moseley E, Moses C, Ryan P, Somai M, Stone DJ, Tang K (2014) From pharmacovigilance to clinical care optimization. Big Data 2(3):134–141. doi:10.1089/big.2014.0008
7. Badawi O, Brennan T, Celi LA, Feng M, Ghassemi M, Ippolito A, Johnson A, Mark RG, Mayaud L, Moody G, Moses C, Naumann T, Pimentel M, Pollard TJ, Santos M, Stone DJ, Zimolzak AJ (2014) Making big data useful for health care: a summary of the inaugural MIT critical data conference. J Med Internet Res Med Inform 2(2):e22. doi:10.2196/medinform.3447
8. Moskowitz A, McSparron J, Stone DJ, Celi LA (2015) Preparing a new generation of clinicians for the era of big data. Harvard Med Student Rev 2(1):24–27
9. Ghassemi M, Celi LA, Stone DJ (2015) The data revolution in critical care. Ann Update Intensive Care Emerg Med 2015(2015):573–586
10. Stone DJ, Csete ME, Celi LA (2015) Engineering control into medicine. J Crit Care. Published Online: January 29, 2015. doi:10.1016/j.jcrc.2015.01.019
11. Health Information Technology for Economic and Clinical Health (HITECH) Act, Title XIII of Division A and Title IV of Division B of the American Recovery and Reinvestment Act of 2009 (ARRA), Pub. L. No. 111-5, 123 Stat. 226 (Feb. 17, 2009), codified at 42 U.S.C. §§300jj et seq.; §§17901 et seq
12. Horstmanshoff HFJ, Stol M, Tilburg C (2004) Magic and rationality in ancient near Eastern and Graeco-Roman medicine, pp 97–99. Brill Publishers. ISBN 978-90-04-13666-3
13. Bell SK, Folcarelli PH, Anselmo MK, Crotty BH, Flier LA, Walker J (2014) Connecting Patients and Clinicians: The Anticipated Effects of Open Notes on Patient Safety and Quality of Care. Jt Comm J Qual Patient Saf 41(8):378–384(7)
14. Celi LA, Marshall JD, Lai Y, Stone DJ, Physician documentation and decision making in the digital era. J Med Internet Res Med Inform (Forthcoming)
15. Longhurst CA, Harrington RA, Shah NH (2014) A 'green button' for using aggregate patient data at the point of care. Health Aff 33(7):1229–1235
16. Friedman C, Rubin J, Brown J et al (2014) J Am Med Inform Assoc 0:1–6. doi:10.1136/amiajnl-2014-002977
17. Institute of Medicine (2012) Best care at lower cost: the path to continuously learning health care in America. Retrieved from: http://iom.nationalacademies.org/Reports/2012/Best-Care-at-Lower-Cost-The-Path-to-Continuously-Learning-Health-Care-in-America.aspx
18. The Joint Commission (2014) National Patient safety goal on alarm management 2013
19. www.etiometry.com
20. www.heroscore.com
21. Personal communication, Randall Moorman, MD
22. Sodickson A, Opraseuth J, Ledbetter S (2011) Outside imaging in emergency department transfer patients: CD import reduces rates of subsequent imaging utilization. Radiology 260 (2):408–413

23. A Robust Health Data Infrastructure. (Prepared by JASON at the MITRE Corporation under Contract No. JSR-13-700). Agency for Healthcare Research and Quality, Rockville, MD. April 2014. AHRQ Publication No. 14-0041-EF
24. Health Level Seven® International. HL7 Launches Joint Argonaut Project to Advance FHIR. N.p., 4 Dec. 2014. Web. 31 Aug. 2015. http://www.hl7.org/documentcenter/public_temp_32560CB2-1C23-BA17-0CBD5D492A8F70CD/pressreleases/HL7_PRESS_20141204.pdf
25. Austin JM et al (2015) National hospital ratings systems share few common scores and may generate confusion instead of clarity. Health Aff 34(3):423–430. doi:10.1377/hlthaff.2014.0201
26. Elton GEBM, Ripcsak GEH (2005) Automated detection of adverse events using natural language processing of discharge summaries. J Am Med Inform Assoc 12:448–458
27. Kohn LT, Corrigan JM, Donaldson MS (eds) (2000) To err is human: building a safer health system, vol 2. National Academy Press, Washington, DC

Chapter 5
The Story of MIMIC

Roger Mark

Take Home Messages

- MIMIC is a Medical Information Mart for Intensive Care and consists of several comprehensive data streams in the intensive care environment, in high levels of richness and detail, supporting complex signal processing and clinical querying that could permit early detection of complex problems, provide useful guidance on therapeutic interventions, and ultimately lead to improved patient outcomes.
- This complicated effort required a committed and coordinated collaboration across academic, industry, and clinical institutions to provide a radically open access data platform accessible by researchers around the world.

5.1 The Vision

Patients in hospital intensive care units (ICUs) are physiologically fragile and unstable, generally have life-threatening conditions, and require close monitoring and rapid therapeutic interventions. They are connected to an array of equipment and monitors, and are carefully attended by the clinical staff. Staggering amounts of data are collected daily on each patient in an ICU: multi-channel waveform data sampled hundreds of times each second, vital sign time series updated each second or minute, alarms and alerts, lab results, imaging results, records of medication and fluid administration, staff notes and more. In early 2000, our group at the Laboratory of Computational Physiology at MIT recognized that the richness and detail of the collected data opened the feasibility of creating a new generation of monitoring systems to track the physiologic state of the patient, employing the power of modern signal processing, pattern recognition, computational modeling, and knowledge-based clinical reasoning. In the long term, we hoped to design

MIT Critical Data, *Secondary Analysis of Electronic Health Records*,
DOI 10.1007/978-3-319-43742-2_5

monitoring systems that not only synthesized and reported all relevant measurements to clinicians, but also formed pathophysiologic hypotheses that best explained the observed data. Such systems would permit early detection of complex problems, provide useful guidance on therapeutic interventions, and ultimately lead to improved patient outcomes.

It was also clear that although petabytes of data are captured daily during care delivery in the country's ICUs, most of these data were not being used to generate evidence or to discover new knowledge. The challenge, therefore, was to employ existing technology to collect, archive and organize finely detailed ICU data, resulting in a research resource of enormous potential to create new clinical knowledge, new decision support tools, and new ICU technology. We proposed to develop and make public a "substantial and representative" database gathered from complex medical and surgical ICU patients.

5.2 Data Acquisition

In 2003, with colleagues from academia (Massachusetts Institute of Technology), industry (Philips Medical Systems), and clinical medicine (Beth Israel Deaconess Medical Center, BIDMC) we received NIH (National Institutes of Health) funding to launch the project "Integrating Signals, Models and Reasoning in Critical Care", a major goal of which was to build a massive critical care research database. The study was approved by the Institutional Review Boards of BIDMC (Boston, MA) and MIT (Cambridge, MA). The requirement for individual patient consent was waived because the study would not impact clinical care and all protected health information was to be de-identified.

We set out to collect comprehensive clinical and physiologic data from all ICU patients admitted to the multiple adult medical and surgical ICUs of our hospital (BIDMC). Each patient record began at ICU admission and ended at final discharge from the hospital. The data acquisition process was continuous and invisible to staff. It did not impact the care of patients or methods of monitoring. Three categories of data were collected: *clinical data*, which were aggregated from ICU information systems and hospital archives; high-resolution *physiological data* (waveforms and time series of vital signs and alarms obtained from bedside monitors); and *death data* from Social Security Administration Death Master Files (See Fig. 5.1).

5.2.1 Clinical Data

Bedside clinical data were downloaded from archived data files of the CareVue Clinical Information System (Philips Healthcare, Andover, MA) used in the ICUs. Additional clinical data were obtained from the hospital's extensive digital archives. The data classes included:

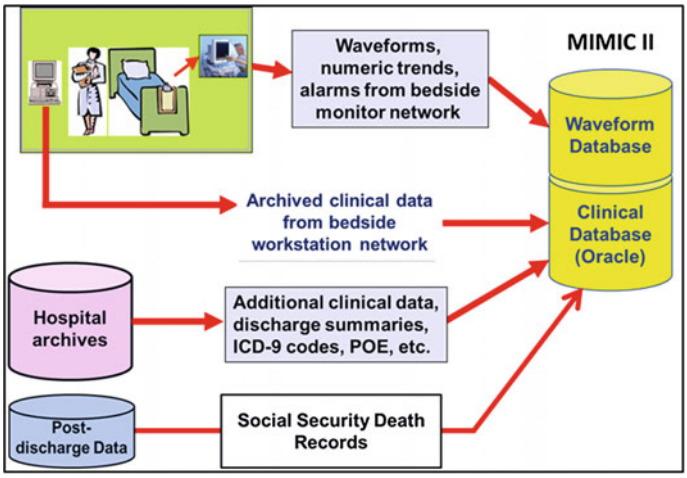

Fig. 5.1 MIMIC II data sources

- **Patient demographics**
- **Hospital administrative data**: admission/discharge/death dates, room tracking, billing codes, etc.
- **Physiologic**: hourly vital signs, clinical severity scores, ventilator settings, etc.
- **Medications**: IV medications, physician orders
- **Lab tests**: chemistry, hematology, ABGs, microbiology, etc.
- **Fluid balance data**
- **Notes and reports**: Discharge summaries; progress notes; ECG, imaging and echo reports.

5.2.2 Physiological Data

Physiological data were obtained with the technical assistance of the monitoring system vendor. Patient monitors were located at every ICU patient bed. Each monitor acquired and digitized multi-parameter physiological waveform data, processed the signals to derive time series (trends) of clinical measures such as heart rate, blood pressures, and oxygen saturation, etc., and also produced bedside monitor alarms. The waveforms (such as electrocardiogram, blood pressures, pulse plethysmograms, respirations) were sampled at 125 Hz, and trend data were updated each minute. The data were subsequently stored temporarily in a central database server that typically supported several ICUs. A customized archiving agent created and stored permanent copies of the physiological data. The data were physically transported from the hospital to the laboratory every 2–4 weeks where they were de-identified, converted to an open source data format, and incorporated into the MIMIC II waveform database. Unfortunately, limited capacity and

intermittent failures of the archiving agents limited waveform collection to a fraction of the monitored ICU beds.

5.2.3 *Death Data*

The Social Security Death Master files were used to document subsequent dates of death for patients who were discharged alive from the hospital. Such data are important for 28-day and 1-year mortality studies.

5.3 Data Merger and Organization

A major effort was required in order to organize the diverse collected data into a well-documented relational database containing integrated medical records for each patient. Across the hospital's clinical databases, patients are identified by their unique Medical Record Numbers and their Fiscal Numbers (the latter uniquely identifies a particular hospitalization for patients who might have been admitted multiple times), which allowed us to merge information from many different hospital sources. The data were finally organized into a comprehensive relational database. More information on database merger, in particular, how database integrity was ensured, is available at the MIMIC-II web site [1]. The database user guide is also online [2].

An additional task was to convert the patient waveform data from Philips' proprietary format into an open-source format. With assistance from the medical equipment vendor, the waveforms, trends, and alarms were translated into WFDB, an open data format that is used for publicly available databases on the National Institutes of Health-sponsored *PhysioNet* web site [3].

All data that were integrated into the MIMIC-II database were de-identified in compliance with Health Insurance Portability and Accountability Act standards to facilitate public access to MIMIC-II. Deletion of protected health information from structured data sources was straightforward (e.g., database fields that provide the patient name, date of birth, etc.). We also removed protected health information from the discharge summaries, diagnostic reports, and the approximately 700,000 free-text nursing and respiratory notes in MIMIC-II using an automated algorithm that has been shown to have superior performance in comparison to clinicians in detecting protected health information [4]. This algorithm accommodates the broad spectrum of writing styles in our data set, including personal variations in syntax, abbreviations, and spelling. We have posted the algorithm in open-source form as a general tool to be used by others for de-identification of free-text notes [5].

5.4 Data Sharing

MIMIC-II is an unprecedented and innovative open research resource that grants researchers from around the world free access to highly granular ICU data and in the process substantially accelerates knowledge creation in the field of critical care medicine. The MIMIC Waveform Database is freely available to all via the PhysioNet website, and no registration is required. The MIMIC Clinical Database is also available without cost. To restrict users to legitimate medical researchers, access to the clinical database requires completion of a simple data use agreement (DUA) and proof that the researcher has completed human subjects training [6].

The MIMIC-II clinical database is available in two forms. In the first form, interested researchers can obtain a flat-file text version of the clinical database and the associated database schema that enables them to reconstruct the database using a database management system of their choice. In the second form, interested researchers can gain limited access to the database through QueryBuilder, a password-protected web service. Database searches using QueryBuilder allow users to familiarize themselves with the database tables and to program database queries using the Structured Query Language. Query output, however, is limited to 1000 rows because of our laboratory's limited computational resources. Accessing and processing data from MIMIC-II is complex. It is recommended that studies based on the MIMIC-II clinical database be conducted as collaborative efforts that include clinical, statistical, and relational database expertise. Detailed documentation and procedures for obtaining access to MIMIC-II are available at the MIMIC-II web site [1]. The current release of MIMIC-II is version 2.6, containing approximately 36,000 patients, including approximately 7000 neonates, and covering the period 2001–2008. At the present time approximately 1700 individuals worldwide in academia, industry, and medicine have been credentialed to access MIMIC-II and are producing research results in physiologic signal processing, clinical decision support, predictive algorithms in critical care, pharmacovigilance, natural language processing, and more.

5.5 Updating

In 2008 the hospital made a major change in the ICU information system technology and in ICU documentation procedures. The Philips CareVue system was replaced with iMDsoft's MetaVision technology. In 2013 we began a major update to MIMIC to incorporate adult ICU data for the period 2008–2012. The effort required learning the entirely new data schema of MetaVision, and merging the new data format with the existing MIMIC design. The new MetaVision data included new data elements such as physician progress notes, oral and bolus medication administration records, etc. Updated data were extracted from hospital archives and from the SSA death files for the newly added patients. Almost two years of effort was invested to acquire, organize, debug, normalize and document the new database before releasing it.

MIMIC-III includes 20,000 new adult ICU admissions, bringing the total to approximately 60,000. The new database is known as MIMIC-III, and the acronym has been recast as "**M**edical **I**nformation **M**art for **I**ntensive **C**are" [7].

5.6 Support

Support of the MIMIC databases includes: credentialing new users, administration of the authorized user list (i.e. users who have signed the DUA and have been granted permission to access MIMIC-II), user account creation, password resets and granting/revoking permissions. The servers providing MIMIC-II include authentication, application, database and web servers. All systems must be monitored, maintained, upgraded and backed up; the maintenance burden continues to increase as the number of database users grows. The engineering staff at LCP attempt to answer user queries as needed. Common questions are added to list of frequently asked questions on the MIMIC website and we regularly update our online documentation.

5.7 Lessons Learned

Building and distributing MIMIC-like databases is challenging, complex, and requires the cooperation and support of a number of individuals and institutions. A list of some of the more important requirements follows (Table 5.1).

Table 5.1 Health data requirements

1. The availability of digitized ICU and hospital data including structured and unstructured clinical data and high resolution waveform and vital sign data
2. A cooperative and supportive hospital IT department to assist in data extraction
3. A supportive IRB and hospital administration to assure both protection of patient privacy and release of de-identified data to the research community
4. Adequate engineering and data science capability to design and implement the database schema and to de-identify the data (including the unstructured textual data)
5. Sophisticated signal processing expertise to reformat and manage proprietary waveform data streams
6. Cooperation and technical support of equipment vendors
7. Adequate computational facilities for data archiving and distribution
8. Adequate technical and administrative personnel to provide user support and credentialing of users
9. Adequate financial support

5.8 Future Directions

The MIMIC-III database is a powerful and flexible research resource, but the generalizability of MIMIC-based studies is somewhat limited by the fact that the data are collected from a single institution. Multi-center data would have the advantages of including wider practice variability, and of course a larger number of cases. Data from international institutions would add still greater strength to the database owing to the even larger variations in practice and patient populations.

Our long-term goal is to create a public, multi-center, international data archive for critical care research. We envisage a massive, detailed, high-resolution ICU data archive containing complete medical records from patients around the world. The difficulty of such a project cannot be understated; nevertheless we propose to lay the foundation for such a system by developing a scalable framework that can readily incorporate data from multiple institutions, capable of supporting research on cohorts of critically ill patients from around the world.

Acknowledgments The development and maintenance of the MIMIC and PhysioNet resources have been funded by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) and the National Institute of General Medical Sciences (NIGMS) over the period 2003 to present. Grants R01EB1659, R01EB017205, R01GM104987, and U01EB008577.

References

1. MIMIC-II Web Site. http://physionet.org/mimic2
2. MIMIC User Guide. http://physionet.org/mimic2/UserGuide/
3. WaveForm DataBase Data Format. http://www.physionet.org/physiotools/wfdb.shtml
4. Neamatullah I, Douglass M, Lehman LH, Reisner A, Villarroel M, Long WJ, Szolovits P, Moody GB, Mark RG, Clifford GD (2008) Automated de-identification of free-text medical records. BMC Med Inform Decis Mak 8:32. doi:10.1186/1472-6947-8-327
5. Deidentification Software. http://www.physionet.org/physiotools/deid/
6. Accessing MIMIC. http://www.physionet.org/mimic2/mimic2_access.shtml
7. MIMIC-III Website. http://mimic.physionet.org/

Chapter 6
Integrating Non-clinical Data with EHRs

Yuan Lai, Edward Moseley, Francisco Salgueiro and David Stone

Take Home Messages

- Non-clinical factors make a significant contribution to an individual's health and providing this data to clinicians could inform context, counseling, and treatments.
- Data stewardship will be essential to protect confidential health information while still yielding the benefits of an integrated health system.

6.1 Introduction

The definition of "clinical" data is expanding, as a datum becomes clinical once it has a relation to a disease process. For example: the accessibility of one's home would classically be defined as non-clinical data, but in the context of a patient with a disability, this fact may become clinically relevant, and entered into the encounter note much like the patient's blood pressure and body temperature. However, even with this simple example, we can envision some of the problems with traditional non-clinical data being re-classified as clinical data, particularly due to its complexity.

6.2 Non-clinical Factors and Determinants of Health

Non-clinical factors are already significantly linked to health. Many public health policies focusing on transportation, recreation, food systems and community development are based on the relation between health and non-clinical determinants

The original version of this chapter was revised: A chapter author's name Edward Moseley was added. The erratum to this chapter is available at 10.1007/978-3-319-43742-2_30

MIT Critical Data, *Secondary Analysis of Electronic Health Records*,
DOI 10.1007/978-3-319-43742-2_6

such as behavioral, social and environmental factors [1]. Behavioral factors such as physical activity, diet, smoking and alcohol consumption are highly related to epidemic of obesity [2]. Some of this information, such as alcohol and tobacco use, is regularly documented by clinicians. Other information, such as dietary behaviors and physical activity, isn't typically captured, but may be tracked by new technology (such as wearable computers commonly referred to as "wearables") and integrated into electronic health records (EHRs). Such efforts may provide clinicians with additional context with which to counsel patients in an effort to increase their physical activity and reach a desired health outcome.

From a public health perspective, the same data obtained from these devices may be aggregated and used to guide decisions on public health policies. Continuing the prior example, proper amounts of physical activity will contribute to lower rates of mortality and chronic disease including coronary heart disease, hypertension, diabetes, breast cancer and depression across an entire population. Such data can be used to guide public health interventions in an evidence-based, cost-effective manner.

Both social and environmental factors are highly related to health. Social Determinants of Health (SDH) are non-clinical factors that affect the social and economic status of individuals and communities, including such items as their birthplace, living conditions, working conditions and demographic attributes [3]. Also included are social stressors such as crime, violence, and physical disorders, as well as others [4].

Environmental factors (i.e., air pollution, extreme weather, noise and poor indoor environmental quality) are highly related to an individual's health status. Densely built urban regions create air pollution, heat islands and high levels of noise, which have been implicated in causing or worsening a variety of health issues. For example, a study in New York City showed that asthma-related emergency admissions in youth from 5 to 17 years old were highly related to ambient ozone exposure. This annual NYC Community Health Survey also reveals that self-reported chronic health problems are related to extreme heat, suggesting that temperature can effect, or exacerbate, the symptoms of an individual's chronic illness. Social factors such as age and poverty levels also impact health. A study in New York City shows that fine particles ($PM_{2.5}$, a surrogate marker for pollution) attributable asthma hospital admissions are 4.5 times greater in high-poverty neighborhoods [5].

While outdoor environmental conditions merit public health attention, the average American spends only an hour of each day outdoors, and most individuals live, work and rest in an indoor environment, where other concerns reside. Poor indoor quality can cause building related illness and "sick building syndrome" (SBS)—where occupants experience acute health issues and discomfort, while no diagnosable illness can be readily identified [6]. Again in New York City, housing data was combined from multiple agencies in an effort to address indoor pollution concerns—using predictive analytics, the city was able to increase the rate of detection of buildings considered dangerous, as well as improve the timeliness in locating apartments with safety concerns or health hazards [7].

6.3 Increasing Data Availability

For many years scientists and researchers have had to deal with very limited available data to study behavioral, social and environmental factors that exist in cities, as well as the difficulty in evaluating their model with a large pool of urban data [8]. The big data revolution is bringing vast volumes of data and paradigmatic transformations to many industries within urban services and operations. This is particularly true in commerce, security and health care, as more data are systematically gathered, stored, and analyzed. The emergence of urban informatics also coincides with a transition from traditionally closed and fragmented data systems to more fully connected and open data networks that include mass communications, citizen involvement (e.g. social media), and informational flow [9].

In 2008, 3.3 billion of the world's inhabitants lived in cities, representing, for the first time in history the majority of the human population [10]. In 2014, 54 % of population lives in urban area and it is expected to increase to 66 % by 2050 [11]. With the growth of cities, there are rising concerns in public health circles regarding the impact of associated issues such as aging populations, high population densities, inadequate sanitation, environmental degradation, climate change factors, an increasing frequency of natural disasters, as well as current and looming resource shortages. A concomitantly large amount of information is required to plan and provide for the public health of these urban entities, as well as to prevent and react to adverse public events of all types (e.g. epidemiological, natural, criminal and politico-terroristic disasters).

The nature of the city as an agglomeration of inhabitants, physical objects and activities makes it a rich source of urban data. Today, billions of individuals are generating the digital data through their cellphones and use of the Internet including social networks. Hardware like global positioning systems (GPS) and other sensors are also becoming ubiquitous as they become more affordable, resulting in diverse types of data being collected in new and unique ways [12]. This is especially true in cities due to their massive populations, creating hotspots of data generation and hubs of information flow. Such extensive data availability may also provide the substrate for more statistically robust models across multiple disciplines.

An overview of the volume, variety, and format of open urban data is essential to further integration with electronic health records. As more cities begin building their informational infrastructure, the volume of city data increases rapidly. The majority of urban data are in tabular format with location-based information [8]. Data source and collection processes vary based on the nature of urban data. Passive sensors continuously collect environmental data such as temperature, air quality, solar radiation, and noise, and construct an urban sensing infrastructure along with ubiquitous computing [13]. There is also a large amount of city data generated by citizens such as service requests and complaints. Some pre-existing data, like those in the appropriate tabular format, are immediately ready for integration, while other data contained in more complex file types, like Portable

Document Format (PDF) or others, are more difficult to parse. This problem can be compounded if the data are encoded in uncommon character languages.

The fact that many non-clinical data, especially urban data, is geo-located enables clinicians to consider patient health within a broader view. Many environmental, social and behavioral factors link together spatially, and such spatial correlation is a key measurement in epidemiology, as it allows for the facilitation of data integration based on location. Connections and solutions become more visible by linking non-clinical data with EHR on a public health and city planning level. Recently, IBM announced that, by teaming supercomputer Watson's cognitive computing with data from CVS Health (a pharmacy chain with locations across the U.S.), we will have better predictions regarding the prevalence of chronic conditions such as heart disease and diabetes in different cities and locations [14].

6.4 Integration, Application and Calibration

In a summary of all cities in the United States that published open data sets as of 2013, it was found that greater than 75 % of datasets were prepared in tabular format [8]. Tabular data is most amenable for automated integration, as it is already in the final format prior to being integrated into most relational databases (as long as the dataset contains a meaningful attribute, or variable, with which to relate to other data entries). Furthermore, data integration occurs most easily when the dataset is "tidy", or follows the rule of "one observation per row and one variable per column." Any data manipulation process resulting in a dataset that is aggregated or summarized could remove a great deal of utility from that data [15].

For instance, a table that is familiar within one working environment may not be easily decipherable to another individual and may be nearly impossible for a machine to parse without proper context given for what is within the table. An example could be a table of blood pressure over time and in different locations for a number of patients, which may look like (Table 6.1).

Here we see two patients, Patient 1 and Patient 2, presenting to two locations, Random and Randomly, RA, on two different dates. While this table may be easily read by someone familiar with the format, such that an individual would understand that Patient 1 on the 1st of January, 2015, presented to a healthcare setting in Random, RA with a systolic blood pressure of 130 mmHg and a diastolic pressure of 75 mmHg, it may be rather difficult to manipulate these data to a tidy format without understanding the context of the table.

Table 6.1 Example of a table requiring proper context to read

Patient blood pressure chart	Random, RA		Randomly, RA	
	1-Jan-15	7-Jan-15	1-Jan-15	7-Jan-15
Patient 1	130/75	139/83	141/77	146/82
Patient 2	158/95	151/91	150/81	141/84

If this table were to be manipulated in a manner that would make it easily analyzed by a machine (as well as other individuals without requiring an explanation of the context), it would follow the rule of one column per variable and one row per observation, as below (Table 6.2).

There are further limitations imparted due to data resolutions, which refers to the detail level of data in space, time or theme, especially the spatial dimension of the data [16]. Examples include: MM/DD/YY time formats compared to YYYY; or zip codes compared to geographic coordinates. Even with these limitations, one may still be able to draw relevant information from these spatial and temporal data.

One method to provide spatial orientation to a clinical encounter has recently been adopted by the administrators of the Medical Information Mart for Intensive Care (MIMIC) database, which currently contains data from over 37,000 intensive care unit admissions [17]. Researchers utilize the United States Zip Code system to approximate the patients' area of residence. This method reports the first three digits of the patient's zip code, while omitting the last two digits [18]. The first three digits of a zip code contain two pieces of information: the first integer in the code refers to a number of states, the following two integers refer to a U.S. Postal Service Sectional Center Facility, through which the mail for that state's counties is processed [19]. The first three digits of the zip code are sufficient to find all other zip codes serviced by the Sectional Center Facility, and population level data of many types are available by zip code as per the U.S. Government's census [20].

Table 6.2 A tidy dataset that contains a readily machine-readable format of the data in Table 6.1

Patient ID	Place	Date (MM/DD/YYYY)	Pressure (mmHg)	Cycle
1	Random, RA	1/1/2015	130	Systole
1	Random, RA	1/1/2015	75	Diastole
1	Random, RA	1/7/2015	139	Systole
1	Random, RA	1/7/2015	83	Diastole
1	Randomly, RA	1/1/2015	141	Systole
1	Randomly, RA	1/1/2015	77	Diastole
1	Randomly, RA	1/7/2015	146	Systole
1	Randomly, RA	1/7/2015	82	Diastole
2	Random, RA	1/1/2015	158	Systole
2	Random, RA	1/1/2015	95	Diastole
2	Random, RA	1/7/2015	151	Systole
2	Random, RA	1/7/2015	91	Diastole
2	Randomly, RA	1/1/2015	150	Systole
2	Randomly, RA	1/1/2015	81	Diastole
2	Randomly, RA	1/7/2015	141	Systole
2	Randomly, RA	1/7/2015	84	Diastole

Connections and solutions become more visible by linking non-clinical data with EHRs on a public health and city planning level. Although many previous studies show the correlation between air pollution and asthma, it is only recently individuals became able to trace PM2.5, SO_2 and Nickel (Ni) in the air back to the generators in buildings with aged boilers and heating systems, which is due in large part to increasing data collection and integration across multiple agencies and disciplines [21]. As studies reveal additional links between our environment and pathological processes, our ability to address potential health threats will be limited by our ability to measure these environmental factors in sufficient resolution to be able to apply it to patient level, creating truly personalized medicine.

For instance, two variables, commonly captured in many observations, are geo-spatiality and temporality. Since all actions share these conditions, integration is possible among a variety of data otherwise loosely utilized in the clinical encounter. When engaged in an encounter, a clinician can determine, from data collected during the examination and history taking, the precise location of the patient over a particular period of time within some spatial resolution. As a case example, a patient may present with an inflammatory process of the respiratory tract. The individual may live in random, RA, and work as an administrator in Randomly, RA; one can plot these variables over time, and separate them to represent both the individuals' work and home environment—as well as other travel (Fig. 6.1).

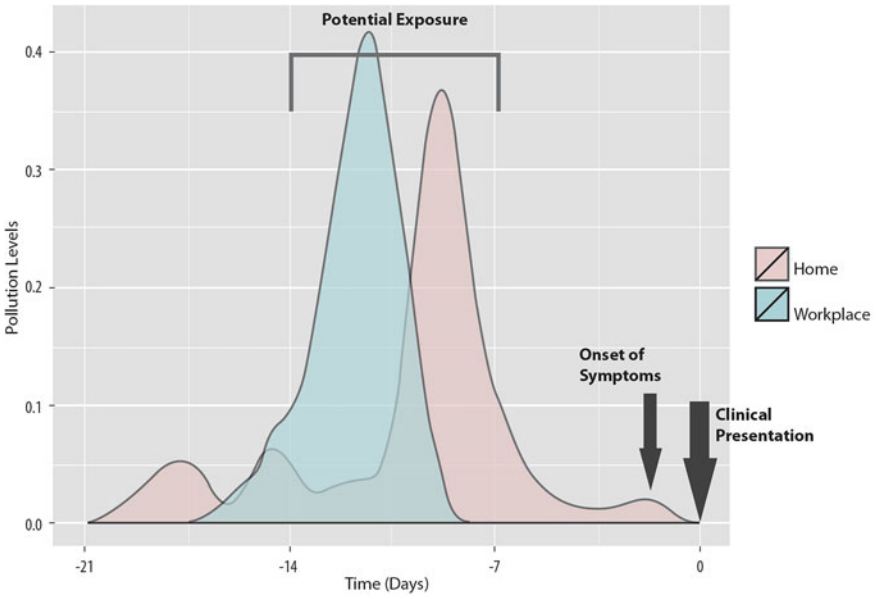

Fig. 6.1 Example of pollution levels over time for a patient's "work" and "home" environment with approximate labels that may provide clinically relevant decision support

This same method may be applied to other variables that could be determined to have statistical correlates of significance during the timeframe prior to the onset of symptoms and then the clinical encounter.

With the increasing availability of information technology, there is less need for centralized information networks, and the opportunity is open for the individual to participate in data collection, creating virtual sensor networks of environmental and disease measurement. Mobile and social web have created powerful opportunities for urban informatics and disaster planning particularly in public health surveillance and crisis response [13]. There are geo-located mobile crowdsourcing applications such as Health Map's Outbreaks Near Me [22] and Sickweather [23] collecting data on a real-time social network.

In the 2014 Ebola Virus Disease outbreak, self-reporting and close contact reporting was essential to create accurate disease outbreak maps [24]. The emergence of wearables is pushing both EHR manufacturers to develop frameworks that integrate data from wearable devices, and third party companies to provide cloud storage and integration of data from different wearables for greater analytic power.

Attention and investment in digital health and digital cities continues to grow rapidly. In digital health care, investors' funding has soared from $1.1 Billion in 2011 to $5.0 Billion in 2014, and big data analytics ranks as the #1 most active subsector of digital healthcare startups in both amount of investment and number of deals [25]. Integration will be a long process requiring digital capabilities, new policies, collaboration between the public and private sectors, and innovations from both industry leaders and research institutions [26]. Yet we believe with more interdisciplinary collaborations in data mining and analytics, we will gain new knowledge on the health-associated non-clinical factors and indicators of disease outcomes [27]. Furthermore, such integration creates a feedback loop, pushing cities to collect better and larger amounts of data. Integrating non-clinical information into health records remains challenging. Ideally the information obtained from the patient would flow into the larger urban pool and vice versa. Challenges remain on protecting confidentiality at a single patient level and determining applicability of macroscopic data to the single patient.

6.5 A Well-Connected Empowerment

Disease processes can result and be modified by interactions of the patient and his or her environment. Understanding this environment is of importance to clinicians, hospitals, public health policy makers and patients themselves. With this information we can preempt patients at risk for disease (primary prevention), act earlier in minimizing morbidity from disease (secondary prevention) and optimize therapeutics.

A good example of the use of non-clinical data for disease prevention is the use of geographical based information systems (GIS) for preemptive screening of

populations at risk for sexually transmitted diseases (STDs). Geographical information systems are used for STD surveillance in about 50 % of state STD surveillance programs is the U.S. [28]. In Baltimore (Maryland, U.S.) a GIS based study identified core groups of repeat gonococcal (an STD) infection that showed geographical clustering [29]. The authors hinted at the possibility of increased yield when directing prevention to geographically restricted populations.

A logical next step is the interaction between public health authority systems and electronic medical records. As de-identified geographical health information becomes publically available, an electronic medical record would be able to download this information from the cloud, apply it to the patient's zip code, sex, age and sexual preference (if documented) and warn/cue the clinician that would decide if an intervention is required based on a calculated risk to acquire a STD.

6.6 Conclusion

Good data stewardship will be essential for protecting confidential health information from unintended and illegal disclosure. For patients, the idea of increasing empowerment in their health is essential [8]. Increasing sensor application and data visualization make our own behavior and surroundings more visible and tangible, and alert us about potential environmental risks. More importantly, it will help us to better understand and gain power over our own lives.

The dichotomy of addressing population health versus individual health must be addressed. Researchers should ask: what information is relevant to the target which I'm addressing, and what data do we feed from this patient's record into the public health realm? The corollary to that question is: how can we balance the individual's right to privacy with the benefit of non-clinical data applicable to the individual and to the large populations? Finally: how can we create systems that select relevant data from a single patient and present it to the clinician in a population-health context? In this chapter, we have attempted to provide an overview of the potential use of traditionally non-clinical data in electronic health records, in addition to mapping some of the pitfalls and strategies to using such data, as well as highlighting practical examples of the use of these data in a clinical environment.

References

1. Barton H, Grant M (2013) Urban planning for health cities, a review of the progress of the european healthy cities program. J Urban Health Bull NY Acad Med 90:129–141
2. Badland HM, Schofield GM, Witten K, Schluter PJ, Mavoa S, Kearns RA, Hinckson EA, Oliver M, Kaiwai H, Jensen VG, Ergler C, McGrath L, McPhee J (2009) Understanding the relationship between activity and neighborhoods (URBAN) study: research design and methodology. BMC Pub Health 9:244
3. Osypuk TL, Joshi P, Geronimo K, Acevedo-Garcia D (2014) Do social and economic policies influence health? Rev Curr Epidemiol Rep 1:149–164
4. Shmool JLC, Kubzansky LD, Newman OD, Spengler J, Shepard P, Clougherty JE (2014) Social stressors and air pollution across New York City communities: a spatial approach for assessing correlations among multiple exposures. Environ Health 13:91
5. Kheirbek I, Wheeler K, Walters S, Kass D, Matte T (2013) PM2.5 and ozone health impacts and disparities in New York City: sensitivity to spatial and temporal resolution. Air Qual Atmos Health 6:473–486
6. Indoor Air Facts No. 4 sick building syndrome. United States Environmental Protection Agency, Research and Development (MD-56) (1991)
7. Goldstein B, Dyson L (2013) Beyond transparency: open data and the future of civic innovation. Code for America Press, San Francisco
8. Barbosa L, Pham K, Silva C, Vieira MR, Freire J (2014) Structured open urban data: understanding the landscape. Big Data 2:144–154
9. Shane DG (2011) Urban design since 1945—a global perspective. Wiley, New York, p 284
10. National Intelligence Council (2012) Global trends 2030: alternative worlds. National Intelligence Council
11. World Urbanization Prospects, United Nations (2014)
12. Goldsmith S, Crawford S (2014) The responsive city: engaging communities through data-smart governance. Wiley, New York
13. Boulos M, Resch B, Crowley D, Breslin J, Sohn G, Burtner R, Pike W, Jezierski E, Chuang K (2011) Crowdsourcing, citizen sensing and sensor web technologies for public and environmental health surveillance and crisis management: trends, OGC standards and application examples. Int J Health Geographic 10:67
14. McMullan T. Dr Watson: IBM plans to use Big Data to manage diabetes and obesity. URL: http://www.alphr.com/life-culture/1001303/dr-watson-ibm-plans-to-use-big-data-to-manage-diabetes-and-obesity
15. Wickham H (2014) Tidy data. J Stat Softw 10:59
16. Haining R (2004) Spatial data analysis—theory and practice. Cambridge University Press, Cambridge, p 67
17. MIMIC II Databases. Available from: http://physionet.org/mimic2. Accessed 02 Aug 2015
18. Massachusetts Institute of Technology, Laboratory of Computational Physiology. mimic2 v3.0 D_PATIENTS table. URL: https://github.com/mimic2/v3.0/blob/ad9c045a5a778c6eb283bdad310594484cca873c/_posts/2015-04-22-dpatients.md. Accessed 02 Aug 2015 (Archived by WebCite® at http://www.webcitation.org/6aUNzhW6g)
19. http://pe.usps.com/businessmail101/glossary.htm
20. http://factfinder.census.gov/
21. Jeffery N, McKelvey W, Matte T (2015) using tracking infrastructure to support public health programs, policies, and emergency response in New York City. Pub Health Manag Pract 21(2 Supp):S102–S106
22. http://www.healthmap.org/outbreaksnearme/
23. http://www.sickweather.com

24. Kouadio KI, Clement P, Bolongei J, Tamba A, Gasasira AN, Warsame A, Okeibunor JC, Ota MO, Tamba B, Gumede N, Shaba K, Poy A, Salla M, Mihigo R, Nshimirimana D (2015) Epidemiological and surveillance response to Ebola virus disease outbreak in Lofa County, Liberia (Mar–Sept 2014); lessons learned, edn 1. PLOS Currents Outbreaks. 6 May 2015. doi: 10.1371/currents.outbreaks.9681514e450dc8d19d47e1724d2553a5
25. The re-imagination of healthcare. StartUp Health Insights. www.startuphealth.com/insights
26. Ericsson Networked Society City Index (2014)
27. Corti B, Badland H, Mavoa S, Turrell G, Bull F, Boruff B, Pettit C, Bauman A, Hooper P, Villanueva K, Burt T, Feng X, Learnihan V, Davey R, Grenfell R, Thackway S (2014) Reconnecting urban planning with health: a protocol for the development and validation of national livability indicators associated with non-communicable disease risk behaviors and health outcomes. Pub Health Res Pract 25(1):e2511405
28. Bissette JM, Stover JA, Newman LM, Delcher PC, Bernstein KT, Matthews L (2009) Assessment of geographic information systems and data confidentiality guidelines in STD programs. Pub Health Rep 124(Suppl 2):58–64
29. Bernstein TK, Curriero FC, Jennings JM et al (2004) Defining core gonorrhea transmission utilizing spatial data. Am J Epidemiol 160:51–58

Chapter 7
Using EHR to Conduct Outcome and Health Services Research

Laura Myers and Jennifer Stevens

Take Home Messages

- Electronic Health Records have become an essential tool in clinical research, both as a supplement to existing methods, but also in the growing domains of outcomes research and analytics.
- While EHR data is extensive and analytics are powerful, it is essential to fully understand the biases and limitations introduced when used in health services research.

7.1 Introduction

Data from electronic health records (EHR) can be a powerful tool for research. However, researchers must be aware of the fallibility of data collected for clinical purposes and of biases inherent to using EHR data to conduct sound health outcomes and health services research. Innovative methods are currently being developed to improve the quality of data and thus our ability to draw conclusions from studies that use EHR data.

The United States devotes a large share of the Gross Domestic Product (17.6 % in 2009) to health care [1]. With such a huge financial and social investment in healthcare, important questions are fundamental to evaluating this investment:

> How do we know what treatment works and for which patients?
> How much should health care cost? When is too much to pay? In what type of care should we invest more or less resources?
> How does the health system work and how could it function better?

Health services research is a field of research that lives at the intersection of health care policy, management, and clinical care delivery and seeks to answer

MIT Critical Data, *Secondary Analysis of Electronic Health Records*,
DOI 10.1007/978-3-319-43742-2_7

these questions. Fundamentally, health services research places the health system under the microscope as the organism of study.

To begin to address these questions, researchers need large volumes of data across multiple patients, across different types of health delivery structures, and across time. The simultaneous growth of this field of research in the past 15 years has coincided with the development of the electronic health record and the increasing number of providers who make use of them in their workspace [2]. The EHR provides large quantities of raw data to fuel this research, both at the granular level of the patient and provider and at the aggregated level of the hospital, state, or nation.

Conducting research with EHR data has many challenges. EHR data are riddled with biases, collected for purposes other than research, inputted by a variety of users for the same patient, and difficult to integrate across health systems [See previous chapter "Confounding by Indication"]. This chapter will focus on the attempts to capitalize on the promise of the EHR for health services research with careful consideration of the challenges researchers must address to derive meaningful and valid conclusions.

7.2 The Rise of EHRs in Health Services Research

7.2.1 The EHR in Outcomes and Observational Studies

Observational studies, either retrospective or prospective, attempt to draw inferences about the effects of different exposures. Within health services research, these exposures include both different types of clinical exposures (e.g., does hormone replacement therapy help or hurt patients?) and health care delivery exposures (e.g., does admission to a large hospital for cardiac revascularization improve survival from myocardial infarction over admission to a small hospital). The availability of the extensive health data in electronic health records has fueled this type of research, as data extraction and transcription from paper records has ceased to be a barrier to research. These studies capitalize on the demographic and clinical elements that are routinely recorded as part of an encounter with the health system (e.g., age, sex, race, procedures performed, length of stay, critical care resources used).

We have highlighted a number of examples of this type of research below. Each one is an example of research that has made use of electronic health data, either at the national or hospital level, to draw inferences about health care delivery and care.

Does health care delivery vary? The researchers who compile and examine the Dartmouth Atlas have demonstrated substantial geographic variation in care. In their original article in *Science*, Wennberg and Gittlesohn noted wide variations in the use of health services in Vermont [3]. These authors employed data derived from the use of different types of medical services—home health services, inpatient discharges, etc.—to draw these inferences. Subsequent investigations into national variation in care have been able to capitalize on the availability of such data electronically [4].

Do hospitals with more experience in a particular area perform better? Birkmeyer and colleagues studied the intersection of hospital volume and surgical outcomes with absolute differences in adjusted mortality rates between low volume hospitals and high volume hospitals ranging from 12 % for pancreatic resection to 0.2 % for carotid endarterectomy [5]. Kahn et al. also used data available in over 20,000 patients to demonstrate that mortality associated with mechanical ventilation was 37 % lower in high volume hospitals compared with low volume hospitals [6]. Both of these research groups made use of large volumes of clinical and claims data—Medicare claims data in the case of Birkmeyer and colleagues and the APACHE database from Cerner for Kahn et al.—to ask important questions about where patients should seek different types of care.

How can we identify harm to patients despite usual care? Herzig and colleagues made use of the granular EHR at a single institution and found that the widely-prescribed medications that suppress acid production were associated with an increased risk of pneumonia [7]. Other authors have similarly looked at the EHR found that these types of medications are often continued on discharge from the hospital [8, 9].

To facilitate appropriate modeling and identification of confounders in observational studies, researchers have had to devise methods to extract markers of diagnoses, severity of illness, and patient comorbidities using only the electronic fingerprint. Post et al. [10] developed an algorithm to search for patients who had diuretic-refractory hypertension by querying for patients who had a diagnosis of hypertension despite 6 months treatment with a diuretic. Previously validated methods for reliably measuring the severity of a patient's illness, such as APACHE or SAPS scores [11, 12], have data elements that are not easily extracted in the absence of manual inputting of data. To meet these challenges, researchers such as Escobar and Elixhauser have proposed alternative, electronically derived methods for both severity of illness measures [13, 14] and identification of comorbidities [14]. Escobar's work, with a severity of illness measure with an area under the curve of 0.88, makes use of highly granular electronic data including laboratory values; Elixhauser's comorbidity measure is publically available through the Agency for Healthcare Research and Quality and solely requires billing data [15].

Finally, researchers must develop and employ appropriate mathematical models that can accommodate the short-comings of electronic health data or else they risk drawing inaccurate conclusions. Examples of such modeling techniques are extensive have included propensity scores, causal methods such as marginal structural models and inverse probability weights, and designs from other fields such as instrumental variable analysis [16–19]. The details of these methods are discussed elsewhere in this text.

7.2.2 The EHR as Tool to Facilitate Patient Enrollment in Prospective Trials

Despite the power of the EHR to conduct health services and outcomes research retrospectively, the gold standard in research remains prospective and randomized trials. The EHR has functioned as a valuable tool to screen patients at a large scale

for eligibility. In this instance, research staff uses the data available through the electronic record as a high-volume screening technique to target recruitment efforts to the most appropriate patients. Clinical trials that develop electronic strategies for patient identification and recruitment are at an even greater advantage, although such robust methods have been described as sensitive but not specific, and frequently require coupling screening efforts with manual review of individual records [20]. Embi et al. [21] have proposed using the EHR to simultaneously generate Clinical Trial Alerts, particularly in commercial EHRs such as Epic to leverage the EHR in a point of care strategy. This strategy could expedite enrollment although it must be weighed against the risk of losing patient confidentiality, an ongoing tension between patient care and clinical trial enrollment [22].

7.2.3 The EHR as Tool to Study and Improve Patient Outcomes

Quality can also be tracked and reported through EHRs, either for internal quality improvement or for national benchmarking; the Veterans' Affairs' (VA) healthcare system highlights this. Byrne et al. [23] reported that in the 1990s, the VA spent more money on information technology infrastructure and achieved higher rates of adoption compared to the private sector. Their home-grown EHR, which is called VistA, provided a way to track preventative care processes such as cancer and diabetes screening through electronic pop up messages. Between 2004 and 2007, they found that the VA system achieved better glucose and lipid control for diabetics compared to a Medicare HMO benchmark [23]. While much capital investment was needed during the initial implementation of VistA, it is estimated that adopting this infrastructure saved the VA system $3.09 billion in the long term. It also continues to be a source of quality improvement as quality metrics evolve over time [23].

7.3 How to Avoid Common Pitfalls When Using EHR to Do Health Services Research

We would propose the following hypothetical research study as a case study to highlight common challenges to conducting health services research with electronic health data:

> Proposed research study: Antipsychotic medications (e.g. haloperidol) are prescribed frequently in the intensive care unit to treat patients with active delirium. However, these medications have been associated with their own potential risk of harm [24] that is separate from the overall risk of harm from delirium. The researchers are interested in whether treatment with antipsychotics increases the risk of in-hospital death and increases the cost of care and use of resources in the hospital.

7.3.1 Step 1: Recognize the Fallibility of the EHR

The EHR is rarely complete or correct. Hogan et al. [25] tried to estimate how complete and accurate data are in studies that are conducted on an EHR, finding significant variability in both. Completeness ranged from 31 to 100 % and correctness ranged from 67 to 100 % [25]. Table 7.1 highlights examples of different diagnoses and possible sources of data, which may or may not be present for all patients.

> Proposed research study: The researchers will need to extract which patients were exposed to antipsychotics and which were not. However, there is unlikely to be one single place where this information is stored. Should they use pharmacy dispensing data? Nursing administration data? Should they look at which patients were charged for the medications? What if they need these data from multiple hospitals with different electronic health records?

Additionally, even with a robust data extraction strategy, the fidelity of different types of data is variable [26–33]. For example, many EHR systems have the option of entering free text for a medical condition, which may be spelled wrong or be worded unconventionally. As another example, the relative reimbursement of a particular billing code may influence the incidence of that code in the electronic health record so billing may not reflect the true incidence and prevalence of the disease [34, 35].

7.3.2 Step 2: Understand Confounding, Bias, and Missing Data When Using the EHR for Research

We would highlight the following methodological issues inherent in conducting research with electronic health records: selection bias, confounding, and missing data. These are explored in greater depth in other chapters of this text.

Table 7.1 Examples of the range of data elements that may be used to identify patients with either ischemic heart disease or acute lung injury through the electronic health record

Disease state	Data source	Example
Ischemic heart disease	Billing data	ICD-9 code 410 [48]
	Laboratory data	Positive troponin during admission
	Physician documentation	In the discharge summary: "the patient was noted to have ST elevations on ECG and was taken to the cath lab"…
Acute lung injury	Billing data	ICD-9 code 518.5 and 518.82 with the procedural codes 96.70, 96.71 and 96.72 for mechanical ventilation [49]
	Radiology data	"Bilateral" and "infiltrates" on chest x-ray reads [50]
	Laboratory data	PaO2/FiO2 < 300 mmHg

Selection bias, or the failure of the population of study to represent the generalizable population, can occur if all the patients, including controls, are already seeking medical care within an EHR-based system. For example, in EHR-based studies comparing medical versus surgical approaches to the same condition may not be comparing equivalent patients in each group; patients seeking a surgical correction may fundamentally differ from those seeking a more conservative approach. Hripcsak et al. [36] used a large clinical data set from a tertiary center in 2007 to compare mortality from pneumonia to a hand-collected data set that had been published previously; the different search criteria altered the patient population and the subsequent risk of death. While it is not eliminated entirely, selection bias is reduced when prospective randomization takes place [37].

Confounding bias represents the failure to appropriately account for an additional variable that influences both the dependent and independent variable. In research with electronic health records, confounding represents a particular challenge, as identification of all possible confounding variables is nearly impossible.

> Proposed research study: The researchers in this study are interested in the patient-level outcomes of what happens to those patients exposed to antipsychotics during their stay. But patients who are actively delirious while in the ICU are likely to be sicker than those who are not actively delirious and sicker patients require more hospital resources. As a result, antipsychotics will appear to be associated with a higher risk of in-hospital mortality and use of hospital resources not due to the independent effect of the drug but rather as a result of confounding by indication.

Missing data or unevenly sampled data collected as part of the EHR creates its own complex set of challenges for health services research. For example, restricting the analysis to patients with only a complete set of data may yield very different (and poorly generalizable) inferences. The multidimentionality of this problem often goes unexamined and underestimated. Nearly all conventional analytic software presumes completeness of the matrix of data, leading many researchers to fail to fully address these issues. For example, data can be misaligned due to lack of sampling, missing data, or simple misalignment. In other words, the data could not be measured during a period of time for an intentional reason (e.g., a patient was extubated and therefore no values for mechanical ventilation were documented) and should not be imputed or the data was measured but was unintentionally not recorded and therefore can be imputed. Rusanov et al. studied 10,000 outpatients at a tertiary center who underwent general anesthesia for elective procedures. Patients with a higher risk of adverse outcome going into surgery had more data points including laboratory values, medication orders and possibly admission orders compared to less sick patients [38], making the missing data for less sick patients intentional. Methods for handling missing data have included omitting cases are note complete, pairwise deletion, mean substitution, regression substitution, or using modeling techniques for maximum likelihood and multiple imputation [39].

7.4 Future Directions for the EHR and Health Services Research

7.4.1 Ensuring Adequate Patient Privacy Protection

It is controversial whether using EHR for research goes against our national privacy standard. In large cohorts, many patients may be present with the same health information, thereby rendering the data sufficiently deidentified. Further, Ingelfinger et al. acknowledge that countries with healthcare registries such as Scandinavia have a distinct research advantage [40]. However, health information is a protected class of information under the Health Insurance Portability and Accountability Act, so there is significant awareness among U.S. healthcare professionals and researchers about its proper storage and dissemination. Some argue that patients should be consented (versus just notified) that their information could be used for research purposes in the future. Ingelfinger et al. [40] recommends IRB approval of registries and a rigorous deidentification process.

Public perception on the secondary use of EHR may not be as prohibitive as policymakers may have believed. In a survey of 3300 people, they were more willing to have their information used for research by university hospitals, compared to public health departments or for quality improvement purposes [41]. They were much less willing to contribute to marketing efforts or have the information used by pharmaceutical companies [41].

With the growing amount of information being entered into EHRs across the country, the American Medical Informatics Association convened a panel to make recommendations for how best to use EHR securely for purposes other than direct patient care. In 2006, the panel called for a national standard to deal with the issue of privacy. They described complex situations where there were security breaches due to problems with deidentification or data was being sold by physicians for profit [42]. While the panel demanded that the national framework be transparent, comprehensive and publicly accepted, they did not propose a particular standard at that time [42]. Other groups such as the Patient-Centered Outcomes Research Institute have since addressed the same conflict in a national forum in 2012. Similarly, while visions were discussed, no explicit recommendation was set forth [PCORI]. Controversy continues in this area.

7.5 Multidimensional Collaborations

Going forward, the true power of integrated data can only be harnessed by forming more collaborations, both within institutions and between them. Research on a national scale in the U.S. has been shown to be feasible. The FDA implemented a pilot program in 2009 called the Mini-Sentinel program. It brought together 31 academic and private organizations to monitor for safety events related to

medications and devices currently on the market [43]. Admittedly, merging databases may require significant financial resources, especially if the datasets need to be coded and/or validated, but researchers like Bradley et al. [44] believe this is a cost-effective use of grant money because of the vast potential to make advances in the way we deliver care. Fundamental to the feasibility of multidimensional collaborations is the ability to ensure accuracy of large-scale data and integrate it across multiple health record technologies and platforms. Efforts to ensure data quality and accessibility must be promoted alongside patient privacy.

7.6 Conclusion

Researchers continue to ask fundamental questions of our health system, making use of the deluge of data generated by EHRs. Unfortunately, that deluge is messy and problematic. As the field of health services research with EHRs continues to evolve, we must hold researchers to rigorous standards [45] and encourage more investment in research-friendly clinical databases as well as cross-institutional collaborations. Only then will the discoveries in health outcomes and health services research be one click away [46, 47]. It is time for healthcare to reap the same reward from a rich data source that is already in existence.

References

1. Center for Medicare and Medicaid Services (2015) National health expenditure data fact sheet
2. Jha AK, DesRoches CM, Campbell EG, Donelan K, Rao SR et al (2009) Use of electronic health records in U.S. hospitals. N Engl J Med 360:1628–1638
3. Wennberg J, Gittelsohn (1973) Small area variations in health care delivery. Science 182:1102–1108
4. Stevens JP, Nyweide D, Maresh S, Zaslavsky A, Shrank W et al (2015) Variation in inpatient consultation among older adults in the United States. J Gen Intern Med 30:992–999
5. Birkmeyer JD (2000) Relation of surgical volume to outcome. Ann Surg 232:724–725
6. Kahn JM, Goss CH, Heagerty PJ, Kramer AA, O'Brien CR et al (2006) Hospital volume and the outcomes of mechanical ventilation. N Engl J Med 355:41–50

7. Herzig SJ, Howell MD, Ngo LH, Marcantonio ER (2009) Acid-suppressive medication use and the risk for hospital-acquired pneumonia. JAMA 301:2120–2128
8. Murphy CE, Stevens AM, Ferrentino N, Crookes BA, Hebert JC et al (2008) Frequency of inappropriate continuation of acid suppressive therapy after discharge in patients who began therapy in the surgical intensive care unit. Pharmacotherapy 28:968–976
9. Zink DA, Pohlman M, Barnes M, Cannon ME (2005) Long-term use of acid suppression started inappropriately during hospitalization. Aliment Pharmacol Ther 21:1203–1209
10. Post AR, Kurc T, Cholleti S, Gao J, Lin X et al (2013) The analytic information warehouse (AIW): a platform for analytics using electronic health record data. J Biomed Inform 46:410–424
11. Zimmerman JE, Kramer AA, McNair DS, Malila FM (2006) Acute physiology and chronic health evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. Crit Care Med 34:1297–1310
12. Moreno RP, Metnitz PG, Almeida E, Jordan B, Bauer P et al (2005) SAPS 3—from evaluation of the patient to evaluation of the intensive care unit. Part 2: development of a prognostic model for hospital mortality at ICU admission. Intensive Care Med 31:1345–1355
13. Escobar GJ, Greene JD, Scheirer P, Gardner MN, Draper D et al (2008) Risk-adjusting hospital inpatient mortality using automated inpatient, outpatient, and laboratory databases. Med Care 46:232–239
14. Elixhauser A, Steiner C, Harris DR, Coffey RM (1998) Comorbidity measures for use with administrative data. Med Care 36:8–27
15. Project HCaU (2015) Comorbidity software, Version 3.7
16. Rubin DB, Thomas N (1996) Matching using estimated propensity scores: relating theory to practice. Biometrics 52:249–264
17. Rubin DB (1997) Estimating causal effects from large data sets using propensity scores. Ann Intern Med 127:757–763
18. Howell MD, Novack V, Grgurich P, Soulliard D, Novack L et al (2010) Iatrogenic gastric acid suppression and the risk of nosocomial Clostridium difficile infection. Arch Intern Med 170:784–790
19. Hernan MA, Brumback B, Robins JM (2000) Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. Epidemiology 11:561–570
20. Thadani SR, Weng C, Bigger JT, Ennever JF, Wajngurt D (2009) Electronic screening improves efficiency in clinical trial recruitment. J Am Med Inform Assoc 16:869–873
21. Embi PJ, Jain A, Clark J, Harris CM (2005) Development of an electronic health record-based clinical trial alert system to enhance recruitment at the point of care. AMIA Annu Symp Proc, 231–235
22. PCORnet (2015) Rethinking clinical trials: a living textbook of pragmatic clinical trials
23. Byrne CM, Mercincavage LM, Pan EC, Vincent AG, Johnston DS et al (2010) The value from investments in health information technology at the U.S. Department of Veterans Affairs. Health Aff (Millwood) 29:629–638
24. Ray WA, Chung CP, Murray KT, Hall K, Stein CM (2009) Atypical antipsychotic drugs and the risk of sudden cardiac death. N Engl J Med 360:225–235
25. Hogan WR, Wagner MM (1997) Accuracy of data in computer-based patient records. J Am Med Inform Assoc 4:342–355
26. Lee DS, Donovan L, Austin PC, Gong Y, Liu PP et al (2005) Comparison of coding of heart failure and comorbidities in administrative and clinical data for use in outcomes research. Med Care 43:182–188
27. Iwashyna TJ, Odden A, Rohde J, Bonham C, Kuhn L et al (2014) Identifying patients with severe sepsis using administrative claims: patient-level validation of the angus implementation of the international consensus conference definition of severe sepsis. Med Care 52:e39–e43
28. Jones G, Tarright N, Boelle PY, Marty J, Lalande V et al (2012) Accuracy of ICD-10 codes for surveillance of clostridium difficile infections, France. Emerg Infect Dis 18:979–981
29. Kramer JR, Davila JA, Miller ED, Richardson P, Giordano TP et al (2008) The validity of viral hepatitis and chronic liver disease diagnoses in Veterans Affairs Administrative databases. Aliment Pharmacol Ther 27:274–282

30. van de Garde EM, Oosterheert JJ, Bonten M, Kaplan RC, Leufkens HG (2007) International classification of diseases codes showed modest sensitivity for detecting community-acquired pneumonia. J Clin Epidemiol 60:834–838
31. Movig KL, Leufkens HG, Lenderink AW, Egberts AC (2003) Validity of hospital discharge International classification of diseases (ICD) codes for identifying patients with hyponatremia. J Clin Epidemiol 56:530–535
32. Sickbert-Bennett EE, Weber DJ, Poole C, MacDonald PD, Maillard JM (2010) Utility of international classification of diseases, ninth revision, clinical modification codes for communicable disease surveillance. Am J Epidemiol 172:1299–1305
33. Jhung MA, Banerjee SN (2009) Administrative coding data and health care-associated infections. Clin Infect Dis 49:949–955
34. O'Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF et al (2005) Measuring diagnoses: ICD code accuracy. Health Serv Res 40:1620–1639
35. Richesson RL, Rusincovitch SA, Wixted D, Batch BC, Feinglos MN et al (2013) A comparison of phenotype definitions for diabetes mellitus. J Am Med Inform Assoc 20:e319–e326
36. Hripcsak G, Knirsch C, Zhou L, Wilcox A, Melton G (2011) Bias associated with mining electronic health records. J Biomed Discov Collab 6:48–52
37. Hernan MA, Alonso A, Logan R, Grodstein F, Michels KB et al (2008) Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. Epidemiology 19:766–779
38. Rusanov A, Weiskopf NG, Wang S, Weng C (2014) Hidden in plain sight: bias towards sick patients when sampling patients with sufficient electronic health record data for research. BMC Med Inform Decis Mak 14:51
39. Allison PD (2001) Missing data. Sage Publishers, Thousand Oaks
40. Ingelfinger JR, Drazen JM (2004) Registry research and medical privacy. N Engl J Med 350:1452–1453
41. Grande D, Mitra N, Shah A, Wan F, Asch DA (2013) Public preferences about secondary uses of electronic health information. JAMA Intern Med 173:1798–1806
42. Safran C, Bloomrosen M, Hammond WE, Labkoff S, Markel-Fox S et al (2007) Toward a national framework for the secondary use of health data: an American medical informatics association white paper. J Am Med Inform Assoc 14:1–9
43. Platt R, Carnahan RM, Brown JS, Chrischilles E, Curtis LH et al (2012) The U.S. food and drug administration's mini-sentinel program: status and direction. Pharmacoepidemiol Drug Saf 21(Suppl 1):1–8
44. Bradley CJ, Penberthy L, Devers KJ, Holden DJ (2010) Health services research and data linkages: issues, methods, and directions for the future. Health Serv Res 45:1468–1488
45. Kahn MG, Raebel MA, Glanz JM, Riedlinger K, Steiner JF (2012) A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. Med Care 50(Suppl):S21–S29
46. Weber GM, Mandl KD, Kohane IS (2014) Finding the missing link for big biomedical data. JAMA 311:2479–2480
47. Murdoch TB, Detsky AS (2013) The inevitable application of big data to health care. JAMA 309:1351–1352
48. Birman-Deych E, Waterman AD, Yan Y, Nilasena DS, Radford MJ et al (2005) Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. Med Care 43:480–485
49. Reynolds HN, McCunn M, Borg U, Habashi N, Cottingham C et al (1998) Acute respiratory distress syndrome: estimated incidence and mortality rate in a 5 million-person population base. Crit Care 2:29–34
50. Herasevich V, Tsapenko M, Kojicic M, Ahmed A, Kashyap R et al (2011) Limiting ventilator-induced lung injury through individual electronic medical record surveillance. Crit Care Med 39:34–39

Chapter 8
Residual Confounding Lurking in Big Data: A Source of Error

John Danziger and Andrew J. Zimolzak

Take Home Messages

- Any observational study may have unidentified confounding variables that influence the effects of the primary exposure, therefore we must rely on research transparency along with thoughtful and careful examination of the limitations to have confidence in any hypotheses.
- Pathophysiology is complicated and often obfuscates the measured data with many observations being mere proxies for a physiological process and many different factors progressing to similar dysfunction.

8.1 Introduction

> Nothing is more dangerous than an idea, when you have only one…
>
> —Emile Chartier

Big Data is defined by its vastness, often with large highly granular datasets, which when combined with advanced analytical and statistical approaches, can power very convincing conclusions [1]. Herein perhaps lies the greatest challenge with using big data appropriately: understanding what is not available. In order to avoid false inferences of causality, it is critical to recognize the influences that might affect the outcome of interest, yet are not readily measurable.

Given the difficulty in performing well-designed prospective, randomized studies in clinical medicine, Big Data resources such as the Medical Information Mart for Intensive Care (MIMIC) database [2] are highly attractive. They provide a powerful resource to examine the strength of potential associations and to test

MIT Critical Data, *Secondary Analysis of Electronic Health Records*,
DOI 10.1007/978-3-319-43742-2_8

whether assumed physiological principles remain robust in clinical medicine. However, given their often observational nature, causality can not be established, and great care should be taken when using observational data to influence practice patterns. There are numerous examples [3, 4] in clinical medicine where observational data had been used to determine clinical decision making, only to eventually be disproven, and in the meantime, potentially causing harm. Although associations may be powerful, missing the unseen connections leads to false inferences. The unrecognized effect of an additional variable associated with the primary exposure that influences the outcome of interest is known as confounding.

8.2 Confounding Variables in Big Data

Confounding is often referred to as a "mixing of effects" [5] wherein the effects of the exposure on a particular outcome are associated with an additional factor, thereby distorting the true relationship. In this manner, confounding may falsely suggest an apparent association when no real association exists. Confounding is a particular threat in observational data, as is often the case with Big Data, due to the inability to randomize groups to the exposure. The process of randomization essentially mitigates the influence of unrecognized influences, because these influences should be nearly equally distributed to the groups. However, more frequently observational data is composed of patient groups that have been distinguished based on clinical factors. For example, with critical care observational data, such as MIMIC, such "non-random allocation" has occurred simply by reaching the intensive care unit (ICU). There has been some decision process by an admitting team, perhaps in the Emergency Department, that the patient is ill enough for the ICU. That decision process is likely influenced by a host of factors, some of which are identifiable, as in blood pressure and severity of illness, and others that are not, as in "the patient just looks sick" intuition of the provider.

8.2.1 The Obesity Paradox

As an example of the subtlety of this confounding influence, let's tackle the question of obesity as a predictor of mortality. In most community-based studies [6, 7], obesity is associated with poorer outcomes: obese patients have a higher risk of dying than normal weighted individuals likely mediated by an increased incidence of diabetes, hypertension, and cardiovascular disease. However, amongst patients admitted to the ICU, obesity is a strong survival benefit [8, 9], with multiple studies elucidated better outcomes amongst obese critically ill patients than normal weighted critical ill patients.

There are potentially many explanations for this paradoxical association. On one hand, it is plausible that critically ill obese patients have higher nutritional stores and are better able to withstand the prolonged state of cachexia associated with critical illness than normal weighted patients. However, let's explore some other possibilities. Since obesity is typically defined by the body mass index (BMI) upon admission to the ICU, it is possible that unrecognized influences on body weight prior to hospitalization that independently affect outcome might be the true reason for this paradoxical association. For example, fluid accumulation, as might occur with congestive heart failure, will increase body weight, but not fat mass, resulting in an inappropriately elevated BMI. This fluid accumulation, when resulting in pulmonary edema, is generally considered a marker of illness severity and a warrants a higher level of care, such as the ICU. Thus, this fluid accumulation would prompt the emergency room team to admit the patient to the ICU rather than to the general medicine ward. Now, heart failure is typically a reasonably treatable disease process. Diuretics are an effective widely used treatment, and likely can resolve the specific factor (i.e. fluid overload) that leads to ICU care. Thus, such a patient would seem obese, but might not be, and would have a reasonable chance of survival. Compare that to another such patient, who developed cachexia from metastatic cancer, and lost thirty pounds prior to presenting to the emergency room. That patient's BMI would have dropped significantly over the few weeks prior to illness, and his poor prognosis and illness might lead to an ICU admission, where his prognosis would be poor. In the latter scenario, concluding that a low BMI was associated with a poor outcome may not be strictly correct, since it is often rather the complications of the underlying cancer that lead to mortality.

8.2.2 Selection Bias

Let's explore one last possibility relating to how the obesity paradox in critical care might be confounded. Imagine two genetically identical fraternal twins with the exact same comorbidities and exposures, presenting with cellulitis, weakness, and diarrhea, both of whom will need frequent cleaning and dressing changes. The only difference is that one twin has a normal weight, whereas the other is morbidly obese. Now, the emergency room team must decide which level of care these patients require. Given the challenges of caring for morbidly obese patient (lifting a heavy leg, turning to change), it is plausible that obesity itself might influence the emergency room's choice regarding disposition. In that case, there would be a tremendous selection bias. In essence, the obese patient who would have been generally healthy enough for a general ward ends up in the ICU due to obesity alone, where the observational data begins. Not surprisingly, that patient will do better than other ICU patients, since he was healthier in the first place and was admitted simply because he was obese.

Such selection bias, which can be quite subtle, is a challenging problem in non-randomly allocated studies. Patients groups are often differentiated by their

illness severity, and thus any observational study assessing the effects of related treatments may fail to address underlying associated factors. For example, a recent observational Big Data study attempted to examine whether exposure to proton pump inhibitors (PPI) was associated with hypomagnesemia [10]. Indeed, in many thousands of examined patients, PPI users had lower admission serum magnesium concentrations. Yet, the indication for why the patients were prescribed PPIs in the first place was not known. Plausibly, patients who present with dyspepsia or other related gastrointestinal symptoms, which are major indications for PPI prescription, might have lower intake of magnesium-containing foods. Thus, the conclusion that PPI was responsible for lower magnesium concentrations would be conjecture, since lower dietary intake would be an equally reasonable explanation.

8.2.3 Uncertain Pathophysiology

In addition to selection bias, as illustrated in the obesity paradox and PPI associated hypomagnesemia examples, there is another important source of confounding, particularly in critical care studies. Given that physiology and pathophysiology are such strong determinants of outcomes in critical illness, the ability to fully account for the underlying pathophysiologic pathways is extraordinarily important, but also notoriously difficult. Consider that clinicians caring for patients, standing at the patient's bedside in direct examination of all the details, sometimes cannot explain the physiologic process. Recognizing diastolic heart failure remains challenging. Accurately characterizing organ function is not straightforward. And if the caring physician can't delineate the underlying processes, how can observational data, so removed from the patient? It can't, and this is a huge source of potential mistakes. Let's consider some examples.

In critical care, the frequent laboratory studies that are easily measured with precise reproducibility make a welcoming target for cross sectional analysis. In the literature, almost every common laboratory abnormality has been associated with a poor outcome, including abnormalities of sodium, potassium, chloride, bicarbonate, blood urea nitrogen, creatinine, glucose, hemoglobin, etc. Many of these cross sectional studies have led to management guidelines. The important question however is whether the laboratory abnormality itself leads to a poor patient outcome, or whether instead, the underlying patient pathophysiology that leads to the laboratory abnormality is the primary cause.

Take for example hyponatremia. There is extensive observational data linking hyponatremia to mortality. In response, there have been extensive treatment guidelines on how to correct hyponatremia through a combination of water restriction and sodium administration [11]. However, the mechanistic explanation for how chronic and/or mild hyponatremia might cause a poor outcome is not totally convincing. Some data might suggest that potential subtle cerebral edema might lead to imbalance and falls, but this is not a completely convincing explanation for the association of admission hyponatremia with in-hospital death.

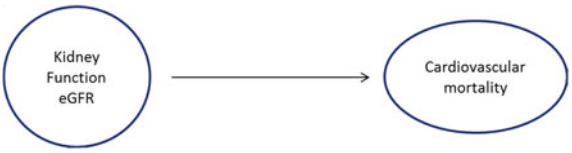

Fig. 8.1 Concept map of the association of kidney function, as determined by the glomerular filtration rate, as a determinant of cardiovascular morality

Many cross-sectional studies have not addressed the underlying reason for hyponatremia in the first place. Most often, hyponatremia is caused by sensed volume depletion, as might occur in liver disease and heart disease. Sensed volume is a concept describing the body's internal measure of intravascular volume, which directly affects the body's sodium avidity, and which under certain conditions affects its water avidity. Sensed volume is quite difficult to determine clinically, and there are no billing or diagnostic codes to describe it. Therefore, even though sensed volume is the strongest determinant of serum sodium concentrations in large population studies, it is not a capturable variable, and thus it cannot be included as a covariate in adjusted analyses. Its absence likely leads to false conclusions. As of now, despite a plethora of studies showing that hyponatremia is associated with poor outcomes, we collectively can not conclude whether it is the water excess itself, or the underlying cardiac or liver pathophysiologic abnormalities that cause the hyponatremia, that is of greater importance.

Let us consider another very important example. There have been a plethora of studies in the critical care literature linking renal function to a myriad of outcomes [12, 13]. One undisputed conclusion is that impaired renal function is associated with increased cardiovascular mortality, as illustrated in Fig. 8.1.

However, this association is really quite complex, with a number of important confounding issues that undermine this conclusion. The first issue is how accurately a serum creatinine measurement reflects the glomerular filtration rate (GFR). Calculations such as the Modification of Diet in Renal Disease (MDRD) equation were developed as epidemiologic tools to estimate GFR [14] but do not accurately define underlying renal physiology. Furthermore, even if one considers the serum creatinine as a measure of GFR, there are multiple other aspects of kidney functions beyond the GFR, including sodium and fluid balance, erythropoietin and activated vitamin D production, and tubular function, none of which are easily measurable, and thus cannot be accounted for.

However, in addition to confounding due to an inability to accurately characterize "renal function," significant residual confounding due to unaccounted pathophysiology is equally problematic. In relation to the association of renal function with cardiovascular mortality, there are many determinants of cardiac function that simultaneously and independently influence both the serum creatinine

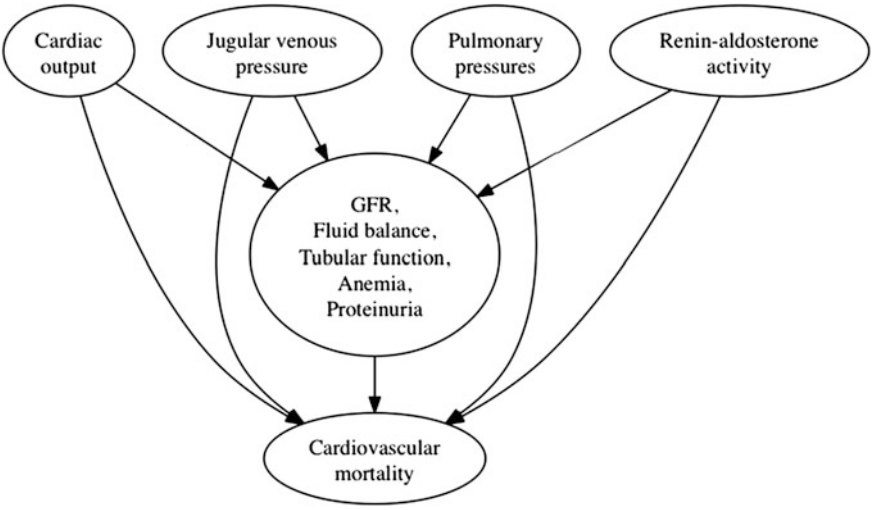

Fig. 8.2 Concept map of the association of renal function and cardiovascular mortality revealing more of the confounding influences

concentration and cardiovascular outcomes. For example, increased jugular venous pressures are a strong determinant of cardiac outcome and influence renal function through renal vein congestion. Cardiac output, pulmonary artery pressures, and activation of the renin-angiotensin-aldosterone axis also likely influence both renal function and cardiac outcomes. The concept map is likely more similar to Fig. 8.2.

Since many of these variables are rarely measured or quantified in large epidemiologic studies, significant residual confounding likely exists, and potential bias by failing to appreciate the complexity of the underlying pathophysiology is likely.

Multiple statistical techniques have been developed to account for residual confounding to non-randomization and to underlying severity of illness in critical care. Propensity scores, which attempt to better capture the factors that lead to the non-randomized allocation (i.e. the factors which influence the decision to admit to the ICU or to expose to a PPI) are used widely to minimize selection bias [15]. Adjustment using variables that attempt to capture severity of illness, such as the Simplified Acute Physiology Score (SAPS) [16], or the Sequential/ Sepsis-related Organ Failure Assessment (SOFA) score [17], or comorbidity adjustment scores, such as Charlson or Elixhauser [18, 19], remain imprecise, as does risk adjustment with area under the receiver operating characteristic curve (AUROC). Ultimately, significant confounding cannot be adjusted away by the most sophisticated statistical techniques, and thoughtful and careful examination of the limitations of any observational study must be transparent.

8.3 Conclusion

In summary, tread gently when harvesting the power of Big Data, for what is not seen is exactly what may be of most interest. Be clear about the limitations of using observational data, and suggest that most observational studies are hypothesis generating and require more well designed studies to better address the question at hand.

References

1. Bourne PE (2014) What big data means to me. J Am Med Inf Assoc. 21(2):194–194
2. Saeed M, Villarroel M, Reisner AT, Clifford G, Lehman L-W, Moody G et al (2011) Multiparameter intelligent monitoring in intensive care II: a public-access intensive care unit database. Crit Care Med 39(5):952–960
3. Patel CJ, Burford B, Ioannidis JPA (2015) Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. J Clin Epidemiol 68 (9):1046–1058
4. Tzoulaki I, Siontis KCM, Ioannidis JPA (2011) Prognostic effect size of cardiovascular biomarkers in datasets from observational studies versus randomised trials: meta-epidemiology study. BMJ 343:d6829
5. Greenland S (2005) Confounding. In: Armitage P, Colton T (eds) Encyclopedia of biostatistics, 2nd edn.
6. National Task Force on the Prevention and Treatment of Obesity (2000) Overweight, obesity, and health risk. Arch Intern Med 160(7):898–904
7. Berrington de Gonzalez A, Hartge P, Cerhan JR, Flint AJ, Hannan L, MacInnis RJ et al (2010) Body-mass index and mortality among 1.46 million white adults. N Engl J Med 363 (23):2211–2219
8. Hutagalung R, Marques J, Kobylka K, Zeidan M, Kabisch B, Brunkhorst F et al (2011) The obesity paradox in surgical intensive care unit patients. Intensive Care Med 37(11):1793–1799
9. Pickkers P, de Keizer N, Dusseljee J, Weerheijm D, van der Hoeven JG, Peek N (2013) Body mass index is associated with hospital mortality in critically ill patients: an observational cohort study. Crit Care Med 41(8):1878–1883
10. Danziger J, William JH, Scott DJ, Lee J, Lehman L-W, Mark RG et al (2013) Proton-pump inhibitor use is associated with low serum magnesium concentrations. Kidney Int 83(4): 692–699
11. Verbalis JG, Goldsmith SR, Greenberg A, Korzelius C, Schrier RW, Sterns RH et al (2013) Diagnosis, evaluation, and treatment of hyponatremia: expert panel recommendations. Am J Med 126(10 Suppl 1):S1–S42

12. Apel M, Maia VPL, Zeidan M, Schinkoethe C, Wolf G, Reinhart K et al (2013) End-stage renal disease and outcome in a surgical intensive care unit. Crit Care 17(6):R298
13. Matsushita K, van der Velde M, Astor BC, Woodward M, Levey AS et al (2010) Chronic kidney disease prognosis consortium. Association of estimated glomerular filtration rate and albuminuria with all-cause and cardiovascular mortality in general population cohorts: a collaborative meta-analysis. Lancet 375(9731):2073–2081
14. Levey AS, Bosch JP, Lewis JB, Greene T, Rogers N, Roth D (1999) A more accurate method to estimate glomerular filtration rate from serum creatinine: a new prediction equation. Modification of diet in renal disease study group. Ann Intern Med 130(6):461–470
15. Gayat E, Pirracchio R, Resche-Rigon M, Mebazaa A, Mary J-Y, Porcher R (2010) Propensity scores in intensive care and anaesthesiology literature: a systematic review. Intensive Care Med 36(12):1993–2003
16. Le Gall JR, Lemeshow S, Saulnier F (1993) A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. JAMA 270(24):2957–2963
17. Vincent JL, Moreno R, Takala J, Willatts S, De Mendonça A, Bruining H et al (1996) The SOFA (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. On behalf of the working group on sepsis-related problems of the European society of intensive care medicine. Intensive Care Med 22(7):707–710
18. Charlson ME, Pompei P, Ales KL, MacKenzie CR (1987) A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chronic Dis 40 (5):373–383
19. Elixhauser A, Steiner C, Harris DR, Coffey RM (1998) Comorbidity measures for use with administrative data. Med Care 36(1):8–27

Part II
A Cookbook: From Research Question Formulation to Validation of Findings

The first part of this textbook has given the reader a general perspective about Electronic Health Records (EHRs), their potential for medical research and use for retrospective data analyses. Part II focuses on the use of one particular EHR, the Medical Information Mart for Intensive Care (MIMIC) database, curated by the Laboratory for Computational Physiology at MIT. The readers will have an opportunity to develop their analytical skills for clinical data mining while following a complete research project, from the initial definition of a research question to the assessment of the final results' robustness. This part is designed like a cookbook, with each chapter comprising some theoretical concepts, followed by worked examples using MIMIC. Part III of this book will be dedicated to a variety of different case studies to further your understanding of more advanced analysis methods.

This part is subdivided into nine chapters that follow the common process of generating new medical evidence using clinical data mining. In Chap. 9, the reader will learn how to transform a clinical question into a pertinent research question, which includes defining an appropriate study design and select the exposure and outcome of interest. In Chap. 10, the researcher will learn how to define which patient population is most relevant for investigating the research question. Owing to the essential and often challenging aspect of analysis of EHRs, it will be described in the following four chapters elaborately. Chapters 11 and 12 deal with the essential task of data preparation and pre-processing, which is mandatory before any data can be fed into a statistical analysis tool. Chapter 11 explains how a database is structured, what type of data they can contain and how to extract the variables of interest using queries; Chap. 12 presents some common methods of data pre-processing, which usually implies cleaning, integrating, then reducing the data; Chap. 13 provides various methods for dealing with missing data; Chap. 14 discusses techniques to identify and handle outliers. In Chap. 15, common methods for exploring the data are presented, both numerical and graphical. Exploration data analysis gives the researcher some invaluable insight into the features and potential