MEDINFO 2015: eHealth-enabled Health
I.N. Sarkar et al. (Eds.)

doi:10.3233/978-1-61499-564-7-448

Big Data Clinical Research: Validity, Ethics, and Regulation

E. Andrew Balas[a], Marlo Vernon[a], Farah Magrabi [b], Lynne Thomas Gordon[c], Joanne Sexton[a]

[a] *Georgia Regents University, Augusta, GA, USA*
[b] *Australian Institute of Health Innovation, Macquarie University, NSW, Australia*
[c] *American Health Information Management Association, Chicago, IL, USA*

Abstract

Electronic Health Records (EHR) promise improvement for patient care and also offer great value for biomedical research including clinical, public health, and health services research. Unfortunately, the full potential of EHR big data research has remained largely unrealized.

The purpose of this study was to identify rate limiting factors, and develop recommendations to better balance unrestricted extramural EHR access with legitimate safeguarding of EHR data in retrospective research. By exploring primary, secondary, and tertiary sources, this review identifies external constraints and provides a comparative analysis of social influencers in retrospective EHR-based research.

Results indicate that EHRs have the advantage of reflecting the reality of patient care but also show a frequency of between 4.3-86% of incomplete and inaccurate data in various fields. The rapid spread of alternative analytics for health data challenges traditional interpretations of confidentiality protections. A confusing multiplicity of controls creates barriers to big data EHR research.

More research on the use of EHR big data is likely to improve accuracy and validity. Information governance and research approval processes should be simplified. Comprehensive regulatory policies that do not exclusively cover health care entities, are needed. Finally, new computing safeguards are needed to address public concerns, like research access only to aggregate data and not to individually identifiable information.

Keywords:

Electronic Health Records; Clinical Research; Public Health; Health Services Research.

Introduction

The vast amount of clinical data accumulating in Electronic Health Records (EHRs), or big data EHRs, represents an unprecedented opportunity to discover unrecognized risk factors, study the epidemiology of diseases, calculate life expectancy, distinguish best practices from superior outcomes and recognize opportunities for better health care.

The review of patient charts has been the cornerstone of clinical research for centuries. Historically, many landmark discoveries have originated from analyses of retrospective data. The relationship between smoking and lung cancer was first discovered by Müller in 1941 based on an analysis of patient records [1]. From these simple beginnings, recognition of smoking as a health hazard continued to evolve, ultimately becoming one the greatest public health achievements of the 20th century [2]. More contemporary examples include the relationship between thalidomide and birth defects [3], cancer epidemiology and pathophysiology, and vaccine development [4-6].

Facilitating biomedical research is one of the most important, but unrealized promises of introducing EHRs [7]. Researchers interested in conducting needs analysis, process, and outcome evaluations utilizing EHRs often run into seemingly insurmountable barriers.

Extramural access to big data represents a particular challenge (e.g., researcher of institution A trying to study big data of institution B). An illustrative case from a recent personal communication:

> *A professor of one of the world's top ranked universities wanted to get a large EHR data set for research from a leading hospital nearby. The request for an anonymized data set and the study plan was approved by the institutional IRB. In spite of regulatory compliance, ethics clearance, and stellar personal scientific track record, the professor was unable to obtain the EHR data over a period of 18 months and finally gave up.*

With the advancement of computer hardware and software, access to and analysis of clinical data is no longer a primarily technical issue. Today, the principal obstacles to EHR use in research are essentially social, ethical and regulatory.

Significant gaps exist for researchers in requesting, accessing, analyzing, and applying EHR data [8]. The Institute of Medicine reports that disappointment in the lack of EHR improvements has tempered enthusiasm for continuing research efforts [9]. Patients/consumers are increasingly participating in their own care and in care decisions, and most report being open to sharing their data for research [10, 11]. Their priorities regarding use and re-use of data will need to be taken into consideration.

Ideally, EHR-based research should meet the simultaneous but somewhat conflicting requirements of both unrestricted extramural access to the EHR by meritorious, innovative biomedical researchers with minimal administrative requirements and guaranteed zero access to EHRs by unauthorized, unnecessary or potentially harmful users of health data. Obviously, unnecessary and unjustified limitations on the access to EHR big data also represents a serious ethics violation in terms of denied care, lost public health improvement, and unrealized research discoveries benefiting patients.

The purpose of this study is to identify rate limiting social factors and develop recommendations to facilitate research on EHR big data. This study focuses on three dimensions: validity, ethical considerations, and security risks of EHR data use in biomedical research, focusing on the US context.

Methods

The project explores relevant studies and methods originating in biomedicine, health informatics, historical research, bioethics, health administration, computer science, public health and other fields. Eligibility criteria: EHR-based original research project of at least 1000 records, or thematic exploration of EHR big data management in research. Peer reviewed literature and policy documents were explored along the following hierarchy:

1. Primary sources (data): original research publications in the peer-reviewed scientific literature on EHR projects, national and international statistical databases, national surveys, and historical documents illustrating important aspects of EHR use in research.
2. Secondary sources (management): thematic scholarly explorations in the scientific literature, critiques, pertinent scholarly books, government documents, and statements of national and international organizations on social actors in big data research.
3. Tertiary sources (pointers): newspaper articles, speeches, videos, encyclopedias, web pages and other popular publications that help to identify primary and secondary sources of information.

In processing and synthesizing the information, tertiary sources were used solely to identify primary and secondary sources of information, which serve as the backbone; meanwhile, secondary sources were used to elaborate and enhance the ideas and themes of the primary sources.

PubMed and Web of Science were utilized for full text searches. Search terms included: electronic health records (EHR), electronic medical record (EMR), retrospective health research, ethics of EHR data use, confidentiality of EHR data and HIPAA. We then investigated similar themes and ethical dimensions presented in the literature. A PRISMA framework was applied to the larger systematic review resulting from this analysis [12]. The comparative and discerning analyses generated the list of factors that create external and internal influences on big data EHR use.

Results

Patient data from large EHR databases are increasingly available from multiple sources and often for a price. Efforts to evaluate the availability of these data from a practical and ethical standpoint is ongoing (see Table 1).

Validity - Issues of data integrity

Data integrity is defined as the validity, accuracy, reliability, timeliness, and consistency of the data. It remains the first question of recorded EHR data use in biomedical research. Retrospective analyses need to consider the limitations and appropriate use of data, including potential risks of inaccuracies [8]. Retrospective data are collected in variable circumstances, recorded with inconsistent data definitions, missing data, and without standardized testing. Table 2 lists the frequency of some of the reported deficiencies in EHR-based research.

On the other hand, this patient care data represents the reality of actual practice, as opposed to results of sterile research protocols. In many cases, real data can fill gaps in current evidence and provide evidence in areas where clinical trials will never be carried out. Illustrative and appropriate research uses of retrospective clinical data include exploration of risk factors, cost-effectiveness of care, selection of best practices, and the epidemiology of diseases and health conditions.

Table 1 - Selected EHR data aggregators

Source	Type of Data
CMS	Medicare and Medicaid
Blue Health Intelligence	Claims data on 210 million individuals, available longitudinally
Aetna – Accountable Care Solutions	Claims data on Aetna subscribers
Validic	Commercial firm, data aggregator for physicians and health systems
Kaiser Permanente Health Connect (Northern CA)	9.1 million patients Subscriber health claims data
Massachusetts Health Quality Partners	Intramural, work with Department of Public Health and others
OCHIN	Members of 70 health system across 19 states
IMS® Disease Analyzer	EHR are contributed by a representative panel of more than 2.500 physicians in Germany
Humana Health Care – Anvita Health	11.2 million members health data
Cerner Health Facts	Since 2000, EHR collected from 480 contributing facilities throughout USA
Vestrum	EHR data from private physicians
MS HealthVault	Personal health information of "far more" than the tens of thousands of users
Express Scripts and CVS Pharmacies	Sale of prescription information which is “match-backed” by third parties, and linked to website databases

Table 2. – Frequency of deficiencies in EHR-based research

Source	Estimate	Reference
Incompleteness	24%	[11, 13, 14]
	86%	[15]
	65%	[16]
	86%	[17]
CPOE Errors	51.4-91.5	[18]
Inaccuracies, errors	4.3 %	[19]
Inconsistencies	variable	[11, 13, 14]

EHR-based analyses provide the opportunity to evaluate individual outcomes and compare results to larger populations [20]. Reduction in costs comes with streamlined processes and improved practice efficiency. The National Patient-Centered Clinical Research Network hopes to provide unparalleled access by producing national EHR data sets available to researchers [21].

Research on rare diseases is difficult because of small sample sizes, which may be geographically distant from each other. Expanded EHR networks would enable easier access for both patient and provider, and in turn increase opportunities for research [22, 23].

The gold standard of evidence based medicine, multi-center randomized controlled clinical trials (RCT), are enormously expensive and time consuming, particularly when the sample size is very large to support high power analyses [24]. Due to

special resources and arrangements, RCTs often represent centrally-controlled practices that are not representative of current practices; and also, may never be fully replicated in general use. Furthermore, prospective randomized studies can evaluate only beneficial interventions, as opposed to allocation of patients to a harmful intervention, such as smoking, which would be unethical.

In the assessment of biological phenomena and therapeutic potential, the multicenter RCT cannot be replaced by anything less methodologically rigorous. However, due to the very high rate of negative clinical trial results, the use of retrospective data from EHR is recommended for more effective filtering prior to the evaluation of safety and efficacy of new treatments in prospective studies [25].

The use of large databases requires special statistical techniques as enormous sample sizes can overpower results (i.e., practically meaningless minutia can appear statistically significant). Additionally, appropriate scrutiny of data quality and accuracy variability is needed with reasonable logistical and statistical checking prior to analyses.

Ethics: Issues of Privacy, Beneficence, and Non-maleficence

Ethics concerns in big data biomedical research are twofold: wrongdoing in research, and ineffective administration of human subject reviews. In this study, we referred to the DuBois taxonomy of misbehavior in medical research [26]. Out of 15 fundamental kinds of wrongdoings in medical research, two stand out as particularly relevant to big data research (Table 3).

Table 3. Taxonomy of Misbehavior in Medical Research

Application Area	Examples
Violation of privacy or confidentiality	• wrongful disclosure • wrongfully obtaining information • wrongful use of health information • failure to safeguard health information
Failure of informed consent	• no need for such consent regarding archived data • advance blanket research approval forms to facilitate future use

Privacy is the basic human right of limited access by others to aspects of their person, including thoughts, identifying information, and even information in bodily tissues and fluids [27]. Patients are increasingly playing an active role in their care and are often unaware that their health information, de-identified or not, may be used for clinical or community health research in the future. This directly impacts the autonomy of the patient, even if they would have given consent.

Confidentiality is the mandate to protect information that an individual has disclosed in a relationship of trust [27]. In certain circumstances, personal information may be analyzed without consent when the benefits to society outweigh the individual's interest in keeping the information confidential [28]. Typically, big data researchers are not involved in data collection and, therefore, confidentiality and security of protected data become the foremost concerns.

Extramural research is much harder to conduct than collaborating with an intramural colleague or being an inside user of local databases. Institutional Review Boards are tasked to protect human subjects in research. Unfortunately, variable interpretations and a lack of coordination among multi-site IRBs creates a challenging health research environment [7]. This has been recognized by the NIH as an area in need of improvement. The recently published draft policy for public comment noted "there is no evidence that multiple IRB reviews enhance protections for human subjects" [29].

Regulation: Data Security and Alternative Analytics

In the US, the Health Insurance Portability and Accountability Act (HIPAA) and subsequent regulations direct covered entities (e.g., a healthcare institutions) in the protection of individually identifiable protected health information in research. Violations of the Privacy Rule can become the basis for both civil and criminal penalties, including fines and possible time in jail.

A particularly controversial part of the HIPAA provisions is use of de-identified data in research. For facilitated research access, the Privacy Rule requires de-identification (i.e., removal of 18 identifiers including names, social security numbers, telephone numbers, and others). However, de-identification creates obstacles to many research projects, particularly cause-effect and time series studies.

Table 4. Examples of alternative analytics

Source	Type of Data	Alternative Use
28% of US hospitals	Patient wealth screening	Grateful Patient Program
Target	Consumer data	Use of shopping pattern identifies marketing strategies, including based on health behaviors: pregnancy, diabetes
Garmin Connect	Athletic performance data	4 billion miles of performance information
CRM Healthgrades	Aggregate health data	Sells patient lists based on diagnosis, evaluates hospital patient data for non-compliance and QC
Carolinas HealthCare	Consumer data on 2 million people	Identify high-risk patients. Data aggregated through public records, store loyalty program transactions, and credit card purchases.
LexisNexis	Medicaid recipients and consumer data publicly available (vehicle registration, property records, etc.)	Identify Medicaid Fraud and Abuse

Without an authorization, a covered entity can use and disclose protected health information for treatment, payment and health care operations (TPO). Recently, the American Medical Informatics Association started advocating for the inclusion of observational or non-interventional data research as an appropriate operational use of protected health information [30, 31]. Ultimately, EHR big data should be available not just in the present, but also for improved future decision-making.

The rapid spread of alternative analytics for health records also challenges traditional interpretations of covered entity and confidentiality protections. For example, the recently announced Qualcomm Tricorder XPRIZE is a $10 million global competition to accurately diagnose a set of diseases independent of a healthcare professional or facility. Corporations have discovered algorithms in shopping and browsing patterns, utilizing shopping cards and credit card transactions, to identify health diagnosis or needs – without the need to access a person's medical record (Table 4). While this is skillful marketing and consumer targeting, it appears that little thought has been given to the ethics of such analysis.

Data matching with publicly available records is also becoming possible [32]. While the use of this same consumer data when legitimately coupled with a person's health record may have genuine positive outcomes, such as warning that purchased food may interact poorly with current medications or reminders to refill prescriptions, the negative consequences are not far-fetched.

The "Grateful Patient Program" model is gaining more support across the country as a way to increase donations to both non-profit and for-profit hospitals [33]. Offices of giving or communication match EHR data with income/wealth data to identify prospective donor patients and families, and then send targeted information, or even organize special visits and contacts when the patients are admitted [34]. Currently, alternative analytics has far fewer regulatory obstacles than big data biomedical research.

In addition to the ever increasing flow of health information with data stored on hard drives and cloud computing, human tissues and cells also challenge the current system of protections as they are also carriers of personally identifiable health information.

The American Health Information Management Association recommends comprehensive information governance to ensure accuracy, reliability, integrity, timeliness, accessibility, and security of data and information impacting patient care, research studies and public policy. Health care data and information must be governed to meet these imperatives.

Discussion

Big data EHRs have proved to be not only an irreplaceable source of clinical, public health and health services research but also an epicenter of confusing expectations and restrictions. The balance between access to EHRs, validity concerns, ethics safeguards and regulatory protections remains elusive.

To unlock the full potential of big data EHR research, a series of actions are needed:

1) More research on the use of EHR big data is not only a matter of new scientific knowledge but also immediate public interest as more use is likely to discover more errors and stimulate corrective efforts [35]. More EHR big data research is likely to improve accuracy and validity through improved error detections and control mechanisms.

2) The confusing multiplicity of controls creates hindrances in big data EHR research (e.g., IRB for human subject protection, institutional privacy officers for regulatory compliance, multiple IRBs for inter-institutional collaborations). Therefore, information governance and the research approval processes should be integrated and streamlined to be a one-stop research approval process.

3) Considering the rapidly expanding array of health databases outside the health care system, and alternative analytics, there is the risk that academic research and also patient care by licensed clinicians will be outpaced and major ethical and security concerns will be unaddressed. Comprehensive policies are needed for secondary use of all electronic health data, not just those in currently covered health care entities.

4) The many rate-limiting privacy and information security requirements call for new computing safeguards to address public concerns (e.g., creation of access only to aggregate data but not to individually identifiable information, tools for matching big data from diverse sources without revealing individual data).

Trust in the protection and utilization of health data is essential to continued public support. The public must see and believe that their data security is taken seriously and used responsibly. It is reasonable to expect that the larger availability of EHRs and the opportunities to match data from multiple databases should lead to an accelerated rate of valuable research discoveries.

References

[1] Müller F. Tabakmißbrauch und Lungencarcinom. Z Krebs-forsch. 1940 1940/01/01;49(1):57-85. German.

[2] Centers for Disease Control and Prevention. Ten great public health achievements--United States, 1900-1999. MMWR Morbidity and mortality weekly report. 1999;48(12):241.

[3] Lenz W, Knapp K. Thalidomide embryopathy. Archives of environmental health. 1962 Aug;5:100-5. Epub 1962/08/01. eng.

[4] Chin L, Andersen JN, Futreal PA. Cancer genomics: from discovery science to personalized medicine. Nature medicine. 2011;17(3):297-303.

[5] DeVita Jr VT, Rosenberg SA. Two hundred years of cancer research. New England Journal of Medicine. 2012;366(23):2207-14.

[6] zur Hausen H. Papillomaviruses in the causation of human cancers—a brief historical account. Virology. 2009;384(2):260-5.

[7] AHRQ. A Robust Health Data Infrastructure. Rockville, MD: Agency for Healthcare Resarch and Quality,, Director HI; 2014 April 2014. Report No.: 14-0041-EF,.

[8] Roth CP, Lim Y-W, Pevnick JM, Asch SM, McGlynn EA. The Challenge of Measuring Quality of Care From the Electronic Health Record. American Journal of Medical Quality. 2009 September 1, 2009;24(5):385-94.

[9] Institute of Medicine. Clinical Data as the Basic Staple of Health Learning: Creating and Protecting a Public Good: Workshop Summary. Washington DC: National Academy of Sciences; 2010.

[10] Alston C, Paget L, Halvorson G, Novelli B, Guest J, McCabe P, et al. Communicating with patients on health care evidence. Institute of Medicine, Washington [DC]. 2012.

[11] Walker J, Meltsner M, Delbanco T. US experience with doctors and patients sharing clinical notes. BMJ. 2015;350(g7785).

[12] Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JP, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare

interventions: explanation and elaboration. Bmj. 2009;339:b2700.

[13] Botsis T, Hartvigsen G, Chen F, Weng C. Secondary use of EHR: data quality issues and informatics opportunities. AMIA summits on translational science proceedings. 2010;2010:1.

[14] Denham CR, Classen DC, Swenson SJ, Henderson MJ, Zeltner T, Bates DW. Safe use of electronic health records and health information technology systems: trust but verify. Journal of patient safety. 2013 Dec;9(4):177-89. Epub 2013/11/22. eng.

[15] Thiru K, Hassey A, Sullivan F. Systematic review of scope and quality of electronic patient record data in primary care. Bmj. 2003 May 17;326(7398):1070.

[16] Kopcke F, Trinczek B, Majeed RW, Schreiweis B, Wenk J, Leusch T, et al. Evaluation of data completeness in the electronic health record for the purpose of patient recruitment into clinical trials: a retrospective analysis of element presence. BMC medical informatics and decision making. 2013;13:37.

[17] McGinnis KA, Skanderson M, Levin FL, Brandt C, Erdos J, Justice AC. Comparison of two VA laboratory data repositories indicates that missing data vary despite originating from the same source. Medical care. 2009 Jan;47(1):121-4.

[18] Koppel R, Metlay JP, Cohen A, Abaluck B, Localio AR, Kimmel SE, et al. Role of computerized physician order entry systems in facilitating medication errors. Jama. 2005 Mar 9;293(10):1197-203.

[19] Weiss J, Kumata J, Galar A, Turkish L, editors. Postoperative Eye Drop Documentation Omissions With EHRs After Resident Cataract Surgery: An Underrecognized Source of Error. American Academy of Opthalmology Annual Meeting; 2014 October 14, 2014; Chicago, IL.

[20] Schroeder EB, Goodrich GK, Newton KM, Schmittdiel JA, Raebel MA. Implications of different laboratory-based incident diabetic kidney disease definitions on comparative effectiveness studies. Journal of comparative effectiveness research. 2014 Jul;3(4):359-69.

[21] Terry K. Giant EHR-based network will compare treatments. The National Patient-Centered Clinical Research Institute's program may allow physicians to compare how treatments work. Medical economics. 2014 Jun 25;91(12):38-41.

[22] Bowles KH, Potashnik S, Ratcliffe SJ, Rosenberg M, Shih NW, Topaz M, et al. Conducting Research Using the Electronic Health Record Across Multi-Hospital Systems: Semantic Harmonization Implications for Administrators. The Journal of nursing administration. 2013 Jun;43(6):355-60.

[23] Safran C, Bloomrosen M, Hammond WE, Labkoff S, Markel-Fox S, Tang PC, et al. Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper. Journal of the American Medical Informatics Association : JAMIA. 2007 Jan-Feb;14(1):1-9. Epub 2006/11/02. eng.

[24] Concato J. When to randomize, or 'Evidence-based medicine needs Medicine-based evidence'. Pharmacoepidemiology and Drug Safety. 2012;21:6-12.

[25] Olsen L, McGinnis JM. Redesigning the Clinical Effectiveness Research Paradigm:: Innovation and Practice-Based Approaches: Workshop Summary: National Academies Press; 2010.

[26] DuBois JM, Kraus E, Vasher M. The development of a taxonomy of wrongdoing in medical practice and research. American journal of preventive medicine. 2012 Jan;42(1):89-98.

[27] Plaza J, Fischbach R. Current Issues in Research Ethics : Privacy and Confidentiality [Online Learning]. New York, NY: Columbia University, Columbia Center for New Media Teaching and Learning,; [cited 2014 December 15]. Available from: http://ccnmtl.columbia.edu/projects/cire/pac/introduction/index.html.

[28] Lowrance WW. Learning from experience: privacy and the secondary use of data in health research. The journal of biolaw & business. 2003;6(4):30-60.

[29] National Institutes of Health. Request for Comments on the Draft NIH Policy on the Use of a Single Institutional Review Board for Multi-Site Research 2014 [cited 2014 December 8]. Available from: http://grants.nih.gov/grants/guide/notice-files/NOT-OD-15-026.html.

[30] Middleton B, Martin R, AMIA. HIPAA Changes Needed to Permit Researcher Access to Patient Records, Letter to Rep. Fred Upton re: Request for Comments on Energy and Commerce Digital Health White Paper, 2014 [updated 10/15/2014; cited 2014 December 12]. Available from: http://www.amia.org/sites/amia.org/files/AMIA-21st-Century-Cures-Comments-to-House-EC%20-%202014-10-15.pdf.

[31] Safran C, Bloomrosen M, Hammond WE, Labkoff S, Markel-Fox S, Tang PC, et al. Toward a National Framework for the Secondary Use of Health Data: An American Medical Informatics Association White Paper. Journal of the American Medical Informatics Association. 2007 1//;14(1):1-9.

[32] Wicks P, Massagli M, Frost J, Brownstein C, Okun S, Vaughan T, et al. Sharing Health Data for Better Outcomes on PatientsLikeMe. Journal of Medical Internet Research. 2010 Apr-Jun;12(2):e19.

[33] Stewart R, Wolfe L, Flynn J, Carrese J, Wright SM. Success in Grateful Patient Philanthropy: Insights from Experienced Physicians. The American Journal of Medicine.124(12):1180-5.

[34] Elj T. Grateful patient programs: current trends, strategies and tactics. AHP journal/Association for Healthcare Philanthropy. 2006:8-9, 11, 3 passim.

[35] Miriovsky BJ, Shulman LN, Abernethy AP. Importance of Health Information Technology, Electronic Health Records, and Continuously Aggregating Data to Comparative Effectiveness Research and Learning Health Care. Journal of Clinical Oncology. 2012 December 1, 2012;30(34):4243-8.

Address for correspondence

E. Andrew Balas, MD, PhD, Georgia Regents University, 987 St. Sebastian Way, EC 3423, Augusta, GA 30912, Office: 706-721-2621, Fax: 706-721-7312, Email: andrew.balas@gru.edu.