Frontiers in Public Health Services and Systems Research

Volume 5 | Number 5 Article 3

October 2016

Electronic Health Records and Population Health Research

Joan A. Casey
University of California, Berkeley, joanacasey@berkeley.edu

Brian S. Schwartz
Bloomberg School of Public Health, bschwar1@jhu.edu

Walter F. Stewart
Sutter Health, Research, Development and Dissemination, stewarwf@sutterhealth.org

Nancy E. Adler
University of California, San Francisco, nancy.adler@ucsf.edu

Recommended Citation
Casey JA, Schwartz BS, Stewart WF, Adler NE. Electronic health records and population health research. Front Public Health Serv Sys Res 2016; 5(5):15–22. DOI: https://doi.org/10.13023/FPHSSR.0505.03.

Electronic Health Records and Population Health Research

ABSTRACT

Adoption of electronic health records (EHRs) by clinical practices and hospitals in the US has increased substantially since 2009, and offers opportunities for population health researchers to access rich structured and unstructured clinical data on large, diverse, and geographically distributed populations. However, because EHRs are intended for clinical and administrative use, the data must be curated for effective use in research. We describe EHRs, examine their use in population health research, and compare the strengths and limitations of these applications to traditional epidemiologic methods.

To date, EHR data have primarily been used to validate prior findings, to study specific diseases and population subgroups, to examine environmental and social factors and stigmatized conditions, to develop and implement predictive models, and to evaluate natural experiments. Although primary data collection may provide more reliable data and better population retention, EHR-based studies are less expensive and require less time to complete. In addition, large patient samples that can be readily identified from EHR data enable researchers to evaluate simultaneously multiple risk factors and/or outcomes while maintaining study power.

In addition to current advantages, improved capture of social, behavioral, environmental, and genetic data, and use of natural language processing, clinical biobanks, and personal sensing via smartphone should further enable EHR researchers to understand complex diseases with multifactorial etiologies. Integrating emerging technologies with clinical care could lead to innovative approaches to *precision public health*, reduce health care spending on individuals, and directly improve population health.

Keywords
electronic health records, population health, public health research

Cover Page Footnote
This Frontiers article is a shorter version of the following article: Using Electronic Health Records for Population Health Research: A Review of Methods and Applications by Joan A. Casey, Brian S. Schwartz, Walter F. Stewart, and Nancy E. Adler Click here to access the full article in the Annual Review of Public Health: http://arevie.ws/2dEakwW. No competing financial or editorial interests were reported by the authors of this paper.

BACKGROUND

Epidemiologic research design and inference are constrained by the cost and availability of data and shaped by prevailing theories of disease causation. Until the mid-20th century the lack of longitudinal, individual-level data delayed identification of the causes of diseases and reduced certainty of causal inference. Government funding in the second half of the 20th century enabled a dramatic growth in the study and long-term follow-up of population cohorts, which were foundational to our present understanding of the causes of diseases. However, research funding has declined in the 21st century. Concurrently, lower participation rates in prospective studies have increased cost and raised concerns about selective participation. Fortuitously, health systems and electronic health records (EHRs) offer a promising alternative for population health research.

In the U.S., adoption of EHRs has been motivated, in part, by the 2009 Health Information Technology for Economic and Clinical Health (HITECH) Act, which provided financial incentives to professionals and hospitals that meet EHR "meaningful use" requirements. By 2012 nearly three fourths of primary care physicians were using EHRs for clinical care encounters.[1]

Electronic health records provide a low-cost means of accessing rich longitudinal data on large populations, and are linkable to contextual data via geographic information systems (GIS). EHR data have already made considerable contributions to research. In this Frontiers article—an abbreviated version of the original article in the *Annual Review of Public Health*[2] —we describe the features of EHRs and related data, summarize their use in epidemiologic research, and contrast traditional and EHR-based studies with the goal of informing future research.

TRANSLATING CLINICAL TO EPIDEMIOLOGIC

In using EHR data for research (Figure 1, attached as an Additional File), it is important to understand how it came to be. Structured and unstructured data are documented in EHRs for clinical care and billing purposes. In contrast to conventional cohort studies with standardized protocols, EHR data collection is driven by the needs and perspectives of patients, physicians, and health systems, and reflects patient health status and how and when they seek care. A given entry (e.g., diagnostic codes, imaging and laboratory orders, and medication orders and dosing) can reflect a variety of considerations including a patient's health status, patients' provider concerns, and/or differences in physician and practice documentation.

Electronic health records capture data on an open cohort in which patients may enter or leave care at any time. As in traditional epidemiology, individuals can only contribute person-time when they are under observation and at risk for the outcome of interest. The notion of being "under observation" must be operationalized and requires consideration of documented patient contact with the health system during a specified time period. Patients in closed health systems must be members with the system's plan, whereas open health systems serve patients with and without their health plans. Most health systems in the U.S. are open or a blend of open and closed systems. Research conducted in open systems is more generalizable; the primary care population (i.e., patients who regularly see a primary care provider in the system) is often representative of the region's general population.

ELECTRONIC HEALTH RECORDS EPIDEMIOLOGY VS TRADITIONAL COHORT STUDIES

Traditional longitudinal studies offer comprehensive and precise protocols for data collection and may

more readily retain study populations than research with secondary EHR data (Table 1). However, EHR-based studies require less funding and time to complete and generally include substantially larger, more generalizable populations. Future expansion of EHR technology will also enable greater tracking of individuals for research as they seek care from multiple providers.

TABLE 1. Comparison of traditional and EHR epidemiology studies

Study feature	Traditional study	EHR study
Original purpose of data collection	Research; requires primary data collection.	Clinical care; research relies on secondary data.
Cost	More expensive, primarily government-funded.	Less expensive; data collection is funded by health care system; research can be funded with a variety of sources or may not require funding at all.
Access	Open to all researchers at a minimal cost.	Central repositories in Europe are open to all researchers; access to US health care data is constrained.
Common study design	Prospective cohort, nested case–control, cross-sectional.	Retrospective or prospective cohort, nested case–control; cross-sectional less common because longitudinal data are available.
Time frame	Further follow-up restricted by funding; must wait for health outcomes to occur for prospective studies.	Retrospective data availability restricted by date of EHR implementation; additional years of data available at low cost.
Study population	Based on recruitment; may involve incentives or suffer from healthy volunteer effects; fewer participants than EHR.	Based on patient use of a specific health system, and the system's opt-in or opt-out participation; many more participants are available; can use EHR data to prescreen patients for eligibility; various population designs are available, e.g., primary care patients, specialty cohorts.
Data on family members	Sometimes available.	Not linked owing to confidentiality but possible to reconstruct relationships with EHR data; no restrictions on future capture in EHR as part of a research study.
Follow-up	Scheduled; continues as long as funding supports, often with standardized timing between visits.	Occurs during health care encounters; in general, will have more unique encounters, with variable timing between visits.
Data collection and storage	Established protocol; generally robust approach to data collection; often with primary focus in one area of epidemiology with specialized measurements, e.g., exposure assessment, genetics; biosamples stored for future analysis.	Recorded during health care encounter with varying levels of detail based on provider practices; stored in clinical diagnoses, laboratory results, current medications and medication orders, problem list, and notes; biosamples rarely banked.
Conditions captured	Any outcomes and all severities as specified at the beginning of the study by investigators as long as ascertainment can be validly operationalized.	Only those outcomes requiring care by a physician; data missing on mild, self-resolving, or short-lived conditions.
Outcome ascertainment	Consistent outcome definitions, identified in the same way for each participant; investigators can specify in advance outcomes to study and how to measure.	Based on physician-specific clinical diagnosis, identified from a variety of locations in EHR, diagnosis enriched with other clinical information, e.g., laboratory tests, medications.
Clinical covariate ascertainment	Prespecified variables.	Entire health record, tests, and treatments are available, but not random, and perhaps confounded by disease severity and other factors.

https://uknowledge.uky.edu/frontiersinphssr/vol5/iss5/3
DOI: https://doi.org/10.13023/FPHSSR.0505.03

Nonclinical covariate ascertainment	Prespecified variables.	Limited or missing data on social and behavioral domains; GIS-based variables can substitute for some missing data.
Environmental exposures	Can capture exposures based on specific strategies in study design; more expensive; more labor-intensive; better specificity.	Can measure surrogates using GIS-based strategies with varying levels of quality and relevance; relies on temporal and spatial variability of exposures of interest.
Community conditions e.g., social, built, and food environments	Measured with GIS, or sometimes by direct observation if a small number of communities are under study.	Assigned based GIS, generally for a large number of participants in many communities spanning large geographies.
Internal validity	**Attrition:** participants must return for study visits. **Statistical regression:** participants with extreme initial values will regress toward the mean on subsequent visits. **Data collection:** standardized across sites; participation in study and barrage of health tests may affect subsequent health. **Nonparticipation bias:** systematic error related to participation, related to attrition bias where participants with certain characteristics are more likely to drop out.	**Attrition:** participants will continue to contribute as long as they remain in the health care system and seek care. **Statistical regression:** possible, but ameliorated by large sample size. **Data collection:** outcomes may be measured or recorded differently by different health care providers. **Nonparticipation bias:** systematic error related to participation, related to the population with access to, or that chooses to seek, care. **Recall bias:** reduced by using longitudinal EHR data prior to events.
External validity	**Representative sample:** participants must agree to join the study, participation rates are declining overall; past strategies to identify population-representative samples, e.g., random digit dialing, are becoming obsolete.	**Representative sample:** participants must be enrolled in the system and receiving care; documented care is more likely for more serious or troublesome conditions and less so for mild conditions; most HMORN members can identify subsets of their cared-for patients that represent the general population in their regions.

EHR, electronic health record; GIS, geographic information systems; HMORN, Health Maintenance Organization Research Network

USES OF EHRS FOR EPIDEMIOLOGIC RESEARCH

Electronic health record studies to date have drawn from de-identified health system data. In the UK, researchers can assemble study populations from central repositories of anonymized data including the Clinical Data Analysis Report System. This system gathers data from over 500 general practitioners to provide data on over 5 million patients. Increasingly, U.S. researchers are collaborating to assemble multisystem cohorts. For example, a study from four healthcare systems that make up the Chronic Hepatitis Cohort documented large underestimates of the role of hepatitis C on mortality.[3]

The strengths of EHRs have enabled researchers to:

1. confirm or challenge prior findings;
2. study multiple risk factors and/or outcomes, subpopulations, rare outcomes;
3. incorporate data on physical, built, and social environments; and
4. more effectively study stigmatized conditions.

Researchers are also capitalizing on the widespread, rapid capture of EHR data to conduct predictive modeling and studies of natural experiments.

Social and environmental epidemiology, in particular, benefits from EHR data since patients are distributed across space and time. Routinely updated addresses allow linkage of patients to location-specific data and use of GIS to study an individual's proximity to disease-related hazards. For example, EHR data on nearly 2 million patients provided estimates of associations between area-level socioeconomic deprivation and a dozen cardiovascular disease presentations.[4] Another study using EHR data established that living near high-density livestock production was associated with increased odds of antibiotic-resistant infection.[5]

DATA ACCESS AND PATIENT PRIVACY AND AUTONOMY

Typically, U.S. healthcare systems, clinics, and providers own property rights to patient data and often restrict access to system affiliates. In contrast, federally funded cohort studies require data sharing requirements and can provide free access for researchers. While U.S. providers generally bear responsibility for data misuse (e.g., breaches) and associated financial penalties, researchers typically pay for data extraction, transfer, and cleaning, a consideration in study design and budgeting. In the UK, researchers can pay to or freely access large databases containing de-identified nationally representative samples of individuals. These databases contain comprehensive EHR and other data (e.g., area deprivation).[4]

Electronic health record researchers must pay close attention to ethical use and privacy and security of protected health information. EHR's electronic format lends itself to new forms of data breach—laptop theft or inadvertent emailing of data—but also allows additional safeguards—data encryption and computer algorithms rather than manual chart reviews—to protect patient privacy and confidentiality. In many cases, patients must opt-out if they want to restrict access to their data for research applications, rather than opt-in. Some providers are adopting a dynamic consent model, where patients can monitor how their data is used and change consent over time.

IMPLICATIONS

Recent EHR research has studied less commonly investigated risk factors like intimate partner violence, sexual abuse, abandoned coalmines, and fracking. Additional technological advances, including improved capture of social/behavioral, environmental, and genetic data, natural language processing, clinical biobanks, personal sensing via smartphone, and social media—when linked to EHRs—should enable researchers to disentangle the complex, multifactorial etiologies of disease and to inform epidemiologic theory.

Electronic health record epidemiology can help bridge the divide between individual healthcare and public health. New *precision medicine* efforts might include population health data to advance clinical care. Imagine a child who presents with shortness of breath, wheezing, and cough. Diagnosis and treatment could be individualized and optimized if the clinician was aware, through real-time geocoding, linkage to secondary data sources, and messaging through the EHR, that the patient lived near a major industrial park with elevated sulfur dioxide levels in the vicinity. More generally, EHR-based research can evolve the concept behind and implementation of *precision medicine* to include occupational, environmental, social, and behavioral determinants of health, enabling what we hope will become innovative approaches to *precision public health*.

https://uknowledge.uky.edu/frontiersinphssr/vol5/iss5/3
DOI: https://doi.org/10.13023/FPHSSR.0505.03

SUMMARY BOX

What is already known about this topic? In an era of declining research funding for traditional cohort studies, EHRs offer an alternative with low-cost sources of rich longitudinal health data on large geographically, socioeconomically, and culturally diverse populations for research.

What is added by this report? We find that (1) Studies using secondary EHR data for epidemiologic research differ from traditional cohort studies in important ways and have complementary strengths and weaknesses; (2) EHR-based research has helped reevaluate prior findings; study of subgroups, rare diseases, multiple diseases and stigmatized conditions; and (3) EHR-based research aids social and environmental epidemiology, improves predictive modeling and can exploit natural experiments.

What are the implications for public health practice, policy, and research? Moving forward, improved capture of social and behavioral determinants of health, better standardization, and linkage with emerging technologies and data streams to EHR data should increase data quality and expand research opportunities to improve public health.

REFERENCES

1. Adler-Milstein J, DesRoches CM, Furukawa MF, et al. More than half of US hospitals have at least a basic EHR, but stage 2 criteria remain challenging for most. Health Aff (Millwood) 2014;33(9):1664–71. PMID: 25104826; DOI: 10.1377/hlthaff.2014.0453.

2. Casey JA, Schwartz BS, Stewart WF, Adler N. Using electronic health records for population health research: a review of methods and applications. Ann Rev Publ Health 2015;37:61–81. PMID: 26667605; DOI: 10.1146/annurev-publhealth-032315-021353.

3. Mahajan R, Xing J, Liu SJ, et al. Mortality among persons in care with hepatitis C virus infection: the Chronic Hepatitis Cohort Study (CHeCS), 2006–2010. Clin Infect Dis 2014;58(8):1055–61. PMID: 24523214; DOI: 10.1093/cid/ciu077.

4. Pujades-Rodriguez M, Timmis A, Stogiannis D, et al. Socioeconomic deprivation and the incidence of 12 cardiovascular diseases in 1.9 million women and men: implications for risk prediction and prevention. PLoS One. 2014;9(8):e104671. PMID: 25144739; PMCID: PMC4140710; DOI: 10.1371/journal.pone.0104671.

5. Casey JA, Curriero FC, Cosgrove SE, Nachman KE, Schwartz BS. High-density livestock operations, crop field application of manure, and risk of community-associated methicillin-resistant *Staphylococcus aureus* infection in Pennsylvania. JAMA Intern Med. 2013;173(21):1980–90. PMID: 24043228; PMCID: PMC4372690; DOI: 10.1001/jamainternmed.2013.10408.

FIGURE 1 (attached as an Additional File). Schematic summary depicting the process followed in epidemiologic research using EHR data. Healthcare providers collect information in real-time – inputting it into the EHR – during patient encounters with the health system. This data then becomes available to researchers who use it to conduct studies. We provide descriptions of activities during each step of the research process and notes on aspects unique to EHR research. Abbreviations: EHR, electronic health record; GIS, geographic information systems; IRB, institutional review board.

https://uknowledge.uky.edu/frontiersinphssr/vol5/iss5/3
DOI: https://doi.org/10.13023/FPHSSR.0505.03