Title

Using Electronic Health Records for Population Health Research: A Review of Methods and Applications

Journal

Annual Review of Public Health, 37(1)

ISSN

0163-7525

Authors

Casey, Joan A
Schwartz, Brian S
Stewart, Walter F
et al.

Publication Date

2016-03-18

DOI

10.1146/annurev-publhealth-032315-021353

Peer reviewed

HHS Public Access

Author manuscript

Annu Rev Public Health. Author manuscript; available in PMC 2019 September 04.

Published in final edited form as:

Annu Rev Public Health. 2016 ; 37: 61–81. doi:10.1146/annurev-publhealth-032315-021353.

Using Electronic Health Records for Population Health Research: A Review of Methods and Applications

Joan A. Casey[1], **Brian S. Schwartz**[2,3], **Walter F. Stewart**[4], **Nancy E. Adler**[5]

[1]Robert Wood Johnson Foundation Health and Society Scholars Program at the University of California, San Francisco, and the University of California, Berkeley, Berkeley, California 94720-7360

[2]Departments of Environmental Health Sciences and Epidemiology, Bloomberg School of Public Health, and the Department of Medicine, School of Medicine, Johns Hopkins University, Baltimore, Maryland 21205

[3]Center for Health Research, Geisinger Health System, Danville, Pennsylvania 17822

[4]Research, Development and Dissemination, Sutter Health, Walnut Creek, California 94596

[5]Center for Health and Community and the Department of Psychiatry, University of California, San Francisco, California 94118

Abstract

The use and functionality of electronic health records (EHRs) have increased rapidly in the past decade. Although the primary purpose of EHRs is clinical, researchers have used them to conduct epidemiologic investigations, ranging from cross-sectional studies within a given hospital to longitudinal studies on geographically distributed patients. Herein, we describe EHRs, examine their use in population health research, and compare them with traditional epidemiologic methods. We describe diverse research applications that benefit from the large sample sizes and generalizable patient populations afforded by EHRs. These have included reevaluation of prior findings, a range of diseases and subgroups, environmental and social epidemiology, stigmatized conditions, predictive modeling, and evaluation of natural experiments. Although studies using primary data collection methods may have more reliable data and better population retention, EHR-based studies are less expensive and require less time to complete. Future EHR epidemiology with enhanced collection of social/behavior measures, linkage with vital records, and integration of emerging technologies such as personal sensing could improve clinical care and population health.

Keywords

electronic health records; EHR; environmental epidemiology; social epidemiology; geographic information systems; health determinants

joanacasey@berkeley.edu.

INTRODUCTION

Epidemiologic research design and inference are shaped by prevailing theories, by available measures of risk factors, and by the cost of obtaining relevant data. Prior to the 1950s, researchers commonly used vital statistics to conduct cross-sectional and time series studies of noninfectious disease. The lack of longitudinal data limited causal inference. In the second half of the twentieth century, funding allowed researchers to develop cohorts of individuals who were followed over time. However, in the twenty-first century, declining research support and participation rates (42) complicate the conduct of traditional costly and time-consuming prospective studies.

The recent rise in the use of electronic health records (EHRs) offers a timely alternative. These databases provide a low-cost means of accessing rich longitudinal data on large populations for epidemiologic research. Not simply a digital version of a paper record (127), EHRs can be linked to contextual data using geographic information systems (GIS) and combined with self-reported data to address questions about complex networks of causation. Such work has the potential to evolve epidemiologic theory in the twenty-first century (69, 86).

In this review we describe the nature of EHRs and how they have been used in epidemiologic research. Since its recent inception, EHR data have made considerable contributions to a broad population health scholarship, from infectious disease research to social epidemiology. We summarize this literature and then contrast traditional and EHR-based studies to highlight specific strengths and weaknesses of each with the goal of informing future research.

EHR ADOPTION AND FUNCTIONS

EHRs were originally developed for billing purposes. However, their purview has expanded, motivated by meaningful use requirements expressed in the Health Information Technology for Economic and Clinical Health (HITECH) Act, part of the 2009 American Recovery and Reinvestment Act. Financial incentives to professionals and hospitals for EHR use are tied to existing and emerging requirements. Requirements include standard capture of vital statistics, an up-to-date problem list, and others relevant to patient engagement and data sharing (34, 127). The implementation of meaningful use will likely accelerate capture and standardization of data and benefit epidemiologic research (14).

In 2012, 69% of primary care physicians in the United States reported using EHRs, an increase of 32% from 2010 (3). Parallel changes have unfolded in other industrialized countries, and current usage ranges from lower levels in China and South Korea (115,134) to nearly universal adoption in Australia, New Zealand, and northern Europe (110). Although the focus of this article is primarily on the use of EHR data for research in the United States, we draw on relevant research elsewhere.

EHR DATA AND DEFINING EPIDEMIOLOGIC PARAMETERS

Data included in EHRs are intended for clinical and administrative use. As discussed below, these data can be used effectively for research purposes, but doing so requires some caution and creativity.

EHR Data Collection and Content

Unlike standardized primary data collection in epidemiologic research, EHR data are collected for the purposes of the clinical encounter. Rather than being driven by research needs, the data collected are directly influenced by patient health status, by how and when they seek care, and by variation in physician care practices and documentation. Accordingly, the patient and physician, not the researcher, stipulate the amount of time a patient is under observation (person-time), which impacts calculation of prevalence, incidence, and risk ratios.

EHRs used by different health systems vary in the number of domains (e.g., vital signs, laboratory data) of health care data that they collect. Over time, systems tend to add functionality to their EHR and expand the number of domains collected (Table 1). Longitudinal research is made possible by using the dates associated with specific EHR entries. Doing so allows researchers to study not only disease onset, but also disease severity and progression.

Diagnostic codes warrant special consideration in EHR research. Physicians use codes to depict a patient's condition, to document indications for orders (i.e., medications, laboratory tests, imaging), and to justify the levels of service and billing. The location of a code in the EHR can also provide useful information. Image and laboratory order codes indicate what the physician suspects about the patient's condition that requires validation or what the physician knows about the patient (e.g., hypercholesterolemia) and is monitoring (e.g., low-density lipoprotein). Medication orders/dosages or scheduling of return visits may represent the degree of physician concern, reflected in the explicit action required to manage the health condition. Even though diagnostic codes provide critical information on an individual's health status, providers may not use them consistently, and the meaning of any given code may vary among providers and across time.

Study Design and Study Population Assembly

EHR-based studies involve predominately case series, nested case-control studies, and prospective and retrospective cohorts. Researchers can use EHR data to rapidly identify cases and assess eligibility for individual or frequency matching in nested case-control studies (130). EHRs capture data on an open cohort in which patients may enter or leave care at any time. A patient can contribute person-time only if they are under observation and are at risk for the outcome of interest. Although the notion "under observation" will vary, at a minimum it requires that a patient be documented as having an encounter with a qualified provider (e.g., primary care physician). Researchers may find it difficult to interpret gaps in care in the EHR. When a patient lacks data, one cannot distinguish between patients who have left care, who have been well and have not sought care, or who have missed routine

visits for other reasons. This ambiguity in whether patients are under observation is relevant to the person-time documentation required for estimating incidence rates. If patients enter care before an EHR has been implemented in a given system, some domains or events may not be captured and available for study (i.e., left-censored). Conversely, if they exit care, EHR data will lack information on events occurring after that time (i.e., right-censored).

Constructing Epidemiologic Variables

Outcomes and exposures.—EHRs can be used to define disease onset and outcomes and to determine case and control status on a selected outcome, exposure measures, and covariates. For numerous reasons, the single appearance of a diagnostic code does not necessarily indicate that a patient has a disease. For example, in identifying chronic rhinosinusitis, Hsu and colleagues found that the positive predictive value (PPV) for the ICD-9 (International Classification of Diseases) code 471.x for nasal polyps was 85%, whereas 473.x for chronic sinusitis had a PPV of only 34% (54). With additional information—evaluation by an otorhinolaryngologist, for example—the PPV rose to 91%. The accuracy of disease definition is often improved by using ICD-9 codes and other information over time and is often better in relation to more severe disease (e.g., myocardial infarction). Aspects of the EHR may enhance data validity. For example, alerts, commonly used in clinical decision support (122), can also be used to notify clinicians of input errors to support real-time data correction.

Clinical text is also captured in the EHR, often in a notes section. It includes discharge summaries, treatment plans, and progress notes, which can contain information about patients that is useful for research purposes. However, this information may be inconsistently recorded. For example, Wasserman et al. (129) searched text notes for 465 children and found fever reported in 278 different ways (e.g., “fever,” “pyrexia,” “elevated temp”). One approach to deal with nuanced clinical text is to use open source natural language processing tools. These can extract text relevant to defining disease stage, severity, and progression or symptoms (6, 124), which may not be well captured by diagnostic codes. For instance, Andersen et al. (6) used natural language processing to extract suicidal ideation from clinical notes on >3 million Americans from 1700 primary care physicians and found that only 3% of patients with recorded suicidal ideation had a corresponding ICD-9 code.

Disease etiology.—Whereas disease status is often well documented in EHRs, disease etiology, including fundamental causes of disease (70) (e.g., social, behavioral, environmental factors), is often not well documented. Some data are not retained, including, for example, residential addresses over time (only the current address is used for billing). Researchers have used health insurance status (e.g., commercial versus Medicaid) as a proxy for individual socioeconomic status (SES) (13, 18, 36, 44, 67, 83) and have assigned neighborhood SES on the basis of the median income or an index of deprivation in patients’ communities (13,18,29,35,38). Although data on physical activity and other important behaviors and social risks are not routinely captured (2, 16), the Institute of Medicine has recommended that these and other domains be integrated into routine EHR data collection, including four existing (i.e., race/ethnicity, current address, alcohol use, and tobacco use) and eight new domains (e.g., stress, social isolation, physical activity) (27).

DIVERSE USES OF EHR DATA FOR EPIDEMIOLOGIC RESEARCH

Researchers have applied extract, transform, and load algorithms to EHRs to assemble study populations from a variety of settings (Table 2). The most successful EHR research to date has used deidentified databases in UK and US health care systems whose patient populations receive most or all of their care within the system. Researchers initially used EHRs for comparative effectiveness and health services research, pharmacoepidemiology and genetics epidemiology [e.g., the Electronic Medical Records and Genomics (eMERGE) Network], and disease surveillance. These efforts have been summarized elsewhere (12, 49, 66, 81, 95, 120) and are not covered in this review.

Assembling Research Cohorts from EHR Data

Researchers can use EHRs to form standard cohorts and to assemble groups of patients with specific diseases. Kaiser Permanente in the United States has several EHR-based cohorts (7, 30, 31, 89), including the Diabetes Study of Northern California (DISTANCE) study (85). DISTANCE involves 20,000 patients with diabetes and has addressed wide-ranging issues, including diabetes outcomes among Asians and Pacific Islanders (59), the impact of neighborhood deprivation on cardiometabolic health indicators (68), and the relationship of SES to risk of hypoglycemia (9).

Researchers from two or more health systems are increasingly collaborating and assembling multisystem cohorts; the HMO Research Network has been a leader in this type of research since 1994 (112). Three other US examples are the Consortium on Safe Labor (28, 78, 83, 106), which uses EHR delivery and birth data from 19 hospitals; the Clinical Assessment, Reporting, and Tracking system in Veterans Administration (VA) hospitals (128); and the Chronic Hepatitis Cohort Study, which combined data from four health care systems on more than 1.6 million adults to identify a cohort of hepatitis B and hepatitis C patients (87). With Chronic Hepatitis Cohort data, Mahajan et al. (77) found that only 30% of hepatitis C-positive patients who died with documented liver disease had hepatitis C on their death certificate, uncovering huge underestimates of the role of hepatitis C on mortality in the United States (77).

Researchers have also assembled study populations from central repositories of anonymized data including the Clinical Data Analysis Report System in Hong Kong (22) and the Clinical Practice Research Datalink (CPRD) (4, 50, 102), the Health Improvement Network (THIN) (35, 93), and QResearch in the United Kingdom (125). The CPRD, which gathers data from more than 500 UK general practitioners, has data on more than 5 million active pediatric and adult patients. Repositories provide researchers with normalized, longitudinal data, enabling greater opportunities for research, as evidenced by the >1,000 peer-reviewed published papers using CPRD data.

Parallel rise in available EHR data and concern about obesity spurred some of the first population health research with EHRs (13, 52, 57, 64, 72, 96, 107, 111, 131). Weight and height used to calculate body mass index (BMI) is recorded during many clinical encounters. Additionally, BMI data in EHRs have relatively low error rates; notably, errors in child BMI are generally < 1% (119). Not surprisingly, few studies have focused on cancer (46, 62, 114,

121), given the availability of cancer registries worldwide. In the following sections, we provide specific examples of EHR research and their major areas of contribution to date.

Reevaluating Prior Findings

Researchers have employed large EHR data sets to reevaluate conclusions drawn from smaller studies. For example, many small studies reported positive or inconsistent associations between midlife BMI and later-life dementia. Qizilbash et al. (102) used longitudinal CPRD data on 2 million people and found that higher midlife BMI was associated with a decreased risk for dementia, which suggested that obesity could be protective for dementia or that weight loss may result from early dementia, both important areas for future research. In another study, Hibbard and colleagues (28) used Consortium on Safe Labor EHR data (N = 233,844 deliveries) to control for factors missing from prior birth outcome studies (e.g., maternal medical conditions). They found that late preterm birth compared with birth at term was associated with increased respiratory morbidity, but the association was smaller than reported in prior studies (28). Similarly, studies with small samples from fertility clinics had previously linked celiac disease to infertility. Dhalwani et al. (35) calculated incidence rates of infertility in >2 million UK women and found no evidence of such a connection.

Multiple Risks, Subgroup Differences, and Rare Outcomes

The large patient samples from EHRs enable researchers to evaluate multiple risk factors and/or outcomes simultaneously, to test associations in subpopulations, and to study rare outcomes.

For example, researchers in the Netherlands evaluated access to green space in relation to disease diagnoses with >10% prevalence. Using 12 months of EHR data on more than 300,000 patients from 195 general practitioners (74), this team found that green space was protective in 15 of 24 disease clusters, including musculoskeletal and neurological clusters with the strongest associations for anxiety and depression, especially among children and individuals of low SES.

In a subgroup analysis, Scherrer et al. (109) used seven years of VA EHR data and found that major depressive disorder and type 2 diabetes alone each increased the risk of myocardial infarction by about 30%. However, evidence of having both health problems increased risk by more than 80%, with important clinical implications. Rapsomaniki et al. (103) studied > 1 million UK adults using CALIBER (CArdiovascular research using LInked Bespoke studies and Electronic Health Record) data to evaluate age category–specific risk of 12 acute and chronic cardiovascular diseases (CVDs) related to systolic and diastolic blood pressure. The study was able to provide an adequate sample size to evaluate important subgroups (e.g., those with low blood pressure or who were on blood pressure-lowering drugs) and found varying associations across subgroups between systolic and diastolic blood pressure and CVD end points (e.g., between systolic, but not diastolic, blood pressure and stable angina). Because the health data covered the majority of the UK population, these findings had excellent external validity (103) and were in contrast to prior studies that evaluated fewer CVDs (132) across narrower age and blood pressure ranges.

Rare disease research can also benefit from EHR data, which help alleviate methodological constraints. Thomas et al. (125) used four UK EHR databases to study chickenpox as a risk factor for stroke, a rare event in children. Using patients as their own controls, they observed a fourfold increase in risk of pediatric stroke in the first 0–6 months after chickenpox. This study identified avenues for future research on links between infections and vascular injury and their role in stroke.

Environmental and Social Epidemiology

EHR data sets have allowed environmental and social epidemiologists to leverage data on patients distributed across a wide range of physical, built, and social environments. Because patient addresses are routinely checked and updated at each encounter for billing and communication purposes, researchers can readily link geocoded addresses to location-specific data and use GIS to study an individual's proximity to hazards related to disease. This process can be used to study negative health impacts from both direct exposure, e.g., air pollution and contextual exposure, e.g., residential zip code poverty rates.

Physical environment.—EHR studies have evaluated exposures to risks and resources in the physical environment (e.g., air pollution, green space) and health outcomes (e.g., hypertension, diabetes, migraines) (72, 74, 78, 80, 88, 106). For example, in a novel study of exposure to acute air pollution, Mannisto et al. (78) used EHR data on 151,276 deliveries from 19 hospitals across the United States from the Consortium on Safe Labor and found elevated odds of high blood pressure at delivery in women exposed to higher levels of 4 air pollutants in the 4 hours preceding hospitalization. Casey et al. (19) obtained data from the Geisinger EHR on more than 10,000 births to evaluate objectively recorded health risks associated with unconventional natural gas development. They identified significantly increased odds of preterm birth in women exposed to more unconventional natural gas development activity during their pregnancies.

Built environment.—Studies of the built environment have focused on land use (e.g., street connectivity, population density, agriculture), food (e.g., density of fast-food restaurants, food deserts), and physical activity environments (e.g., access parks, diversity of physical activity establishments) (18, 19, 38, 71, 72, 107, 111, 118). Duncan et al. (38) found greater increases over time in BMI *z*-scores for 50,000 children and adolescents who were residing in less walkable neighborhoods versus those in more walkable neighborhoods, after controlling for age, sex, race/ethnicity, and neighborhood median household income. Casey et al. (18) reported that living near high-density, industrial livestock production or the crop fields to which manure was applied increased the risk for methicillin-resistant *Staphylococcus aureus*; this study provided the first evidence of agricultural risk for antibiotic-resistant infections in a general population sample.

Social environment.—Social epidemiology's rich history of studying the influence of neighbor-hoods and communities on health (75,123) has expanded through the use of EHR data. EHR-based studies have generally used an administratively defined surrogate for neighborhoods, such as census tracts, and then used census data to link community-level exposures to EHR data through geocoded patient addresses (21, 43, 92, 100, 107, 111, 126).

For example, Nau et al. (92) used data on Geisinger Clinic children and adolescents (N = 163,473) and found that community socioeconomic deprivation was associated with steeper BMI trajectories. Pujades et al. (100) used CALIBER data on nearly 2 million patients and confirmed prior associations between socioeconomic deprivation and myocardial infarction and CVD mortality, with new evidence of heterogeneity by age groups, CVD types, and sex. Most EHR social epidemiology has evaluated associations of community SES (e.g., median household income or education level) and health, but some have studied other exposures, including intimate partner violence (98, 104), sexual abuse (24), and community violence (116).

Predictive Modeling

The convergence of machine learning tools and big data methods is motivating development of predictive models that can readily use diverse, high-volume EHR data to guide decision making for individual patients (56). Researchers have used EHR data to assign Framingham Heart scores (45) and QScores (http://www.qresearch.org/), which predict the risk of outcomes such as cancer, diabetes, and stroke. Better cardiac prediction has been achieved by adding variables available in the EHR that were not included in the traditional prediction models (32, 63, 93, 133). Osborn et al. (93) developed an algorithm to predict CVD in patients with severe mental illness and found that the Framingham model overpredicted events in mentally ill men by 32%. Other algorithms have been developed to predict treatment failure among HIV-positive patients to better target interventions (101, 105). Most algorithms have utilized only EHR data; the addition of place-based predictors of patient health (i.e., social and environmental variables) could improve performance.

Research on Stigmatized Conditions

EHRs can be used to study stigmatized conditions, such as mental health outcomes or HIV, where patient recruitment and follow-up can sometimes pose challenges. For example, McCoy et al. (82) used EHR data to classify psychiatric inpatients using Research Domain Criteria Project criteria. Loadings on cognitive, arousal, negative valence, and social domains predicted the length of hospital stay and readmission, whereas ICD-9 codes did not, exemplifying the promise that information extracted from EHRs can improve diagnosis and predict health outcomes (55). In Rwanda, Betancourt et al. (10) used EHR data and information from community health workers to compare mental health outcomes in HIV-positive children, children living with HIV-positive parents, and HIV-unaffected children. They demonstrated that children living with HIV-positive parents require the same mental health services as do children who are themselves infected (10).

Natural Experiments

The widespread use of EHRs enables the rapid collection of data when natural experiments occur. Johnson & Beal (58) exploited the isolation of Altru Health System in North Dakota, where a comprehensive smoke-free ordinance went into effect. Using EHR data from the only acute care center in a 70-mile radius, the investigators found a significant decrease in the heart attack rate after the ban. In the Netherlands, Dirkzwager et al. (36) assessed health problems one year prior and two years after a fireworks disaster using data from family medical practices. In addition to finding poorer health overall postdisaster, they identified

groups in need of priority postdisaster care: those with preexisting mental illness and those forced to relocate.

EPIDEMIOLOGIC PRINCIPLES: COMPARING TRADITIONAL AND EHR STUDIES

Compared with studies using primary data collection, EHR-based studies are considerably less expensive, require less time to complete, and involve substantially larger and more generalizable populations with fewer limitations to follow-up (Table 3). However, traditional studies offer more comprehensive and precise protocols for data collection and better study population retention. Below, we consider the comparative strengths and weaknesses of the two approaches.

Study Population Selection

Investigators who directly recruit study participants encounter several limitations to obtaining truly representative samples. One limitation is that the interest of individuals and groups with salient characteristics in participating in research varies (42). Women, married individuals, those of higher SES, and those to whom the research topic is most relevant are more likely to enroll, whereas those with risk behaviors such as smoking, drinking, and drug use are less likely to do so (42). In combination with the declining participation rates in recent decades, this occurrence raises concerns about selection bias and external validity of traditional population health studies.

EHR studies also experience challenges with representativeness and missing data. On the one hand, that EHR studies can include in the analysis every person who receives care reduces selection bias. However, patients enrolled in a given health care system may differ in meaningful ways from the general population. To test representativeness, researchers can compare the age, sex, race/ethnicity, and other relevant characteristics of their patient population to census data in the matching region. Missing data may introduce bias into all studies. Since EHR data collection is less standardized, missing data may be especially problematic. For example, Qizilbash et al. (102) began with CPRD data on 6.1 million individuals, but were forced to exclude 48% of eligible participants because of missing BMI data.

Issues of generalizability pose less of a problem for regional environmental or social epidemiology than for general disease surveillance efforts. Increasing standardization and interoperability of EHR records should allow for pooling of data from multiple systems, thereby increasing representativeness and strengthening external validity. In addition, efforts to implement the use of structured templates in EHR may improve data completeness (20).

Study Population Attrition

Both traditional and EHR cohorts suffer from attrition, which can be problematic for longitudinal research. Traditional cohort studies experience attrition if people withdraw from the study or are lost owing to a move, although actively managed studies can reduce loss. For example, participation rates in the four follow-up exams in the Multi-Ethnic Study of

Atherosclerosis (MESA) over an 11-year period following the initial assessment were 91%, 87%, 84%, and 68% (84); this pattern is representative of retention in large cohort studies.

Attrition in EHR studies arises primarily because of patient disenrollment. Study subjects may leave care for a variety of reasons. Some instances of disenrollment may be due to patient or disease characteristics, whereas others may reflect modifications to insurance coverage due to changes in employment, legislation, or regulations. If researchers use sequential cohorts, there may be changes in composition. If individual patients are followed, those who disenroll will be lost to follow-up. For example, a study of 20–39.9-year-olds enrolled in Kaiser Permanente from 2007 to 2009 found that 68% of active members from 2007 were retained at the end of 2009 (65). Retention increased with age; 76% of those 35–39.9 years old remained.

Recall Bias

Disease diagnosis may skew a patient's recall of prior events. This lack of reliable information may be especially problematic in controversial areas such as childhood vaccination and autism onset. Because EHRs can specify timing and risk, they may reduce recall bias and other types of information bias. For example, two studies used longitudinal data from the CPRD to assess the measles, mumps, and rubella vaccine as a risk factor for future autism diagnosis (61, 117), which assured no recall bias in vaccination reports. An additional advantage to using EHR data for social and environmental research is a reduction in possible diagnostic and reporting biases. Outcome data are obtained from reports of physicians and patients who are unlikely to be aware of the exposure of interest. EHR data can similarly reduce Hawthorne effects and social desirability bias.

Time, Cost, and Size

Because they use existing data, EHR studies require less time and money to conduct and can involve more participants than studies that require primary data collection. We contrast these factors in three traditional studies of CVD risk factors compared with a cohort drawn from a health system's EHR.

The traditional studies are the Framingham Heart Study (FHS) (n = 5,209 adults aged 30–59 years from Framingham, Massachusetts, enrolled in 1948 and followed up since); the Atherosclerosis Risk in Communities (ARIC) (n = 16,000 adults 45–64 years in 4 communities, 4 follow-up visits, one every 3 years); and the Multi-Ethnic Study of Atherosclerosis (MESA) (n = 6,800 adults aged 45–84 years in 6 communities, 5 follow-up visits to date over 12-year period) (91).

The EHR data were constituted from a retrospective data pull from the Geisinger EHR for the years 2006–2013. With institutional review board approval, we selected 138,514 patients aged ≥45 years at baseline. The data contained 12/13 domains (no imaging files) highlighted in Table 1.

Cost.—Compared with direct recruitment and follow-up in traditional studies, obtaining data from the EHR is much less expensive. For example, as of 2012, the FHS had received $140 million, ARIC $189 million, and MESA $121 million in funding from the National

Heart, Lung, and Blood Institute (NHLBI) (91). In contrast, using an extract, transform, and load algorithm on the Geisinger EHR data cost about $50,000. The approximate average cost per participant in the EHR sample was $0.11 for 8 years of data compared with $17,750 and $11,800 per participant for 12 years of MESA and ARIC data, respectively, and $2,732 per participant for 67 years of FHS data (unadjusted for inflation).

Time.—The strongest cohort studies are prospective and wait for outcomes to develop; the FHS began in 1948 but did not have its first important research finding until 1960 (41). EHR-based retrospective cohort studies can produce results within a year or two. Since its origin in 2011, researchers have used the CALIBER database, which combines the UK's nationwide CPRD data with CVD procedure registries (33) to evaluate risk factors for 12 different CVDs in 1.3–1.9 million patients (99, 100, 103, 113).

Traditional prospective studies must recontact participants and may face difficulties with maintaining study samples, which can impact the length and/or depth of follow-up. For example, recent budget cuts have forced the FHS to eliminate in-person exams. EHR data could provide a solution. The NHLBI and others have identified health information technology as a way to rework large cohort studies to decrease costs and increase enrollment (69, 79).

As noted earlier, it may be difficult to follow patients for long periods of time within a specific EHR. However, there is increasing emphasis on greater linkage and record sharing across systems. In addition to allowing prior clinical information to follow patients to wherever they seek care, these advances will also enable greater tracking of individuals for research.

Variables Available for Analysis

Traditional epidemiologic studies obtain data designed to address a specific research question. In an EHR, the same information may not be universally available or collected in a standard way. For instance, the MESA study used a standard intake form to assess smoking status, whereas the capture of smoking status in EHRs can be sporadic and varied in quality and detail. Relevant data on smoking may appear in different parts of the record. The social history section may contain time-varying data on pack-years, and encounters may provide diagnostic codes (e.g., ICD-9 305.1, tobacco use disorder), cessation counseling referral, and medications orders (e.g., varenicline) relevant to documenting smoking status (18).

Traditional studies can collect information that is not routinely included in EHRs. For example, the MESA study gathered sleep, psychosocial, employment, physical activity, and dietary data and biospecimens for biobanking at many of the follow-up visits, and participants completed computed tomography scans, magnetic resonance imaging, and carotid intima-media thickness tests (11). Such data are available on only a subset of patients in the EHR when tests are completed for diagnostic or treatment purposes.

In addition, EHR research can study only conditions that are routinely captured. EHR data collection is particularly weak for mild or remitting conditions (e.g., mild asthma, early diabetes, and sprains/strains) for which many patients do not seek care.

Finally, primary data collection often includes data on family members. Owing to confidentiality concerns, family members' EHRs are not directly linked. However, matching algorithms based on names, dates, birth weight, or other information may be used to link, for example, mothers to neonates (76).

Access to Data and Issues of Privacy and Security

Regulations requiring data sharing for federally funded research studies give researchers free access to the data from large cohorts such as the FSA, ARIC, and MESA. Whereas any researcher can pay to use the United Kingdom's CPRD, THIN, and QResearch databases, access to US-based EHR data is more difficult. In the United States, health care systems, not patients, typically own the property rights to EHR data. Systems then must decide who can access the data, which they generally limit to system affiliates. If access is granted to researchers, they usually bear the costs of data extraction and transfer and must develop data use agreements. However, given that sample size does not contribute to costs associated with using EHR data, large EHR studies can still remain inexpensive compared with similar studies using primary data collection.

When access is granted to use EHRs, special attention is needed to assure ethical use of the information. Population health research that relies on protected health information (PHI) may risk violating individuals' privacy rights (90). EHRs can both accentuate and ameliorate such risks. One issue is the nature of consent. Rather than obtaining active consent, many health systems require patients to opt out if they do not want their EHR data used for research purposes. As a result, some patients may unknowingly contribute their personal health data to research. Although this method has generally not been a problem, it can be if the research aims conflict with patients' moral or other values (108). Opt-in participation models protect privacy but require more time and funding and can lead to selection bias (51). Recently, a dynamic consent model has been proposed where patients can monitor how their data is used and change their consent over time (60). A second issue is maintenance of patient-provider confidentiality. This factor is especially relevant when researchers merge data from multiple health systems, and it requires that reliable deidentification and security methods are in place. Third, because providers record PHI in many different formats, it may persist in free text despite efforts at deidentification. Finally, although not unique to EHRs, electronic data storage may lend itself to new forms of data breach, including laptop loss or inadvertent emailing of data.

At the same time, digitally stored data also offer safeguards. EHR data can be encrypted and require role-based access and authentication. Additionally, extracting EHR data with computer algorithms results in less researcher exposure to PHI and fewer opportunities for privacy breaches than do manual chart reviews or traditional data collection (17). While many patients support the use of EHRs in research (73), it is incumbent on researchers, clinicians, and policy makers to balance the benefits of having representative and informative samples with protecting individual privacy and confidentiality.

CONCLUSIONS AND FUTURE DIRECTIONS

EHR-based epidemiology has already produced a large body of important research and will only grow as EHR use expands, costs fall, linkage to vital or other records increases, and accessibility improves. Furthermore, opportunities will increase as new technologies allow patient data capture without input from medical professionals. For example, patients can self-report data on a patient portal that links to their EHR. Portal use has been associated with better care adherence, improved patient–provider relationships, and improved patient autonomy and self-efficacy (94). Empowered patients should not only make more informed health decisions, but may more readily participate in research to the benefit of both clinical care and population health (34).

Other advances in combination with EHRs can enable researchers to understand complex diseases with multifactorial etiologies. These could include improved capture of social/behavioral (2), environmental, and genetic data (56); phenotyping (37); clinical biobanks; improved natural language processing; personal sensing via smartphone; and social media. Such advances may enable researchers to incorporate variables such as racial segregation, exercise, and social networks into their studies and extend and modify epidemiologic theory.

In addition to informing population health research, EHR epidemiology and social-behavioral studies can advance clinical care and new precision medicine efforts (26). Imagine a child who presents with shortness of breath, wheezing, and cough. Diagnosis and treatment could be individualized and optimized if the clinician were aware, through real-time geocoding, linkage to secondary data sources, and messaging through the EHR, that the patient lived near a major industrial park [which had been shown, via EHR research, to be linked to higher risks of asthma (5)] and that sulfur dioxide levels in the vicinity are elevated. More generally, EHR research can help to evolve the concept behind and implementation of precision medicine to include important predictors of individual variability that lie outside the body and include occupational, environmental, and social determinants of health (1). EHR research can move such work forward in what we hope will become innovative approaches to precision public health.

As population health research with EHR makes use of new technologies, the work will raise ethical and practical issues. Privacy agreements and security must keep pace with research to achieve the full promise of such research. Although EHRs are designed and used for clinical care, their research utility goes beyond the hospital walls. Stage 3 of HITECH recognizes this potential; a proposed objective requires meaningful use participants to share health data with public health agencies or clinical registries (34). EHR epidemiology can help bridge the divide between individual health and public health and reduce health care spending on individuals while leading to direct improvements in population health.

ACKNOWLEDGMENTS

Funding was provided by the Robert Wood Johnson Health and Society Scholars program. The authors thank Lara Cushing for her suggestions on the manuscript.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

LITERATURE CITED

1. Adler NE, Prather AA. 2015 Risk for Type 2 diabetes mellitus: person, place, and precision prevention. JAMA Intern. Med. 175:1321–22 [PubMed: 26120971]
2. Adler NE, Stead WW. 2015 Patients in context—EHR capture of social and behavioral determinants of health. N. Engl. J. Med. 372:698–701 [PubMed: 25693009]
3. Adler-Milstein J, DesRoches CM, Furukawa MF, Worzala C, Charles D, et al. 2014 More than half of US hospitals have at least a basic EHR, but stage 2 criteria remain challenging for most. Health Aff. 33:1664–71
4. Alonso A, Jick SS, Hernán MA. 2006 Allergy, histamine 1 receptor blockers, and the risk of multiple sclerosis. Neurology 66:572–75 [PubMed: 16505314]
5. Alwahaibi A, Zeka A. 2015 Respiratory and allergic health effects in a young population in proximity of a major industrial park in Oman. J. Epidemiol. Community Health. doi: 10.1136/jech-2015-205609. In press
6. Anderson HD, Pace WD, Brandt E, Nielsen RD, Allen RR, et al. 2015 Monitoring suicidal patients in primary care using electronic health records. J. Am. Board Fam. Med. 28:65–71 [PubMed: 25567824]
7. Armstrong-Wells J, Johnston SC, Wu YW, Sidney S, Fullerton HJ. 2009 Prevalence and predictors of perinatal hemorrhagic stroke: results from the Kaiser Pediatric Stroke Study. Pediatrics 123:823–28 [PubMed: 19255009]
8. Baillargeon J, Paar D, Wu H, Giordano T, Murray O, et al. 2008 Psychiatric disorders, HIV infection and HIV/hepatitis co-infection in the correctional setting. AIDS Care 20:124–29 [PubMed: 18278623]
9. Berkowitz SA, Karter AJ, Lyles CR, Liu JY, Schillinger D, et al. 2014 Low socioeconomic status is associated with increased risk for hypoglycemia in diabetes patients: the Diabetes Study of Northern California (DISTANCE). J. Health Care Poor Underserved 25:478–90 [PubMed: 24858863]
10. Betancourt T, Scorza P, Kanyanganzi F, Fawzi MC, Sezibera V, et al. 2014 HIV and child mental health: a case-control study in Rwanda. Pediatrics 134:e464–72 [PubMed: 25049342]
11. Bild DE, Bluemke DA, Burke GL, Detrano R, Diez Roux AV, et al. 2002 Multi-ethnic study of atherosclerosis: objectives and design. Am.J. Epidemiol. 156:871–81 [PubMed: 12397006]
12. Birkhead GS, Klompas M, Shah NR. 2015 Uses of electronic health records for public health surveillance to advance public health. Annu. Rev. Public Health 36:345–59 [PubMed: 25581157]
13. Black MH, Smith N, Porter AH, Jacobsen SJ, Koebnick C. 2012 Higher prevalence of obesity among children with asthma. Obesity 20:1041–47 [PubMed: 22252049]
14. Blumenthal D, Tavenner M. 2010 The “meaningful use” regulation for electronic health records. N. Engl. J. Med. 363:501–4 [PubMed: 20647183]
15. Botros N, Concato J, Mohsenin V, Selim B, Doctor K, Yaggi HK. 2009 Obstructive sleep apnea as a risk factor for type 2 diabetes. Am. J. Med. 122:1122–27 [PubMed: 19958890]
16. Brown JL. 2012 A piece of my mind: the unasked question. JAMA 308:1869–70 [PubMed: 23150007]
17. Carrell DS, Halgrim S, Tran D-T, Buist DS, Chubak J, et al. 2014 Carrell et al. respond to “Observational Research and the EHR.” Am. J. Epidemiol. 179:762–63 [PubMed: 24488509]
18. Casey JA, Curriero FC, Cosgrove SE, Nachman KE, Schwartz BS. 2013 High-density livestock operations, crop field application of manure, and risk of community-associated methicillin-resistant Staphylococcus aureus infection in Pennsylvania. JAMA Intern. Med. 173:1980–90 [PubMed: 24043228]
19. Casey JA, Savitz DA, Rassmusen SG, Ogburn EL, Pollak J, Schwartz BS. 2015 Unconventional natural gas development and birth outcomes in Pennsylvania, USA. Epidemiology. doi: 10.1097/EDE.0000000000000387. In press

20. Castillo EG, Olfson M, Pincus HA, Vawdrey D, Stroup TS. 2015 Electronic health records in mental health research: a framework for developing valid research methods. Psychiatr. Serv. 66:193–96 [PubMed: 25642614]
21. Chang TS, Gangnon RE, David Page C, Buckingham WR, Tandias A, et al. 2015 Sparse modeling of spatial environmental variables associated with asthma. J. Biomed. Inform. 53:320–29 [PubMed: 25533437]
22. Cheuk BL, Cheung GC, Cheng SW. 2004 Epidemiology of venous thromboembolism in a Chinese population. Br. J. Surg. 91:424–28 [PubMed: 15048741]
23. Choi SK, Park YG, Park IY, Ko HS, Shin JC. 2014 Impact of antenatal depression on perinatal outcomes and postpartum depression in Korean women. J. Res. Med. Sci. 19:807–12 [PubMed: 25535492]
24. Clark MM, Hanna BK, Mai JL, Graszer KM, Krochta JG, et al. 2007 Sexual abuse survivors and psychiatric hospitalization after bariatric surgery. Obes. Surg. 17:465–69 [PubMed: 17608258]
25. Cohen HA, Blau H, Hoshen M, Batat E, Balicer RD. 2014 Seasonality of asthma: a retrospective population study. Pediatrics 133:e923–32 [PubMed: 24616356]
26. Collins FS, Varmus H. 2015 Anew initiative on precision medicine. N. Engl. J. Med. 372:793–95 [PubMed: 25635347]
27. Comm. on the Recomm. Soc. and Behav. Domains and Meas. for Electron. Health Rec., IOM (Inst. Med.) 2014 Capturing Social and Behavioral Domains and Measures in Electronic Health Records: Phase 2. Washington, DC: Natl. Acad. Press
28. Consort. on Safe Labor, Hibbard JU, Wilkins I, Sun L, Gregory K, et al. 2010 Respiratory morbidity in late preterm births. JAMA 304:419–25 [PubMed: 20664042]
29. Court H, McLean G, Guthrie B, Mercer SW, Smith DJ. 2014 Visual impairment is associated with physical and mental comorbidities in older adults: a cross-sectional study. BMC Med. 12:181 [PubMed: 25603915]
30. Croen LA, Grether JK, Yoshida CK, Odouli R, Van de Water J. 2005 Maternal autoimmune diseases, asthma and allergies, and childhood autism spectrum disorders: a case-control study. Arch. Pediatr. Adolesc. Med. 159:151–57 [PubMed: 15699309]
31. Croen LA, Yoshida CK, Odouli R, Newman TB. 2005 Neonatal hyperbilirubinemia and risk of autism spectrum disorders. Pediatrics 115:e135–38 [PubMed: 15687420]
32. Dai W, Brisimi TS, Adams WG, Mela T, Saligrama V, Paschalidis ICh. 2015 Prediction of hospitalization due to heart diseases by supervised learning methods. Int. J. Med. Inform. 84:189–97 [PubMed: 25497295]
33. Denaxas SC, George J, Herrett E, Shah AD, Kalra D, et al. 2012 Data resource profile: cardiovascular disease research using linked bespoke studies and electronic health records (CALIBER). Int. J. Epidemiol. 41:1625–38 [PubMed: 23220717]
34. Dep. Health Hum. Serv. 2015 Medicare and Medicaid programs; electronic health record incentive program—stage 3 and modifications to meaningful use in 2015 through 2017. Fed. Regist. https://federalregister.gov/a/2015-25595
35. Dhalwani NN, West J, Sultan AA, Ban L, Tata LJ. 2014 Women with celiac disease present with fertility problems no more often than women in the general population. Gastroenterology 147:1267–74 [PubMed: 25157666]
36. Dirkzwager AJ, Grievink L, van der Velden PG, Yzermans CJ. 2006 Risk factors for psychological and physical health problems after a man-made disaster. Prospective study. Br. J. Psychiatry 189:144–49 [PubMed: 16880484]
37. Doshi-Velez F, Ge Y, Kohane I. 2014 Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis. Pediatrics 133:e54–63 [PubMed: 24323995]
38. Duncan DT, Sharifi M, Melly SJ, Marshall R, Sequist TD, et al. 2014 Characteristics of walkable built environments and BMI *z*-scores in children: evidence from a large electronic health record database. Environ. Health Perspect. 122:1359–65 [PubMed: 25248212]
39. Esteban-Vasallo MD, Domínguez-Berjón MF, Gil-Prieto R, Astray-Mochales J, Gil de Miguel Á. 2014 Sociodemographic characteristics and chronic medical conditions as risk factors for herpes zoster: a population-based study from primary care in Madrid (Spain). Hum. Vaccin. Immunother. 10:1650–60 [PubMed: 24805130]

40. Forcey DS, Hocking JS, Tabrizi SN, Bradshaw CS, Chen MY, et al. 2014 Chlamydia detection during the menstrual cycle: a cross-sectional study of women attending a sexual health service. PLOS ONE 9:e85263
41. Framingham Heart Study. Research milestones. Framingham Heart Study, Framingham, MA http://www.framinghamheartstudy.org/about-fhs/research-milestones.php
42. Galea S, Tracy M. 2007 Participation rates in epidemiologic studies. Ann. Epidemiol. 17:643–53 [PubMed: 17553702]
43. Geraghty EM, Balsbaugh T, Nuovo J, Tandon S. 2010 Using geographic information systems (GIS) to assess outcome disparities in patients with type 2 diabetes and hyperlipidemia. J. Am. Board Fam. Med. 23:88–96 [PubMed: 20051547]
44. Goyal NK, Fiks AG, Lorch SA. 2011 Association of late-preterm birth with asthma in young children: practice-based study. Pediatrics 128:e830–38 [PubMed: 21911345]
45. Green BB, Anderson ML, Cook AJ, Catz S, Fishman PA, et al. 2012 Using body mass index data in the electronic health record to calculate cardiovascular risk. Am. J. Prev. Med. 42:342–47 [PubMed: 22424246]
46. Halfdanarson TR, Bamlet WR, McWilliams RR, Hobday TJ, Burch PA, et al. 2014 Risk factors for pancreatic neuroendocrine tumors: a clinic-based case-control study. Pancreas 43:1219–22 [PubMed: 25291526]
47. Hata A, Kuniyoshi M, Ohkusa Y. 2011 Risk of Herpes zoster in patients with underlying diseases: a retrospective hospital-based cohort study. Infection 39:537–44 [PubMed: 21800108]
48. Hawkins MA, Callahan CM, Stump TE, Stewart JC. 2014 Depressive symptom clusters as predictors of incident coronary artery disease: a 15-year prospective study. Psychosom. Med. 76:38–43
49. Hennessy S 2006 Use of health care databases in pharmacoepidemiology. Basic Clin. Pharmacol. Toxicol. 98:311–13 [PubMed: 16611207]
50. Hesdorffer DC, Ishihara L, Mynepalli L, Webb DJ, Weil J, Hauser WA. 2012 Epilepsy, suicidality, and psychiatric disorders: a bidirectional association. Ann. Neurol. 72:184–91 [PubMed: 22887468]
51. Hill EM, Turner EL, Martin RM, Donovan JL. 2013 “Let’s get the best quality research we can”: public awareness and acceptance of consent to use existing data in health research: a systematic review and qualitative study. BMC Med. Res. Methodol. 13:72 [PubMed: 23734773]
52. Hillier TA, Pedula KL, Schmidt MM, Mullen JA, Charles MA, Pettitt DJ. 2007 Childhood obesity and metabolic imprinting: the ongoing effects of maternal hyperglycemia. Diabetes Care 30:2287–92 [PubMed: 17519427]
53. Hinkle SN, Albert PS, Mendola P, Sjaarda LA, Boghossian NS, et al. 2014 Differences in risk factors for incident and recurrent small-for-gestational-age birthweight: a hospital-based cohort study. BJOG 121:1080–88 [PubMed: 24702952]
54. Hsu J, Pacheco JA, Stevens WW, Smith ME, Avila PC. 2014 Accuracy of phenotyping chronic rhinosinusitis in the electronic health record. Am. J. Rhinol. Allergy 28:140–44 [PubMed: 24717952]
55. Insel TR, Cuthbert BN. 2015 Medicine. Brain disorders? Precisely. Science 348:499–500 [PubMed: 25931539]
56. Jensen PB, Jensen LJ, Brunak S. 2012 Mining electronic health records: towards better research applications and clinical care. Nat. Rev. Genet. 13:395–405 [PubMed: 22549152]
57. Jick SS, Lieberman ES, Rahman MU, Choi HK. 2006 Glucocorticoid use, other associated factors, and the risk of tuberculosis. Arthritis Rheum. 55:19–26 [PubMed: 16463407]
58. Johnson EL, Beal JR. 2013 Impact of a comprehensive smoke-free law following a partial smoke-free law on incidence of heart attacks at a rural community hospital. Nicot. Tob. Res. 15:745–47
59. Kanaya AM, Adler N, Moffet HH, Liu J, Schillinger D, et al. 2011 Heterogeneity of diabetes outcomes among Asians and Pacific Islanders in the US: the Diabetes Study of Northern California (DISTANCE). Diabetes Care 34:930–37 [PubMed: 21350114]
60. Kaye J, Curren L, Anderson N, Edwards K, Fullerton SM, et al. 2012 From patients to partners: participant-centric initiatives in biomedical research. Nat. Rev. Genet. 13:371–76 [PubMed: 22473380]

61. Kaye JA, del Mar Melero-Montes M, Jick H. 2001 Mumps, measles, and rubella vaccine and the incidence of autism recorded by general practitioners: a time trend analysis. BMJ 322:460–63 [PubMed: 11222420]
62. Keegan TH, Kurian AW, Gali K, Tao L, Lichtensztajn DY, et al. 2015 Racial/ethnic and socioeconomic differences in short-term breast cancer survival among women in an integrated health system. Am. J. Public Health 105:938–46 [PubMed: 25790426]
63. Kennedy EH, Wiitala WL, Hayward RA, Sussman JB. 2013 Improved cardiovascular risk prediction using nonparametric regression and electronic health record data. Med. Care 51:251–58 [PubMed: 23269109]
64. Koebnick C, Smith N, Black MH, Porter AH, Richie BA, et al. 2012 Pediatric obesity and gallstone disease. J. Pediatr. Gastroenterol. Nutr. 55:328–33 [PubMed: 22314396]
65. Koebnick C, Smith N, Huang K, Martinez MP, Clancy HA, et al. 2012 OBAYA (Obesity and Adverse Health Outcomes in Young Adults): feasibility of a population-based multiethnic cohort study using electronic medical records. Popul. Health Metr. 10:15 [PubMed: 22909293]
66. Kohane IS. 2011 Using electronic health records to drive discovery in disease genomics. Nat. Rev. Genet. 12:417–28 [PubMed: 21587298]
67. Kristal RB, Blank AE, Wylie-Rosett J, Selwyn PA. 2015 Factors associated with daily consumption of sugar-sweetened beverages among adult patients at four federally qualified health centers, Bronx, New York, 2013. Prev. Chronic Dis. 12:E02 [PubMed: 25569695]
68. Laraia BA, Karter AJ, Warton EM, Schillinger D, Moffet HH, Adler N. 2012 Place matters: neighborhood deprivation and cardiometabolic risk factors in the Diabetes Study of Northern California (DISTANCE). Soc. Sci. Med. 74:1082–90 [PubMed: 22373821]
69. Lauer MS. 2012 Time for a creative transformation of epidemiology in the United States. JAMA 308:1804–5 [PubMed: 23117782]
70. Link BG, Phelan J. 1995 Social conditions as fundamental causes of disease. J. Health Soc. Behav. Spec. No:80–94
71. Liu AY, Curriero FC, Glass TA, Stewart WF, Schwartz BS. 2013 The contextual influence of coal abandoned mine lands in communities and type 2 diabetes in Pennsylvania. Health Place 22:115–22 [PubMed: 23689181]
72. Liu GC, Wilson JS, Qi R, Ying J. 2007 Green neighborhoods, food retail and childhood overweight: differences by population density. Am. J. Health Promot. 21:317–25 [PubMed: 17465177]
73. Luchenski SA, Reed JE, Marston C, Papoutsi C, Majeed A, Bell D. 2013 Patient and public views on electronic health records and their uses in the United Kingdom: cross-sectional survey. J. Med. Internet Res. 15:e160 [PubMed: 23975239]
74. Maas J, Verheij RA, de Vries S, Spreeuwenberg P, Schellevis FG, Groenewegen PP. 2009 Morbidity is related to a green living environment. J. Epidemiol. Community Health 63:967–73 [PubMed: 19833605]
75. Macintyre S, Maciver S, Sooman A. 1993 Area, class and health: Should we be focusing on places or people? J. Soc. Policy 22:213–34
76. Mack C 2014 PS1–13: Probabilistic linkage (also known as “fuzzy matching”): the theoretical foundations of modern record linkage. Clin. Med. Res. 12:95
77. Mahajan R, Xing J, Liu SJ, Ly KN, Moorman AC, et al. 2014 Mortality among persons in care with hepatitis C virus infection: the Chronic Hepatitis Cohort Study (CHeCS), 2006–2010. Clin. Infect. Dis. 58:1055–61 [PubMed: 24523214]
78. Männistö T, Mendola P, Liu D, Leishear K, Sherman S, Laughon SK.2015 Acute air pollution exposure and blood pressure at delivery among women with and without hypertension. Am. J. Hypertens. 28:58–72 [PubMed: 24795401]
79. Manolio TA, Weis BK, Cowie CC, Hoover RN, Hudson K,et al.2012 New models for large prospective studies: Is there a better way? Am.J. Epidemiol. 175:859–66 [PubMed: 22411865]
80. May L, Carim M, Yadav K.2011 Adult asthma exacerbations and environmental triggers: a retrospective review of ED visits using an electronic medical record. Am. J. Emerg. Med. 29:1074–82 [PubMed: 20708875]

81. McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, et al. 2011 The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Med. Genomics 4:13 [PubMed: 21269473]
82. McCoy TH, Castro VM, Rosenfield HR, Cagan A, Kohane IS, Perlis RH. 2015 A clinical perspective on the relevance of research domain criteria in electronic health records. Am. J. Psychiatry 172:316–20 [PubMed: 25827030]
83. Mendola P, Mumford SL, Männistö TI, Holston A, Reddy UM, Laughon SK. 2015 Controlled direct effects of preeclampsia on neonatal health after accounting for mediation by preterm birth. Epidemiology 26:17–26 [PubMed: 25437315]
84. MESA (Multi-Ethnic Study of Atheroscler.). 2015 Study timeline and procedures. MESA Coord. Cent., Univ. Wa. Seattle http://www.mesanhlbi.org/aboutMESAStudyTime.aspx
85. Moffet HH, Adler N, Schillinger D, Ahmed AT, Laraia B, et al. 2009 Cohort profile: The Diabetes Study of Northern California (DISTANCE)—objectives and design of a survey follow-up study of social health disparities in a managed care population. Int. J. Epidemiol. 38:38–47 [PubMed: 18326513]
86. Mooney SJ, Westreich DJ, El-Sayed AM. 2015 Commentary: Epidemiology in the era of big data. Epidemiology 26:390–94 [PubMed: 25756221]
87. Moorman AC, Gordon SC, Rupp LB, Spradling PR, Teshale EH, et al. 2013 Baseline characteristics and mortality among people in care for chronic viral hepatitis: the chronic hepatitis cohort study. Clin. Infect. Dis. 56:40–50 [PubMed: 22990852]
88. Mukamal KJ, Wellenius GA, Suh HH, Mittleman MA. 2009 Weather and air pollution as triggers of severe headaches. Neurology 72:922–27 [PubMed: 19273827]
89. Musser ED, Hawkey E, Kachan-Liu SS, Lees P, Roullet JB, et al. 2014 Shared familial transmission of autism spectrum and attention-deficit/hyperactivity disorders. J. Child Psychol. Psychiatry 55:819–27 [PubMed: 24444366]
90. Nass SJ, Levit LA, Gostin LO, eds. 2009 Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research. Washington, DC: Natl. Acad. Press
91. Natl. Heart, Lung, Blood Inst. 2013 Research and development contracts In Fact Book Fiscal Year 2012, pp. 101–10. Bethesda, MD: US Dep. Health Hum. Serv., Natl. Heart Lung Blood Inst http://www.nhlbi.nih.gov/about/documents/factbook/2012/chapter10
92. Nau C, Schwartz BS, Bandeen-Roche K, Liu A, Pollak J, et al. 2015 Community socioeconomic deprivation and obesity trajectories in children using electronic health records. Obesity 23:207–12 [PubMed: 25324223]
93. Osborn DP, Hardoon S, Omar RZ, Holt RI, King M, et al.2015 Cardiovascularrisk prediction models for people with severe mental illness: results from the Prediction and Management of Cardiovascular Risk in People With Severe Mental Illnesses (PRIMROSE) research program. JAMA Psychiatry 72:143–51 [PubMed: 25536289]
94. Otte-Trojel T, de Bont A, Rundall TG, van de Klundert J. 2014 How outcomes are achieved through patient portals: a realist review. J. Am. Med. Inform. Assoc. 21:751–57 [PubMed: 24503882]
95. Pace WD, Cifuentes M,Valuck RJ, Staton EW, Brandt EC, West DR.2009 An electronic practice-based network for observational comparative effectiveness research. Ann. Intern. Med. 151:338–40 [PubMed: 19638402]
96. Palaniappan LP, Wong EC, Shin JJ, Fortmann SP, Lauderdale DS. 2011 Asian Americans have greater prevalence of metabolic syndrome despite lower body mass index. Int. J. Obes. 35:393–400
97. Park S, Kim JW, Kim BN, Bae JH, Shin MS, et al. 2015 Clinical characteristics and precipitating factors of adolescent suicide attempters admitted for psychiatric inpatient care in South Korea. Psychiatry Investig. 12:29–36
98. Pavey AR, Gorman GH, Kuehn D, Stokes TA, Hisle-Gorman E. 2014 Intimate partner violence increases adverse outcomes at birth and in early infancy. J. Pediatr. 165:1034–39 [PubMed: 25128162]
99. Pujades-Rodriguez M, George J, Shah AD, Rapsomaniki E, Denaxas S, et al. 2015 Heterogeneous associations between smoking and a wide range of initial presentations of cardiovascular disease

in1937 360 people in England: lifetime risks and implications for risk prediction. Int. J. Epidemiol. 44:129–41 [PubMed: 25416721]

100. Pujades-Rodriguez M, Timmis A, Stogiannis D, Rapsomaniki E, Denaxas S, et al. 2014 Socioeconomic deprivation and the incidence of 12 cardiovascular diseases in 1.9 million women and men: implications for risk prediction and prevention. PLOS ONE 9:e104671
101. Puttkammer N, Zeliadt S, Balan JG, Baseman J, Destine R, et al. 2014 Development of an electronic medical record based alert for risk of HIV treatment failure in a low-resource setting. PLOS ONE 9:e112261
102. Qizilbash N, Gregson J, Johnson ME, Pearce N, Douglas I, et al. 2015 BMI and risk of dementia in two million people over two decades: a retrospective cohort study. Lancet Diabetes Endocrinol. 3:431–36 [PubMed: 25866264]
103. Rapsomaniki E, Timmis A, George J, Pujades-Rodriguez M, Shah AD, et al. 2014 Blood pressure and incidence of twelve cardiovascular diseases: lifetime risks, healthy life-years lost, and age-specific associations in 1.25 million people. Lancet 383:1899–911 [PubMed: 24881994]
104. Reis BY, Kohane IS, Mandl KD. 2009 Longitudinal histories as predictors of future diagnoses of domestic abuse: modelling study. BMJ 339:b3677 [PubMed: 19789406]
105. Robbins GK, Johnson KL, Chang Y, Jackson KE, Sax PE, et al. 2010 Predicting virologic failure in an HIV clinic. Clin. Infect. Dis. 50:779–86 [PubMed: 20121574]
106. Robledo CA, Mendola P, Yeung E, Männistö T, Sundaram R, et al. 2015 Preconception and early pregnancy air pollution exposures and risk of gestational diabetes mellitus. Environ. Res. 137:316–22 [PubMed: 25601734]
107. Roth C, Foraker RE, Payne PR, Embi PJ. 2014 Community-level determinants of obesity: harnessing the power of electronic health records for retrospective data analysis. BMC Med. Inform. Decis. Mak. 14:36 [PubMed: 24886134]
108. Rothstein MA. 2010 Is deidentification sufficient to protect health privacy in research? Am. J. Bioeth. 10:3–11
109. Scherrer JF, Garfield LD, Chrusciel T, Hauptman PJ, Carney RM, et al. 2011 Increased risk of myocardial infarction in depressed patients with type 2 diabetes. Diabetes Care 34:1729–34 [PubMed: 21680721]
110. Schoen C, Osborn R, Squires D, Doty M, Rasmussen P, et al. 2012 A survey of primary care doctors in ten countries shows progress in use of health information technology, less in other areas. Health Aff. 31:2805–16
111. Schwartz BS, Stewart WF, Godby S, Pollak J, DeWalle J, et al. 2011 Body mass index and the built and social environments in children and adolescents using electronic health records. Am. J. Prev. Med. 41:e17–28 [PubMed: 21961475]
112. Selby JV. 1997 Linking automated databases for research in managed care settings. Ann. Intern. Med. 127:719–24 [PubMed: 9382386]
113. Shah AD, Langenberg C, Rapsomaniki E, Denaxas S, Pujades-Rodriguez M, et al. 2015 Type 2 diabetes and incidence of cardiovascular diseases: a cohort study in 1.9 million people. Lancet Diabetes Endocrinol. 3:105–13 [PubMed: 25466521]
114. Shephard EA, Neal RD, Rose P, Walter FM, Litt EJ, Hamilton WT. 2015 Quantifying the risk of multiple myeloma from symptoms reported in primary care patients: a large case-control study using electronic records. Br. J. Gen. Pract. 65:e106–13 [PubMed: 25624306]
115. Shu T, Liu H, Goss FR, Yang W, Zhou L, et al. 2014 EHR adoption across China's tertiary hospitals: a cross-sectional observational study. Int. J. Med. Inform. 83:113–21 [PubMed: 24262068]
116. Sivarajasingam V, Page N, Morgan P, Matthews K, Moore S, Shepherd J. 2014 Trends in community violence in England and Wales 2005–2009. Injury 45:592–98 [PubMed: 23867145]
117. Smeeth L, Cook C, Fombonne E, Heavey L, Rodrigues LC, et al. MMR vaccination and pervasive developmental disorders: a case-control study. Lancet 364:963–69 [PubMed: 15364187]
118. Smit LA, van der Sman-de Beer F, Opstal-van Winden AW, Hooiveld M, Beekhuizen J, et al. 2012 Q fever and pneumonia in an area with a high livestock density: a large population-based study. PLOS ONE 7:e38843

119. Smith N, Coleman KJ, Lawrence JM, Quinn VP, Getahun D, et al. 2010 Body weight and height data in electronic medical records of children. Int. J. Pediatr. Obes. 5:237–42 [PubMed: 19961272]
120. Sox HC, Goodman SN. 2012 The methods of comparative effectiveness research. Annu. Rev. Public Health 33:425–45 [PubMed: 22224891]
121. Stark A, Stahl M, Kirchner H, Krum S, Prichard J, Evans J. 2010 Body mass index at the time of diagnosis and the risk of advanced stages and poorly differentiated cancers of the breast: findings from a case-series study. Int.J. Obes. 34:1381–86
122. Stewart WF, Shah NR, Selna MJ, Paulus RA, Walker JM. 2007 Bridging the inferential gap: the electronic health record and clinical evidence. Health Aff. 26:w181–91
123. Susser M, Susser E. 1996 Choosing a future for epidemiology: II. From black box to Chinese boxes and eco-epidemiology. Am. J. Public Health 86:674–77 [PubMed: 8629718]
124. Tamang S, Patel MI, Blayney DW, Kuznetsov J, Finlayson SG, et al. 2015 Detecting unplanned care from clinician notes in electronic health records. J. Oncol. Pract. 11(3):e313–19 [PubMed: 25980019]
125. Thomas SL, Minassian C, Ganesan V, Langan SM, Smeeth L. 2014 Chickenpox and risk of stroke: a self-controlled case series analysis. Clin. Infect. Dis. 58:61–68 [PubMed: 24092802]
126. Tomayko EJ, Flood TL, Tandias A, Hanrahan LP. 2015 Linking electronic health records with community-level data to understand childhood obesity risk. Pediatr. Obes. doi: 10.1111/ijpo.12003. In press
127. Trotter F, Uhlman D. 2011 Hacking Healthcare: A Guide to Standards, Workflows, and Meaningful Use. Sebastopol, CA: O'Reilly Media
128. Vigen R, O'Donnell CI, Baron AE, Grunwald GK, Maddox TM, et al. 2013 Association of testosterone therapy with mortality, myocardial infarction, and stroke in men with low testosterone levels. JAMA 310:1829–36 [PubMed: 24193080]
129. Wasserman RC. 2011 Electronic medical records (EMRs), epidemiology, and epistemology: reflections on EMRs and future pediatric clinical research. Acad. Pediatr. 11:280–87 [PubMed: 21622040]
130. Weng C, Appelbaum P, Hripcsak G, Kronish I, Busacca L, et al. 2012 Using EHRs to integrate research with patient care: promises and challenges. J. Am. Med. Inform. Assoc. 19:684–87 [PubMed: 22542813]
131. Whitaker RC, Pepe MS, Wright JA, Seidel KD, Dietz WH. 1998 Early adiposity rebound and the risk of adult obesity. Pediatrics 101:E5
132. Wilkins JT, Ning H, Berry J, Zhao L, Dyer AR, Lloyd-Jones DM. 2012 Lifetime risk and years lived free of total cardiovascular disease. JAMA 308:1795–801 [PubMed: 23117780]
133. Wu J, Roy J, Stewart WF. 2010 Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. Med. Care 48:S106–13 [PubMed: 20473190]
134. Yoon D, Chang BC, Kang SW, Bae H, ParkRW. 2012 Adoption of electronic health records in Korean tertiary teaching and general hospitals. Int. J. Med. Inform. 81:196–203 [PubMed: 22206619]

EHR: a software platform that contains individual-level patient-provider data captured during health care encounters. Epic, eClinicalWorks, McKesson, and Cerner are examples

Geographic information systems (GIS): a tool that allows researchers to combine and visualize spatial data and export analytic variables for merging with EHR data

Meaningful use: providers demonstrate they are meaningfully using their EHRs by meeting increasing thresholds for specific objectives, services, and activities

Primary data collection: new data collected for a specific research purpose, not for clinical care

Incidence rate: the number of disease onsets divided by the person-time at risk; health care encounters determine if a patient is contributing person-time

ICD-9: International Classification of Diseases code

Natural language processing: A technology that extracts information from free text, e.g., detecting sentence boundaries, segmenting text into meaningful groups, inferring temporal relationships

SES: socioeconomic status

Extract, transform, and load: a tool that reads desired clinical EHR data, converts it into a usable form, and then writes it into the research database

Deidentified data: deidentification of protected health information occurs when all the HIPAA identifiers are removed from the data set

Normalized (data): consistently structured and bounded data that link logically with other data available in the system

BMI: body mass index (kg/m^2)

CVD: cardiovascular disease

Geocode: the process of taking a patient address and assigning it to a spatial location with geographic coordinates

Machine learning: algorithms used to predict outcomes based on features of the data; methods include support vector machines and regression trees

Hawthorne effect: changes in reports or behaviors due to awareness of being studied

Social desirability bias: reporting behaviors and beliefs believed to be more acceptable or valued by others

PHI: data generated in the health context, which relates to health and can be used to identify individuals, for instance names and addresses

Patient portal: an online application where patients can access their health information and communicate with their provider electronically

SUMMARY POINTS LIST

1. EHRs provide researchers with low-cost sources of rich longitudinal health data on large geographically, socioeconomically, and culturally diverse populations.
2. EHRs offer particular advantages for environmental and social epidemiology where patient addresses can be linked to individual and contextual exposures that vary spatially.
3. The use of EHRs for epidemiology requires consideration of unique issues related to study population definition, population attrition, disease/case definition, and privacy concerns.
4. Major areas of population health EHR research include reevaluating prior findings; capitalizing on large sample sizes to analyze subgroups and to study rare diseases or multiple diseases simultaneously; social and environmental epidemiology; research on stigmatized conditions; predictive modeling; and exploiting natural experiments.
5. Future developments in EHRs including increased use and sophistication, improved capture of social and behavioral determinants of health, better standardization to allow data merging across health systems, and linkage to vital records and to other emerging technologies (e.g., personal sensing) and data streams (e.g., air pollution data, clinical biobanks) will improve data quality and expand research opportunities to improve public health.

Table 1

Data domains available from electronic health records[a]

Domain	Examples	Utility to epidemiologic research
Demographics[b]	Age, sex, race/ethnicity, residential address	Exposures, confounders, effect modifiers and/or mediators; address used to link to environmental and community data for individual-level or contextual exposures
Health behavior[b]	Tobacco, alcohol, and injection drug use	Outcomes, exposures, confounders, effect modifiers and/or mediators
Vital signs[b]	Pulse, systolic and diastolic blood pressure, height, weight (used to derive BMI)	Outcomes, exposures, confounders, effect modifiers and/or mediators
Outpatient encounters[b]	ICD-9 codes for a wide variety of diagnoses, including diabetes, hypertension, asthma, kidney failure, migraine	Diagnostic codes used to construct variables; encounter type can indicate disease severity; timing of diagnoses in relation to one another and interval between visits may provide signals about the disease course
Inpatient encounters[b]		
Emergency department encounters[b]		
Laboratory data[b]	Lipid panel, basic metabolic panel, microbiologic culture with antibiotic resistance tests, liver function tests, microalbuminuria, hemoglobin A1c	Laboratory orders and results used to identify primary outcome or as covariates and can improve diagnostic accuracy of ICD-9 codes and to evaluate disease progression, severity, and control
Medication order[b]	Type, dose, frequency, duration	Medication orders provide information about disease course and severity, control, and prevention (e.g., hypercholesterolemia, statins, and cardiovascular disease)
Procedures[b]	Electrocardiogram, pulmonary function tests	Procedural data can improve diagnostic accuracy of ICD-9 codes and evaluate disease severity and control
Problem list	ICD-9 code for depression, heart failure, hypertension	Ongoing patient health problems are used to confirm diagnoses in other locations and can also be helpful in defining disease onset
Free text	Encounter notes, imaging notes	Text analysis can provide information on symptoms, onset, duration, and severity; notes can also have information not available elsewhere, e.g., Apgar scores; labor and delivery notes can also be used to link mothers and infants
Imaging	Echocardiogram, magnetic resonance imaging, CT scan	Data used to verify diagnosis and subtype and detection of other health problems

[a] Abbreviations: BMI, body mass index; CT, computerized tomography; ICD-9, International Classification of Diseases.

[b] Longitudinal data or repeated measures can be used to construct time-dependent variables.

Table 2

Selected examples of electronic health record study population data sources from cohort studies

Data source	Sample size	References
Single psychiatric inpatient unit	728–2,010	82, 97
Specialized center/clinic	544–10,017	15, 40
Prison network	370, 511	8
Single hospital	467–55,492	23, 47
Multiple hospitals	1,074–25,241	53, 105
Multiple primary care practices	7,925–345,143	44, 74
Health care system	2,537–919,873	25, 48
Consortium	8,709–233,844	28, 83
Centralized anonymized repository	923–5,244,402	39, 101

Table 3

Comparison of traditional and EHR epidemiology studies[a]

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.
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 on 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.	**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.

Study feature	Traditional study	EHR study
	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.	**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.

[a]Abbreviations: EHR, electronic health record; GIS, geographic information systems; HMORN, Health Maintenance Organization Research Network.