







SPECIAL REPORT

Establishing a National Cardiovascular Disease Surveillance System in the United States Using Electronic Health Record Data: Key Strengths and Limitations

Brent A. Williams , PhD; Stephen Voyce, MD; Stephen Sidney, MD; Véronique L. Roger , MD, MPH; Timothy B. Plante, MD; Sharon Larson, PhD; Michael J. LaMonte , PhD; Darwin R. Labarthe, MD; Bailey M. DeBarmore , MHS; Alexander R. Chang, MD; Alanna M. Chamberlain , PhD; Catherine P. Benziger , MD

ABSTRACT: Cardiovascular disease surveillance involves quantifying the evolving population-level burden of cardiovascular outcomes and risk factors as a data-driven initial step followed by the implementation of interventional strategies designed to alleviate this burden in the target population. Despite widespread acknowledgement of its potential value, a national surveillance system dedicated specifically to cardiovascular disease does not currently exist in the United States. Routinely collected health care data such as from electronic health records (EHRs) are a possible means of achieving national surveillance. Accordingly, this article elaborates on some key strengths and limitations of using EHR data for establishing a national cardiovascular disease surveillance system. Key strengths discussed include the: (1) ubiquity of EHRs and consequent ability to create a more “national” surveillance system, (2) existence of a common data infrastructure underlying the health care enterprise with respect to data domains and the nomenclature by which these data are expressed, (3) longitudinal length and detail that define EHR data when individuals repeatedly patronize a health care organization, and (4) breadth of outcomes capable of being surveilled with EHRs. Key limitations discussed include the: (1) incomplete ascertainment of health information related to health care-seeking behavior and the disconnect of health care data generated at separate health care organizations, (2) suspect data quality resulting from the default information-gathering processes within the clinical enterprise, (3) questionable ability to surveil patients through EHRs in the absence of documented interactions, and (4) the challenge in interpreting temporal trends in health metrics, which can be obscured by changing clinical and administrative processes.

Key Words: cardiovascular disease ■ electronic health records ■ population surveillance

Despite dramatic improvements in cardiovascular disease (CVD)-related mortality over the past several decades, CVD remains a leading cause of death and morbidity in the United States.¹⁻⁴ Quantifying the evolving population-level burden of CVD and its risk factors over time are the sphere of surveillance, although a more complete definition goes beyond measurement by including information dissemination, prioritization, intervention, and remeasurement in a continuous cycle (Figure 1).^{5,6} The essence of surveillance can be perceived graphically

with time along the horizontal axis and some health metric(s) along the vertical (Figure 2): absolute values of a metric quantify the state of an issue, while contemporary values interpreted in the context of historical trends describe its trajectory (ie, improving, worsening, or stable).⁷ As a recent exemplar of surveillance in the CVD arena, improving age-adjusted death rates attributable to heart disease over time were shown to be attenuating in recent years, while the overall number of deaths attributable to heart disease increased (Figure 3).¹

Correspondence to: Brent A. Williams, PhD, Geisinger Health System, 100 North Academy Avenue, Danville, PA 17822. Email: bawilliams2@geisinger.edu
For Disclosures, see page 11.

© 2022 The Authors and Mayo Clinic. Published on behalf of the American Heart Association, Inc., by Wiley. This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

JAHA is available at: www.ahajournals.org/journal/jaha

Nonstandard Abbreviations and Acronyms

| | |
|----------------|---|
| ARIC | Atherosclerosis Risk in Communities |
| BRFSS | Behavioral Risk Factor Surveillance System |
| CDC | Centers for Disease Control and Prevention |
| EHR | Electronic Health Record |
| GBD | Global Burden of Disease |
| HCPCS | Healthcare Common Procedure Coding System |
| HCSRN | Healthcare Systems Research Network |
| LOINC | Logical Observation Identifiers Names and Codes |
| MESA | Multi-Ethnic Study of Atherosclerosis |
| NCHS | National Center for Health Statistics |
| NDC | National Drug Code |
| NHANES | National Health and Nutrition Examination Survey |
| OMOP | Observational Medical Outcomes Partnership |
| PCORnet | National Patient-Centered Clinical Research Network |
| REGARDS | Reasons for Geographic and Racial Differences in Stroke |

IMPORTANCE OF CVD SURVEILLANCE

Despite multiple exhortations for a dedicated national CVD surveillance system over the past 15 years, such a system does not currently exist.^{8–13} CVD inflicts a major physical and economic burden on the country.¹⁴ Population-level surveillance of the appropriate metrics serves to quantify this burden at the national level and enables contrasts between metrics such that prioritizations can be made and impactful public health and/or clinical interventions can be planned and applied.^{5,6,8,10,15,16} Continuous surveillance efforts allow assessing the collective impact of applied interventions on the metrics they are designed to improve.^{8,10,13,15,16}

IDEAL NATIONAL CVD SURVEILLANCE SYSTEM

The ideal national CVD surveillance system would cost-efficiently follow a large, representative set of US residents over extended periods while tracking a broad range of metrics such that a comprehensive picture of the nation's cardiovascular health emerges. The ideal system would have wide geographic coverage; include

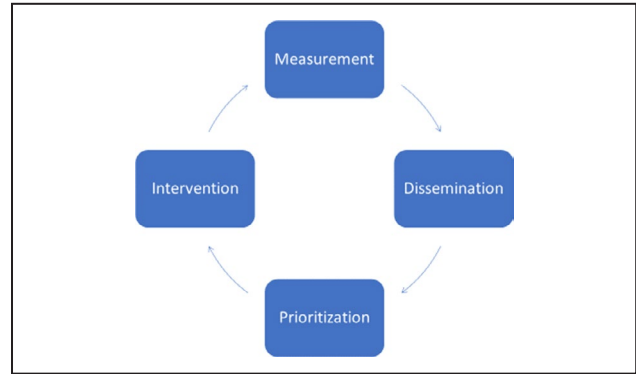


Figure 1. The cycle of surveillance.

Surveillance begins with measurement, followed by dissemination of information, prioritization of metrics for remediation, intervention designed to improve high-priority metrics, and then quantification of improvements through remeasurement in a continuous cycle.

all sociodemographic subgroups including those related to age (ie, very young and very old), race and ethnicity, and socioeconomic status; and measure a wide range of cardiovascular-related metrics including those related to primary, secondary, and tertiary prevention. The information gleaned from the system would be disseminated to interested stakeholders who would prioritize metrics for remediation, plan and apply widespread corrective interventions to improve high-priority metrics, and remeasure important metrics on a regular basis to determine progress.

CURRENT STATE OF CVD SURVEILLANCE IN THE UNITED STATES

Several ongoing efforts provide valuable surveillance metrics related to CVD, many of which have been recently summarized.⁸ Each has strengths and limitations and only a subset is mentioned here. Prospective epidemiologic cohort studies such as the ARIC (Atherosclerosis Risk in Communities), MESA (Multi-Ethnic Study of Atherosclerosis), and REGARDS (Reasons for Geographic and Racial Differences in Stroke) studies, employ rigorous methodology with extended follow-up, but often enroll closed cohorts, have limited geographic reach, tend to focus on primary prevention, and are costly to execute.^{17–19} The Centers for Disease Control and Prevention (CDC) sponsors multiple public health surveillance projects including the National Health and Nutrition Examination Survey (NHANES) and the Behavioral Risk Factor Surveillance System (BRFSS).^{20–22} These initiatives employ standardized, rigorous methodology; measure several important exposures relevant to CVD (eg, diet and exercise); and intend national representativeness. However, they involve repeat cross-sectional assessments

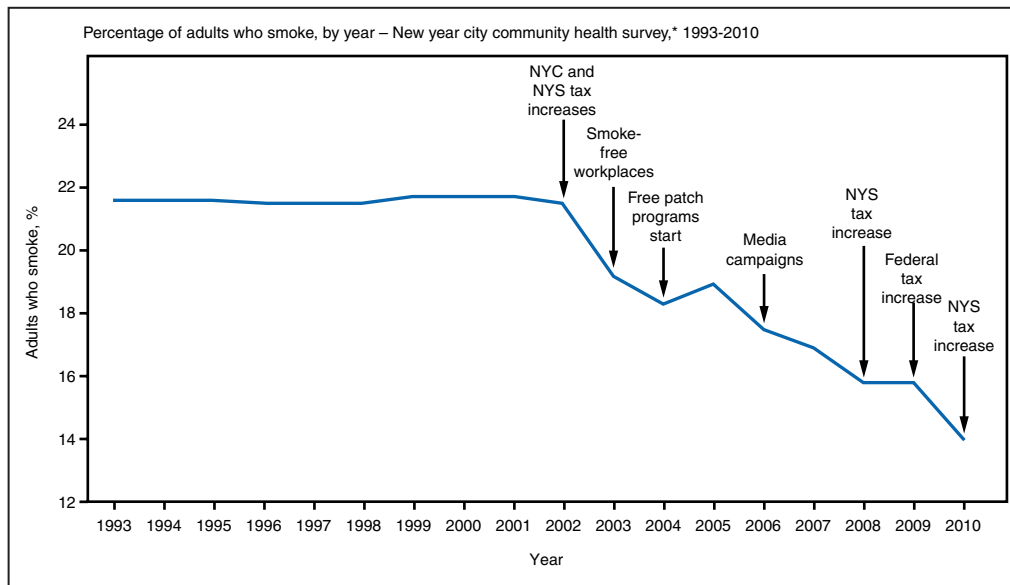


Figure 2. Surveillance depicted graphically.

The essence of surveillance can be perceived graphically with time along the horizontal axis and some health metric along the vertical: absolute values of a metric quantify the state of an issue, while contemporary values interpreted in the context of historical trends describe its trajectory (ie, improving, worsening, or stable). In this surveillance graph, smoking rates in New York City (NYC) were largely constant from 1993 to 2002, but then consistently decreased through 2010. Superimposed on the surveillance graph are the timing of various interventions designed to improve the surveilled metric.⁷ NYS indicates New York State.

without concerted follow-up of individuals across assessments, have limited sample size, have declining participation rates, and lack exclusive focus on CVD, collecting only a limited number of CVD metrics. The GBD (Global Burden of Disease) study is a noteworthy international effort aggregating data from multiple sources that permits surveillance of cardiovascular and other disease metrics in the United States and beyond.²³ Insurance claims databases serve as another possible means of CVD surveillance.^{24–27} Such databases passively track insurance plan members for extended periods with presumably exhaustive capture of clinically relevant health events during enrollment periods, but typically represent limited patient populations and collect a limited number of elements relevant to CVD.²⁸ Finally, electronic health records (EHRs) have recently emerged, with the National Patient-Centered Clinical Research Network (PCORnet) and the Health Care Systems Research Network (HCSRN) being multi-institution consortiums capable of surveillance with EHRs.^{24,29} Both insurance claims and EHRs as avenues to surveillance are discussed in greater detail below.

SURVEILLANCE WITH EHRs

Although the expected value of a national CVD surveillance system has not been questioned, financial and logistical challenges have likely been impediments to implementing such a system prospectively.^{8–13} Indeed, active surveillance requires prospective enrollment and

tracking, which, for a national CVD surveillance system with the desired geographic coverage capturing the many important facets of cardiovascular health, are likely cost-prohibitive. Passive surveillance through the repurposing of existing data sources has obvious appeal by significantly contracting the time and cost requirements. Such existing data sources capable of achieving national surveillance with long-term person tracking include insurance claims and EHRs. Both data sources originate from patients receiving services through health care organizations. Such data are often referred to as routinely collected health care data to signify their origin in the routine operations of health care delivery organizations.

Insurance claims data have existed in electronic form for multiple decades and are increasingly being used for research and public health.³⁰ Multiple features of insurance claims data are appealing for national surveillance. First, claims databases often consist of large patient numbers with few geographic constraints. Indeed, the most popular insurance claims data sources include Medicare and large, aggregated employer-based insurance databases spanning large segments of the country. Enrollment periods clearly define surveillance intervals in claims data, and identification of salient health events within that interval is presumed exhaustive. However, claims databases also have important limitations with respect to surveillance. Being largely restricted to information

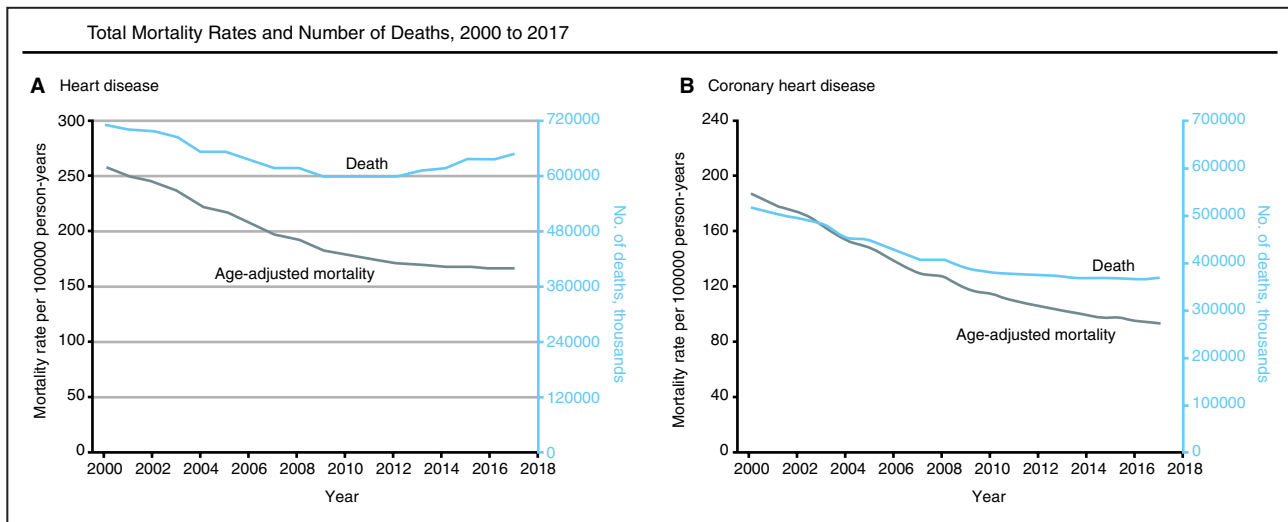


Figure 3. Cardiovascular disease surveillance.

A, Age-adjusted mortality rates per 100,000 person-years and absolute number of deaths attributable to heart disease. **B**, Age-adjusted mortality rates per 100,000 person-years and absolute number of deaths attributable to coronary heart disease. Reproduced with permission from Sidney et al.¹ Copyright ©2019, American Medical Association. All rights reserved.

necessary for reimbursement, claims data are often limited to diagnostic, procedural, and medication codes, while frequently omitting certain metrics relevant to CVD such as blood pressure (BP), smoking, body mass index, laboratory test results, and sometimes death. The availability of claims data for analysis can be subject to significant lag times, sometimes ≥ 2 years, which impedes timely delivery of corrective interventions suggested by the data—an important feature of surveillance. Finally, individual claims data sources reflect limited patient populations. Medicare primarily reflects individuals aged ≥ 65 years, while employer-based claims databases reflect working individuals who are typically younger and healthier than the broader population. The latter often have relatively brief average enrollment intervals given the rate at which Americans change employers.

The recent proliferation of EHRs in the United States has generated enthusiasm for using EHR data for research, assessing quality improvement initiatives, and monitoring public health, including establishment of a national CVD surveillance system.^{8,11,15,28,31–39} As with claims data, EHR data are a byproduct of the documented, routine interactions of patients with health care delivery organizations. While claims data are often limited to elements necessary for billing and reimbursement, EHR data tend to possess richer clinical detail including vital signs (eg, BP), lifestyle information (eg, smoking), laboratory test results (eg, lipids), and, in some cases, death—key components of a comprehensive CVD surveillance system.³¹ EHR data also tend to represent broader patient populations than isolated claims data sources and are more available in

real time. Although EHR data possess some obvious strengths as a potential resource for surveillance—and some work has been done in this area—some inherent features of EHR data cast uncertainty on its value as a surveillance data source.^{40–46} The remainder of this article goes into detail on some of the more relevant strengths and limitations of using EHR data for establishing a national CVD surveillance system (Table).

STRENGTHS OF EHRs FOR NATIONAL CVD SURVEILLANCE

Strength #1: EHRs Are Ubiquitous

In 2009, federal legislation in the United States provided financial incentives to health care organizations to implement EHRs in a meaningful way or face penalties for nonadoption.²⁸ EHR utilization rates subsequently rose to the point where EHR penetrance is universal or nearly so (Figure 4).⁴⁷ The pervasive implementation of EHRs within the health care enterprise has obvious implications for a potential national CVD surveillance system. First, the sheer volume of individuals with EHR data is immense, exceeding 100 million persons—several-fold more than could be feasibly enrolled in any active surveillance system.⁴³ Furthermore, the geographic reach of EHR-based surveillance could extend well beyond any current surveillance efforts, thus making such a system more “national.” This extended reach should enable better surveillance of subpopulations that have been more challenging to account for in volunteer-based surveillance systems. For instance, an EHR-based surveillance system in Colorado captured

Table. Strengths and Limitations of Using EHRs for National CVD Surveillance

| Strengths | Limitations |
|---|---|
| <p>EHRs are ubiquitous</p> <ul style="list-style-type: none"> In 2009, federal legislation in the United States provided financial incentives for health care organizations to implement EHRs in a meaningful way Currently, nearly all health care organizations document clinical care in an EHR; >100 million US residents have EHR data available EHR-based surveillance may generalize well to the entire country and accurately reflect the nation's demographic diversity | <p>Incomplete ascertainment of health information</p> <ul style="list-style-type: none"> In the United States, features of the health care delivery system and varying levels of patient engagement with this system affect data availability and ability to surveil To the extent health information from separate organizations cannot be linked, health profiles based on a single organization's EHR may be incomplete Surveillable subsets must be derived by organizations according to geography, insurance coverage, receipt of primary care, or other factors |
| <p>A common data infrastructure exists</p> <ul style="list-style-type: none"> Generally, a common set of data domains are documented in the medical record for describing a patient's medical profile and services rendered Data are often expressed according to universal coding systems such as <i>ICD</i>; thus, a common data infrastructure underlies the health care enterprise A common data machinery can be implemented across surveillance sites; data models developed through PCORnet and HCSRn may serve as a starting point | <p>Data quality</p> <ul style="list-style-type: none"> The nature of health care service provision within the United States creates significant interpatient variation in how much, when, and what data are collected and recorded Default information-gathering processes in usual clinical care will generate data fraught with measurement error, misclassification, and missing information More frequent health care utilizers have better data quality; EHR data reflect patient health but also how patients interact with health care organizations |
| <p>Longitudinal length and detail</p> <ul style="list-style-type: none"> A health care organization's EHR data collectively reflect a dynamic cohort—individuals enter and exit the cohort according to EHR-documented encounters Many patients within EHR systems have dense, longitudinal data; this detail can be capitalized on for achieving robust surveillance The longitudinal nature of EHRs enables measurement of certain surveillance metrics difficult to estimate through cross-sectional surveys, eg, incidence rates | <p>Vague denominators</p> <ul style="list-style-type: none"> Confidence in denominator tracking with EHRs is limited as care may have been received at outside organizations between documented encounters Younger, healthier individuals are more likely to have long, encounter-free time intervals, making them more challenging to surveil with EHRs Assumptions regarding patient observability between encounters may be necessary for a surveillance system to take advantage of the EHR's most valuable strengths |
| <p>Breadth of outcomes</p> <ul style="list-style-type: none"> EHR-based surveillance is constrained by what is measured during clinical care, yet an extensive list of outcomes and risk factors are surveillance candidates The large size of a national, EHR-based surveillance system may allow surveillance of less common conditions unachievable with current methodology EHR-based surveillance could also track more clinically oriented metrics, such as uptake of new medications, procedure use, and health care utilization | <p>Deciphering trends</p> <ul style="list-style-type: none"> Temporal trends in metrics derived from EHR-based surveillance will be sensitive to parallel changes in clinical and administrative processes Several factors could affect documentation of diagnoses over time irrespective of true changes in disease properties, eg, changing diagnostic criteria Changes in case-mix over time could affect interpretation of outcome trends, eg, more subclinical disease leading to improved outcomes |

EHR indicates electronic health record; HCSRn, Healthcare Systems Research Network; *ICD*, *International Classification of Diseases*; and PCORnet, National Patient-Centered Clinical Research Network.

greater proportions of Hispanics and people living in high-poverty neighborhoods relative to traditional survey methods.⁴⁰ In general, EHR-based surveillance cohorts should be more similar to the general population than cohorts identified for active surveillance, enhancing generalizability of findings to the entire country, and more accurately reflecting the nation's demographic diversity while providing ample sample size for less common subgroups.

Strength #2: A Common Data Infrastructure Exists

EHR data reflect the documented, routine interactions of patients with health care organizations. Generally, a common set of data domains are documented in the medical record for describing a patient's medical profile and services rendered.^{48,49} These common data domains include demographics, vital signs, clinical signs and symptoms, social history (eg, smoking),

encounter diagnoses (primary and secondary), problem lists, medications, diagnostic and interventional procedures, and laboratory test results, among others. EHR data types can be broadly divided into *structured* versus *unstructured*. Structured data refer to elements represented quantitatively (eg, systolic BP) or according to some taxonomy (see below); unstructured data refer to elements gleaned from text such as clinical notes. To facilitate administrative and other organizational functions, nonquantitative structured data are typically expressed according to universal coding systems such as the *International Classification of Diseases (ICD)* system for diagnoses and inpatient procedures, the Current Procedural Terminology (CPT) system and Healthcare Common Procedure Coding System (HCPCS) for procedures, the Logical Observation Identifiers Names and Codes (LOINC) system for medical laboratory observations, and the National Drug Code (NDC) and RxNorm systems for medications.^{28,48,49} Such structured data permit consistency

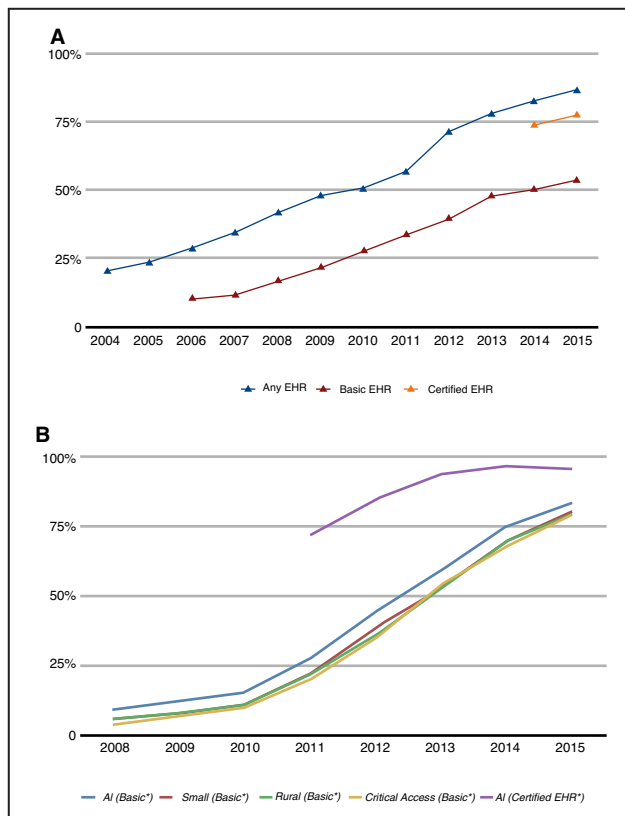


Figure 4. Implementation of electronic health records (EHRs) in the United States over time.

A, EHR uptake among office-based physicians in the United States, 2004–2015. **B**, EHR uptake among nonfederal acute care hospitals in the United States, 2008–2015.⁴⁶

of expression; thus, to the extent this expression is applied broadly across diverse health care organizations, in a sense, a common data infrastructure underlies the entire health care enterprise.^{48,50}

A national CVD surveillance system could leverage this near-universal language provided by structured EHR data into a common data machinery that could be largely standardized across surveillance sites. This data machinery would consist of, among other things, case definitions (or “computable phenotypes”)—the set of rules applied to EHR data for defining the presence of medical diagnoses. Case definitions typically incorporate structured data from ≥ 1 of the aforementioned data domains to (indirectly) identify disease positives. The most common data domain applied is diagnostic codes, but case definitions could also include some combination of medications, laboratory test results, and/or procedures specific to the diagnosis.^{51–54} For case definitions, diagnostic code lists must be developed for which there may not be universal agreement.^{55–58} Furthermore, the care setting from which information was drawn (eg, office versus hospital) and diagnostic position (primary versus secondary diagnosis) may also affect diagnostic confidence.

As apparent from the above, the development of case definitions and other data machinery components involves many decisions large and small, but the common set of data elements available among health care organizations makes creating a scalable data machinery a tenable task. Several health care organizations in the United States already apply common data machineries (also known as data models) to facilitate multisite research with EHRs, the most notable being those employed by PCORnet and the HCSRN.^{29,59,60} Other popular common data models are widely employed, including the Observational Medical Outcomes Partnership (OMOP) data model. Certain components of these existing machineries may serve as a starting point for a national EHR-based surveillance system. Notably, applying a data machinery to an organization’s EHR data enables creating a post hoc surveillance system dating back nearly to EHR inception. A common data machinery also facilitates adding future surveillance sites not involved in initial development. Importantly, developing the appropriate data machinery for achieving national surveillance is expected to require substantial effort at initial development but much less effort over time as adjustments are needed.

Strength #3: Longitudinal Length and Detail

A desirable feature of a surveillance system accommodated by many organizations’ EHRs is the longitudinal tracking of individuals over extended periods. This feature contrasts with many existing surveillance-capable systems limited to repeat cross-sectional assessments where overlap of individuals across assessments may be minimal (eg, NHANES). Any health care organization’s EHR data collectively reflect a dynamic cohort—individuals enter the cohort at an initial post-EHR inception encounter with the organization and exit at the last EHR-documented encounter. Subsequent encounters with the organization effectively serve as follow-up updates, akin to postbaseline reexaminations in epidemiologic cohort studies. When patients repeatedly patronize a particular health care organization, a longitudinal data stream develops that increases in value over time as encounters and data accrue.⁴⁸ A surveillance system can be executed by applying a specified data machinery to this data stream.⁶¹ Index dates (ie, beginning surveillance dates) are typically assigned 6 to 24 months after the first EHR-documented encounter so a more complete assessment of medical history at the index date can be attained. EHR-based surveillance is then achieved through information acquired at postindex date encounters. Many patients within EHR systems have dense, longitudinal data, including repeat measurement of vital signs and laboratory tests, newly developed diagnoses, and repeatedly

documented medications via orders and medication reconciliations. All of this detail can be capitalized on for achieving a robust, national surveillance system.

The longitudinal nature of person-level EHR data enables measurement of several valuable surveillance metrics that are difficult or impossible to track through cross-sectional evaluations such as incidence rates, repeat event rates (eg, recurrent myocardial infarction following a first myocardial infarction), and individual-level risk factor trajectories (eg, serial BP readings).⁶² The calculation of incidence rates showcases a particular strength of longitudinal EHR data. Disease incidence is often considered one of the most valuable surveillance metrics but is costly and laborious to measure through prospective studies and challenging to estimate through cross-sectional surveys. Tracking incident disease through EHRs first requires identifying a “disease-free” interval of time and encounters where there is no documentation of the disease of interest.^{61,63,64} Multiple authors have demonstrated the sensitivity of incidence rates to disease-free interval length, with the optimal interval suggested to be ≥ 2 years.^{63–66} A surveillance interval for incident disease begins where the disease-free interval ends, with incident events identified through the end of the surveillance interval. Clearly, the number of identified incident cases and thus statistical power are directly associated with the longitudinal length of the data source.

Strength #4: Breadth of Outcomes

Prioritizing cardiovascular outcomes and risk factors for interventional action via surveillance requires tracking a wide range of each so informed decisions can be made. Although EHR-based surveillance is constrained by what is measured during usual clinical care, an extensive list of outcomes (ie, “diagnoses” in EHR nomenclature) and risk factors (ie, “diagnoses,” “vital signs,” and “laboratory tests”) remain candidates for surveillance. Notably, several metrics measurable through EHR-based surveillance are also captured by existing surveillance systems such as NHANES (eg, systolic BP); comparing numerical findings across different surveillance data sources might be informative for understanding biases inherent to different methods. Importantly, the large anticipated size of a national, EHR-based surveillance system may allow surveillance of less common conditions unachievable with existing surveillance methodology. Furthermore, although insurance claims–based surveillance could match the cohort sizes achievable with EHRs, several important cardiovascular-related metrics are typically not available through claims databases, as discussed earlier, including smoking, body mass index, BP, laboratory test results such as low-density lipoprotein cholesterol, and sometimes death. Finally, in EHR-based surveillance, outcomes do not need to be prespecified from

the outset as recommended for an active, prospective surveillance system. Outcomes can be added to the system ad hoc as priorities shift, and historical trends for the added outcomes can be quantified fairly expeditiously.

By nature of the data source, an EHR-based surveillance system could also track more clinically oriented metrics typically outside the purview of conventional public health surveillance systems, such as uptake of new medications, procedure use, and health care utilization (eg, emergency department visits). Indeed, an interesting feature of the suggested surveillance system may be the ability to quantify adherence to appropriate or guideline-based care on a large scale (eg, β -blocker use following myocardial infarction). Such metrics might be valuable to a wider range of stakeholders in the clinical and policy arenas as they affect outcomes, drive costs, and reflect care delivery processes and quality of care.⁵⁰ Also worth noting are some key cardiovascular risk factors and outcomes that are not routinely tracked through EHRs, including dietary patterns, physical activity, and cause of death. Notably, cause of death is currently the only cardiovascular-related health metric surveilled nationwide through the National Center for Health Statistics (NCHS).^{1,2}

LIMITATIONS OF EHRs FOR NATIONAL CVD SURVEILLANCE

Limitation #1: Incomplete Ascertainment of Health Information

The ideal national CVD surveillance system would monitor a representative sample of the US population while thoroughly describing their cardiovascular risk factors and outcomes in a manner that generalizes to the entire nation. Multiple phenomena render this ideal a challenging goal with EHR data. EHR data inherently encompass patient populations receiving services through health care–providing organizations, and, as such, cannot surveil individuals never patronizing such organizations. Some estimates suggest up to 50% of the population has no contact with the health care industry in a given year.⁶⁷ Logically, this underrepresented group is younger and healthier.^{68,69} Then, among individuals who do utilize formal health care services, interpatient variation persists in the frequency and types of services sought based on the presence and severity of medical conditions, personal preferences, insurance coverage, ability to access health care, and other factors. Some seek services through formal health care channels infrequently or only under dire circumstances; EHR-derived information from such individuals may not contain the requisite detail for proper surveillance. Ultimately, these varying levels of engagement with the health care enterprise affect

information availability and, consequently, the ability to surveil. Thus, not only will EHR-based surveillance capture just a fraction of the underlying population of interest, the captured subset may differ importantly from the target population. This reality may create challenges when attempting to generalize EHR-based surveillance findings to the broader US population.

Another relevant impediment to a high-functioning, EHR-based, national CVD surveillance system is the disconnect of information generated at separate health care organizations. In the United States, EHR data are typically confined to the organization generating the data. Data sharing with outside organizations may be discouraged by lack of data transfer mechanisms, for privacy or competitive reasons or other factors.^{32,70–72} Importantly, individuals often seek care at multiple health care organizations (ie, care fragmentation), and, thus, to the extent health information generated at separate organizations cannot be linked, an individual's medical profile based on a single organization's EHR may be incomplete.^{32,48,72–75} This concern may be more relevant in urban areas where people naturally have more health care options. For instance, a recent study conducted at 2 urban health care centers noted that only 20% to 30% of potentially linkable health care encounters documented in Medicare claims data were observed in their respective EHRs.⁷⁶ Importantly, study patients with a smaller proportion of their health care encounters documented in their local EHR had increased EHR-based misclassification of certain diseases (ie, undercounting) compared with a Medicare gold standard.^{76,77} Although the percentage of total encounters captured by the local EHRs was disappointingly low in this study, nationally, this percentage likely varies greatly according to contextual factors, especially the number of distinct health care organizations in a geographic region, which also relates to population density. Unfortunately, most health care organizations cannot easily quantify how frequently their clientele seek care at outside organizations. Also of relevance, certain health care organizations provide only a limited set of services (eg, a standalone hospital), and such organizations' EHRs have limited capacity to fully describe a patient's medical profile. Studies have quantified the loss of information when patients are tracked through an incomplete array of care settings.^{78–81} All of these factors place limitations on which health care organizations can meaningfully participate in a national EHR-based CVD surveillance system.

As suggested above, certain features of the health care delivery system in the United States, and variability in how individuals interact with this system, create challenges for EHR-based surveillance. An “all-comers” approach to EHR surveillance (ie, include every patient in an organization's EHR)

is difficult to endorse, as many patients likely have critical data shortcomings. Thus, at issue is identifying EHR-derived “surveillable” subsets, which adequately reflect the entire nation. As surveillable implies some minimal level of EHR data availability, and, perhaps synonymously, data completeness, no definition is without limitation as completeness associates with certain patient attributes such as health consciousness and degree of illness—factors that may introduce selection biases.^{82–86} Although data completeness is a nebulous construct in the context of EHR data, some generalities may prove useful, as demonstrated here by example. Health care organizations have created empirical rules for identifying patient subsets with presumed greater data completeness. For example, the Rochester Epidemiology Project identifies individuals residing in proximity to an exhaustive group of health care organizations within a circumscribed geographic region of the Upper Midwest.⁸⁷ Kaiser Permanente, a health insurer and provider, supplies health insurance to many patients seeking care at its organization, which allows better longitudinal tracking.⁸⁸ Geisinger Health System researchers frequently employ a receipt of primary care criterion as a means of identifying patients in a closer relationship with their organization.^{89–91} Notably, the aforementioned organizations are all integrated health care delivery networks—organizations capable of providing the entire spectrum of health care services from office to hospital care and other specialized services.⁴⁸ Furthermore, 2 of these organizations predominantly provide services in relatively isolated geographic areas, where health care options are fewer and outmigration rates tend to be lower. Although such features—integrated delivery networks and geographic isolation—have obvious appeal from a surveillance standpoint, they also limit generalizability.

The underlying concepts suggested above—geographic proximity, insurance coverage through the organization, receipt of primary care—may be just a few of several definitions that organizations might apply to identify surveillable subsets. More generally, however, many organizations may need to identify this subset by applying some sort of EHR footprint rule to its patient population. That is, patients with a larger EHR footprint—more encounters, more data, more information—should be closer to the data completeness ideal and thus more appropriate for surveillance. Unfortunately, this strategy also identifies a subset predictably biased toward a less healthy group, creating difficulties when generalizing findings to the entire country.^{82–86} Ultimately, any functional EHR-derived surveillance system will need to make imperfect trade-offs between data completeness and external validity.⁸⁶

Limitation #2: Data Quality

Beyond the data completeness concerns related to care (data) fragmentation and infrequent health care utilizers discussed earlier, other EHR data quality issues pose challenges for surveillance.^{31,32,37,49,83,84,92,93} Two distinct aspects of data quality are considered related to: (1) measurement and documentation within the clinical enterprise, and (2) their implications for operationalizing a surveillance system.

The intrinsic nature of service provision within US health care organizations creates significant interpatient variation in how much, when, and what information is collected and recorded.^{32,37,94} This reality is far from the epidemiologic ideal of standardized measurement according to a structured protocol and time schedule, applied uniformly across all study participants. Much of the care provided by US health care organizations revolves around addressing specific complaints (eg, symptoms) or managing chronic disease. Accordingly, the information-gathering tactics of providers often focus on these specific needs and are thus limited.^{37,86,94} Providers vary in how information is elicited, as do patients in their willingness to disclose certain disorders.³² Clinical documentation tends to reflect only *presence* of certain conditions, while *absence* is seldom documented, making *disease absence* versus *disease status unknown* indistinguishable.^{74,94,95} Quantitative measurements made in routine practice can be subject to meaningful degrees of measurement error when clinical protocols are not adhered to. For example, systolic BP measurements made in clinical practice have greater variability and are often biased high when protocols maximizing validity are not followed, while quantitative blood glucose measurements are highly sensitive to fasting status.^{96–98} Furthermore, laboratory tests are often ordered based on demographic characteristics, disease(s) present, disease suspicion, or for monitoring treatment efficacy.⁹⁹ Thus, missing data are seldom missing completely at random, complicating application of common imputation techniques that extrapolate from nonmissing data patterns.^{31,99,100}

In short, the default information-gathering processes within the clinical enterprise will generate data fraught with measurement error, misclassification, and missingness—all of which create inferential challenges.^{54,82,86,94,101} Inevitably, more frequent health care utilizers will have better data quality—especially less misclassification and missing data—simply by having more interaction with the health care enterprise. Likewise, more ill patients will tend to have better data quality given their greater need for health care services.⁶⁹ This variable data quality directly related to health care utilization frequency can impact quantitative findings derived from the resulting data.^{102,103}

For instance, valid identification of true diagnoses is partially dependent on the frequency of health care encounters—the sensitivity and specificity of case definitions are simply enhanced when more encounters can be incorporated into the case-finding process.^{64,65,77,104} Thus, health care encounter frequency is positively associated with the number of documented diagnoses, which can distort associations between diagnoses sensitive to this phenomenon.^{103,105} How to account for the effect of these health care processes analytically remains a significant challenge, as data reflect not only patient health but also the ways patients interact with health care organizations.^{86,102}

Limitation #3: Vague Denominators

In surveillance parlance, *point denominator* refers to the set of patients purportedly being surveilled at a single point in time. Extending this definition, a surveillance interval is a continuous time interval where a surveilled patient appropriately contributes person-time to a denominator for rate calculations. These concepts are requisite components of 2 valuable surveillance metrics—prevalence proportions and incidence rates—as both rely on accurate enumeration of denominators (and numerators) for proper calculation. Correct denominator enumeration with EHRs is seemingly achieved at instances of actual patient contact with health care organizations but vaguely achieved otherwise. Confidence that accurate denominator tracking has been achieved through an EHR is weakened by the care fragmentation phenomenon described earlier—patient-level time intervals in an organization's EHR devoid of documented patient contact may not be properly attributed to the surveillance interval as care may have been received at outside organizations.

The issue of vague denominators is of greater concern when calculating metrics relying on extended surveillance intervals such as incidence rates. Notably, the bookends of an extended surveillance interval can be objectively defined within EHRs according to actual patient encounters, but whether exhaustive surveillance has been achieved during intervening periods is less clear. In EHR-based cohort studies, continuous follow-up intervals are typically applied as there is rarely a suitable empirical argument for allowing multiple, discontinuous follow-up intervals.⁹⁰ This characteristic of EHR-based epidemiology contrasts with insurance claims-based epidemiology, where beginning and ending enrollment dates define a clear surveillance interval wherein capture of salient health events within that interval is presumably exhaustive. Long, encounter-free time intervals in EHRs may be more likely among younger, healthier individuals who utilize health care services less frequently, and decrease confidence in

the ability to surveil this important subset with EHRs. Indeed, confident denominator enumeration will depend partially on context—eg, tracking metrics among patients with heart failure might be more easily accomplished as heart failure is a resource-intensive disorder requiring frequent interaction with health care organizations, but tracking rates of incident acute myocardial infarction in the young will be less easily accomplished as the relevant denominator will be undercounted with EHRs.

At many health care organizations, establishing surveillance interval criteria with EHR data might be closely tied to the surveillable subset principle previously described. Again, such criteria might involve geographic proximity to, insurance coverage or receipt of primary care through, or having a significant EHR “footprint” within the health care organization; or other criteria implying a closer, ongoing relationship between patient and the organization.^{65,106} In many instances, however, some strong assumptions regarding patient observability between encounters may be necessary if a national surveillance system is to take full advantage of the EHR’s most valuable strengths.⁶⁵

Limitation #4: Deciphering Trends

A fundamental goal of surveillance is tracking health metrics over time. Measuring trends quantifies trajectories and serves to evaluate the population-level effects of deployed interventions. Ideally, when a surveillance system reveals an adverse trend in some metric over time, a true worsening attributable to some biologically plausible underlying cause(s) is operating. Likewise, when surveillance detects improvements over time, ideally the trend has a rational explanation, such as improved risk factor control or uptake of effective therapies. Unfortunately, in EHR-based surveillance, multiple artifacts could obscure these desired interpretations.

EHR-based surveillance is a byproduct of ever-evolving clinical and administrative processes, thus quantitative findings derived from an EHR-based surveillance system will necessarily be sensitive to such process changes.¹⁰⁷ Specifically, several factors could cause increased documentation of diagnoses over time irrespective of a change in the underlying disease properties, including changing diagnostic criteria, increased utilization and/or access to diagnostic technology, more sensitive diagnostic testing such as through imaging or biomarkers, and/or simply increased coding of a condition as a result of greater awareness or for administrative (eg, financial) reasons.^{108,109} For example, a recent study reporting increasing atrial fibrillation incidence rates over time with EHR data observed a parallel increase in the use of short-term ECG, which may have partially explained the trend.⁸⁹ In the United

States, temporal trends in metrics may also be affected by the transition from *International Classification of Diseases, Ninth Revision (ICD-9)*, to *International Classification of Diseases, Tenth Revision (ICD-10)*, coding schemes in October 2015. Furthermore, medical history profiles of more recently identified disease cases may appear worse as more historical information is available for determining such profiles relative to cases identified in prior years (ie, more recent cases have more documented conditions because of longer lookback periods). All of these factors may have a true or perceived impact on changing case-mix over time, which could subsequently cloud interpretation of outcome trends. For instance, increasing disease incidence over time attributable to more subclinical (ie, less severe) cases identified via sensitive diagnostic technology could logically translate into improved outcome trends.

WHAT NEXT?

Although this document expresses optimism for EHRs in achieving national CVD surveillance, its ultimate realization will require resolution of some of the more critical limitations outlined here. Unfortunately, many of the limitations discussed have no viable solution (eg, using EHRs to surveil nonhealth care seekers) or can be addressed analytically but only after data procurement (eg, accounting for changing clinical processes when interpreting temporal trends in a health metric). Thus, in our view, the most important potentially resolvable outstanding issue remains the challenge of combining EHR data from disparate health care organizations at the person level such that comprehensive pictures of individual medical profiles can be formed. A detailed discussion of this topic was provided earlier. Apart from limiting surveillance to sites with “good enough” data, multiple possible solutions warrant consideration. First, some of the larger EHR vendors have begun to utilize their data for research.^{43,110,111} As, presumably, all data generated through a particular EHR vendor’s system are available through the particular vendor, different health care organizations using the same vendor could link their data. Having all this data housed within a single entity would also facilitate application of a common data machinery as previously described. Furthermore, linkage of EHR data at the vendor- or health care organization-level with other data sources such as insurance claims has been accomplished. Linking data sources would fill some gaps that result when limiting to a single organization’s EHR. Second, multiple health information exchanges have been formed around the country as a means to share electronic data across different health care organizations in a restricted geographic region. Although health information exchanges were developed primarily to support clinical care, the

ability to aggregate data across health care organizations participating in a health information exchange may facilitate surveillance in that region. Finally, multiple members of PCORnet and the HCSRN likely possess sufficient criteria for adequate surveillance as outlined here, with the additional benefit of an established common data infrastructure. These and other potential avenues to national surveillance could be investigated. Any proposed methodology investigated at a subset of candidate surveillance sites would ideally be compared with some gold standard such as a prospective epidemiologic cohort study tracking individuals for extended periods.

CONCLUDING COMMENTS

The promise of a national CVD surveillance system built around EHR data lies in the ubiquity of EHRs in the United States, the large number of US residents with available EHR data, wide geographic reach, and cost-efficiency relative to prospective alternatives. Many organizations' EHRs can track patients for extended time intervals with rich clinical detail, and the range of possible surveillance metrics is large. The inescapable imperfections of EHR-based surveillance include reliance on noisy data generated during usual clinical operations, exclusive inclusion of health care seekers, the current difficulties in linking EHR data across separate health care organizations, and the imposing of minimum data requirements that may overselect a less healthy subpopulation. While these limitations may restrict which health care organizations can meaningfully participate in EHR-based surveillance, a scalable system nonetheless appears feasible given the common data infrastructure across diverse health care organizations. The proposed system may ultimately consist of a restricted subset of geographically diverse yet representative health care organizations meeting necessary criteria, some of which have been suggested here. Such a system has enormous potential toward achieving its laudable aims of quantifying the cardiovascular health of our nation, providing stimulus for widespread judicious actions, and, ultimately, improving the cardiovascular health of our country.

ARTICLE INFORMATION

Received October 21, 2021; accepted February 9, 2022.

Affiliations

Geisinger Health System, Danville, PA (B.A.W., S.V., A.R.C.); Kaiser Permanente Northern California, Oakland, CA (S.S.); National Heart, Lung, and Blood Institute, Bethesda, MD (V.L.R.); University of Vermont, Burlington, VT (T.B.P.); Main Line Health, Wynnewood, PA (S.L.); State University of New York at Buffalo, NY (M.J.L.); Northwestern University, Chicago, IL (D.R.L.); University of North Carolina, Chapel Hill, NC (B.M.D.); Mayo Clinic, Rochester, MN (A.M.C.); and Essentia Health, Duluth, MN (C.P.B.).

Disclosures

None.

REFERENCES

1. Sidney S, Go AS, Jaffe MG, Solomon MD, Ambrosy AP, Rana JS. Association between aging of the US population and heart disease mortality from 2011 to 2017. *JAMA Cardiol.* 2019;4:1280–1286. doi: 10.1001/jamacardio.2019.4187
2. Sidney S, Quesenberry CP Jr, Jaffe MG, Sorel M, Nguyen-Huynh MN, Kushi LH, Go AS, Rana JS. Recent trends in cardiovascular mortality in the United States and public health goals. *JAMA Cardiol.* 2016;1:594–599. doi: 10.1001/jamacardio.2016.1326
3. US Burden of Disease Collaborators. The state of US health, 1990–2010: burden of diseases, injuries, and risk factors. *JAMA.* 2013;310:591–608. doi: 10.1001/jama.2013.13805
4. Centers for Disease Control and Prevention (CDC). Decline in deaths from heart disease and stroke—United States, 1900–1999. *MMWR Morb Mortal Wkly Rep.* 1999;48:649–656.
5. Chioloro A, Santschi V, Paccaud F. Public health surveillance with electronic medical records: at risk of surveillance bias and overdiagnosis. *Eur J Public Health.* 2013;23:350–351. doi: 10.1093/eurpub/ckt044
6. Teutsch SM, Churchill RE (Eds.) *Principles and Practice of Public Health Surveillance.* Oxford University Press; 2000.
7. Centers for Disease Control and Prevention (CDC). CDC Grand Rounds: the million hearts initiative. *MMWR Morb Mortal Wkly Rep.* 2012;61:1017–1021.
8. Roger VL, Sidney S, Fairchild AL, Howard VJ, Labarthe DR, Shay CM, Tiner AC, Whitsel LP, Rosamond WD; American Heart Association Advocacy Coordinating Committee. Recommendations for cardiovascular health and disease surveillance for 2030 and beyond: a policy statement from the American Heart Association. *Circulation.* 2020;141:e104–e119. doi: 10.1161/CIR.0000000000000756
9. Huffman MD. Maturing methods for cardiovascular disease and stroke surveillance in the United States. *JAMA Cardiol.* 2018;3:390. doi: 10.1001/jamacardio.2018.0812
10. Ford ES, Roger VL, Dunlay SM, Go AS, Rosamond WD. Challenges of ascertaining national trends in the incidence of coronary heart disease in the United States. *J Am Heart Assoc.* 2014;3:e001097. doi: 10.1161/JAHA.114.001097
11. Sidney S, Rosamond WD, Howard VJ, Luepker RV. The “heart disease and stroke statistics—2013 update” and the need for a National Cardiovascular Surveillance System. *Circulation.* 2013;127:21–23. doi: 10.1161/CIRCULATIONAHA.112.155911
12. Institute of Medicine. *A Nationwide Framework for Surveillance of Cardiovascular and Chronic Lung Diseases.* National Academies Press; 2011.
13. Goff DC Jr, Brass L, Braun LT, Croft JB, Flesch JD, Fowkes FG, Hong Y, Howard V, Huston S, Jencks SF, et al. Essential features of a surveillance system to support the prevention and management of heart disease and stroke: a scientific statement from the American Heart Association Councils on Epidemiology and Prevention, Stroke, and Cardiovascular Nursing and the Interdisciplinary Working Groups on Quality of Care and Outcomes Research and Atherosclerotic Peripheral Vascular Disease. *Circulation.* 2007;115:127–155. doi: 10.1161/CIRCULATIONAHA.106.179904
14. Virani SS, Alonso A, Aparicio HJ, Benjamin EJ, Bittencourt MS, Callaway CW, Carson AP, Chamberlain AM, Cheng S, Delling FN, et al; American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee. Heart disease and stroke statistics-2021 update: a report from the American Heart Association. *Circulation.* 2021;143:e254–e743. doi: 10.1161/CIR.0000000000000950
15. Roger VL. Heart failure epidemic: it's complicated. *Circulation.* 2018;138:25–28. doi: 10.1161/CIRCULATIONAHA.118.028478
16. Stevens GA, Alkema L, Black RE, Boerma JT, Collins GS, Ezzati M, Grove JT, Hogan DR, Hogan MC, Horton R, et al. Guidelines for accurate and transparent health estimates reporting: the GATHER statement. *Lancet.* 2016;388:e19–e23. doi: 10.1016/S0140-6736(16)30388-9
17. Blaha MJ, DeFilippis AP. Multi-Ethnic Study of Atherosclerosis (MESA): JACC focus seminar 5/8. *J Am Coll Cardiol.* 2021;77:3195–3216. doi: 10.1016/j.jacc.2021.05.006

18. Wright JD, Folsom AR, Coresh J, Sharrett AR, Couper D, Wagenknecht LE, Mosley TH Jr, Ballantyne CM, Boerwinkle EA, Rosamond WD, et al. The ARIC (Atherosclerosis Risk In Communities) Study: JACC focus seminar 3/8. *J Am Coll Cardiol*. 2021;77:2939–2959. doi: 10.1016/j.jacc.2021.04.035
19. Howard VJ, Cushman M, Pulley L, Gomez CR, Go RC, Prineas RJ, Graham A, Moy CS, Howard G. The reasons for geographic and racial differences in stroke study: objectives and design. *Neuroepidemiology*. 2005;25:135–143. doi: 10.1159/000086678
20. Hsia J, Zhao G, Town M, Ren J, Okoro CA, Pierannunzi C, Garvin W. Comparisons of estimates from the behavioral risk factor surveillance system and other National Health Surveys, 2011–2016. *Am J Prev Med*. 2020;58:e181–e190. doi: 10.1016/j.amepre.2020.01.025
21. Muntner P, Hardy ST, Fine LJ, Jaeger BC, Wozniak G, Levitan EB, Colantonio LD. Trends in blood pressure control among US adults with hypertension, 1999–2000 to 2017–2018. *JAMA*. 2020;324:1190–1200. doi: 10.1001/jama.2020.14545
22. Foti K, Wang D, Appel LJ, Selvin E. Hypertension awareness, treatment, and control in US adults: trends in the hypertension control cascade by population subgroup (National Health and Nutrition Examination Survey, 1999–2016). *Am J Epidemiol*. 2019;188:2165–2174. doi: 10.1093/aje/kwz177
23. Roth GA, Mensah GA, Johnson CO, Addolorato G, Ammirati E, Baddour LM, Barengo NC, Beaton AZ, Benjamin EJ, Benziger CP, et al; GBD-NHLBI-JACC Global Burden of Cardiovascular Diseases Writing Group. Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the GBD 2019 study. *J Am Coll Cardiol*. 2020;76:2982–3021. doi: 10.1016/j.jacc.2020.11.010
24. Reynolds K, Go AS, Leong TK, Boudreau DM, Cassidy-Bushrow AE, Fortmann SP, Goldberg RJ, Gurwitz JH, Magid DJ, Margolis KL, et al. Trends in incidence of hospitalized acute myocardial infarction in the Cardiovascular Research Network (CVRN). *Am J Med*. 2017;130:317–327. doi: 10.1016/j.amjmed.2016.09.014
25. Piccini JP, Hammill BG, Sinner MF, Jensen PN, Hernandez AF, Heckbert SR, Benjamin EJ, Curtis LH. Incidence and prevalence of atrial fibrillation and associated mortality among Medicare beneficiaries, 1993–2007. *Circ Cardiovasc Qual Outcomes*. 2012;5:85–93. doi: 10.1161/CIRCOUTCOMES.111.962688
26. Go AS, Magid DJ, Wells B, Sung SH, Cassidy-Bushrow AE, Greenlee RT, Langer RD, Lieu TA, Margolis KL, Masoudi FA, et al. The Cardiovascular Research Network: a new paradigm for cardiovascular quality and outcomes research. *Circ Cardiovasc Qual Outcomes*. 2008;1:138–147. doi: 10.1161/CIRCOUTCOMES.108.801654
27. Magid DJ, Gurwitz JH, Rumsfeld JS, Go AS. Creating a research data network for cardiovascular disease: the CVRN. *Expert Rev Cardiovasc Ther*. 2008;6:1043–1045. doi: 10.1586/14779072.6.8.1043
28. Shortliffe EH, Cimino JJ. *Biomedical Informatics: Computer Applications in Health Care and Biomedicine*. Springer; 2014.
29. Pletcher MJ, Fontil V, Carton T, Shaw KM, Smith M, Choi S, Todd J, Chamberlain AM, O'Brien EC, Faulkner M, et al. The PCORnet Blood Pressure Control Laboratory: a platform for surveillance and efficient trials. *Circ Cardiovasc Qual Outcomes*. 2020;13:e006115. doi: 10.1161/CIRCOUTCOMES.119.006115
30. Jollis JG, Ancukiewicz M, DeLong ER, Pryor DB, Muhlbaier LH, Mark DB. Discordance of databases designed for claims payment versus clinical information systems: implications for outcomes research. *Ann Intern Med*. 1993;119:844–850. doi: 10.7326/0003-4819-119-8-199310150-00011
31. Hemingway H, Asselbergs FW, Danesh J, Dobson R, Maniadakis N, Maggioni A, van Thiel GJ, Cronin M, Brobert G, Vardas P, et al. Big data from electronic health records for early and late translational cardiovascular research: challenges and potential. *Eur Heart J*. 2018;39:1481–1495. doi: 10.1093/eurheartj/ehx487
32. Raman SR, Curtis LH, Temple R, Andersson T, Ezekowitz J, Ford I, James S, Marsolo K, Mirhaji P, Rocca M, et al. Leveraging electronic health records for clinical research. *Am Heart J*. 2018;202:13–19. doi: 10.1016/j.ahj.2018.04.015
33. Birkhead GS. Successes & continued challenges of electronic health records for chronic disease surveillance. *Am J Public Health*. 2017;107:1365–1367. doi: 10.2105/AJPH.2017.303938
34. Perlman SE, McVeigh KH, Thorpe LE, Jacobson L, Greene CM, Gwynn RC. Innovations in population health surveillance: using electronic health records for chronic disease surveillance. *Am J Public Health*. 2017;107:853–857. doi: 10.2105/AJPH.2017.303813
35. Vasan RS, Benjamin EJ. The future of cardiovascular epidemiology. *Circulation*. 2016;133:2626–2633. doi: 10.1161/CIRCULATIONAHA.116.023528
36. Paul MM, Greene CM, Newton-Dame R, Thorpe LE, Perlman SE, McVeigh KH, Gourevitch MN. The state of population health surveillance using electronic health records: a narrative review. *Popul Health Manag*. 2015;18:209–216. doi: 10.1089/pop.2014.0093
37. Roger VL, Boerwinkle E, Crapo JD, Douglas PS, Epstein JA, Granger CB, Greenland P, Kohane I, Psaty BM. Strategic transformation of population studies: recommendations of the working group on epidemiology and population sciences from the National Heart, Lung, and Blood Advisory Council and Board of External Experts. *Am J Epidemiol*. 2015;181:363–368. doi: 10.1093/aje/kwv011
38. Sorlie PD, Bild DE, Lauer MS. Cardiovascular epidemiology in a changing world—challenges to investigators and the National Heart, Lung, and Blood Institute. *Am J Epidemiol*. 2012;175:597–601. doi: 10.1093/aje/kws138
39. Safran C, Bloomrosen M, Hammond WE, Labkoff S, Markel-Fox S, Tang PC, Detmer DE. Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper. *J Am Med Inform Assoc*. 2007;14:1–9. doi: 10.1197/jamia.M2273
40. Scott KA, Bacon E, Kraus EM, Steiner JF, Budney G, Bondy J, McEwen LD, Davidson AJ. Evaluating population coverage in a regional distributed data network: implications for electronic health record-based public health surveillance. *Public Health Rep*. 2020;135:621–630. doi: 10.1177/0033354920941158
41. Bacon E, Budney G, Bondy J, Kahn MG, McCormick EV, Steiner JF, Tabano D, Waxmonsky JA, Zucker R, Davidson AJ. Developing a regional distributed data network for surveillance of chronic health conditions: the Colorado health observation regional data service. *J Public Health Manag Pract*. 2019;25:498–507. doi: 10.1097/PHH.0000000000000810
42. Horth RZ, Wagstaff S, Jeppson T, Patel V, McClellan J, Bissonette N, Friedrichs M, Dunn AC. Use of electronic health records from a statewide health information exchange to support public health surveillance of diabetes and hypertension. *BMC Public Health*. 2019;19:1106. doi: 10.1186/s12889-019-7367-z
43. Tarabichi Y, Goyden J, Liu R, Lewis S, Sudano J, Kaelber DC. A step closer to nationwide electronic health record-based chronic disease surveillance: characterizing asthma prevalence and emergency department utilization from 100 million patient records through a novel multisite collaboration. *J Am Med Inform Assoc*. 2020;27:127–135. doi: 10.1093/jamia/ocz172
44. Newton-Dame R, McVeigh KH, Schreibstein L, Perlman S, Lurie-Moroni E, Jacobson L, Greene C, Snell E, Thorpe LE. Design of the New York City MacroScope: innovations in population health surveillance using electronic health records. *EGEMS (Wash DC)*. 2016;4:1265.
45. Bernstein JA, Friedman C, Jacobson P, Rubin JC. Ensuring public health's future in a national-scale learning health system. *Am J Prev Med*. 2015;48:480–487. doi: 10.1016/j.amepre.2014.11.013
46. Kho AN, Hynes DM, Goel S, Solomones AE, Price R, Hota B, Sims SA, Bahroos N, Angulo F, Trick WE, et al. CAPriCORN: Chicago Area Patient-Centered Outcomes Research Network. *J Am Med Inform Assoc*. 2014;21:607–611. doi: 10.1136/amiainjnl-2014-002827
47. Office of the National Coordinator for Health Information Technology. 'Non-federal Acute Care Hospital Electronic Health Record Adoption,' Health IT Quick-Stat #47. September 2017. Available at: dashboard.healthit.gov/quickstats/pages/FIG-Hospital-EHR-Adoption.php. Accessed February 10, 2020.
48. Abdelhak M, Hanken MA, eds. *Health Information: Management of a Strategic Resource*. 5th ed. Elsevier; 2016.
49. Hoyt RE, Yoshihashi A. *Health Informatics: Practical Guide for Healthcare and Information Technology Professionals*. 6th ed. 2014. 9781304791108. Available at: <http://lulu.com>. Accessed May 3, 2014.
50. Califf RM, Platt R. Embedding cardiovascular research into practice. *JAMA*. 2013;310:2037–2038. doi: 10.1001/jama.2013.282771
51. Hripicask G, Albers DJ. Next-generation phenotyping of electronic health records. *J Am Med Inform Assoc*. 2013;20:117–121. doi: 10.1136/amiainjnl-2012-001145
52. Pathak J, Kho AN, Denny JC. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. *J Am Med Inform Assoc*. 2013;20:e206–e211. doi: 10.1136/amiainjnl-2013-002428

53. Richesson RL, Hammond WE, Nahm M, Wixted D, Simon GE, Robinson JG, Bauck AE, Cifelli D, Smerek MM, Dickerson J, et al. Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH health care systems collaboratory. *J Am Med Inform Assoc.* 2013;20:e226–e231. doi: 10.1136/amiajnl-2013-001926
54. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nat Rev Genet.* 2012;13:395–405. doi: 10.1038/nrg3208
55. Tate AR, Dungey S, Glew S, Beloff N, Williams R, Williams T. Quality of recording in diabetes in the UK: how does the GP's method of coding clinical data affect incidence estimates? Cross-sectional study using the CPRD database. *BMJ Open.* 2017;7:e012905. doi: 10.1136/bmjopen-2016-012905
56. Watson J, Nicholson BD, Hamilton W, Price S. Identifying clinical features in primary care electronic health record studies: methods for codelist development. *BMJ Open.* 2017;7:e019637. doi: 10.1136/bmjopen-2017-019637
57. Williams R, Kontopantelis E, Buchan I, Peek N. Clinical code set engineering for reusing EHR data for research: a review. *J Biomed Inform.* 2017;70:1–13. doi: 10.1016/j.jbi.2017.04.010
58. Springate DA, Kontopantelis E, Ashcroft DM, Olier I, Parisi R, Chamapiwa E, Reeves D. ClinicalCodes: an online clinical codes repository to improve the validity and reproducibility of research using electronic medical records. *PLoS One.* 2014;9:e99825. doi: 10.1371/journal.pone.0099825
59. Forrest CB, McTigue KM, Hernandez AF, Cohen LW, Cruz H, Haynes K, Kaushal R, Kho AN, Marsolo KA, Nair VP, et al. PCORnet® 2020: current state, accomplishments, and future directions. *J Clin Epidemiol.* 2021;129:60–67. doi: 10.1016/j.jclinepi.2020.09.036
60. Rahm AK, Ladd I, Burnett-Hartman AN, Epstein MM, Lowery JT, Lu CY, Pawloski PA, Sharaf RN, Liang SY, Hunter JE. The Healthcare Systems Research Network (HCSRN) as an environment for dissemination and implementation research: a case study of developing a multi-site research study in precision medicine. *EGEMS (Wash DC).* 2019;7:16.
61. Schneeweiss S, Rassen JA, Brown JS, Rothman KJ, Happe L, Arlett P, Dal Pan G, Goettisch W, Murk W, Wang SV. Graphical depiction of longitudinal study designs in health care databases. *Ann Intern Med.* 2019;170:398–406. doi: 10.7326/M18-3079
62. Shah BR, Drozda J, Peterson ED. Leveraging observational registries to inform comparative effectiveness research. *Am Heart J.* 2010;160:8–15. doi: 10.1016/j.ahj.2010.04.012
63. Roberts AW, Dusetzina SB, Farley JF. Revisiting the washout period in the incident user study design: why 6–12 months may not be sufficient. *J Comp Eff Res.* 2015;4:27–35. doi: 10.2217/ce.14.53
64. Griffiths RI, O'Malley CD, Herbert RJ, Danese MD. Misclassification of incident conditions using claims data: impact of varying the period used to exclude pre-existing disease. *BMC Med Res Methodol.* 2013;13:32. doi: 10.1186/1471-2288-13-32
65. Rassen JA, Bartels DB, Schneeweiss S, Patrick AR, Murk W. Measuring prevalence and incidence of chronic conditions in claims and electronic health record database. *Clin Epidemiol.* 2019;11:1–15. doi: 10.2147/CLEP.S181242
66. Chen G, Lix L, Tu K, Hemmelgarn BR, Campbell NR, McAlister FA, Quan H. Influence of using different databases and 'look back' intervals to define comorbidity profiles for patients with newly diagnosed hypertension: implications for health services researchers. *PLoS One.* 2016;11:1–11. doi: 10.1371/journal.pone.0162074
67. Fuchs VR. Major concepts of health care economics. *Ann Intern Med.* 2015;162:380–383. doi: 10.7326/M14-1183
68. Dixon BE, Gibson PJ, Frederickson Comer K, Rosenman M. Measuring population health using electronic health records: exploring biases and representativeness in a community health information exchange. *Stud Health Technol Inform.* 2015;216:1009.
69. Rusanov A, Weiskopf NG, Wang S, Weng C. Hidden in plain sight: bias towards sick patients when sampling patients with sufficient electronic health record data for research. *BMC Med Inform Decis Mak.* 2014;14:51. doi: 10.1186/1472-6947-14-51
70. Everson J, Patel V, Adler-Milstein J. Information blocking remains prevalent at the start of 21st Century Cures Act: results from a survey of health information exchange organizations. *J Am Med Inform Assoc.* 2021;28:727–732. doi: 10.1093/jamia/ocaa323
71. Greene JA, Lea AS. Digital futures past—the long arc of big data in medicine. *N Engl J Med.* 2019;381:480–485. doi: 10.1056/NEJMp1817674
72. D'Avolio LW, Farwell WR, Fiore LD. Comparative effectiveness research and medical informatics. *Am J Med.* 2010;123:e32–e37. doi: 10.1016/j.amjmed.2010.10.006
73. Frisse ME, Misulis KE. *Essentials of Clinical Informatics.* Oxford University Press; 2019.
74. Xian Y, Hammill BG, Curtis LH. Data sources for heart failure comparative effectiveness research. *Heart Fail Clin.* 2013;9:1–13. doi: 10.1016/j.hfc.2012.09.001
75. Wei WQ, Leibson CL, Ransom JE, Kho AN, Caraballo PJ, Chai HS, Yawn BP, Pacheco JA, Chute CG. Impact of data fragmentation across healthcare centers on the accuracy of a high-throughput clinical phenotyping algorithm for specifying subjects with type 2 diabetes mellitus. *J Am Med Inform Assoc.* 2012;19:219–224. doi: 10.1136/amiajnl-2011-000597
76. Lin KJ, Rosenthal GE, Murphy SN, Mandl KD, Jin Y, Glynn RJ, Schneeweiss S. External validation of an algorithm to identify patients with high data-completeness in electronic health records for comparative effectiveness research. *Clin Epidemiol.* 2020;12:133–141. doi: 10.2147/CLEP.S232540
77. Lin KJ, Glynn RJ, Singer DE, Murphy SN, Lii J, Schneeweiss S. Out-of-system care and recording of patient characteristics critical for comparative effectiveness research. *Epidemiology.* 2018;29:356–363. doi: 10.1097/EDE.0000000000000794
78. Kalbaugh CA, Kucharska-Newton A, Wruck L, Lund JL, Selvin E, Matsushita K, Bengtson LGS, Heiss G, Loehr L. Peripheral artery disease prevalence and incidence estimated from both outpatient and inpatient settings among Medicare fee-for-service beneficiaries in the Atherosclerosis Risk in Communities (ARIC) Study. *J Am Heart Assoc.* 2017;6:e003796. doi: 10.1161/JAHA.116.003796
79. Camplain R, Kucharska-Newton A, Cuthbertson CC, Wright JD, Alonso A, Heiss G. Misclassification of incident hospitalized and outpatient heart failure in administrative claims data: the Atherosclerosis Risk in Communities (ARIC) study. *Pharmacoepidemiol Drug Saf.* 2017;26:421–428. doi: 10.1002/pds.4162
80. Herrett E, Shah AD, Boggon R, Denaxas S, Smeeth L, van Staa T, Timmis A, Hemingway H. Completeness and diagnostic validity of recording acute myocardial infarction events in primary care, hospital care, disease registry, and national mortality records: cohort study. *BMJ.* 2013;346:f2350. doi: 10.1136/bmj.f2350
81. Robitaille C, Bancej C, Dai S, Tu K, Rasali D, Blais C, Plante C, Smith M, Svenson LW, Reimer K, et al. Surveillance of ischemic heart disease should include physician billing claims: population-based evidence from administrative health data across seven Canadian provinces. *BMC Cardiovasc Disord.* 2013;13:88. doi: 10.1186/1471-2261-13-88
82. Weber GM, Adams WG, Bernstam EV, Bickel JP, Fox KP, Marsolo K, Raghavan VA, Turchin A, Zhou X, Murphy SN, et al. Biases introduced by filtering electronic health records for patients with "complete data". *J Am Med Inform Assoc.* 2017;24:1134–1141. doi: 10.1093/jamia/ocx071
83. Weiskopf NG, Rusanov A, Weng C. Sick patients have more data: the non-random completeness of electronic health records. *AMIA Annu Symp Proc.* 2013;2013:1472–1477.
84. Weiskopf NG, Hripcsak G, Swaminathan S, Weng C. Defining and measuring completeness of electronic health records for secondary use. *J Biomed Inform.* 2013;46:830–836. doi: 10.1016/j.jbi.2013.06.010
85. Albers DJ, Hripcsak G. A statistical dynamics approach to the study of human health data: resolving population scale diurnal variation in laboratory data. *Phys Lett.* 2010;374:1159–1164. doi: 10.1016/j.physleta.2009.12.067
86. Stewart WF, Shah NR, Selna MJ, Paulus RA, Walker JM. Bridging the inferential gap: the electronic health record and clinical evidence: emerging tools can help physicians bridge the gap between knowledge they possess and knowledge they do not. *Health Aff.* 2007;26:w181–w191. doi: 10.1377/hlthaff.26.2.w181
87. St Sauver JL, Grossardt BR, Leibson CL, Yawn BP, Melton LJ III, Rocca WA. Generalizability of epidemiological findings and public health decisions: an illustration from the Rochester Epidemiology Project. *Mayo Clin Proc.* 2012;87:151–160. doi: 10.1016/j.mayocp.2011.11.009
88. Palzes VA, Weisner C, Chi FW, Kline-Simon AH, Satre DD, Hirschtritt ME, Ghadiali M, Sterling S. The Kaiser Permanente Northern California Adult Alcohol Registry, an electronic health records-based registry of patients with alcohol problems: development and implementation. *JMIR Med Inform.* 2020;8:e19081. doi: 10.2196/19081
89. Williams BA, Chamberlain AM, Blankenship JC, Hylek EM, Voyce S. Trends in atrial fibrillation incidence rates within an integrated

- health care delivery system, 2006 to 2018. *JAMA Netw Open*. 2020;3:e2014874. doi: 10.1001/jamanetworkopen.2020.14874
90. Williams BA, Chagin KM, Bash LD, Boden WE, Duval S, Fowkes FGR, Mahaffey KW, Patel MD, D'Agostino RB, Peterson ED, et al. External validation of the TIMI risk score for secondary cardiovascular events among patients with recent myocardial infarction. *Atherosclerosis*. 2018;272:80–86. doi: 10.1016/j.atherosclerosis.2018.03.026
 91. Williams BA, Honushefsky AM, Berger PB. Temporal trends in the incidence, prevalence, and survival of patients with atrial fibrillation from 2004–2016. *Am J Cardiol*. 2017;120:1961–1965. doi: 10.1016/j.amjcard.2017.08.014
 92. Brown JS, Kahn M, Toh S. Data quality assessment for comparative effectiveness research in distributed data networks. *Med Care*. 2013;51:S22–S29. doi: 10.1097/MLR.0b013e31829b1e2c
 93. Kahn MG, Batson D, Schilling LM. Data model considerations for clinical effectiveness researchers. *Med Care*. 2012;50(suppl):S60–S67. doi: 10.1097/MLR.0b013e318259bfff
 94. Saczynski JS, McManus DD, Goldberg RJ. Commonly used data-collection approaches in clinical research. *Am J Med*. 2013;126:946–950. doi: 10.1016/j.amjmed.2013.04.016
 95. Wells BJ, Nowacki AS, Chagin K, Kattan MW. Strategies for handling missing data in electronic health record derived data. *EGEMS (Wash DC)*. 2013;1:1035.
 96. Muntner P, Einhorn PT, Cushman WC, Whelton PK, Bello NA, Drawz PE, Green BB, Jones DW, Juraschek SP, Margolis KL, et al. Blood pressure assessment in adults in clinical practice and clinic-based research. *J Am Coll Cardiol*. 2019;73:317–335. doi: 10.1016/j.jacc.2018.10.069
 97. Powers BJ, Olsen MK, Smith VA, Woolson RF, Bosworth HB, Oddone EZ. Measuring blood pressure for decision making and quality reporting: where and how many measures? *Ann Intern Med*. 2011;154:781–788. doi: 10.7326/0003-4819-154-12-201106210-00005
 98. Gore MO, McGuire DK. A test in context: hemoglobin A_{1c} and cardiovascular disease. *J Am Coll Cardiol*. 2016;68:2479–2486. doi: 10.1016/j.jacc.2016.08.070
 99. Petersen I, Welch CA, Nazareth I, Walters K, Marston L, Morris RW, Carpenter JR, Morris TP, Pham TM. Health indicator recording in UK primary care electronic health records: key implications for handling missing data. *Clin Epidemiol*. 2019;11:157–167. doi: 10.2147/CLEP.S191437
 100. Schneeweiss S, Rassen JA, Glynn RJ, Myers J, Daniel GW, Singer J, Solomon DH, Kim S, Rothman KJ, Liu J, et al. Supplementing claims data with outpatient laboratory test results to improve confounding adjustment in effectiveness studies of lipid-lowering treatments. *BMC Med Res Methodol*. 2012;12:180. doi: 10.1186/1471-2288-12-180
 101. Coorevits P, Sundgren M, Klein GO, Bahr A, Claerhout B, Daniel C, Dugas M, Dupont D, Schmidt A, Singleton P, et al. Electronic health records: new opportunities for clinical research. *J Intern Med*. 2013;274:547–560. doi: 10.1111/joim.12119
 102. Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *BMJ*. 2018;361:k1479. doi: 10.1136/bmj.k1479
 103. Goldstein BA, Bhavsar NA, Phelan M, Pencina MJ. Controlling for informed presence bias due to the number of health encounters in an electronic health record. *Am J Epidemiol*. 2016;184:847–855. doi: 10.1093/aje/kww112
 104. Wei WQ, Leibson CL, Ransom JE, Kho AN, Chute CG. The absence of longitudinal data limits the accuracy of high-throughput clinical phenotyping for identifying type 2 diabetes mellitus subjects. *Int J Med Inform*. 2013;82:239–247. doi: 10.1016/j.ijmedinf.2012.05.015
 105. Wennberg JE, Staiger DO, Sharp SM, Gottlieb DJ, Bevan G, McPherson K, Welch HG. Observational intensity bias associated with illness adjustment: cross sectional analysis of insurance claims. *BMJ*. 2013;346:f549. doi: 10.1136/bmj.f549
 106. Bellin E. *How to Ask and Answer Questions Using Electronic Medical Record Data*. CreateSpace Independent Publishing Platform; 2017.
 107. Ehrenstein V, Nielsen H, Pedersen AB, Johnsen SP, Pedersen L. Clinical epidemiology in the era of big data: new opportunities, familiar challenges. *Clin Epidemiol*. 2017;9:245–250. doi: 10.2147/CLEP.S129779
 108. van Oostrom SH, Gijsen R, Stirbu I, Korevaar JC, Schellevis FG, Picavet HS, Hoeymans N. Time trends in prevalence of chronic diseases and multimorbidity not only due to aging: data from general practices and health surveys. *PLoS One*. 2016;11:e0160264. doi: 10.1371/journal.pone.0160264
 109. Glasziou P, Moynihan R, Richards T, Godlee F. Too much medicine; too little care. *BMJ*. 2013;347:f4247. doi: 10.1136/bmj.f4247
 110. Domino JS, Lundy P, Glynn EF, Partington M. Estimating the prevalence of neurosurgical interventions in adults with spina bifida using the Health Facts data set: implications for transition planning and the development of adult clinics. *J Neurosurg Pediatr*. 2021;1–8. doi: 10.3171/2021.10.PEDS21293
 111. Rodriguez F, Lee DJ, Gad SS, Santos MP, Beetel RJ, Vasey J, Bailey RA, Patel A, Blais J, Weir MR, et al. Real-world diagnosis and treatment of diabetic kidney disease. *Adv Ther*. 2021;38:4425–4441. doi: 10.1007/s12325-021-01777-9