SCHOOL OF LAW

CASE WESTERN RESERVE UNIVERSITY

The Use and Misuse of Biomedical Data: Is Bigger Really Better?

Sharona Hoffman
and Andy Podgurski

Case Research Paper Series in Legal Studies
Working Paper 2013-10
November (revised) 2013

American Journal of Law & Medicine, 39 (2013): 497-538

The Use and Misuse of Biomedical Data: Is Bigger Really Better?

Sharona Hoffman[†] and Andy Podgurski[††]

CONTENTS

I. INTRODUCTION 498
II. BACKGROUND 502
A. Ongoing Initiatives to Create Biomedical Databases 503
B. Using Biomedical Databases and Data Networks 506
1. Scientific Discovery 506
2. Quality Assessment and Improvement 509
3. Post-Marketing Surveillance of Drugs and Devices 510
4. Public Health Initiatives 512
5. Litigation 513
III. LIMITATIONS OF BIOMEDICAL DATABASES 515
A. Data Entry Errors 515
B. Incomplete or Fragmented Data 517
C. Data Coding, Standardization, and Extraction 518
D. Errors Due to Software Failures 520
IV. THE CHALLENGES OF BIAS AND CAUSAL INFERENCE 521
A. Selection Bias 521
B. Confounding Bias 523
C. Measurement Bias 525
V. BIOMEDICAL DATABASES AND PERSONAL INFERENCE 525
VI. SOLUTIONS 527
A. Technology Improvements 527
B. Human Hands 530
1. Data Quality Assessment 530
2. Causal Inference Techniques 532
C. Education and Prevention of Research Misuse 536

[†] Edgar A. Hahn Professor of Law and Professor of Bioethics, Co-Director of Law-Medicine Center, Case Western Reserve University School of Law; B.A., Wellesley College; J.D., Harvard Law School; LL.M. in Health Law, University of Houston.

[††] Professor of Electrical Engineering and Computer Science, Case Western Reserve University. B.S., M.S., Ph.D., University of Massachusetts. The authors wish to thank Jessie Hill, Maxwell Mehlman, Dale Nance, Andrew Pollis, and Cassandra Robertson for their very helpful comments on drafts of this paper. The paper was presented at workshops at the Centers for Disease Control and Prevention, MetroHealth Hospital in Cleveland, and Case Western Reserve University School of Law and School of Medicine. The authors appreciate the helpful input they received through these presentations. They are also grateful for the skilled research assistance of Corbin Santo.

VII. CONCLUSION 537

Very large biomedical research databases, containing electronic health records (EHR) and genomic data from millions of patients, have been heralded recently for their potential to accelerate scientific discovery and produce dramatic improvements in medical treatments. Research enabled by these databases may also lead to profound changes in law, regulation, social policy, and even litigation strategies. Yet, is "big data" necessarily better data?

This paper makes an original contribution to the legal literature by focusing on what can go wrong in the process of biomedical database research and what precautions are necessary to avoid critical mistakes. We address three main reasons for approaching such research with care and being cautious in relying on its outcomes for purposes of public policy or litigation. First, the data contained in biomedical databases is surprisingly likely to be incorrect or incomplete. Second, systematic biases, arising from both the nature of the data and the preconceptions of investigators, are serious threats to the validity of research results, especially in answering causal questions. Third, data mining of biomedical databases makes it easier for individuals with political, social, or economic agendas to generate ostensibly scientific but misleading research findings for the purpose of manipulating public opinion and swaying policymakers.

In short, this paper sheds much-needed light on the problems of credulous and uninformed acceptance of research results derived from biomedical databases. An understanding of the pitfalls of big data analysis is of critical importance to anyone who will rely on or dispute its outcomes, including lawyers, policymakers, and the public at large. The Article also recommends technical, methodological, and educational interventions to combat the dangers of database errors and abuses.

I. INTRODUCTION

In 2009, the *Journal of Psychiatric Research* published an article that linked abortion to psychiatric disorders.[1] The researchers examined "national data sets with reproductive history and mental health variables" to formulate their findings.[2] The study was widely cited among abortion opponents,[3] and several states enacted legislation requiring that women seeking abortions receive counseling that includes warnings about potential long-term mental health problems.[4] In 2012, however, the study was discredited by scientists who scrutinized its design and found that it was severely flawed.[5] The original researchers neglected to compare women with

[1] Priscilla K. Coleman et al., *Induced Abortion and Anxiety, Mood, and Substance Abuse Disorders: Isolating the Effects of Abortion in the National Comorbidity Survey*, 43 J. PSYCHIATRIC RES. 770, 773 (2009), *available at* http://www.journalofpsychiatricresearch.com/article/S0022-3956(08)00238-0/abstract.

[2] *Id.* at 770.

[3] Sharon Begley, *Journal Disavows Study Touted by U.S. Abortion Foes*, REUTERS (Mar. 7, 2012, 3:11 P.M.), http://www.reuters.com/article/2012/03/07/us-usa-abortion-psychiatry-idUSTRE8261UD20120307 (stating that the study had been "widely cited by legislators and advocates to argue that abortion raises a woman's risk of mental illness and to push for laws requiring providers" to inform women of this danger).

[4] *Counseling and Waiting Periods for Abortion*, STATE POLICIES IN BRIEF (Guttmacher Inst., New York, N.Y.), May 1, 2012, at 1, 3, *available at* http://www.guttmacher.org/statecenter/spibs/spib_MWPA.pdf.

[5] Ronald C. Kessler & Alan F. Schatzberg, Reply to Letter to the Editor, *Commentary on Abortion Studies of Steinberg and Finer (Social Science & Medicine 2011; 72:72–82) and Coleman*

unplanned pregnancies who did have abortions to those who did not and failed to focus only on mental health problems that manifested after terminated pregnancies.[6] Thus, what appeared to be solid scientific evidence turned out not to be so, but not before having significant impact on some state legislatures.

The accelerating transition from paper medical files to electronic health records (EHR) systems[7] is facilitating the creation of large health information databases.[8] In the future, these may include significant genetic information because many EHRs will contain or be linked to genetic data about patients.[9] In addition, scientists are constructing large databases from genome sequencing projects.[10] Biomedical databases can serve as invaluable resources for researchers. There is justified enthusiasm about the potential for research using them to yield improved treatments and beneficial policy changes, and we have elaborated on the promise of such research in prior work.[11] Computer processing of digitized records permits fast and relatively inexpensive data analysis and synthesis, which can enable scientific discoveries and ultimately affect public policy and law.[12] Notably, the size and scope of integrated biomedical databases may allow researchers to overcome certain problems they encounter with smaller-scale studies, such as unrepresentative study groups and insufficient statistical power or precision.[13]

EHR-based research is likely to become increasingly important because of several federally sponsored initiatives. These include comparative effectiveness research that is promoted by the Patient Protection and Affordable Care Act of 2010[14] and post-marketing surveillance authorized by the Food and Drug Administration Amendments Act of 2007.[15]

Anyone considering the outcomes of record-based studies, however, must recognize the shortcomings of contemporary EHR and genomic data and the

---

(Journal of Psychiatric Research 2009;43:770–6 & Journal of Psychiatric Research 2011;45:1133–4), 46 J. PSYCHIATRIC RES. 410, 410-11 (2012).

[6] *Id.* at 410.

[7] David Blumenthal & Marilyn Tavenner, *The "Meaningful Use" Regulation for Electronic Health Records*, 363 NEW ENG. J. MED. 501, 501 (2010). Others may call EHRs electronic *medical* records (EMR). For the sake of simplicity, we use "EHR" consistently throughout and do not believe there is a substantive distinction between the two terms. *See* Peter Garrett &Joshua J. Seidman, *EMR vs EHR–What Is the Difference?*, HEALTHITBUZZ (Jan. 4, 2011, 12:07 P.M.), http://www.healthit.gov/buzz-blog/electronic-health-and-medical-records/emr-vs-ehr-difference/ ("Some people use the terms 'electronic medical record' and 'electronic health record' (or 'EMR' and 'EHR') interchangeably. But here at the Office of the National Coordinator for Health Information Technology (ONC), you'll notice we use electronic health record or EHR almost exclusively.").

[8] Sharona Hoffman & Andy Podgurski, *Balancing Privacy, Autonomy, and Scientific Needs in Electronic Health Records Research*, 65 SMU L. REV. 85, 91-94 (2012).

[9] M.A. Hoffman, *The Genome-Enabled Electronic Medical Record*, 40 J. BIOMEDICAL INFORMATICS 44, 44 (2006); Isaac S. Kohane, *Using Electronic Health Records to Drive Discovery in Disease Genomics*, 12 NATURE REV. GENETICS 417, 417 (2011).

[10] ARTHUR M. LESK, INTRODUCTION TO GENOMICS 104-05 (2d ed. 2012).

[11] Hoffman & Podgurski, *supra* note 8, at 97-102.

[12] *See* Abel N. Kho et al., *Electronic Medical Records for Genetic Research: Results of the eMERGE Consortium*, 3 SCI. TRANSL. MED. 78re1, 5 (2011); Charles Safran, *Toward a National Framework for the Secondary Use of Health Data: An American Medical Informatics Association White Paper*, 14 J. AM. MED. INFORMATICS ASS'N 1, 2 (2007); Mark G. Weiner & Peter J. Embi, *Toward Reuse of Clinical Data for Research and Quality Improvement: The End of the Beginning?*, 151 ANN. INTERN. MED. 359, 359-60 (2009).

[13] *See infra* notes 225-226.

[14] 42 U.S.C. § 1320e (Supp IV. 2010); *see infra* notes 97-100.

[15] Food and Drug Administration Amendments Act of 2007, Pub. L. No. 110-85, 121 Stat. 823 (codified as amended in scattered sections of 21 U.S.C.); *see infra* Part II.B.3.

challenges of inferring causal effects correctly.[16] Much has been written about EHR privacy risks, but this paper makes a different contribution to the legal literature by focusing on what can go wrong in the process of biomedical data analysis and what precautions must be taken to avoid critical mistakes. It sheds much-needed light on the problems of naïve or irresponsible use of biomedical databases, and these problems are likely to become much more common and pressing in the near future. The data-use pitfalls we discuss are familiar to competent biomedical researchers but must be understood by lawyers, bioethicists, policymakers, and anyone else who will rely on research results.

We use the term "biomedical databases" to mean databases of EHRs and/or genomic information as well as decentralized, federated database systems.[17] Thus, in this paper, we address non-interventional research, that is, research that is based on review of records, which we also call "records-based research" or "observational research."[18] We do not intend to comment on clinical studies in which investigators conduct experiments using human subjects[19] or on research involving the administration of questionnaires or surveys.

Observational studies are relevant to the law because their outcomes can lead to regulatory enforcement actions or to legislative changes, and they can be used as evidence in litigation. For example, observational studies may reveal that use of a medication or device causes patients to suffer serious adverse events, and this discovery may induce the Food and Drug Administration (FDA) to intervene.[20] Observational studies may also uncover statistical associations between illnesses and exposure to certain substances or between diseases and genetic variations.[21] Reports of these associations may be used in litigation by both plaintiffs and defendants.[22] Plaintiffs may file tort cases against product manufacturers, and toxic tort defendants

---

[16] Pamela N. Peterson & Paul D. Varosy, *Observational Comparative Effectiveness Research: Comparative Effectiveness and Caveat Emptor*, 5 CIRCULATION CARDIOVASC. QUALITY & OUTCOMES 150, 151 (2012) (warning that "a primary determinant of the quality of any study is the quality of the data" and that "how the results of observational studies are interpreted and used" is of critical importance).

[17] See *infra* note 37 and accompanying text for definition of federated database system. In other contexts, biomedical databases can also consist of data collected from large-scale clinical studies. Prakash M. Nadkarni, *Managing Attribute—Value Clinical Trials Data Using the ACT/DB Client—Server Database System*, 5 J. AM. MED. INFORMATICS ASS'N 139, 139 (1998) (stating that "complex trials need sophisticated database expertise not readily available to individual investigators").

[18] CHARLES P. FRIEDMAN & JEREMY C. WYATT, EVALUATION METHODS IN BIOMEDICAL INFORMATICS 369 (Kathryn Hannah & Marion Ball eds., 2d ed. 2006) (defining observational studies as involving an "[a]pproach to study design that entails no experimental manipulation"); BRYAN F. J. MANLY, THE DESIGN AND ANALYSIS OF RESEARCH STUDIES 1 (1992) (explaining that observational studies involve the collection of data "by observing some process which may not be well-understood"); PAUL R. ROSENBAUM, OBSERVATIONAL STUDIES vii (2d ed. 2001) (stating that an observational study is "an empiric investigation of treatments, policies, or exposures and the effects they cause, but it differs from an experiment in that the investigator cannot control the assignment of treatments to subjects"). When using the term "observational studies," we refer only to studies involving the review of existing records or data.

[19] MANLY, *supra* note 18, at 1 (explaining that experimental clinical studies involve "the collection of data on a process when there is some manipulation of variables that are assumed to affect the outcome of a process, keeping other variables constant as far as possible"); Hoffman & Podgurski, *supra* note 8, at 98-102 (contrasting clinical trials and observational studies).

[20] *See infra* notes 115-119 and accompanying text.

[21] DAVID L. FAIGMAN ET AL., MODERN SCIENTIFIC EVIDENCE: STANDARDS, STATISTICS, AND RESEARCH METHODS 338-42 (student ed. 2008).

[22] *Id.*

may in turn use scientific evidence to attack plaintiffs' claims and argue that something other than their products caused the plaintiffs' illnesses.[23]

News outlets frequently report new research findings. Press reports often trumpet the discovery that factor *A* is statistically associated with or "linked" to condition *B*. The availability of large biomedical databases greatly facilitates the discovery of such associations. However, the nature of such data can complicate the determination of whether factor *A* actually *causes* or *contributes to* condition *B*. We address three main reasons for a cautious approach to incorporating record-based research into the law.

First, the data contained in biomedical databases may be of poor quality, incomplete, or even deliberately distorted.[24] For example, a recent *New York Times* article reported that the automated features of EHR systems make it easy for doctors to exaggerate the care they provided for purposes of Medicare reimbursement.[25] Doctors can simply click on menu items or copy and paste narrative in order to justify billing, and some lack scruples with respect to overstating or even fictionalizing what occurred during clinical encounters.[26] Such practices not only defraud Medicare, but also compromise the accuracy of EHRs. Moreover, they can systematically bias research results.

Second, valid causal analysis is much more difficult with observational data than with data from well-designed and well-executed randomized experiments or clinical trials.[27] Unfortunately, having large amounts of data ("big data") does not necessarily ameliorate this problem. The challenges of properly analyzing observational data and making appropriate causal inferences[28] are illustrated in a paper entitled "Does Obesity Shorten Life? The Importance of Well-Defined Interventions to Answer Causal Questions."[29] The researchers critique previous observational studies of obesity and mortality and conclude that they were flawed because they failed to specify what interventions were used to reduce body mass index (BMI). Different methods of changing BMI (e.g., surgery, diet, exercise) are associated with different risk levels for patients, and mortality may actually be associated with the treatment rather than the underlying obesity in some cases.[30] Thus, researchers cannot reach meaningful conclusions about the benefits of reducing BMI without knowing what interventions were used to achieve this goal in each instance.[31]

Third, individuals with political, social, or economic agendas may "mine" or "dredge" biomedical databases to find links (statistical associations) between

---

[23] *Id.* at 339-40 (explaining that epidemiological evidence has already played an important role in many mass tort cases); Steve C. Gold, *The More We Know, the Less Intelligent We Are?–How Genomic Information Should, and Should Not, Change Toxic Tort Causation Doctrine*, 34 HARV. ENVTL. L. REV. 369, 412-17 (2010) (discussing genes and other toxins as alternate causes of plaintiffs' injuries).

[24] *See infra* Part III.

[25] Reed Abelson et al., *Medicare Bills Rise as Records Turn Electronic*, N.Y. TIMES, Sept. 21, 2012, at A1, A3, http://www.nytimes.com/2012/09/22/business/medicare-billing-rises-at-hospitals-with-electronic-records.html?_r=0.

[26] *Id.*

[27] *See infra* Part IV.

[28] Samantha Kleinberg & George Hripcsak, *A Review of Causal Inference for Biomedical Informatics*, 44 J. BIOMED. INFORMATICS 1102, 1102 (2011) (defining causal inference as "the process of uncovering causal relationships from data").

[29] *See* Miguel A. Hernán & Sarah L. Taubman, *Does Obesity Shorten Life? The Importance of Well-Defined Interventions to Answer Causal Questions*, 32 INT'L J. OBESITY S8 (2008).

[30] *Id.* at S13.

[31] *Id.*

actions, behaviors, or policies, on the one hand, and outcomes of public interest, on the other hand, for the purpose of manipulating public opinion and swaying policy decisions.[32] The risk of misinterpretation of such results by interested parties is high if they are not well-trained and scrupulous researchers. Research about the purported link between abortion and psychiatric disorders, discussed above, demonstrates this potential danger.[33] Pro-life advocates used questionable scientific data to promote a controversial legislative agenda.

The paper proceeds as follows. Part II provides background information. It describes ongoing efforts to build biomedical databases and analyzes the relevance of observational studies to law and public policy. Part III analyzes common shortcomings of biomedical data that should give analysts and the public pause. These include input errors, incomplete or fragmented records, and flaws in data coding or standardization.

Part IV provides an in-depth discussion of causal inference and of biases affecting observational studies. It analyzes the challenges of inferring causation in observational studies, including the problems of selection bias, confounding bias, and measurement bias. Indeed, confounding bias and selection bias will likely be fundamental concepts in legal reasoning in big data environments. Part V addresses the potential use of observational study outcomes for purposes of furthering political, social, and economic agendas.

Finally, Part VI analyzes the factors that contribute to sound research and provides guidance for policymakers and litigants seeking to determine whether particular research outcomes are reliable. The quality of digitized research databases and the studies that grow out of them will depend not only on good technology, but also on persistent human efforts to safeguard the integrity of research projects. Technological advances are needed to enhance interoperability, data capture, data-extraction capabilities, and system usability. In addition, clinicians and patients can partner to assess the validity of the data contained in EHRs, and investigators must be scrupulous about study design, analysis, and publication. This Part also describes and critiques the use of causal inference diagrams, which have received little attention in the legal literature but is increasingly common in other fields.[34]

Equally important is ensuring that the legal community, journal editors, and the public at large are not misled by those who appear to engage in scientific endeavors but who in truth misuse evidence to promote their own political, social, or economic agendas. Legal practitioners must understand the complex issues raised by big data in order to play a useful role in protecting the public's interests. To this end, we recommend the development of law school and other educational programs about the challenges of observational data analysis and causal inference.

II. BACKGROUND

Researchers and other analysts may gain access to large-scale collections of biomedical data in two primary ways. First, health information can be collected into

---

[32] *See infra* Part V.

[33] *See supra* notes 1-6 and accompanying text.

[34] *See* Nancy C. Staudt & Tyler J. VanderWeele, *Methodological Advances and Empirical Legal Scholarship: A Note on Cox and Miles's Voting Rights Act Study* 109 COLUM. L. REV. SIDEBAR 42, 43 (2009) (asserting that by 2009 the methodology of causal diagrams had "become popular in a number of disciplines – including statistics, biostatistics, epidemiology, and computer science . . . [but had yet] to appear in the empirical law literature").

large databases and de-identified to protect patient privacy.[35] Such databases could be limited to particular hospital systems, be expanded to cover entire regions, or even be national in scope.[36] In the alternative, researchers may use a "federated system" by which medical institutions manage and maintain control of their own databases, but they allow researchers to submit statistical queries through a standard web service in order to obtain summary statistics for a study population.[37] Trusted third-party aggregators can operate the query service.[38]

Many large biomedical databases and federated systems already exist and are used for non-treatment purposes.[39] The term "secondary use" refers to the utilization of health information outside the clinical setting.[40] This Part describes a sample of data-collection initiatives. It also discusses how experts in the biomedical research, quality assessment, public health, and litigation arenas may utilize EHR data.

A. ONGOING INITIATIVES TO CREATE BIOMEDICAL DATABASES

The Federal Government has clearly recognized the usefulness of biomedical databases and enthusiastically supports database projects. The Obama Administration has announced an overarching effort called the "Big Data Research and Development Initiative" ("Big Data").[41] The initiative's purposes are to advance cutting-edge technologies needed to gather and process "huge quantities of data;" to employ those technologies to promote scientific discovery, improved national security, and education; and to expand the workforce skilled in these technologies.[42] Big Data will involve six federal agencies and departments and is estimated to cost $200 million.[43] As part of Big Data, the National Institutes of Health (NIH) will make data from its 1000 Genomes Project publicly available through cloud computing.[44]

---

[35] Hoffman & Podgurski, *supra* note 8, at 128-30.

[36] WILSON D. PACE ET AL., AGENCY FOR HEALTH CARE RES. & QUALITY, DISTRIBUTED AMBULATORY RESEARCH IN THERAPEUTIC NETWORK (DARTNET): SUMMARY REPORT ii (2009), *available at* http://www.effectivehealthcare.ahrq.gov/ehc/products/53/151/2009_0728DEcIDE_DARTNet.pdf.

[37] Griffin M. Weber et al., *The Shared Health Research Information Network (SHRINE): A Prototype Federated Query Tool for Clinical Data Repositories*, 16 J. AM. MED. INFORMATICS ASS'N 624, 624 (2009). A federated network can be defined as one that "links geographically and organizationally separate databases to allow a single query to pull information from multiple databases while maintaining the privacy and confidentiality of each database." PACE, *supra* note 36, at ii.

[38] Hoffman & Podgurski, *supra* note 8, at 131-33.

[39] *Id.* at 91.

[40] Jessica S. Ancker et al., *Root Causes Underlying Challenges to Secondary Use of Data*, AMIA ANNUAL SYMPOSIUM PROCEEDINGS 57, 57 (2011); Taxiarchis Botsis et al., *Secondary Use of EHR: Data Quality Issues and Informatics Opportunities*, AMIA JOINT SUMMITS ON TRANSL. SCI. 1, 1 (2010).

[41] Press Release, Office of Sci. & Tech. Policy, Exec. Office of the President, Obama Administration Unveils "Big Data" Initiative: Announces $200 Million in New R & D Investments (Mar. 29, 2012), *available at* http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf.

[42] *Id.*

[43] *Id.* The agencies are the Office of Science and Technology Policy, National Science Foundation, National Institutes of Health, Department of Defense, Department of Energy, and U.S. Geological Survey.

[44] *Id.* The international 1000 Genome Project "aims to find most genetic variants that have frequencies of at least 1 percent in the populations studied." According to the National Institutes of Health, it is the world's largest human genetic variation data set, with 200 terabytes – "the equivalent of 16 million file cabinets filled with text, or more than 30,000 standard DVDs." The information is

At the same time, many federal entities are independently building health information databases.[45] For example, the Department of Veterans Affairs (VA) is registering volunteers for its Million Veteran Program to construct a large research framework that will link anonymized blood samples and health information.[46] The VA plans to study how genes affect health and disease.[47]

The Centers for Medicare & Medicaid Services created a research database called the Chronic Condition Data Warehouse.[48] The database provides researchers with information about Medicare and Medicaid beneficiaries, claims for services, and assessment data.[49]

In May of 2008 the FDA launched the Sentinel System in order to facilitate post-marketing surveillance and early detection of medical products' safety problems.[50] The Sentinel initiative aims to enable the FDA to access health information from 100,000,000 individuals.[51] Sentinel is a federated system that will allow the FDA to send queries concerning potential product-safety problems to data holders such as Medicare, the VA, and major medical centers.[52] Using special analysis programs, the data holders will assess their records and send summary responses to the FDA.[53]

A large number of private-sector initiatives are ongoing as well. Geisinger Health Systems operates MedMining, a company that extracts EHR data, de-identifies it, and offers it to researchers.[54] The data sets that MedMining delivers to its customers include "lab results, vital signs, medications, procedures, diagnoses, lifestyle data, and detailed costs" from inpatient and outpatient facilities.[55]

Explorys has formed a large healthcare database derived from financial, administrative, and medical records.[56] It has partnered with major healthcare organizations such as the Cleveland Clinic Foundation and Summa Health System to aggregate and standardize health information from ten million patients and over

---

available on the Amazon Web Services cloud. Jeannie Baumann, *White House Initiative Aims to Improve Use of Large Digital Databases for R & D*, 11 MED. RES. L. & POL'Y 217, 217-18 (2012).

[45] *See, e.g.*, Ctrs. for Medicare & Medicaid Servs. (CMS), *About Chronic Conditions Data Warehouse*, CHRONIC CONDITIONS DATA WAREHOUSE, https://www.ccwdata.org/web/guest/about-ccw (last visited Oct. 16, 2013); *FDA's Sentinel Initiative,* U.S. FOOD & DRUG ADMIN. (Sept. 4, 2013), http://www.fda.gov/safety/FDAsSentinelInitiative/ucm2007250.htm; *Million Veteran Program: A Partnership with Veterans*, U.S. DEP'T OF VETERANS AFFAIRS (Mar. 6, 2013), http://www.research.va.gov/mvp/veterans.cfm.

[46] *Million Veteran Program*, *supra* note 45.

[47] *Id.*

[48] CMS, *supra* note 45.

[49] *Id.* CCW was created pursuant to section 723 of the Medicare Modernization Act of 2003. Medicare Modernization Act of 2003 § 723, 42 U.S.C. § 1395b-8 (2006).

[50] Deven McGraw et al., *A Policy Framework for Public Health Uses of Electronic Health Data*, 21(S1) PHARMACOEPI. & DRUG SAFETY 18, 18 (2012); *FDA's Sentinel Initiative*, *supra* note 45. The Sentinel initiative was authorized by Congress in the Food and Drug Administration Amendments Act of 2007. *FDA's Sentinel Initiative-Background*, U.S. FOOD & DRUG ADMIN. (Sept. 22, 2013), http://www.fda.gov/Safety/FDAsSentinelInitiative/ucm149340.htm.

[51] McGraw et al., *supra* note 50, at 18.

[52] *Id.* at 19. *See* Hoffman & Podgurski, *supra* note 8, at 131-33 (discussing distributed databases).

[53] McGraw et al., *supra* note 50, at 19.

[54] *Welcome to MedMining*, MEDMINING, http://www.medmining.com/index.html (last visited Oct. 13, 2013).

[55] *Id.*

[56] *Explorys Overview*, EXPLORYS, https://www.explorys.com/docs/data-sheets/explorys-overview.pdf (last visited Oct. 16, 2013).

thirty billion clinical events.[57] Using a cloud-computing platform, it provides customers with big data to use for research and quality improvement purposes.[58]

The electronic Medical Records and Genomics Network (eMERGE) is a consortium of five institutions with DNA repositories linked to EHRs that supply relevant clinical data.[59] The National Human Genome Research Institute supports eMERGE, and the National Institute of General Medical Sciences provides it with additional funding.[60] Each eMERGE center will study "the relationship between genome-wide genetic variation and a common disease/trait," using genome-wide association analysis.[61] A primary purpose of eMERGE is to develop approaches to conducting large-scale genetic research using DNA biobanks that are connected to EHR systems.[62]

The Distributed Ambulatory Research in Therapeutics Network Institute (DARTNet) is a collaboration among nine research networks, including 85 healthcare organizations and over 3,000 clinicians across the United States.[63] The first DARTNet federated network, eNQUIRENet, was created in 2007 and funded by the Agency for Healthcare Research and Quality.[64] DARTNet members allow data from their EHRs to be captured, de-identified, coded, standardized, and stored in a Clinical Data Repository (CDR) within each entity that also connects to billing, lab, hospital, and prescription databases.[65] CDR data are then transferred to a second database that makes de-identified information available to researchers through a secure web portal.[66]

Other agencies and organizations are building electronic registries and databases that focus on specific disease categories in an effort to promote research and quality improvement endeavors. These include the Cancer Biomedical Informatics Grid,[67]

---

[57] *Id.*

[58] *Id.*

[59] EMERGE NETWORK, http://emerge.mc.vanderbilt.edu/ (last visited Oct. 16, 2013). The seven sites are: Group Health Cooperative with the University of Washington, Geisinger, Marshfield Clinic, Mayo Clinic, Mount Sinai School of Medicine, Northwestern University, and Vanderbilt University. National Human Genome Research Institute, *Electronic Medical Records and Genomics (eMERGE) Network*, GENOME.GOV, http://www.genome.gov/27540473 (last updated Aug. 29, 2013).

[60] EMERGE NETWORK, *supra* note 59; *Appropriations Subcommittee Statement on the Fiscal Year 2013 Budget* (Mar. 23, 2012), NAT'L INST. GEN. MED. SCIS., http://www.nigms.nih.gov/About/Budget/Statements/March23_2012.htm.

[61] Catherine A. McCarty et al., *The eMERGE Network: A Consortium of Biorepositories Linked to Electronic Medical Records Data for Conducting Genomic Studies*, 4 BMC MED. GENOMICS 13, 14 (2011).

[62] *Id.* A recent study found that data captured from EHRs could identify disease characteristics with sufficient accuracy to be used in genome-wide association studies. Kho et al., *supra* note 12, at 4-5.

[63] DARTNET INSTITUTE: INFORMING PRACTICE, IMPROVING CARE, http://www.dartnet.info/ (last visited Oct. 16, 2013); *About DARTNet*, DARTNET INST., http://www.dartnet.info/AboutDI.htm (last visited Oct. 15, 2013).

[64] *History of the Organization*, DARTNET INST., http://www.dartnet.info/organization.htm (last visited Oct. 15, 2013); *Networks*, DARTNET INST., http://www.dartnet.info/networks.htm (last visited Oct. 15, 2013).

[65] *See About DARTNet*, *supra* note 63.

[66] *Technology*, DARTNET INST., http://www.dartnet.info/Technology.htm (last visited Oct. 15, 2013).

[67] *Software Tools*, NAT'L CANCER INST., http://www.cancer.gov/clinicaltrials/international/answers/softwaretools (last visited Oct. 15, 2013) (stating that the initiatives' goal is to "[b]uild or adapt tools for collecting, analyzing, integrating, and disseminating information associated with cancer research and care").

the Interagency Registry for Mechanically Assisted Circulatory Support,[68] the Extracorporeal Life Support Organization,[69] and the United Network for Organ Sharing.[70]

B. USING BIOMEDICAL DATABASES AND DATA NETWORKS

Large-scale biomedical databases may be used for many purposes. This section addresses a variety of ways in which they are likely to be used by researchers, regulators, public health officials, commercial entities, and lawyers. As we have indicated, biomedical databases constitute an important tool for medical researchers. They are also used by healthcare providers who conduct quality assessment and improvement activities, and they assist the FDA in monitoring the safety of drugs and devices on an ongoing basis. In addition, biomedical databases can support public health initiatives and allow litigants in tort cases to develop evidence concerning causation and harm.

1. Scientific Discovery

Biomedical databases can enable researchers to conduct large-scale observational studies that will fill existing knowledge gaps. Even today, clinicians practice medicine with an unsettling degree of uncertainty.[71] According to some estimates, doctors know that the treatments they prescribe will be effective in only twenty to twenty-five percent of cases.[72] Database proponents believe that records-based research could contribute substantially to the resolution of these uncertainties.[73]

Biomedical databases could allow researchers to access a vast quantity of information about millions of patients who are treated in varied clinical settings,

---

[68] *INTERMACS Description*, Interagency Registry for Mechanically Assisted Circulatory Support (INTERMACS), http://www.uab.edu/ctsresearch/intermacs/description.htm (last visited Oct. 22, 2013) (explaining that analysis of the collected data is expected to improve patient care and "influence future research").

[69] *ELSO Registry Information Data Policy*, ECMO REGISTRY EXTRACORPOREAL LIFE SUPPORT ORG., http://www.elso.med.umich.edu/DataRequests.html (last updated Oct. 12, 2010) (providing details concerning the collection of data with most identifiers removed, submission of queries, and release of query results to members in aggregate form).

[70] *Data*, UNITED NETWORK FOR ORGAN SHARING (UNOS), http://www.unos.org/donation/index.php?topic=data (last visited Oct. 15, 2013) (discussing the creation of UNet, an online database system that "contains data regarding every organ donation and transplant event occurring in the United States since 1986").

[71] *See* David A. Hyman & Charles Silver, *The Poor State of Health Care Quality in the U.S.: Is Malpractice Liability Part of the Problem or Part of the Solution?* 90 CORNELL L. REV. 893, 952 (2005) (observing that a "great deal of uncertainty exists about the 'best' treatment for particular clinical conditions, and about the 'best' way to perform those treatments" and that the "efficacy of most medical treatments has never been proven"); Walter F. Stewart et al., *Bridging the Inferential Gap: The Electronic Health Record and Clinical Evidence*, 26 HEALTH AFF. w181, w181 (2007) (discussing the "inferential gap" between "the paucity of what is proved to be effective for selected groups of patients versus the infinitely complex clinical decisions required for individual patients").

[72] John Carey, *Medical Guesswork*, BUSINESSWEEK, May 29, 2006, at 73, *available at* http://www.businessweek.com/stories/2006-05-28/medical-guesswork (asserting that many physicians "say the portion of medicine that has been proven effective is still outrageously low – in the range of 20% to 25%").

[73] Hoffman & Podgurski, *supra* note 8, at 97-102 (discussing the benefits of EHR-based research).

have diverse attributes, and live in different regions of the country.[74] Available information could include patients' medical histories over their entire lifetimes. The data reviewed in database studies, consequently, may be far more abundant and comprehensive than the data generated by clinical trials,[75] which are rigorously controlled and often involve fewer than 3000 patients.[76] Large-scale studies have the potential to better reflect the entire population and expose how treatments are actually used in a large variety of medical facilities.[77] They also tend to enhance the precision of statistical analyses.[78]

If the researchers aim to show whether a specific treatment achieves the desired benefits, they may reasonably choose to conduct a randomized clinical trial to ensure that uncontrolled variables that influence outcomes, such as age or drug interactions, do not confound the study.[79] However, observational studies may be needed to determine whether the results of randomized clinical trials that involved only a few thousand patients can be generalized to the patient population at large and to realistic treatment situations rather than carefully controlled ones.[80] Furthermore, observational research based on medical records will often be sufficient to determine a treatment's adverse effects.[81] It is also useful for generating and testing speculative hypotheses that could lead to important insights.[82] Observational studies are often

---

[74] Lynn M. Etheredge, *A Rapid-Learning Health System*, 26 HEALTH AFF. w107, w111 (2007), *available at* http://content.healthaffairs.org/cgi/content/full/26/2w107; Hoffman & Podgurski, *supra* note 8, at 97-102; Louise Liang, *The Gap Between Evidence and Practice,* 26 HEALTH AFF. w119, w120 (2007) (asserting that EHRs "have the potential to take over where clinical trials and evidence-based research leave off, by providing real-world evidence of drugs' and treatments' effectiveness across subpopulations and over longer periods of time"); *see* James H. Ware & Mary Beth Hamel, *Pragmatic Trials – Guides to Better Patient Care?,* 364 NEW. ENG. J. MED. 1685, 1685 (2011) (discussing the shortcomings of clinical trials).

[75] Clinical studies involve "the collection of data on a process when there is some manipulation of variables that are assumed to affect the outcome of a process, keeping other variables constant as far as possible." BRYAN F. J. MANLY, THE DESIGN AND ANALYSIS OF RESEARCH STUDIES 1 (1992). Thus, they involve actual experimentation on human subjects rather than just review of their medical records.

[76] Sheila Weiss Smith, *Sidelining Safety–The FDA's Inadequate Response to the IOM*, 357 NEW. ENG. J. MED. 960, 961 (2007).

[77] *See* John P.A. Ioannidis, *Why Most Published Research Findings Are False*, 2 PLOS MED. 696, 700 (2005).

[78] *See* David Moher et al., *Statistical Power, Sample Size, and Their Reporting in Randomized Controlled Trials*, 272 J. AM. MED. ASS'N 122, 122-24 (1994).

[79] Hoffman & Podgurski, *supra* note 8, at 98-99; Jan P. Vandenbroucke, *The HRT Controversy: Observational Studies and RCTs Fall in Line*, 373 LANCET 1233, 1234 (2009).

[80] Stuart L. Silverman, *From Randomized Controlled Trials to Observational Studies*, 122 AM. J. MED. 114, 114 (2009) (explaining that "[o]bservational studies may be an important addition to the clinician's resources by complementing randomized controlled trial data with information on efficacy, safety, and patient compliance in a population of real-world patients"); Stewart et al., *supra* note 71, at 73 (stating that analysis of EHR data should help bridge the "inferential gap" between "the paucity of what is proved to be effective for selected groups of patients versus the infinitely complex clinical decisions required for individual patients").

[81] Jan P. Vandenbroucke, *Observational Research, Randomised Trials, and Two Views of Medical Science*, 5 PLOS MED. 339, 341 (2008), *available at* http://www.plosmedicine.org/article/info%3Adoi%2F10.1371%2Fjournal.pmed.0050067 (explaining that adverse effects are generally unexpected and unpredictable, and therefore are not subject to "confounding by indication" and can be determined through observational studies). *See infra* notes 244-247 and accompanying text for discussion of confounding by indication.

[82] Vandenbroucke, *supra* note 81, at 343 (asserting that "[m]uch good can come from going down the wrong alley and detecting why it is wrong, or playing with a seemingly useless hypothesis; the real breakthrough might come from that experience").

less costly and time-consuming than experimental research, especially when researchers obtain the required data from existing databases.[83]

The benefits of observational studies are illustrated by the highly publicized controversy concerning an alleged association between vaccination and autism. In 1998, Dr. Andrew J. Wakefield and colleagues published a study in the *Lancet* that suggested a link between autism and the measles, mumps, rubella (MMR) vaccination.[84] The findings were based on testing of twelve children with developmental disorders.[85] In 2004 most of the authors "retracted the interpretation placed upon these findings in the paper"[86] after large-scale observational research involving the review of hundreds of records of autistic children in the United Kingdom found no causal association between the MMR vaccine and autism.[87] Consequently, the Centers for Disease Control and Prevention (CDC) now reassures the public on its website that there is no link between autism and vaccines.[88]

For purposes of genetic research, EHRs can be coupled with genetic samples and data so that analysts can obtain detailed and comprehensive characterizations of study subjects.[89] An increasingly common form of big-data observational research is genome-wide association studies (GWASs).[90] GWASs compare the DNA of individuals with a particular disease or condition to the DNA of unaffected individuals in order to find the genes involved in the disease.[91] A government website catalogues published GWASs and on October 24, 2013, listed 1,727 studies that had been conducted since 2005.[92] Critics have noted that although GWASs led to the discovery of many genetic variants that are statistically associated with disease; thus far, most of the variants appear to have a minimal effect on disease and

---

[83] Kjell Benson & Arthur J. Hartz, *A Comparison of Observational Studies and Randomized, Controlled Trials*, 342 NEW ENG. J. MED. 1878, 1878 (2000) (mentioning "greater timeliness" as an advantage of observational studies); David Kaelber et al., *Patient Characteristics Associated with Venous Thromboembolic Events: A Cohort Study Using Pooled Electronic Health Record Data*, 19 J. AM. MED. INFORMATICS ASS'N 965, 966 (2012); Friedrich K. Port, *Role of Observational Studies Versus Clinical Trials in ESRD Research*, 57 KIDNEY INT'L S3, S4 (2000), *available at* http://www.nature.com/ki/journal/v57/n74s/full/4491615a.html. For further details about the benefits of observational studies, see Hoffman & Podgurski, *supra* note 8, at 97-102.

[84] Andrew J. Wakefield et al., *Ileal-Lymphoid-Nodular Hyperplasia, Non-Specific Colitis, and Pervasive Developmental Disorder in Children*, 351 LANCET 637, 641 (1998).

[85] *Id.* at 637.

[86] Simon H. Murch et al., *Retraction of an Interpretation*, 363 LANCET 750, 750 (2004). Dr. Wakefield did not join the retraction.

[87] Brent Taylor et al., *Autism and Measles, Mumps, and Rubella Vaccine: No Epidemiological Evidence for a Causal Association*, 353 LANCET 2026, 2026-29 (1999). Furthermore, the British General Medical Council found that Dr. Wakefield was guilty of multiple transgressions, including dishonesty, financial misconduct, and "callous disregard" of the suffering of the children who were his research subjects. General Medical Council, *Dr. Andrew Jeremy Wakefield, Determination on Serious Professional Misconduct (SPM) and Sanction* 8 (May 24, 2010), *available at* http://www.gmc-uk.org/Wakefield_SPM_and_SANCTION.pdf_32595267.pdf.

[88] *Measles, Mumps, and Rubella (MMR) Vaccine*, CTRS. FOR DISEASE CONTROL AND PREVENTION, http://www.cdc.gov/vaccinesafety/Vaccines/MMR/MMR.html (last updated Feb. 7, 2011).

[89] Kohane, *supra* note 9, at 417.

[90] Brian D. Juran & Konstantinos N. Lazaridis, *Genomics in the Post-GWAS Era*, 31 SEM. IN LIVER DISEASE 215, 215 (2011); Christophe G. Lambert & Laura J. Black, *Learning From Our GWAS Mistakes: From Experimental Design to Scientific Method*, 13 BIOSTATISTICS 195, 196 (2012).

[91] *Dictionary of Cancer Terms*, NAT'L CANCER INST., http://www.cancer.gov/dictionary?cdrid=636779 (last visited Oct. 25, 2013).

[92] NAT'L HUMAN GENOME RES. INST., *A Catalog of Published Genome-Wide Association Studies*, GENOME.GOV, http://www.genome.gov/gwastudies/ (last visited Oct. 26, 2013).

explain only a small percentage of heritability.[93] Others assert that many GWASs to date have been compromised by serious design flaws.[94] However, GWASs remain an important scientific endeavor and will likely lead to significant discoveries in the future.

A different method of scanning the genome is genome-wide linkage studies (GWLSs). Researchers perform GWLSs when they are focusing on biologically related individuals and a phenotype, such as breast cancer, that some but not all of the family members have.[95] Based on patterns of correlation between alleles[96] and disease found within families, researchers attempt to detect broad DNA regions in which disease susceptibility loci are most likely to be found.[97]

The Federal Government and many medical experts have embraced the objective of conducting extensive comparative effectiveness research (CER).[98] The Patient Protection and Affordable Care Act of 2010 defines CER as "research evaluating and comparing health outcomes and the clinical effectiveness, risks, and benefits of 2 or more medical treatments, services, and items . . . ."[99] CER can be conducted in part through observational studies, which can be particularly illuminating because they reflect actual usage of treatments.[100] The outcomes of CER and other observational studies may ultimately enable the healthcare community to alleviate human suffering more effectively, reduce medical costs, and save patients' lives.[101]

2. Quality Assessment and Improvement

Healthcare providers routinely collect quality measures concerning the services they provide.[102] Increasingly, they will use EHR databases to obtain necessary

---

[93] David J. Hunter, *Lessons from Genome-Wide Association Studies for Epidemiology*, 23 EPIDEMIOLOGY 363, 363 (2012) (stating that "GWAS-discovered variants are relatively 'weak' risk factors" and "are not modifiable factors with direct potential to reduce disease incidence" but will improve "understanding of disease mechanisms" and perhaps facilitate "identification of persons at higher or lower risk of specific diseases"); Juran & Lazaridis, *supra* note 90, at 215-16.

[94] Lambert & Black, *supra* note 90, at 196-97.

[95] P.A. Holmans et al., *Genomewide Linkage Scan of Schizophrenia in a Large Multicenter Pedigree Sample Using Single Nucleotide Polymorphisms*, 14 MOLECULAR PSYCH. 786, 786-87 (2009).

[96] An allele "is one of two or more versions of a gene." Thus, the term "allele" is used when there is "variation among genes." NAT'L HUMAN GENOME RES. INST., *Allele*, GENOME.GOV, http://www.genome.gov/glossary/?id=4 (last visited Oct. 25, 2013).

[97] Holmans et al., *supra* note 95, at 787.

[98] 42 U.S.C. § 1320e (Supp IV. 2010); INST. OF MED., INITIAL NATIONAL PRIORITIES FOR COMPARATIVE EFFECTIVENESS RESEARCH (2009), *available at* http://www.iom.edu/Reports/2009/ComparativeEffectivenessResearchPriorities.aspx (emphasizing the need for CER and proposing initial CER priorities).

[99] 42 U.S.C. § 1320e(a)(2)(A) (Supp. IV 2010).

[100] *See id.* § 1320e(d)(2)(A). *See* John Concato et al., *Observational Methods in Comparative Effectiveness Research*, 123 AM. J. MED. e16, e16 (2010); S. Schneeweiss et al., *Assessing the Comparative Effectiveness of Newly Marketed Medications: Methodological Challenges and Implications for Drug Development*, 90 CLIN. PHARMACOLOGY & THERAPEUTICS 777, 777 (2011) (discussing the use of "secondary health-care data, including electronic medical records" for purposes of CER); Vandenbroucke, *supra* note 81, at 340.

[101] *See* 42 U.S.C. § 1320e(d)(2)(A) (Supp. IV 2010); L. Manchikanti et al., *Facts, Fallacies, and Politics of Comparative Effectiveness Research: Part 1. Basic Consideration*, 13 PAIN PHYSICIAN E23, E39 (2010); Adam G. Elshaug & Alan M. Garber, *How CER Could Pay for Itself – Insights from Vertebral Fracture Treatments*, 364 NEW ENG. J. MED. 1390, 1392-93 (2011).

[102] Kitty S. Chan et al., *Electronic Health Records and the Reliability and Validity of Quality Measures: A Review of the Literature*, 67 MED. CARE RES. & REV. 503, 504 (2010).

information.[103]

Medical facilities and government authorities conduct a variety of oversight activities. Providers may seek data for internal quality assessment purposes in order to judge the success of particular initiatives.[104] Insurers may require facilities to submit process and outcome information in the context of pay-for-performance programs.[105] In addition, the Centers for Medicare and Medicaid Services (CMS) and many state governments require quality measurements and public reporting.[106] A prime example is CMS's Hospital Compare, which features publicly-available data about the quality of care at over 4000 hospitals.[107]

3. Post-Marketing Surveillance of Drugs and Devices

EHR databases could assist the FDA in regulating drugs and devices.[108] The Food and Drug Administration Amendments Act of 2007 (FDAAA)[109] expanded the FDA's authority to monitor medical products after they have been approved and deployed in the marketplace.[110] Evidence concerning drug safety in the post-marketing period will be developed in significant part through observational studies.[111] Such studies will be made possible through the Sentinel System, the national health data network discussed previously.[112]

While clinical trials constitute the gold standard for purposes of FDA drug approval,[113] they have important weaknesses that have been documented elsewhere

---

[103] *Id.*; Amanda Parsons et al., *Validity of Electronic Health Record-Derived Quality Measurement for Performance Monitoring*, 19 J. AM. MED. INFORMATICS ASS'N 604, 609 (2012) (finding that "EHR-derived quality measurement has limitations due to several factors, most notably variations in EHR content, structure and data format, as well as local data capture and extraction procedures"); Joachim Roski & Mark McClellan, *Measuring Health Care Performance Now, Not Tomorrow: Essential Steps to Support Effective Health Reform*, 30 HEALTH AFF. 682, 683 (2011).

[104] *See* Monica M. Horvath et al., *The DEDUCE Guided Query Tool: Providing Simplified Access to Clinical Data for Research and Quality Improvement*, 44 J. BIOMED. INFORMATICS 266, 273 (2011) (stating that Duke University Hospital sought data in order to evaluate the effects of new health information technology that it had implemented).

[105] *See* Chan et al., *supra* note 102, at 504; Paul C. Tang et al., *Comparison of Methodologies for Calculating Quality Measures Based on Administrative Data Versus Clinical Data from an Electronic Health Record System: Implications for Performance*, 14 J. AM. MED. INFORMATICS ASS'N 10, 10 (2007).

[106] Joseph S. Ross et al., *State-Sponsored Public Reporting of Hospital Quality: Results Are Hard to Find and Lack Uniformity*, 29 HEALTH AFF. 2317, 2318-19 (2010); HANYS QUALITY INST., UNDERSTANDING PUBLICLY REPORTED HOSPITAL QUALITY MEASURES: INITIAL STEPS TOWARD ALIGNMENT, STANDARDIZATION, AND VALUE, 1-3 (2007), *available at* http://www.hanys.org/publications/upload/hanys_quality_report_card.pdf.

[107] *See* Ross et al., *supra* note 106, at 2318; *What Is Hospital Compare?*, U.S. DEP'T. HEALTH & HUMAN SERVS., http://www.hospitalcompare.hhs.gov/About/WhatIs/What-Is-HOS.aspx (last visited Oct. 22, 2013).

[108] *See* Barbara J. Evans, *Seven Pillars of a New Evidentiary Paradigm: The Food, Drug, and Cosmetic Act Enters the Genomic Era*, 85 NOTRE DAME L. REV. 419, 479–85 (2010) (discussing "infrastructure to develop evidence for postmarket observational studies").

[109] FDAAA of 2007, Pub. L. No. 110-85, 121 Stat. 823 (codified as amended in scattered sections of 21 U.S.C.).

[110] 21 U.S.C. § 355(o)(3) (Supp. IV 2010) (discussing post-approval studies).

[111] *See id*. § 355(o)(3)(D) (stating that clinical trials should be conducted only if other types of studies would be inadequate).

[112] *See supra* notes 45, 50-52 and accompanying text.

[113] Friedrich K. Port, *Role of Observational Studies Versus Clinical Trials in ESRD Research*, 57 KIDNEY INT'L S3, S3 (2000), *available at* http://www.nature.com/ki/journal/v57/n74s/full/4491615a.html (stating that "[r]andomized controlled clinical trials have been considered by many to be the only reliable source for information in health services research"). *See also* Sharona Hoffman, *The Use of Placebos in Clinical Trials: Responsible*

in the literature.[114] These include the studies' relatively short duration, small size, and limited generalizability.[115] Congress thus opted to supplement pre-approval clinical trials with post-marketing surveillance. Emerging evidence concerning drug safety problems may not only be illuminating for physicians, but also may lead to legal interventions. The FDA may implement regulatory measures to manage drug risks through Risk Evaluation and Mitigation Strategies,[116] or it may require changes in drug labeling.[117] In cases of imminent public danger, the FDA may also withdraw or suspend its approval of the drug[118] or ask manufacturers to remove drugs voluntarily from the market, as the Agency did in 2010 in the case of the pain medication propoxyphene (Darvocet).[119]

Like the United States, the European Union is pursuing initiatives to enhance drug safety monitoring.[120] The European Commission has funded the "Exploring and Understanding Adverse Drug Reactions by Integrative Mining of Clinical Records and Biomedical Knowledge" project (EU-ADR), which launched in 2008.[121] EU-ADR is designed to exploit data from over thirty million patients' EHRs in the Netherlands, Denmark, the United Kingdom, and Italy.[122] Experts will use computational techniques to analyze EHRs in order to identify possible drug-related adverse events that signal a need for further scrutiny.[123]

Thus far, the United States' Sentinel System projects have focused on drugs, but this data network or other electronic resources could be used to monitor devices as well.[124] Medical devices are often complex and delicate, and failures in any of their many components can significantly endanger patient lives.[125] One example of such a failure is the erosion of insulation in leads for implantable cardioverter-defibrillators manufactured by St. Jude Medical that were recalled in 2011.[126] Commentators have called for intensified post-marketing surveillance of devices through analysis of electronically available data.[127] Registries of high-risk medical devices, such as the

---

Research or Unethical Practice?, 33 CONN. L. REV. 449, 452-54 (2001) (describing different designs of clinical trials).

[114] *See* Evans, *supra* note 108, at 439-50; Vandenbroucke, *supra* note 81, at 339.

[115] Evans, *supra* note 108, at 439-50 (arguing that observational research and randomized clinical trials are each preferable in different circumstances, depending on the particulars of the research hypothesis).

[116] 21 U.S.C. § 355-1 (Supp. IV 2010).

[117] *Id.* § 355(o)(4).

[118] *Id.* § 355(e).

[119] *Xanodyne Agrees to Remove Propoxyphene from U.S. Market*, U.S. FOOD & DRUG ADMIN., (Nov. 19, 2010), http://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm234350.htm (stating that the FDA based its request in part on a review of "postmarketing safety databases").

[120] Preciosa M. Coloma et al., *Combining Electronic Healthcare Databases in Europe to Allow for Large-Scale Drug Safety Monitoring: The EU-ADR Project*, 20 PHARMACOEPI. & DRUG SAFETY 1 (2011); Gianluca Trifirò et al., *Data Mining on Electronic Health Record Databases for Signal Detection in Pharmacovigilance: Which Events to Monitor*? 18 PHARMACOEPI. & DRUG SAFETY 1176, 1177 (2009).

[121] *See* EU-ADR PROJECT, http://ec.europa.eu/digital-agenda/en/news/exploring-and-understanding-adverse-drug-reactions-integrative-mining-clinical-records-and (last visited Oct. 22, 2013).

[122] *See* WELCOME TO THE EU-ADR WEBSITE, *available at* http://euadr-project.org/drupal/ (last visited Nov. 9, 2013).

[123] *See id.*; Coloma et al., *supra* note 120, at 2; Trifirò et al., *supra* note 120, at 1177.

[124] Frederic S. Resnic & Sharon-Lise T. Normand, *Postmarketing Surveillance of Medical Devices – Filling in the Gaps*, 366 NEW ENG. J. MED. 875, 875 (2012).

[125] *See id.*

[126] *See id.*; Robert G. Hauser, *Here We Go Again – Another Failure of Postmarketing Device Surveillance*, 366 NEW ENG. J. MED. 873, 873-74 (2012).

[127] Hauser, *supra* note 126, at 874.

Interagency Registry for Mechanically Assisted Circulatory Support (INTERMACS), are a valuable tool because they collect clinical data about patients after their devices have been implanted.[128] Ideally, automated surveillance systems would trigger alerts if the frequency of adverse events related to a device exceeded a designated threshold[129] so that healthcare providers could react appropriately and prevent further harm to patients.

4. Public Health Initiatives

Federal regulations and public health projects demonstrate that biomedical databases will also be used to promote public health goals. Healthcare providers who wish to receive government incentive payments to support EHR system implementation efforts must comply with "Meaningful Use" regulations that specify the EHR functions they must be able to perform.[130] These include sending certain lab results and reports electronically to public health agencies and providing electronic information to immunization registries.[131] Public health authorities are meant to collect the submitted information in databases and use it to conduct disease surveillance and respond to public health threats.[132]

Some public health entities have already launched programs that use electronic data.[133] Examples are programs that track information about vaccine-related adverse events, sexually transmitted diseases (STDs), and HIV/AIDS.[134]

The Centers for Disease Control and Prevention is collaborating with eight large health maintenance organizations to detect adverse events associated with vaccinations.[135] The Vaccine Safety Datalink (VSD) has access to large clinical data repositories that are linked together and provide information about almost 2.5 percent of the U.S. population.[136] Information garnered by the VSD could potentially lead to changes in state vaccination laws.[137]

New York City implemented an EHR system in 2004-05 for its ten Department of Health and Mental Hygiene public clinics that treat patients with STDs.[138] The

---

[128] Resnic & Normand, *supra* note 119, at 876; *Welcome to INTERMACS*, UAB SCHOOL MED., http://www.uab.edu/medicine/intermacs/ (last visited Oct. 22, 2013).

[129] Resnic & Normand, *supra* note 119, at 877.

[130] Leslie Lenert & David Sundwall, *Public Health Surveillance and Meaningful Use Regulations: A Crisis of Opportunity*, 102 AM. J. PUB. HEALTH e1, e1 (2012).

[131] Resnic & Normand, *supra* note 124, at 876; *Welcome to INTERMACS*, *supra* note 128.

[132] Lenert & Sundwall, *supra* note 130, at e1-e2 (arguing that the infrastructure of contemporary public health authorities is inadequate for the task of receiving and processing such large amounts of information).

[133] Sharona Hoffman & Andy Podgurski, *Big Bad Data: Law, Public Health, and Biomedical Databases*, 41 J.L. MED. & ETHICS 56, 56 (2013).

[134] *See, e.g.*, *eHealth in Public Health*, CAL. DEPT. HEALTH, http://www.cdph.ca.gov/data/informatics/Pages/eHealth.aspx (last visited Oct. 25, 2013); *Disease Prevention*, MONROE CNTY., http://www2.monroecounty.gov/health-diseases.php (last visited Oct. 25, 2013); *HIV/AIDS Program Home*, IOWA DEPT. PUB. HEALTH, http://www.idph.state.ia.us/HivStdHep/HIV-AIDS.aspx?prog=Hiv&pg=HivHome (last visited Oct. 25, 2013).

[135] Brian Hazlehurst et al., *Detecting Possible Vaccine Adverse Events in Clinical Notes of the Electronic Medical Record*, 27 VACCINE 2077, 2077 (2009).

[136] *Id.* at 2081.

[137] *See State Vaccination Requirements*, CTRS. FOR DISEASE CONTROL & PREVENTION, http://www.cdc.gov/vaccines/vac-gen/laws/state-reqs.htm (last modified Sept. 30, 2011); *State Law and Vaccine Requirements*, NAT'L VACCINE INFO. CTR., http://www.nvic.org/vaccine-laws/state-vaccine-requirements.aspx (last visited Oct. 22, 2013).

[138] *See* Rachel Paneth-Pollak et al., *Using STD Electronic Medical Record Data to Drive Public Health Program Decisions in New York City*, 100 AM. J. PUB. HEALTH 586, 586 (2010).

EHRs have enabled the Department to analyze the city's clinical services.[139] Several evaluations led the city to alter its policies in order to increase opportunities for STD testing and access to care.[140]

The Louisiana Public Health Information Exchange (LaPHIE) links statewide public health surveillance information with individual EHR data.[141] LaPHIE alerts clinicians when an HIV-positive patient who has not received HIV care for over twelve months presents at any healthcare facility for any reason, so that providers may pursue HIV care with that patient.[142] Such information exchange networks can constitute a valuable tool for combating infectious disease and assist states in fulfilling their public health responsibilities.[143]

5. Litigation

If databases of de-identified EHR information become publicly available for non-clinical purposes, litigants who seek to prove causation or harm in mass tort cases may mine them for evidence.[144] Several such databases already exist. For example, California, Texas, and Vermont have databases of inpatient hospital discharge data.[145] Selected datasets that do not directly identify patients but include diagnoses, treatments, drug intake, and other details are available for purchase by the public.[146] Private sector enterprises, such as MedMining and Strategic Healthcare Programs, also offer requestors access to their health information databases.[147]

Data availability has been promoted in academic circles as well. Professor Marc Rodwin argues that patient data should routinely be treated as public property.[148] He posits that federal law should require clinicians, hospitals, and insurers to report de-identified patient data to public authorities who would create aggregate databases that would be available to private entities, subject to public oversight.[149]

Litigants have already used epidemiological evidence in many mass tort cases, such as those alleging harm from "asbestos, Bendectin, electro-magnetic radiation, IUDs, silicone implants, and tobacco products."[150] Epidemiological data is most

---

[139] *Id.*

[140] *Id.* at 589.

[141] Jane Herwehe et al., *Implementation of an Innovative, Integrated Electronic Medical Record (EMR) and Public Health Information Exchange for HIV/AIDS*, 19 J. AM. MED. INFORMATICS ASS'N 448, 448 (2012).

[142] *Id.* at 448-49.

[143] *Id.* at 452 (Louisiana has developed similar alerts for tuberculosis patients in need of follow-up care.).

[144] Sharona Hoffman & Andy Podgurski, *Finding A Cure: The Case for Regulation and Oversight of Electronic Health Record Systems*, 22 HARV. J.L. & TECH. 103, 124 (2008).

[145] *See Inpatient Hospital Discharge Data*, CAL. DIABETES PROGRAM, http://www.caldiabetes.org/content.cfm?contentID=487&CategoriesID=31&CFID=5020870&CFTOKEN=92167121 (last visited Oct. 22, 2013); *Health Care Information User Manual, Texas Hospital Inpatient Discharge Public Use Data File*, TEX. DEP'T STATE HEALTH SERVS. http://www.dshs.state.tx.us/thcic/hospitals/Inpatientpudf.shtm (last updated Aug. 12, 2013); *VUHDDS Frequently Asked Questions*, VT. DEP'T HEALTH, http://healthvermont.gov/research/hospital-utilization/VHUR_FAQS.aspx (last visited Oct. 25, 2013).

[146] *See* Herwehe et al., *supra* note 141, at 452; Marc A. Rodwin, *Patient Data: Property, Privacy & the Public Interest*, 36 AM. J.L. & MED. 586, 615 (2010). *See also* Hoffman & Podgurski, *supra* note 8, at 95-97, 104-07, 128-31 (discussing de-identification of data).

[147] *See Welcome to MedMining*, MEDMINING, http://www.medmining.com/index.html (last visited Oct. 22, 2013); *Request Data*, STRATEGIC HEALTHCARE PROGRAMS, LLC, https://www.shpdata.com/company/requestdata.aspx (last visited Oct. 22, 2013).

[148] Rodwin, *supra* note 146, at 590.

[149] *Id.* at 589.

[150] FAIGMAN ET AL., *supra* note 20, at 339-40.

often employed with respect to causation, and it is not unusual for the courts to accept it as persuasive.[151] In the future, observational studies based on biomedical databases may frequently aid in developing compelling epidemiological evidence.

While plaintiffs will attempt to prove causation through database analysis, defendants may use the same tool to undermine plaintiffs' claims of causation.[152] For example, defense counsel could argue that plaintiffs' illnesses are linked to genetic factors rather than to defendants' products.[153] Researchers have found that genetic variants influence conditions that are often at the center of legal disputes. Genetic variants may increase individuals' likelihood of being heavy smokers,[154] of developing lung cancer or chronic obstructive pulmonary disease,[155] and of suffering from carpal tunnel syndrome.[156] Commentators predict that defendants will increasingly attempt to defeat plaintiffs' allegations by arguing that "the genes did it."[157]

The Burlington Northern and Santa Fe Railway Company attempted to use this approach more than a decade ago.[158] When several employees claimed that they suffered from carpal tunnel syndrome (CTS) caused by their work, Burlington Northern required them to provide a blood sample that would be tested for a genetic marker believed to be associated with CTS.[159] In addition, a study of tobacco manufacturers' defenses in personal injury cases brought by smokers with cancer revealed that in at least one case, *Mehlman v. Philip Morris, Inc.*, a manufacturer

---

[151] *Id.* at 341; Norris v. Baxter Healthcare Corp., 397 F.3d 878, 882 (10th Cir. 2005) (noting that "the body of epidemiology largely finds no association between silicone breast implants and immune system diseases").

[152] Sharon Milberger et al., *Tobacco Manufacturers' Defence Against Plaintiffs' Claims of Cancer Causation: Throwing Mud at the Wall and Hoping Some of It Will Stick*, 15 TOBACCO CONTROL iv17, iv22 (Supp. IV 2006).

[153] *See* Bowen v. E.I. Du Pont De Nemours & Co., No. Civ.A. 97C-06-194 CH, 2005 WL 1952859, at *4 (Del. Super. Ct. 2005) (involving a claim by defendant that the injuries and condition in question "constitute CHARGE Syndrome, which is generally thought to be genetic, as opposed to environmental, in origin").

[154] Nancy L. Saccone et al., *Multiple Independent Loci at Chromosome 15q25.1 Affect Smoking Quantity: a Meta-Analysis and Comparison with Lung Cancer and COPD*, 8 PLOS GENETICS 1, 3 (2010); Thorgeir E. Thorgeirsson et al., *Sequence Variants at CHRNB3-CHRNA6 and CYP2A6 Affect Smoking Behavior*, 42 NATURE GENETICS 448, 448 (2010).

[155] Paul Brennan et al., *Genetics of Lung-Cancer Susceptibility*, 12 LANCET ONCOLOGY 399, 403-04 (2011); Peter Broderick et al., *Deciphering the Impact of Common Genetic Variation on Lung Cancer Risk: A Genome-Wide Association Study*, 69 CANCER RES. 6633, 6633 (2009); Michael H. Cho et al., *A Genome-Wide Association Study of COPD Identifies A Susceptibility Locus on Chromosome 19q13*, 21 HUMAN MOLECULAR GENETICS 947, 948-49 (2012); Saccone et al., *supra* note 154, at 3; Thorgeirsson et al., *supra* note 154, at 448.

[156] Alan J. Hakim et al., *The Genetic Contribution to Carpal Tunnel Syndrome in Women: A Twin Study*, 47 ARTHRITIS & RHEUMATISM 275, 277 (2002); Santiago Lozano-Calderon et al., *The Quality and Strength of Evidence for Etiology: Example of Carpal Tunnel Syndrome*, 33A J. HAND SURGERY AM. 525, 532-33 (2008).

[157] Gold, *supra* note 23, at 412; Diane E. Hoffman & Karen H. Rothenberg, *Judging Genes: Implications of the Second Generation of Genetic Tests in the Courtroom*, 66 MD. L. REV. 858, 867 (2007); Gary E. Marchant, *Genetic Data in Toxic Tort Litigation*, 14 J.L. & POL'Y 7, 12 (2006); Susan Poulter, *Genetic Testing in Toxic Injury Litigation: The Path to Scientific Certainty or Blind Alley?*, 41 JURIMETRICS J. 211, 217-20 (2001).

[158] EEOC v. Burlington N. and Santa Fe Ry. Co., No. 02-C-0456, 2002 WL 32155386, at *1 (E. D. Wis. 2002).

[159] *Id.* The case settled before trial.

cited "heredity" as one of the factors that caused the plaintiff's cancer.[160] A jury found against the plaintiff in 2001.[161]

It is also possible that self-appointed watchdogs may mine public databases to determine whether exposure to particular products or substances results in adverse health consequences. Based on their findings, they could publicize supposed problems, demand government intervention or encourage lawyers to initiate litigation.

III. LIMITATIONS OF BIOMEDICAL DATABASES

Biomedical databases constitute a potentially invaluable research resource, but researchers, analysts, and other stakeholders must appreciate that existing EHRs and genomic data often contain errors, are incomplete, or suffer from other shortcomings. While any collection of research data may be contaminated by inaccuracies, biomedical databases may be particularly flawed. The information in EHRs is initially collected for clinical and billing purposes, and thus it might be ill-suited for research.[162] Moreover, the sheer volume of information contained in large biomedical databases[163] and the complex analytical methods and tools required to conduct large-scale observational studies[164] create myriad opportunities for the introduction of errors and omissions. While improved technology may remedy many database shortcomings in the future,[165] these deficiencies are currently of serious concern. This section examines a variety of potential data quality problems.

A. DATA ENTRY ERRORS

EHR databases may be tainted by data entry errors. Digitization can prevent some data quality problems, such as those associated with illegible handwriting,[166] but it does not remove the risk that entries in patient records will be incorrect. Inaccurate clinical data will become inaccurate research data if the EHRs are imported to research databases or are accessible to investigators through federated systems.

Clinicians who enter data into records can easily invert numbers or mistype words, select wrong items from a menu, check the wrong box, obtain inaccurate information from patients, and make other data entry mistakes.[167] A variety of researchers and commentators have recognized that computerization itself can contribute to errors involving "loss of concentration, slip of the finger, or

---

[160] Milberger et al., *supra* note 152, at iv22 tbl. 6; Mehlman v. Philip Morris, Inc., No. L-1141-99, (Sup. Ct. N.J. filed Feb. 4, 1999), *available at* http://legacy.library.ucsf.edu/tid/ekz52d00/pdf (Legacy Tobacco Documents Library).

[161] Milberger et al., *supra* note 152, at iv22; Stephen D. Sugarman, Address at the Robert Wood Johnson Foundation's SAPRP Conference: Tobacco Litigation Update (revised as of November 5, 2001) 2 (Nov. 14, 2001), *available at* http://www.law.berkeley.edu/sugarman/tobacco_litigation_upate_october_2001_.doc. The decedent, plaintiff's wife, had stopped smoking 30 years before her death.

[162] *See infra* Part III.C.

[163] *See supra* Part II.A. (discussing database initiatives).

[164] *See infra* Parts III.D., IV (discussing software failures and the challenges of causal inference).

[165] *See infra* Part VI.A.

[166] WIN PHILLIPS & YANG GONG, HUMAN COMPUTER INTERACTION: INTERACTING IN VARIOUS APPLICATION DOMAINS 589, 591 (Julie A. Jacko ed., 2009).

[167] Ancker et al., *supra* note 40, at 61; Botsis et al., *supra* note 40, at 3-4; Sharona Hoffman & Andy Podgurski, *E-Health Hazards: Provider Liability and Electronic Health Record Systems*, 24 BERKELEY TECH. L.J. 1523, 1544-45 (2009) (discussing input errors).

distraction,"[168] and some of these errors are unique to digitized rather than paper records. For example, in order to save time, clinicians may copy and paste relevant narrative from prior visit notes and place it in the wrong location in the chart or fail to carefully edit or update it.[169] They may also neglect to save notes in unclosed charts, or they may misinterpret lab results that are displayed in a confusing fashion and consequently make incorrect notes about patients' progress.[170]

Several studies have attempted to estimate error rates in EHRs. A study of a number of research databases containing information about oncology patients at an academic medical center found error rates of 2.3-26.9%.[171] Errors were attributable to data entry mistakes, misinterpretation of hard-copy documents when information was typed into the database, and perpetuation of errors that were contained in the original paper documents and then copied during the data entry process.[172] Another publication found an average error rate of 9.76%.[173] A small study involving twenty-five Israeli physicians revealed that over sixty percent of participants admitted that they had mistyped information, entered information into the wrong patients' charts, or selected an incorrect item from an electronic menu of choices, though pharmacists often served as a safeguard against actual treatment mistakes.[174]

A different study highlighted the frequency of medication discrepancies in hospitals.[175] The study identified 2066 medication discrepancies relating to 180 patients, of which 939 were unintentional and therefore constituted errors.[176] Among the errors, 257 instances had the potential to harm patients.[177] The majority (72%) of errors stemmed from mistakes made while taking preadmission medication histories, and the remainder was caused by failure to reconcile medication history with discharge orders.[178] The most common reasons for potentially dangerous errors were the patients' own confusion about the medications they took before hospital admission, medication changes during hospitalization, and intern assistance with recording patient histories.[179]

---

[168] Robert E. Hirschtick, *Copy-and-Paste*, 295 J. AM. MED. ASS'N 2335, 2335-36 (2006); Sheila Roszell & Cheryl Stewart, *E-charting Point-of-Care Data Entry Dilemma*, 38 J. NURSING ADMIN. 417, 417 (2008).

[169] PHILLIPS & GONG, *supra* note 166, at 591.

[170] *Id.*

[171] Saveli I. Goldberg et al., *Analysis of Data Errors in Clinical Research Databases*, 2008 AMIA ANN. SYMP. PROC. 242, 244.

[172] *Id.* A second study by the same authors examined weight measurement errors. An algorithm checked the weight records of 25,000 patients, including 420,469 weight entries. It found errors in .58% of entries in the records of "up to 7% of all patients." *See* Saveli Goldberg et al., *A Weighty Problem: Identification, Characteristics and Risk Factors for Errors in EMR Data*, 2010 AMIA ANN. SYMP. PROC. 251, 253-54.

[173] Meredith L. Nahm et al., *Quantifying Data Quality for Clinical Trials Using Electronic Data Capture*, 3 PLOS ONE 1, 1 (2008) (discussing a literature review of "42 articles that provided source-to-database error rates, primarily from registries" and finding that the "average error rate across these publications was 976 errors per 10,000 fields"). *See also* Krystl Haerian et al., *Use of Clinical Alerting to Improve the Collection of Clinical Research Data*, 2009 AMIA ANN. SYMP. PROC. 218, 218 (discussing data error rates pertaining to research databases).

[174] Aviv Shachak et al., *Primary Care Physicians' Use of an Electronic Medical Record System: A Cognitive Task Analysis*, 24 J. GEN. INTERNAL MED. 341, 342-44 (2009).

[175] Jennifer R. Pippins et al., *Classifying and Predicting Errors of Inpatient Medication Reconciliation*, 23 J. GEN. INTERNAL MED. 1414, 1414 (2008).

[176] *Id.* at 1416.

[177] *Id.*

[178] *Id.*

[179] *Id.* at 1417.

Data errors can skew the outcomes of research studies. A study that focused on pneumonia cases emphasized that even a small number of errors can have "a relatively large effect" on mortality estimates.[180] Other researchers have confirmed that even error rates as small as one to five percent could cause significant inaccuracies in mortality and adverse event estimates.[181] Database operators and analysts also cannot ignore the possibility that in the worst-case scenario, hackers could access biomedical databases and intentionally introduce errors or alter records.[182] Data errors that do not occur at random are especially problematic because they may systematically bias research outcomes.[183]

Like EHR databases, genome-sequencing projects have been plagued by genome annotation errors.[184] One study found that the frequency of misannotation has grown during the years 1993 to 2005 and that the protein sequence databases that were studied exhibited levels of annotation errors spanning from zero to over sixty percent.[185] Problems range from plain spelling mistakes to "incorrectly tuned parameters in automatic annotation pipelines,"[186] and they can significantly impact scientists' ability to learn about the evolution, biology, metabolic processes, and other aspects of organisms.[187] Experts have called for the development of guidelines and standards to improve the submission, retrieval, processing, and analysis of genomic data.[188]

B. INCOMPLETE OR FRAGMENTED DATA

Incomplete or fragmented data may also compromise the reliability of EHR database information. At times, EHR data does not include all of the information needed for particular research projects.[189] Clinicians generally do not approach the task of EHR documentation with research studies in mind.[190] To illustrate, in one

---

[180] George Hripcsak et al., *Bias Associated with Mining Electronic Health Records*, 6 J. BIOMED. DISCOVERY & COLLABORATION 48, 52 (2011).

[181] Steve Gallivan & Christina Pagel, *Modelling of Errors in Databases*, 11 HEALTH CARE MGMT. SCI. 35, 39 (2008); Christina Pagel & Steve Gallivan, *Exploring Potential Consequences on Mortality Estimates of Errors in Clinical Databases*, 20 IMA J. MGMT. MATHEMATICS 385, 391 (2009).

[182] *See* Jennifer Dobner, *Fallout Grows from Hacking of Utah Health Database*, REUTERS (Apr. 9, 2012), http://www.reuters.com/article/2012/04/10/us-usa-hackers-utah-idUSBRE83904G20120410 (discussing an incident in which Eastern European hackers gained access to state health records of over 780,000 patients).

[183] Sander Greenland, *Multiple-Bias Modelling for Analysis of Observational Data*, 168 J. ROYAL STAT. SOC'Y: SERIES A (STAT. IN SOC'Y) 267, 267-68 (2005).

[184] Murray P. Cox et al., *SolexaQA: At-A-Glance Quality Assessment of Illumina Second-Generation Sequencing Data*, 11 BMC BIOINFORMATICS 485, 485 (2010); William Klimke et al., *Solving the Problem: Genome Annotation Standards Before the Data Deluge*, 5 STANDARDS IN GENOMIC SCIS. 168, 168 (2011); Alexandra M. Schnoes et al., *Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies*, 5 PLOS COMPUTATIONAL BIOLOGY 1, 1 (Dec. 2009).

[185] Schnoes et al., *supra* note 184, at 2, 6.

[186] Klimke et al., *supra* note 184, at 169.

[187] *Id.* at 168.

[188] *Id.* at 170.

[189] Craig D. Newgard et al., *Electronic Versus Manual Data Processing: Evaluating the Use of Electronic Health Records in Out-of-Hospital Clinical Research*, 19 ACAD. EMERGENCY MED. 217, 224 (2012).

[190] M. Alan Brookhart et al., *Confounding Control in Healthcare Database Research: Challenges and Potential Approaches*, 48 MED. CARE S114, S115 (2010) (explaining that one of the limitations of healthcare databases is that "because the data were not collected as part of [a] designed study, many variables that the researcher might wish to have access to remain unrecorded").

instance, a manual review of EHR data from the New York-Presbyterian Hospital clinical data warehouse revealed that when pneumonia patients died in the emergency department, clinicians "spent little time documenting symptoms so that in the electronic health record, the patient appeared to be healthy other than the death."[191]

Data about treatment outcomes is particularly likely to be missing.[192] For example, a patient discharged from an emergency room may not seek further care at all or may later visit a physician who has a different EHR system, making it impossible to track whether the treatment was effective over the long-term.[193] The absence of information about treatment outcomes in an EHR is difficult to interpret. It could mean that the therapy cured the patient, and she did not report the positive result because she required no follow-up, but it could also mean she experienced no relief, or her condition deteriorated and she went to a specialist or another doctor.

Data fragmentation often exists because different facilities have EHR systems that are not interoperable.[194] Thus, seriously ill patients who are treated at multiple medical centers as their disease progresses may have their records divided among several EHR systems, and these are unlikely to be integrated into a single research database.[195]

C. DATA CODING, STANDARDIZATION, AND EXTRACTION

Medical data in EHRs is often coded, but the coding can be inconsistent, incorrect, or misleading.[196] Healthcare providers code data in accordance with the International Classification of Disease, version 9 (ICD-9), developed by the World Health Organization, and the Current Procedural Terminology, version 4 (CPT-4), formulated by the American Medical Association to record procedures and laboratory tests.[197] By 2014, healthcare providers will be required to switch to ICD-10, which has approximately 155,000 codes rather than ICD-9's 17,000 codes.[198]

---

[191] Hripcsak et al., *supra* note 180, at 50. It is especially challenging to analyze the effects of treatments or exposures in the face of data with missing items if they are not missing completely at random. Such non-random omissions create the potential for biased results. Craig H. Mallinckrodt et al., *Assessing and Interpreting Treatment Effects in Longitudinal Clinical Trials with Missing Data,* 53 BIOLOGICAL PSYCHIATRY 754, 755 (2003).

[192] Newgard et al., *supra* note 189, at 225.

[193] *Id.*

[194] Interoperable systems can communicate with each other, exchange data, and operate seamlessly and in a coordinated fashion across organizations. BIOMEDICAL INFORMATICS: COMPUTER APPLICATIONS IN HEALTH CARE & BIOMEDICINE 952 (Edward H. Shortliffe & James J. Cimino eds., 2006).

[195] Botsis et al., *supra* note 40, at 4 (stating that the EHR system that was mined for purposes of the study did not contain records of patients who were transferred to dedicated cancer centers because of the severity of their disease or who had initially been treated elsewhere).

[196] Naren Ramakrishnan et al., *Mining Electronic Health Records*, COMPUTER 95, 96 (2010), *available at* http://people.cs.vt.edu/ramakris/papers/ehrmining10.pdf (discussing "the lack of data standards").

[197] *Id.* at 95.

[198] *HHS Issues Final ICD-10 Sets and Updated Electronic Transaction Standards Rules*, U.S. DEP'T OF HEALTH & HUMAN SERVS. (Jan. 15, 2009), http://www.hhs.gov/news/press/2009pres/01/20090115f.html; *ICD-10*, CTRS. FOR MEDICARE & MEDICAID SERVS., http://www.cms.gov/Medicare/Coding/ICD10/index.html?redirect=/ICD10 (last modified Sept. 9, 2013) (indicating that HHS published a proposed rule that would delay the compliance date, setting it at October 1, 2014 rather than October 1, 2013); *ICD-10 Code Set to Replace ICD-9*, AM. MED. ASS'N, http://www.ama-assn.org/ama/pub/physician-resources/solutions-managing-your-practice/coding-billing-insurance/hipaahealth-insurance-portability-accountability-act/transaction-code-set-standards/icd10-code-set.page (last visited Oct. 22, 2010).

EHR systems also provide their own selection menus with various codes to facilitate data entry.[199]

Critics have charged that healthcare providers often use coding to maximize billing opportunities, rather than to build the most accurate record possible.[200] Menus and lists built into EHR systems may encourage clinicians to charge for more services by suggesting items for which they can bill and making it easy to click on boxes for billing purposes.[201] One study found that the practice of "upcoding" services provided to Medicare patients was very common and may account for as much as fifteen percent of Medicare's expenditures for general office visits, or $2.13 billion annually.[202] Medicare coding contains many ambiguities that enable doctors' offices to make strategic choices that will enhance their revenues.[203] The Federal Government has recognized the problem and is reportedly investigating Medicare fraud related to upcoding.[204]

Several studies have specifically identified coding inadequacies as an obstacle to secondary use of EHR data.[205] According to one source, ICD-9 codes are not specific enough for cancer to enable researchers to distinguish primary tumors from metastatic ones.[206] ICD-9 coding deficiencies will remain a problem even after the adoption of ICD-10 because existing patient records will contain ICD-9 codes.[207] A British study examined the separate codes offered by a general practice EHR system and concluded that the coding screen did not clarify which designation was appropriate for acute rather than more moderate disease and which range of codes indicate the presence of chronic obstructive pulmonary disease.[208] In addition, the study found that different physician groups use different codes to label the same type of patient, so, for example, some patients receiving medication to combat osteoporosis were not coded as having osteoporosis.[209]

Discrepancies and variability in quality of data may be attributable to a number of factors. According to one source, record-keeping was found to be more meticulous if it was relevant to financial gain, contractual obligations, or external

---

[199] Simon de Lusignan et al., *Routinely-Collected General Practice Data Are Complex, but With Systematic Processing Can Be Used for Quality Improvement and Research*, 14 INFORMATICS IN PRIMARY CARE 59, 62 (2006) (analyzing a "picking list . . . taken from a general practice computer system").

[200] Ramakrishnan et al., *supra* note 196, at 95; Peter V. Jensen et al., *Mining Electronic Health Records: Towards Better Research Applications and Clinical Care*, 13 NATURE REVS. GENETICS 395, 401 (2012) (mentioning "systematic erroneous use of disease terminology codes caused by strategic billing"); Kohane, *supra* note 9, at 424 (asserting that "the primary driver of EHR implementation has been clinical reimbursement rather than the potential for reuse of the clinical data for research").

[201] OFFICE INSPECTOR GEN., U.S. DEPT. HEALTH & HUMAN SERVS., TOP MANAGEMENT & PERFORMANCE CHALLENGES (2012), https://oig.hhs.gov/reports-and-publications/top-challenges/2012/issue09.asp.

[202] Christopher S. Brunt, *CPT Fee Differentials and Visit Upcoding Under Medicare Part B*, 20 HEALTH ECON. 831, 840 (2011). The $2.13 billion figure is in 2007 dollars. *Id.*

[203] *Id.*

[204] Andrea K. Walker, *Medical Billing a Target of Fraud Investigations*, BALT. SUN, Jan. 12, 2012, http://articles.baltimoresun.com/2012-01-12/health/bs-hs-umms-malnutrition-response-2-20120112_1_health-care-fraud-coding-billing.

[205] *See, e.g.*, Siaw-Teng Liaw et al., *Data Quality and Fitness for Purpose of Routinely Collected Data – A General Practice Case Study from an Electronic Practice-Based Research Network (ePBRN)*, 2011 AMIA ANN. SYMP. PROC. 785, 789 (noting a "lack of implemented terminology and coding standards").

[206] Botsis et al., *supra* note 40, at 4.

[207] *See* AM. MED. ASS'N, *supra* note 198.

[208] De Lusignan et al., *supra* note 199, at 62.

[209] *Id.*

viewers, and more sloppy if it was used internally only.[210] In addition, the existence of multiple fields for documentation and variable practices among different personnel (e.g., documenting scheduled date versus actual date of an appointment) can generate irregularities.[211] Yet another complication is medical offices' inconsistent use of terms, phrases, and abbreviations. To illustrate, the abbreviation "MS" can mean "mitral stenosis," "multiple sclerosis," "morphine sulfate," or "magnesium sulfate."[212] If the context is not clear from the EHR, a reader might fail to understand what is meant by "MS" in a particular record.

Further challenges arise from the existence of free text narrative in EHRs. EHR systems allow providers to enter both coded information and unstructured, natural-language notes about patients.[213] Important information that is not captured in structured data may be contained in notes, and such information is much more difficult to extract accurately from EHRs for secondary use.[214] As an example, database mining may fail to reveal a link between worsening asthma and smoking if the progression of asthma is coded but smoking history is described only in free-text clinical notes.[215] Similarly, family history and information about adverse reactions to drugs are likely to be presented in narrative form rather than in coded form.[216] Experts may employ natural-language processing tools to extract data from free-text narrative, but these techniques are still developing and are often imperfect.[217]

If diagnoses, measurements, or medical histories contained in EHRs are not standardized or are inaccessible because they do not appear in structured form, database contents may be inadequate for secondary uses.[218] Similarly, if medical vocabulary is not harmonized, researchers may misunderstand or be unable to make sense out of database records.

D. ERRORS DUE TO SOFTWARE FAILURES

Errors in research data and in the results of computer analysis can also result from incorrect processing by defective software. EHR systems, like any complex software system, may contain unrecognized or unrepaired software defects.[219] Such defects may cause some of the data recorded in a patient's EHR to be incorrect. A value that is incorrect but is nevertheless plausible may not be discovered and hence may remain in a patient's EHR when it is used in a research study.

Even if the raw data is correct, errors may arise during data analysis due to defects in the software used to conduct the analysis.[220] This is particularly likely if

---

[210] Ancker et al., *supra* note 40, at 61.

[211] *Id.*

[212] Christopher G. Chute, *Medical Concept Representation, in* MEDICAL INFORMATICS: KNOWLEDGE MANAGEMENT & DATA MINING IN BIOMEDICINE 163, 170 tbl. 6-1 (Hsinchun Chen et al. eds., 2010).

[213] S. Trent Rosenbloom et al., *Data from Clinical Notes: A Perspective on the Tension Between Structure and Flexible Documentation*, 18 J. AM. MED. INFORMATICS ASS'N 181, 181-82 (2011).

[214] *Id.* at 184 (stating that some physicians prefer the flexibility and expressivity of notes); Ramakrishnan et al., *supra* note 196, at 96-97 (explaining that "much of the relevant data is 'locked up' in free text documents").

[215] Ramakrishnan et al., *supra* note 196, at 97.

[216] Kohane, *supra* note 9, at 420.

[217] Kho et al., *supra* note 12 at 2-4; Ramakrishnan et al., *supra* note 196, at 97.

[218] Andrea L. Benin et al., *How Good Are the Data? Feasible Approach to Validation of Metrics of Quality Derived from an Outpatient Electronic Health Record*, 26 AM. J. MED. QUALITY 441, 441 (2011).

[219] *See* Hoffman & Podgurski, *supra* note 167, at 1552.

[220] Les Hatton, *The Chimera of Software Quality*, 40 COMPUTER 104, 104 (2007).

the software is complex and is developed by scientists or their assistants without the help of skilled software developers.[221] Inexperienced programmers are likely both to create incorrect software and to test it inadequately.[222] Even commercially developed biomedical research software, however, may produce erroneous results.[223] Ideally, scientists should work in close cooperation with software experts to develop and thoroughly validate software used in biomedical research.

IV. THE CHALLENGES OF BIAS AND CAUSAL INFERENCE

As argued above, the availability of large collections of data does not guarantee sound study outcomes.[224] Even if the data itself is unblemished, those analyzing it will face many obstacles to drawing correct conclusions from it. This Part analyzes subtle but important problems of bias affecting observational studies, in particular, selection bias, confounding bias, and measurement bias. An understanding of these issues is crucial for anyone seeking to interpret the results of biomedical database studies.

It is worth noting that one type of problem, namely sampling error, is less likely to be a major concern in observational studies based on large biomedical databases. Sampling error arises when inferences are drawn from observations of a randomly chosen sample of individuals whose statistical characteristics (e.g., smoking rate, average weight, or average duration of illness) differ from those of the source population due to random chance.[225] Conclusions drawn from the particular sample, therefore, cannot be accurately generalized to the population of interest. Sampling error tends to diminish as the sample size increases, and the extent of this error is well characterized by traditional statistical methods such as computation of confidence intervals.[226] Since the research databases contemplated in this Article are large and automatically queried, considerable samples can be processed efficiently to reduce sampling error. However, selection bias can nevertheless result if the individuals whose records are stored in a biomedical database and who satisfy the criteria for a given study do not constitute a random sample of the population targeted by the study.

A. SELECTION BIAS

If data subjects have the opportunity to opt out of inclusion in a database or if certain individuals' records are otherwise excluded, a class of problems often called

---

[221] *See* Diane F. Kelly, *A Software Chasm: Software Engineering and Scientific Computing*, 24 IEEE SOFTWARE118, 118-20 (Nov.-Dec. 2007); Hatton, *supra* note 220, at 104; Rebecca Sanders & Diane Kelly, *Dealing with Risk in Scientific Software Development*, 25 IEEE SOFTWARE 21, 27 (July-Aug. 2008).

[222] *See* Kelly, *supra* note 221, at 118.

[223] *See* Nicole K. Henderson-MacLennan et al., *Pathway Analysis Software: Annotation Errors and Solutions*, 101 MOLECULAR GENETICS & METABOLISM 134, 137-38 (2010); Sanders & Kelly, *supra* note 207, at 25.

[224] *See supra* Part III.

[225] KENNETH J. ROTHMAN ET AL., MODERN EPIDEMIOLOGY 148-9 (3d ed. 2008).

[226] *Id.* at 149. According to one source, a "confidence interval calculated for a measure of treatment effect shows the range within which the true treatment effect is likely to lie (subject to a number of assumptions)." Huw T. O. Davies & Iain K. Crombie, *What Are Confidence Intervals and P-Values?*, WHAT IS…? SERIES (Apr. 2009), *available at* http://www.medicine.ox.ac.uk/bandolier/painres/download/whatis/what_are_conf_inter.pdf.

"selection bias" may arise.[227] Selection bias may occur when the subset of individuals studied is not representative of the patient population of interest.[228] This kind of selection bias could manifest, for example, if a disproportionate number of people of one ancestry or economic class opt out of participating in a database.[229] It can likewise exist if individuals with certain behavior traits that might be important in some studies—such as diet, exercise, smoking status, and alcohol or drug consumption—choose not to participate or cannot access medical facilities in which studies take place.[230] Selection bias can distort assessments of measures such as disease prevalence or exposure risk because study estimates will differ systematically from the true values of these measures for the target population.[231] That is, the estimates will not be generalizable from the research subjects to the larger population about which analysts wish to draw conclusions.[232]

Another, more subtle kind of selection bias, which is also called "collider-stratification bias,"[233] "collider-bias,"[234] or "M-bias,"[235] is specific to causal-effect studies.[236] These studies typically seek to measure the average beneficial effect on patients of a particular treatment or the average harmful effect on individuals of a particular exposure.[237] Collider-stratification bias occurs in analyzing study data when the analysis is conditioned on (e.g., stratified by) one or more levels of a variable that is a common effect (a "collider") of both the treatment/exposure variable and the outcome variable or that is a common effect of a cause of the treatment/exposure and a cause of the outcome.[238]

Consider the following classic example. Commonly, some patients are lost to follow-up, and thus outcome measurements that would be essential for research purposes are unavailable. The data from these patients cannot be included in studies. Both the treatment and outcome at issue may influence which patients stop seeking medical care. Patients may fail to return for follow-up both because the treatment is unpleasant (treatment factor) and because they actually feel better and don't see a need to return to their doctors (an outcome factor). The loss of these study subjects can create a spurious statistical association between the treatment/exposure variable and the outcome variable that becomes mixed with and distorts the true causal effect of the former on the latter.[239] Because collider-stratification bias is associated with

---

[227] *See* DAVID L. FAIGMAN ET AL., MODERN SCIENTIFIC EVIDENCE: THE LAW AND SCIENCE OF EXPERT TESTIMONY § 4:16 (2008). ROTHMAN ET AL., *supra* note 225, at 196.

[228] Franklin G. Miller, *Research on Medical Records Without Informed Consent,* 36 L. MED. & ETHICS 560, 560 (2008); *see* COMM. ON HEALTH RESEARCH & THE PRIVACY OF INFO.: THE HIPAA PRIVACY RULE, INST. OF MED. (IOM), BEYOND THE HIPAA PRIVACY RULE: ENHANCING PRIVACY, IMPROVING HEALTH THROUGH RESEARCH 209 (Sharyl J. Nass et al., 2009) [hereinafter IOM REPORT].

[229] *See* IOM REPORT, *supra* note 228, at 213-14.

[230] *Id.* at 212.

[231] Miguel A. Hernán et al., *A Structural Approach to Selection Bias*, 15 EPIDEMIOLOGY 615, 615 (2004) (explaining that "the common consequence of selection bias is that the association between exposure and outcome among those selected for analysis differs from the association among those eligible").

[232] *See* IOM REPORT, *supra* note 228, at 209.

[233] Stephen R. Cole, *Illustrating Bias Due to Conditioning on a Collider*, 39 INT'L J. EPIDEMIOLOGY 417, 417 (2010).

[234] ROTHMAN ET AL., *supra* note 225, at 185.

[235] Hernán et al., *supra* note 231, at 618.

[236] Miguel A. Hernán, *A Definition of Causal Effect for Epidemiological Research*, 58 J. EPIDEMIOLOGY & COMMUNITY HEALTH 265, 265 (2004).

[237] *Id.*

[238] ROTHMAN ET AL., supra note 225, at 185.

[239] Hernán et al., *supra* note 231, at 617-18.

the exclusion of some patients from a study, it is categorized as a type of selection bias.[240]

B. CONFOUNDING BIAS

In observational causal-effect studies, confounding bias (confounding) may be an even greater concern than selection bias.[241] "Classical" confounding occurs because of the presence of a common cause of the treatment/exposure variable and the outcome variable.[242] Confounding is different from collider-stratification bias because it involves a common *cause* of the treatment/exposure and outcome variables rather than a common *effect* of the variables.[243]

The following hypothetical illustrates classical confounding. Suppose a physician's treatment choices are influenced by the severity or duration of a patient's disease, which also influence the outcome of treatment.[244] Thus, patients at a later stage of a disease may receive one treatment (treatment *A*) and those who are at an earlier stage may receive a different therapy (treatment *B*). At the same time, sicker patients may have worse treatment outcomes than healthier individuals. Unless such a common cause, which is called a "confounding variable" or "confounder," is adjusted for appropriately during statistical data analysis, it may induce a spurious association between the treatment variable and the outcome variable, which distorts estimation of the true causal effects of treatments.[245] In other words, researchers may reach incorrect conclusions regarding the efficacy of the two treatments because of the confounding variable: the degree of sickness suffered by patients receiving the different therapies. Treatment *A* may appear to be less effective than treatment *B* not because it is in fact an inferior therapy but because so many of the patients receiving treatment *A* are in a late stage of the disease and would not do well no matter what treatment they received. This particular form of confounding, called "confounding by indication," is especially challenging to adjust for, because it may involve multiple factors that influence physicians' treatment decisions.[246]

Socioeconomic factors and patient lifestyle choices may also be confounders. Those who lack financial resources or adequate health coverage may select less expensive treatments not because those are the best choices for them but because those are the only affordable options.[247] Low income may also separately lead to poor health for reasons such as poor nutrition or financial stress. In the case of preventive care, a treatment's perceived benefits may be amplified because health-oriented individuals interested in the intervention also pursue exercise, low-fat diets, and other health-promoting behaviors. These patients' impressive outcomes thus would not be associated solely with the preventive measure.[248]

---

[240] Collider-stratification bias may also occur because of a poorly conceived attempt to adjust for confounding bias, discussed below. Hernán et al., *supra* note 231, at 620 (stating that "[a]lthough stratification is commonly used to adjust for confounding, it can have unintended effects").

[241] *See* Sander Greenland, *Quantifying Biases in Causal Models: Classical Confounding vs. Collider-Stratification Bias*, 14 EPIDEMIOLOGY 300, 306 (2003).

[242] *See id.* at 301.

[243] Hernán et al., *supra* note 231, at 615.

[244] *See* Bruce M. Psaty & David S. Siscovick, *Minimizing Bias Due to Confounding by Indication in Comparative Effectiveness Research*, 304 J. AM. MED. ASS'N 897, 897 (2010).

[245] Hernán et al., *supra* note 231, at 618.

[246] *See* Jaclyn L.F. Bosco et al., *A Most Stubborn Bias: No Adjustment Method Fully Resolves Confounding by Indication in Observational Studies*, 63 J. CLINICAL EPIDEMIOLOGY 64, 70 (2010).

[247] *See* Brookhart et al., *supra* note 190, at S115.

[248] *Id.*

To reduce or eliminate confounding bias in an observational study, those conducting it must strive to ascertain, accurately measure, and adjust for all potential confounding variables.[249] In many studies, however, it is by no means clear which variables are potential confounders. Medical care is often dependent on a complex web of variables relating to the healthcare system, clinicians, and patients themselves, and the factors at work in each case may not be obvious.[250]

Ideal randomized experiments, when they are feasible, prevent confounding because randomly assigning treatments to patients (possibly including a placebo) ensures there are no associations between the treatment variable and potential confounding variables.[251] In an observational study, investigators do not control treatment assignment because they review records that reflect treatments that have been previously administered.[252] Researchers must therefore attempt to obtain the values of confounding variables and adjust for them during analysis of the study data.[253]

One option is to restrict the values of the confounding variables—that is, to exclude subjects for whom the values of these variables are outside a chosen range—in order to ensure that the treatment groups are similar to each other.[254] For example, in an observational study of the comparative effectiveness of diuretics and beta blockers for the prevention of heart attacks, researchers could minimize confounding by indication by restricting the analysis to patients without any evidence of clinical cardiovascular disease.[255] This restriction is desirable because among patients with cardiovascular disease, the use of beta blockers rather than diuretics is an indicator of more severe disease and greater pre-treatment risk of heart attack.[256] Pre-existing disease in those taking beta blockers could skew study results and make it difficult to ascertain the relative effectiveness of the two drugs. The disadvantage of this restriction is that the results obtained do not generalize directly to the entire population of patients taking diuretics or beta blockers, because many in fact have cardiovascular disease.[257]

In some studies in which restriction is necessary, it is possible to generalize the results based on existing knowledge.[258] For example, although the causal link between smoking and lung cancer was established mainly through studies with male subjects, experts assumed the association existed in women too, because the lungs of men and women are anatomically similar.[259] In other cases, confounding bias can be controlled, without eliminating any groups of subjects from the analysis, by computing separate causal-effect estimates for each stratum (level) of a confounding variable.[260] Thus, if age is a confounder, analysts could calculate separate estimates for each ten-year interval of patient ages (birth to ten years old, ten to twenty years old, etc.). If the study cohort is representative of the target population, researchers can obtain an estimate of the average causal effect in the population by computing a

---

[249] *See* ROTHMAN ET AL., *supra* note 225, at 158.
[250] Brookhart et al., *supra* note 190, at S114.
[251] Bosco et al., *supra* note 246, at 64 (stating that "confounding is best controlled by a randomized design").
[252] *See, e.g.*, *id.*
[253] *Id.* at 64-65.
[254] *Id.* at 65.
[255] Psaty & Siscovick, *supra* note 244, at 898.
[256] *Id.*
[257] *Id.*
[258] ROTHMAN ET AL., *supra* note 225, at 146-47 (discussing generalizability).
[259] *Id.* at 147.
[260] *Id.* at 266.

weighted average of the stratum-specific effect estimates, where the weight for each stratum is the ratio of its size to the entire cohort's size.[261]

C. MEASUREMENT BIAS

Measurement biases arise from errors in measurement and data collection.[262] Observational study results may be compromised if the biomedical records that are analyzed contain such errors. Measurement errors occur for a variety of reasons. Measurement instruments might not be calibrated properly or might lack sufficient sensitivity to detect differences in relevant variables.[263] Storage time or conditions for biological samples might be different and might affect study results.[264] To the extent that researchers solicit and record patients' own accounts and memories, the subjects' ability to recall details may be influenced by the questioner's competence, patience, and apparent sympathy or by the degree to which the patient perceives the topic to be important and relevant to her life.[265] In addition, patients may have impaired memories or may lie in response to questions if they are embarrassed about the truth.[266] Accurate measurement may be further hindered by incomplete, erroneous, or miscoded EHR data that obfuscates true values.[267]

In causal-effect studies, errors in measurement of the treatment/exposure and the outcome are most problematic when they are associated (dependent) and when they are differential, that is, when the treatment affects the measurement error for the outcome or the outcome affects the measurement error for the treatment.[268] For example, differential measurement error could occur in a study of the effect of treatment *A* on dementia, if the use of *A* was determined only by interviewing study participants, because dementia affects subjects' ability to recall whether and how they were treated.[269] Mismeasurement of confounding variables also impedes adjustments intended to eliminate confounding bias.[270]

V. BIOMEDICAL DATABASES AND PERSONAL AGENDAS

Individuals with political, social, or economic agendas may exploit observational research outcomes to influence public opinion and legislative action. Manufacturers, for example, may fund studies regarding their products in the hope of generating new discoveries concerning the benefits of those products, thereby increasing their sales,[271] though subsequent marketing efforts would be constrained by advertising regulations enforced by federal agencies such as the FDA and the

---

[261] *Id.* at 271.

[262] Gael P. Hammer et al., *Avoiding Bias in Observational Studies*, 106 DEUTSCHES ÄRZTEBLATT INT'L 664, 665 (2009).

[263] *Id.*

[264] *Id.*

[265] *See id.*

[266] *See id.*

[267] Brookhart et al., *supra* note 190, at S116. See *supra* Part III for discussion of deficiencies in EHR documentation.

[268] ROTHMAN ET AL., *supra* note 225, at 137-38.

[269] Miguel A. Hernán & Stephen R. Cole, *Causal Diagrams and Measurement Bias*, 170 AM. J. EPIDEMIOLOGY 959, 960 (2009).

[270] ROTHMAN ET AL., *supra* note 225, at 144-45.

[271] George Davey Smith, *Big Business, Big Science?*, 37 INT'L J. EPIDEMIOLOGY 1, 1 (2008) (stating that "corporate influences can distort the knowledge base of epidemiology" when epidemiologists work "as the hired guns of industry").

Federal Trade Commission.[272] Data mining of biomedical databases may facilitate the discovery of statistical associations between use of certain products and the occurrence of desirable outcomes (e.g., lower disease risk). It is easy for marketers to imply that statistical associations are causal ones, even when the evidence for this is dubious because of confounding or other biases.[273]

Those who are politically motivated may attempt to use data that appears scientific in order to pursue legislative or regulatory goals. This has already happened. As noted in the Introduction, a now-debunked study finding that abortions caused lasting psychological harm to women[274] was used by advocates to influence state legislatures. Consequently, several states enacted legislation that required clinicians to warn individuals who seek abortions that they could suffer future mental health ailments.[275]

Needless to say, even highly trained researchers at academic institutions can produce low-quality work. The pressure to publish for tenure and promotion purposes or a desire for fame may tempt some faculty members to focus excessively on how publishable their work will be, at the expense of its scientific value. At least one study has explored whether publication pressures themselves generate biases among researchers and concluded that the answer is yes.[276] A review of 1316 papers found that United States researchers in competitive academic environments were biased against "negative" results that failed to support the hypothesis that was tested.[277] Researchers believe that "positive" results confirming hypotheses are more likely to be published and subsequently cited.[278] This assumption may influence not only their choices as to which projects to pursue and which results to write up, but also, potentially, their objectivity in conducting research.[279]

The dangers of misinterpretation, deliberate data distortion, flawed studies, and irresponsible use of research outcomes, might be acute if biomedical databases become publicly available. Individuals who have little research training but strong political voices and easy access to the media could deliberately misuse such databases.

As argued throughout this Article, amateurs are unlikely to produce reliable research outcomes.[280] Even leading scientists at times have difficulty interpreting research results and may disseminate confusing or misleading information to the public. A well-known example relates to the effect of postmenopausal hormone replacement therapy (HRT) on heart disease and breast cancer.[281] For several years, experts believed that the outcomes of randomized clinical trials conflicted with

---

[272] *See Prescription Drug Advertising: Questions and Answers*, U.S. FOOD & DRUG ADMIN., http://www.fda.gov/Drugs/ResourcesForYou/Consumers/PrescriptionDrugAdvertising/UCM076768.htm#control_advertisements (last updated Sept. 13, 2012).

[273] *See* Richard S. Rivlin, *Can Garlic Reduce Risk of Cancer?* 89 AM. J. CLINICAL NUTRITION 17, 17 (2009) (asserting that "the very strict criteria required to make a health claim [about the benefits of garlic consumption] may not be met by the limited number of studies conducted to date that are currently available").

[274] *See supra* notes 1-6 and accompanying text.

[275] *Id.*

[276] Daniele Fannelli, *Do Pressures to Publish Increase Scientists' Bias? An Empirical Support from US States* [sic] *Data*, 5 PLOS ONE 1, 4 (2010).

[277] *Id.*

[278] *Id.* at 1.

[279] *Id.*

[280] *See, e.g.*, *supra* Part IV.

[281] *See* Jacques E. Rossouw et al., *Postmenopausal Hormone Therapy and Risk of Cardiovascular Disease by Age and Years Since Menopause*, 297 J. AM. MED. ASS'N 1465, 1465 (2007).

results of observational studies concerning HRT.[282] It was only with reanalysis of the data that the two were reconciled. Experts realized that the effect of HRT changed over time, which explained the apparent discrepancies.[283] In the end, scientists concluded that HRT increased the risk of both heart disease and breast cancer, but the cardiovascular risk was slightly higher for older women and the cancer risk was higher for women closer to menopause.[284]

Yet, regardless of the quality of the work product, anyone can set up a website or blog and post documents that appear to be serious scientific studies. Moreover, these can quickly enjoy worldwide dissemination.

In addition, web-savvy people may employ strategies to increase the visibility of their messages. Even Google searches are vulnerable to manipulation through practices known as "search engine optimization."[285] Google's algorithm includes a popularity metric called PageRank, which considers "inbound links to a website as popularity votes."[286] Website owners can increase traffic to their sites by paying other websites to link to them.[287] They can also try to trick the search engine by showing Google crawlers[288] information that is different from that available to users.[289] Yet, many viewers will likely assume that studies that appear first or often in search-engine results are necessarily more credible and valuable than others.

VI. SOLUTIONS

We do not mean to suggest that biomedical databases are too flawed to be of value or that observational studies based on such databases are a lost cause. Quite to the contrary, we are optimistic about these emerging resources and capabilities.[290] The success of research initiatives, however, will depend on advances in technology as well as on human efforts to validate biomedical data and master the complexities of conducting sound observational studies. Anyone relying on record-based study outcomes, including government officials, attorneys, and the public at large, must know what questions to ask and not be naïve about the studies' validity. In this section we propose technological enhancements, study-design and validation improvements, and educational initiatives to combat potential database and research inadequacies or abuses.

A. TECHNOLOGY IMPROVEMENTS

As EHR system technology matures, its capacity to capture accurate and comprehensive data sets should continue to improve. Health information technology

---

[282] Vandenbroucke, *supra* note 79, at 1233.

[283] *Id.* at 1233-34.

[284] *Id.* at 1235. In addition, the risk of heart disease was found to increase in the first years of HRT use but then waned. *Id.* at 1234.

[285] Viva R. Moffat, *Regulating Search*, 22 HARV. J.L. & TECH. 475, 481 (2009).

[286] Eric Goldman, *Search Engine Bias and the Demise of Search Engine Utopianism*, 8 YALE J.L. & TECH. 188, 193 (2006).

[287] *Facts about Google and Competition*, GOOGLE, http://www.google.com/competition/howgooglesearchworks.html (last visited Oct. 22, 2013).

[288] Google continually acquires new information by sending "automated 'spiders' and 'crawlers' onto the Web." Moffat, *supra* note 285, at 481.

[289] Goldman, *supra* note 286, at 193.

[290] *See* Hoffman & Podgurski, *supra* note 8, at 97-102 (discussing the benefits of EHR-based research).

experts are increasingly likely to recognize that these systems serve not only as clinical and billing tools, but also as data sources for secondary use purposes.

Incompatibility of different EHR systems poses significant challenges for researchers.[291] Reconciling the format and meaning of data that come from different systems constitutes very resource-intensive and burdensome work for analysts.[292] Even the records of individual patients who see doctors at more than one facility can become fragmented, and it may be nearly impossible to put the pieces together into a cohesive whole if the separate EHR systems are not interoperable.[293] The problem of data fragmentation could be cured through semantic interoperability, which would enable "information systems to exchange information on the basis of shared, pre-established and negotiated meanings of terms and expressions."[294] Health information exchange capabilities are essential to the research endeavor.[295]

Many barriers hinder the achievement of semantic interoperability.[296] Among them are: (1) the extremely large number of stakeholders in the United States; (2) the lack of coordination, standardization, trained personnel, and appropriate technology; and (3) the government pressure to transition from paper records to EHR systems as quickly as possible.[297] In addition, semantic interoperability likely does not appeal to EHR vendors. If vendors standardize their products, they will make it easier for customers who have one vendor's product to switch to a competitor's EHR system. Customers could learn to use new systems more easily and transfer existing records to new EHR systems with less difficulty.[298] As of 2012, a workgroup consisting of ten states and twenty-six vendors is collaborating to develop standard specifications to promote interoperability and the exchange of health information, but progress has been slow.[299]

All stakeholders must continue to work together to develop mechanisms to standardize representations of EHR information in order to eliminate data

---

[291] Dipak Kalra et al., *ARGOS Policy Brief on Semantic Interoperability,* 170 STUD. IN HEALTH TECH. & INFORMATICS 1, 5 (2011); *See also supra* Parts III.B.-III.C.

[292] Carole Goble & Robert Stevens, *State of the Nation in Data Integration for Bioinformatics,* 41 J. BIOMED. INFORMATICS 687, 687 (2008) (stating that "the integration of resources—a prerequisite for most bioinformatics analysis—is a perennial and costly challenge").

[293] *See supra* notes 191-192 and accompanying text.

[294] Kim H. Veltman, *Syntactic and Semantic Interoperability: New Approaches to Knowledge and the Semantic Web*, 7 NEW REV. INFO. NETWORKING 159, 167 (2001). *See also* Robert H. Dolin & Liora Alschuler, *Approaching Semantic Interoperability in Health Level Seven*, 18 J. AM. MED. INFORMATICS ASS'N 99, 99-100 (2010) (providing alternative definitions of "semantic interoperability").

[295] *See* Botsis et al., *supra* note 40, at 4 (stating that incompleteness "could be mitigated using health information exchange (HIE) methods"); Herwehe et al., *supra* note 141, at 448 (explaining that "[e]lectronic health information exchange (HIE) offers a provider-acceptable means of utilizing information from multiple sources"); Jensen et al., *supra* note 200, at 403 (stating that "EHR data need to be merged across regional barriers in order to provide the strongest basis for research").

[296] Werner Ceusters & Barry Smith, *Semantic Interoperability in Healthcare State of the Art in the US*, ST. U.N.Y. BUFFALO 1, 4 (2010), http://ontology.buffalo.edu/medo/Semantic_Interoperability.pdf.

[297] *Id.*

[298] M. Alexander Otto, *Despite Small Steps, EHR Interoperability Remains Elusive*, INTERNAL MED. NEWS (Jan. 31, 2011), http://www.internalmedicinenews.com/news/more-top-news/single-view/despite-small-steps-ehr-interoperability-remains-elusive/71b93edeb0.html.

[299] Mike Miliard, *EHR/HIE Interoperability Workgroup Agrees on Connectivity Specs*, HEALTHCARE IT NEWS (Nov. 9, 2011), http://www.healthcareitnews.com/news/ehrhie-interoperability-workgroup-agrees-connectivity-specs; *Official PR: 10 States Now Unified to Standardize Health Data Interoperability*, EHR/HIE INTEROPERABILITY WORKGROUP (Feb. 20, 2012), http://interopwg.org/news/OFFICIAL-PR-10-States-Now-Unified-to-Standardize-Health-Data-Interoperability.html.

ambiguities and facilitate integration of records from EHR systems produced by different vendors.[300] The Federal Government could incentivize the development of semantic interoperability by incorporating increasingly stringent interoperability requirements into the meaningful use regulations.[301] These regulations, which are being implemented in three stages over several years, detail the requirements that healthcare providers must meet in order to receive federal funding to assist them in implementing EHR systems.[302]

Even with interoperability, many EHRs would continue to be incomplete and contain inaccuracies.[303] These deficiencies could be mitigated in part through increased use of electronic means for collecting patient data, such as remote patient monitoring.[304] A variety of devices, including glucometers, implantable cardioverter-defibrillators, and blood pressure monitors, can register clinical measurements at home and report them to patients' healthcare providers.[305]

Several other technological improvements would be useful as well. Enhanced user interface design could make it easier for clinicians to navigate EHR systems and accurately record data.[306] Voice recognition software, when sufficiently advanced, could also reduce the risk of input errors and allow users to operate systems more quickly and to add additional detail to their documentation.[307] As discussed above, improved and widely available natural-language processing tools would also enable analysts to extract more comprehensive data from EHRs.[308]

In some circumstances, clinical alerts could contribute to the accuracy of data collection.[309] If expected measurements can be determined in advance, EHRs could generate alerts when clinicians enter values that deviate significantly from the anticipated figures.[310] In one study focusing on height and weight measures, researchers had an alert pop up when clinicians entered figures with a ten percent or

---

[300] *See* Sharona Hoffman & Andy Podgurski, *Finding A Cure: The Case for Regulation and Oversight of Electronic Health Record Systems*, 22 HARV. J.L. & TECH. 103, 152-53 (2008) (recommending the development of a common exchange representation).

[301] *See* 45 C.F.R. §§ 170.205, 170.207 (2012) (providing current health information exchange standards).

[302] *See EHR Incentive Program*, CTRS. FOR MEDICARE & MEDICAID SERVS., https://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/index.html?redirect=/EHRIncentivePrograms/30_Meaningful_Use.asp (last modified June 26, 2013).

[303] *See supra* Part III.

[304] Kevin D. Blanchet, *Remote Patient Monitoring*, 14 TELEMED. & E-HEALTH 127, 128-30 (2008); *Technologies for Remote Patient Monitoring in Older Adults,* CTR. FOR TECH. & AGING 1, 4 (2009), http://www.techandaging.org/RPMpositionpaperDraft.pdf.

[305] *Technologies for Remote Patient Monitoring*, *supra* note 304, at 4.

[306] *See* Michael E. Wiklund, *Making Medical Device Interfaces More User-Friendly*, *in* DESIGNING USABILITY INTO MEDICAL PRODUCTS 151–60 (Michael E. Wiklund & Stephen B. Wilcox eds., 2005) (discussing user-interface problems and techniques for enhancing the user-friendliness of medical device interfaces); Adrian Williams, *Design for Better Data: How Software and Users Interact Onscreen Matters to Data Quality*, 77 J. AM. HEALTH INFO. MGMT. INST. 56, 56 (2006) (stating that "[p]oorly designed software that confronts the user with confusing screens, excessive data entry fields, or unclear navigational tools . . . threatens the quality of the data that users enter").

[307] *See supra* note 306; Ken Terry, *Voice Recognition Moves Up a Notch: When the Computer Can Type While You Talk, You Save Money and Time*, 81 MED. ECON. TCP11 (2004).

[308] *See supra* note 216 and accompanying text.

[309] Krystl Haerian et al., *Use of Clinical Alerting to Improve the Collection of Clinical Research Data*, 2009 AMIA ANN. SYMP. PROC. 218, 219-20.

[310] *Id.* at 219.

larger variance from those previously recorded.[311] After the alerts were implemented, EHR error rates were reduced from 2.4% to .9%.[312]

As healthcare facilities become increasingly interested in research endeavors that utilize electronic databases, they may demand that EHR vendors build systems that encourage or require clinicians to capture research-relevant data. Similarly, healthcare systems may train their employees to be diligent about collecting data that is needed for secondary use. Healthcare providers have much to gain from observational research using biomedical databases. New discoveries can lead to better patient care, cost savings, and financial profits from the adoption of more effective treatments.[313] Those who wish to enjoy these benefits should be motivated to do their utmost to produce data that is accurate, complete, and easily usable for research purposes.

B. HUMAN HANDS

Technology alone cannot remedy the weaknesses of biomedical databases and observational study outcomes. Continuous human vigilance and human intervention will be critical to the integrity of the research endeavor. Two safeguards of particular importance are data quality assessment and careful application of modern causal inference techniques.

1. Data Quality Assessment

Analysts will need to take steps to assess the validity of biomedical data. This could be accomplished through audits that scrutinize a sample of EHRs.[314] Researchers would select a randomly chosen sample of records, review them, and then interview the patients at issue (and possibly their caregivers) to determine the records' accuracy level. This process would enable analysts to estimate error rates for particular databases or federated systems in order to characterize uncertainty about research results.[315] If data analysts are receiving information from multiple sources and do not have access to patients, their ability to assess data quality will be more limited.[316] However, they may be able to compare data sets from different sources, identify anomalous values, and ask the original data holders to verify their accuracy.[317]

---

[311] *Id.*

[312] *Id.* at 220.

[313] *See supra* Part II.B.1.

[314] U.S. GOV'T ACCOUNTABILITY OFFICE, GAO-06-54, HOSPITAL QUALITY DATA: CMS NEEDS MORE RIGOROUS METHODS TO ENSURE RELIABILITY OF PUBLICLY RELEASED DATA 5 (2006) (discussing the Centers for Medicare and Medicaid Services' process "for ensuring the accuracy of the quality data submitted by hospitals for the APU program"); Leon G. Fine et al., *How to Evaluate and Improve the Quality and Credibility of an Outcomes Database: Validation and Feedback Study on the UK Cardiac Surgery Experience*, 326 BRIT. MED. J. 25, 25-26 (2003).

[315] *See* Douglas Curran-Everett & Dale J. Benos, *Guidelines for Reporting Statistics in Journals Published by the American Physiological Society*, 18 PHYSIOLOGY GENOMICS 249, 250 (2004) (discussing the importance of reporting uncertainty).

[316] *Id.*

[317] Michael G. Kahn et al., *A Pragmatic Framework for Single-Site and Multisite Data Quality Assessment in Electronic Health Record-Based Clinical Research*, 50 MED. CARE S21, S22 (2012).

Experts have also developed methods to validate or assess the quality of annotation in genome records.[318] For example, a guidance document for genetic epidemiologists proposes the creation of an index that assigns grades of "A," B," and "C" to three indicators: amount of evidence, extent of replication, and protection from bias.[319] The index would also generate an overall assessment of epidemiological credibility as being "strong," "moderate," or "weak."[320]

Investigators should employ reputable "best practice" guidelines for conducting observational research using biomedical databases. One example is a 2011 FDA draft guidance document entitled "Best Practices for Conducting and Reporting Pharmacoepidemiologic Safety Studies Using Electronic Healthcare Data Sets."[321] Likewise, the *International Journal of Epidemiology*'s *Assessment of Cumulative Evidence on Genetic Associations: Interim Guidelines* offers guidance to genetic researchers.[322]

Clinicians and researchers could also employ the technique of crowdsourcing to the project of verifying EHR data. Crowdsourcing occurs when an entity enables "a population (crowd) to solve a problem,"[323] using an "open call" to induce an undefined, large network of people rather than employees to perform work.[324] Patients may review their own medical records and often have access to some of their health information through an interactive component of the EHR called a personal health record.[325] Patients who identify EHR errors should be able to report them easily to clinicians. For their part, clinicians should be obligated to read all error alerts, assess them, and make corrections if the patient has in fact found a mistake.[326] Federal regulations already allow patients to review their records and

---

[318] Klimke et al., *supra* note 184, at 168 (describing methods to assess annotation quality, including combining different pieces of evidence "in order to assign confidence levels to a particular annotation").

[319] John P. A. Ioannidis et al., *Assessment of Cumulative Evidence on Genetic Associations: Interim Guidelines*, 37 INT'L J. EPIDEMIOLOGY 120, 122 (2008).

[320] *Id.* at 126.

[321] FOOD & DRUG ADMIN., BEST PRACTICES FOR CONDUCTING AND REPORTING PHARMACOEPIDEMIOLOGIC SAFETY STUDIES USING ELECTRONIC HEALTHCARE DATA SETS (2013), *available at* http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM243537.pdf. *See also* Simon Sanderson et al., *Tools for Assessing Quality and Susceptibility to Bias in Observational Studies in Epidemiology: A Systematic Review and Annotated Bibliography*, 36 INT'L J. EPIDEMIOLOGY 666, 666-74 (2007) (providing guidance concerning observational studies but not specifically about EHR-based research).

[322] *See* Paolo Boffetta et al., *Recommendations and Proposed Guidelines for Assessing the Cumulative Evidence on Joint Effects of Genes and Environments on Cancer Occurrence in Humans*, 41 INT'L J. EPIDEMIOLOGY 686, 686-704 (2012); *see generally* Ioannidis et al., *supra* note 319.

[323] Michael Christopher Gibbons, *Use of Health Information Technology Among Racial and Ethnic Underserved Communities*, 8 PERSP. IN HEALTH INFO. MGMT. 1, 6 (2011), *available at* http://perspectives.ahima.org/PDF/Winter_2011/Use_of_HIT_Among_Racial_and_Ethnic_Underserved_Communities/Use_of_HIT_Among_Racial_and_Ethnic_Underserved_Communities_final.pdf.

[324] Daren C. Brabham, *Crowdsourcing as a Model for Problem Solving: An Introduction and Cases,* 14 CONVERGENCE 75, 76 (2008); Jeff Howe, *Crowdsourcing: A Definition,* CROWDSOURCING (June 2, 2006), http://crowdsourcing.typepad.com/cs/2006/06/crowdsourcing_a.html.

[325] Paul C. Tang et al., *Personal Health Records: Definitions, Benefits, and Strategies for Overcoming Barriers to Adoption*, 13 J. AM. MED. INFORMATICS ASS'N 121, 122 (2006) (citing MARKLE FOUND., CONNECTING FOR HEALTH: THE PERSONAL HEALTH WORKING GROUP FINAL REPORT (2003) (defining a personal health record as "an electronic application through which individuals can access, manage and share their health information, and that of others for whom they are authorized, in a private, secure, and confidential environment")).

[326] In some cases, patients will be wrong about the existence of an error, and thus clinicians must scrutinize error reports before changing EHR entries.

request amendment in case of error.[327] Thus, this approach is novel only in that it would affirmatively encourage patients to scrutinize their medical files. Patients should submit error reports through their EHR systems' secure messaging feature[328] or a dedicated website so that they may create a record of their requests and document their assertions that errors exist.[329]

2. Causal Inference Techniques

In the last two decades, researchers have made substantial progress in the development of a methodology for making causal inferences from observational data. Causal diagrams (also called causal graphs, directed acyclic graphs, or DAGs) are an important component of this methodology and have become a popular tool in the fields of statistics, biostatistics, epidemiology, and computer science.[330] The use of sound causal inference techniques is essential for analysis of large amounts of complex data from biomedical databases. Lawyers and policy officials would be well advised to keep abreast of causal research inference developments and be able to understand causal diagrams.

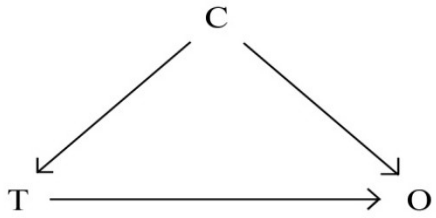

Figure 1: Causal diagram showing causal arrows between treatment variable *T*, outcome variable *O*, and confounder *C*.

A causal diagram consists of points or vertices, each representing a variable.[331] There is an arrow or "edge" $A \rightarrow B$ connecting a variable A to a variable B if A is known or assumed to cause B.[332] The variables typically include a treatment or exposure variable (e.g., indicating which medication a patient received), an outcome variable (e.g., representing a patient's disease status), and a number of covariates representing clinical, demographic, and possibly genetic factors.[333] For each study

[327] *See* HIPAA Privacy Rule, 45 C.F.R. § 164.526 (2012) ("An individual has the right to have a covered entity amend protected health information or a record about the individual in a designated record set.").

[328] *See* Hoffman & Podgurski, *supra* note 157, at 1530, 1549 (describing secure messaging).

[329] *See* Elizabeth Pennisi, *Proposal to 'Wikify' GenBank Meets Stiff Resistance*, 319 SCI. 1598, 1598 (2008) (describing a controversy regarding the process for correcting errors in GenBank, "the U.S. public archive of sequence data").

[330] JUDEA PEARL, CAUSALITY 65-68 (2d ed. 2009); Tyler J. VanderWeele & Nancy C. Staudt, *Causal Diagrams for Empirical Legal Research: Methodology for Identifying Causation, Avoiding Bias, and Interpreting Results*, 10 L. PROBABILITY & RISK 329, 329-30 (2011).

[331] VanderWeele & Staudt, *supra* note 330, at 333; Jeffrey Swanson & Jennifer Ibrahim, *Picturing Public Health Law Research: Using Causal Diagrams to Model and Test Theory*, PUB. HEALTH L. RES. 1, 6 (2011), http://publichealthlawresearch.org/sites/default/files/SwansonIbrahim-CausalDiagrams-March2012.pdf.

[332] *See supra* note 307.

[333] Swanson & Ibrahim, *supra* note 331, at 6.

subject, researchers obtain values for all variables, if possible, from the subject's medical record or other sources.[334] A very simple causal diagram, reflecting the relationships between a treatment variable *T*, and outcome variable *O*, and a confounder *C*, is shown in Figure 1. The causal diagram represents investigators' assumptions about causal relationships between variables in a particular study or about the absence of such relationships.[335] It is intended to be "a map of . . . cause and effect relations," allowing researchers to understand relationships among relevant variables so that they can construct valid statistical models, avoid confounding, and correctly interpret study results.[336] In the process of creating causal diagrams, analysts attempt to specify the causal relationships and dependencies among all relevant factors involved in a particular problem, leaving the ultimate question of the relationship between the exposure of interest and the outcome to be discovered through research.[337]

Causal diagrams can thus assist analysts in determining the measures to be used in a study and in understanding potential sources of bias.[338] They provide a clear, visual means to depict assumptions about the relationships of variables and highlight complexities that researchers might overlook in the absence of the graphs.[339] Moreover, researchers have derived precise conditions, in terms of a correct causal diagram, for selecting a set of confounders to adjust for in order to "identify" (statistically characterize) a given causal effect.[340] Judea Pearl's "Back-Door Criterion" constitutes a well-known example.[341] It characterizes the circumstances in which a set of variables is sufficient, if adjusted for, to eliminate confounding by "blocking," in a technical sense, all "back-door" paths in a causal diagram between the treatment/exposure variable and the outcome variable.[342] Note that an unblocked (open) back-door path permits a confounder to induce a spurious association between the treatment/exposure and the outcome.[343] Such conditions can be checked algorithmically, given a correctly specified causal diagram for a study.[344] In Figure 1, the open back-door path $T \leftarrow C \rightarrow O$ represents the ability of the variable *C* to confound estimation of the causal effect of the treatment variable *T* upon the outcome variable *O*.[345] This back-door path can be blocked by conditioning on (adjusting for) the confounder *C*, e.g., by restricting the analysis to a single level of *C*.[346]

Returning to our previous example, suppose *T* indicates the drug a patient is given to treat high blood pressure (1: diuretic, 2: beta blocker); *C* is an indicator for evidence of clinical cardiovascular disease (1: yes, 0: no); and the outcome variable

---

[334] *Id.*
[335] VanderWeele & Staudt, *supra* note 330, at 332.
[336] *Id.* at 329.
[337] Brookhart et al., *supra* note 190, at S116.
[338] *Id.*; Swanson & Ibrahim, *supra* note 331, at 1.
[339] VanderWeele & Staudt, *supra* note 330, at 335.
[340] PEARL, *supra* note 330, at 65-68; Ilya Shpitser et al., *On the Validity of Covariate Adjustment for Estimating Causal Effects*, 26TH ANN. CONF. ON UNCERTAINTY IN ARTIFICIAL INTELL. (UAI-10) 527, 527-26 (2010); Tyler J. VanderWeele & Ilya Shpitser, *A New Criterion for Confounder Selection*, 67 BIOMETRICS 1406, 1406 (2011).
[341] PEARL, *supra* note 330, at 79-81 (explaining the "back-door criterion").
[342] *Id.*
[343] VanderWeele & Staudt, *supra* note 330, at 335.
[344] PEARL, *supra* note 330, at 72-76 (discussing how the effect of interventions is computed).
[345] *Id.*
[346] *Id.*

O indicates whether the patient had a heart attack after treatment (1: yes, 0: no).[347] Restricting the study to patients with $C = 0$, for whom there is no evidence of clinical cardiovascular disease, prevents distortion of the causal effect of *T* (treatment) on *O* (outcome) by a spurious, noncausal association between *T* and *O*. Such an association could occur because use of beta blockers, rather than diuretics, is an indicator of more severe disease and greater pre-treatment risk of heart attack.[348] Without such an adjustment, beta blockers could appear less effective than they really are because more of the patients taking them would likely suffer heart attacks caused by their existing cardiovascular disease.

Confounding, selection bias, and measurement error may also affect observational studies of possible causal relationships between genes and diseases.[349] To address these challenges, researchers have advocated for the use of causal inference methodology, including causal diagrams, in genetic studies.[350] For example, systems biologists employ graphical models to depict what is known about the molecular networks by which genes regulate (or misregulate) the functioning of other genes.[351] The successful use of causal inference techniques in medical genetics is synergistically related to efforts to elucidate the complex biological networks by which genes influence disease.[352]

If properly used, causal diagrams, and modern causal inference methodology more generally, can promote more accurate research results, which can lead to sound public health policies and regulations. Experts have also noted that causal graphs can be used to portray the relationship between the law and human behavior in order to ascertain that legal interventions are having the desired effect.[353]

---

[347] *See supra* notes 252-254 and accompanying text.

[348] *Id*.

[349] John Attia et al., *How to Use an Article About Genetic Association B: Are the Results of the Study Valid?* 301 J. AM. MED. ASS'N 191, 191 (2009); Sara Geneletti et al., *Assessing Causal Relationships in Genomics: From Bradford-Hill Criteria to Complex Gene-Environment Interactions and Directed Acyclic Graphs*, 8 EMERGING THEMES IN EPIDEMIOLOGY 1, 5 (2011). For example, researchers have found that standard statistical approaches for estimating/testing direct genetic effects may yield biased estimates when there is a non-genetic link between the target phenotype and another phenotype. Stijn Vansteelandt et al., *On the Adjustment for Covariates in Genetic Association Analysis: A Novel, Simple Principle to Infer Direct Causal Effects*, 33 GENETIC EPIDEMIOLOGY 394, 395 (2009).

[350] Nuala A. Sheehan et al., *Mendelian Randomisation: A Tool for Assessing Causality in Observational Epidemiology*, *in* GENETIC EPIDEMIOLOGY 153, 153-66 (M. Dawn Teare ed., 2011); Alexander V. Alekseyenko et al., *Causal Graph-Based Analysis of Genome-Wide Association Data in Rheumatoid Arthritis*, 6 BIOLOGY DIRECT 25, 26 (2011); Steven S. Coughlin, *Quantitative Models for Causal Analysis in the Era of Genome Wide Association Studies*, 4 OPEN HEALTH SERV. POL'Y J. 118, 120 (2011). For example, Geneletti et al. present a framework of assessing causal relationships in clinical genomics that integrates Austin Bradford Hill's influential guidelines for assessing causality, on one hand, with the use of graphical models (depicting both causal and non-causal associations), on the other hand. *See* Geneletti et al., *supra* note 349, at 5-6.

[351] Celine Lefebvre et al., *Reverse-Engineering Human Regulatory Networks*, 4 WILEY INTERDISCIP. REV. SYST. BIOLOGY MED. 311, 311 (2012). Such regulation occurs indirectly, via the products of gene expression, namely RNA and proteins. *Id.* at 312.

[352] Albert-László Barabasi et al., *Network Medicine: A Network-Based Approach to Human Disease*, 12 NATURE REV. GENETICS 56, 56 (2011).

[353] Swanson & Ibrahim, *supra* note 331, at 1; Evan Anderson et al., *Measuring Statutory Law and Regulations for Empirical Research*, PUB. HEALTH L. RES. PROGRAM 1, 12 (2012), http://publichealthlawresearch.org/sites/default/files/MeasuringLawRegulationsforEmpiricalResearch-Monograph-AndersonTremper-March2012.pdf (stating that "[b]y forcing researchers to identify plausible links between the law and health outcomes, causal diagrams help flush out the legal inputs relevant to the question of interest").

Nevertheless, causal diagrams support valid causal inferences only if they include all relevant variables and reflect the true causal relationships among them.[354] Analysts must make subjective decisions in selecting which variables and arrows to include, and their own erroneous assumptions, biases, or carelessness can contaminate the final product.[355] Thus, the causal diagram in Figure 1 would be incorrect if there were in truth another variable S and a path $T \rightarrow S \leftarrow O$ (called a "collider") that was unknown to the researchers. As discussed in a prior example, the values of this variable could indicate whether individual patients are lost to follow-up and cannot be included in the study. Losing such patients could lead to selection bias that exaggerates or diminishes the apparent effectiveness of one of the treatments under study.[356]

The use of genomic data with many thousands of variables, though desirable, also complicates causal inference. As genetic discoveries emerge, researchers routinely need to determine whether to include genetic variables as factors that influence disease vulnerability or treatment success.[357] Geneticists have identified over 6000 single-gene disorders[358] and believe that essentially every human disease has a genetic component.[359] Even vulnerability to infection can be affected by individual genotype, which can render some people resistant to infectious diseases including AIDS.[360]

In the future, expert panels may be able to develop widely accepted causal diagrams for disease-related causal influences about which there is general agreement. Individual researchers would be free to customize them by adding or removing links to reflect their own causal hypotheses, but such diagrams could provide significant guidance to analysts.

Public health officials and legal practitioners with expertise in causal inference methodology will be better equipped to understand study outcomes, ask appropriate

---

[354] In addition, statistical analysis of causal effects based on a causal diagram is valid only if certain strong assumptions hold that relate the diagram to the underlying probability distribution of the variables. A. Philip Dawid, *Beware of the DAG!*, 6 J. MACHINE LEARNING RES. 59, 68 (2008), *available at* http://jmlr.csail.mit.edu/proceedings/papers/v6/dawid10a/dawid10a.pdf.

[355] *See* Brookhart et al., *supra* note 190, at S116 (explaining that "in many studies of medical interventions, the available subject-matter knowledge is inadequate to specify with any degree of certainty the causal connections between variables").

[356] *See supra* notes 231, 233-38 and accompanying text. Assume the "collider" variable S indicates whether a study subject is lost to follow-up (1: yes, 0: no) and is influenced by the disease outcome O (1: cured, 0: not cured) under investigation and by treatment T (1: drug A, 0: drug B). If S was always zero (indicating "not lost to follow up") for study participants, then the path $T \rightarrow S \leftarrow O$ would be open and possibly create a spurious association between T and O resulting in selection bias. For example, suppose that a number of study subjects stopped going to the doctor because of unpleasant side effects of drug A (assume drug B has no side effects) or because they experienced no improvement in their disease symptoms and became discouraged. Among subjects who received drug A, those who completed the treatment regime might have experienced an atypically strong therapeutic effect from A, since they were willing to tolerate its side effects. Consequently, treatment A might appear more effective overall, when compared to treatment B, than it really is.

[357] *Genetic Disease Information – Pronto!*, HUMAN GENOME PROJECT INFO., http://web.archive.org/web/20130430183952/http://www.ornl.gov/sci/techresources/Human_Genome/medicine/assist.shtml (last modified Mar. 07, 2012) (accessed by searching for Human Genome Project Information in the Internet Archive); *Understanding Human Genetic Variation*, NAT'L INSTS. OF HEALTH OFFICE OF SCI. EDUC., http://science.education.nih.gov/supplements/nih1/genetic/guide/genetic_variation1.htm (last visited Oct. 15, 2013).

[358] *Genetic Disease Information – Pronto!*, *supra* note 357 (indicating that many other diseases are multi-factorial, chromosomal, and mitochondrial).

[359] *Understanding Human Genetic Variation*, *supra* note 357.

[360] *Id.*

questions, and differentiate between good and bad science. They thus will be more likely to respond to emerging discoveries appropriately.

Causal diagrams might even be useful as litigation tools. In complex tort cases involving questions of causation, lawyers could ask experts to develop plausible causal diagrams that depict the parties' understanding of how plaintiffs' injuries occurred. Experts on opposing sides are unlikely to agree about the details of causal diagrams and could use causal inference methodology to attack their opponents' assumptions and assertions. Fact-finders would then determine which diagram is most accurate before deciding which party will prevail.

C. EDUCATION AND PREVENTION OF RESEARCH MISUSE

Biomedical databases and federated systems will provide an abundance of new data[361] that can be harnessed to bring great benefits to society but could also be abused.[362] As argued above, only highly skilled researchers who understand data quality problems and causal inference methodology are likely to produce reliable study outcomes.[363] The public and even policy experts may be too credulous with respect to information that is drawn from large databases and appears to be scientific.

Numerous studies have highlighted how difficult it is for many people to understand statistical information.[364] One study surveyed German and U.S. citizens and found that approximately 20% of citizens could not determine whether 1%, 5%, or 10% represented the "biggest risk of getting a disease."[365] Other researchers have gone as far as to declare that modern societies suffer from "collective statistical illiteracy."[366]

Consequently, the development of large-scale biomedical databases should be accompanied by educational programs about their strengths and limitations.[367] Such programs should include warnings that not all data will be error-free and not all research outcomes can be trusted. In his well-known book, *How to Lie with Statistics*, Darrell Huff writes that the language of statistics, is often "employed to sensationalize, inflate, confuse, and oversimplify."[368] He goes on to assert that "crooks already know these tricks" and therefore "honest men must learn them in self-defense."[369]

---

[361] *See supra* Parts II.A, II.B.1.

[362] *See supra* Part V.

[363] *See supra* Parts IV, VI.A.1.

[364] *See, e.g.*, Timur Kuran & Cass R. Sunstein, *Availability Cascades and Risk Regulation*, 51 STAN. L. REV. 683, 685 (1999) (discussing "the *availability heuristic*, a pervasive mental shortcut whereby the perceived likelihood of any given event is tied to the ease with which its occurrence can be brought to mind"); Amos Tversky & Daniel Kahneman, *Availability: A Heuristic for Judging Frequency and Probability*, 5 COGNITIVE PSYCHOLOGY 207, 207 (1973) (proposing "that when faced with the difficult task of judging probability or frequency, people employ a limited number of heuristics which reduce these judgments to simpler ones").

[365] Mirta Galesic & Rocio Garcia-Retamero, *Statistical Numeracy for Health: A Cross-Cultural Comparison with Probabilistic National Samples*, 170 ARCHIVES INTERNAL MED. 462, 467 (2010). In addition, "almost 30% could not answer whether 1 in 10, 1 in 100 or 1 in 1000 represents the largest risk" and nearly 30% "could not state what percentage 20 of 100 is." *Id.*

[366] Gerd Gigerenzer et al., *Helping Doctors and Patients Make Sense of Health Statistics*, 8 PSYCHOLOGY SCI. PUB. INT. 53, 54 (2007).

[367] *See* Hoffman & Podgurski, *supra* note 8, at 140-41 (developing a more detailed proposal for educational programs regarding EHR databases).

[368] DARRELL HUFF, HOW TO LIE WITH STATISTICS 8 (1954).

[369] *Id.* at 9.

To this end, the National Institutes of Health could construct a publicly available website about sound and unsound research practices. Other educational messages can take the form of news stories disseminated via media such as television, radio, magazines, and news websites.

Lawyers, judges, and legislators who must consider biomedical data should receive training focusing on observational research and its legal implications. Law schools should offer courses on modern causal inference methodology, which is becoming essential for understanding and assessing scientific evidence. In addition, continuing education courses on data analysis should be available to legal and policy professionals. Without such preparation, they will be ill-equipped to engage with expert witnesses or policy advocates on the subject.

Peer-review journal editors will also need to be familiar with the complexities of biomedical database research. Reviewers for articles about observational studies that rely on electronic databases should have expertise not only with respect to the underlying subject matter, but also concerning the research methodology. In 2008, The *Journal of Clinical Epidemiology* published *The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement*, which included a detailed checklist of items that researchers should discuss in their reports, such as study design, bias, and generalizability.[370] Many top journals instruct authors to comply with the STROBE statement and presumably scrutinize submitted manuscripts to ascertain that the required items are covered.[371] In the future, the checklist may need to be expanded to account for matters that are specific to biomedical databases.

It will be far more difficult to achieve any quality control for the work product of amateur investigators who access publicly available databases. As a first step, however, access may be conditioned on individuals taking a short (e.g., one hour) online course about observational research followed by a simple quiz at the end to test learning. This obligation may deter individuals who are not serious about their research project and do not want to spend the time taking the session. Others may be convinced that research is a serious and demanding endeavor and decide to partner with experienced researchers or abandon frivolous research pursuits.

VII. CONCLUSION

The advent of large-scale biomedical databases brings with it the prospect of significant advances in the medical and social policy arenas coupled with the risks of confusion, uncertainty, or even deception. In the foreseeable future, computerized observational studies may well be a familiar tool for regulators, public health officials, and litigators. But members of the legal community who will rely on these studies to formulate public policy or to support litigation claims must understand the challenges of conducting valid database research and learn to distinguish good

---

[370] Erik von Elm et al., *The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies*, 61 J. CLINICAL EPIDEMIOLOGY 344, 346-47 (2008).

[371] *See Instructions for Authors*, BMJ OPEN, http://bmjopen.bmj.com/site/about/guidelines.xhtml (last visited Oct. 15, 2013); *JAMA Instructions for Authors*, JAMA NETWORK, http://jama.jamanetwork.com/public/instructionsForAuthors.aspx (last updated Sept. 10, 2013); *Types of Article and Manuscript Requirements*, LANCET, http://www.thelancet.com/lancet-neurology-information-for-authors/article-types-manuscript-requirements (last visited Oct. 15, 2013).

science from bad. This point is made clear in John Ioannidis' provocatively titled essay, "Why Most Published Research Findings Are False."[372]

Awareness of the potential for selection bias, measurement bias, and confounding bias, and of how these biases can be reduced or eliminated is critical to appropriate assessment of research outcomes. Similarly, analysts must understand the threats to study validity that are posed by data entry and processing errors, EHR gaps or fragmentation, and data standardization problems. These pitfalls are recognized by experts in epidemiology, pharmacoepidemiology, bioinformatics, and other fields who generally produce highly skilled and statistically sophisticated work. However, non-scientists must also recognize these challenges. When communicating study results, researchers must frankly explain all such threats and what was done to address them and must characterize as objectively as possible all elements of uncertainty about the results.

Mark Twain popularized the saying "[t]here are three kinds of lies: lies, damned lies, and statistics."[373] A danger exists that if the public is repeatedly duped by false or misleading research outcomes, it will come to scorn the entire medical research endeavor. It is only with appropriate insight and sophisticated approaches to conducting and interpreting observational studies that biomedical databases can fulfill their hoped-for promise and promote much societal good.

[372] Ioannidis, *supra* note 77, at 696.

[373] Mark Twain, *Chapters from my Autobiography—XX*, 186 N. AM. REV. 465, 471 (1907), *reprinted in* MARK TWAIN, CHAPTERS FROM MY AUTOBIOGRAPHY ch. 20, at 471 (Shelley Fisher Fishkin ed., Oxford Univ. Press 1996).