Balancing Privacy, Autonomy, and Scientific Needs In Electronic Health Records Research

Sharona Hoffman

Andy Podgurski

BALANCING PRIVACY, AUTONOMY, AND SCIENTIFIC NEEDS IN ELECTRONIC HEALTH RECORDS RESEARCH

*Sharona Hoffman**
*Andy Podgurski***

ABSTRACT

The ongoing transition from paper medical files to electronic health records will provide unprecedented amounts of data for biomedical research, with the potential to catalyze significant advances in medical knowledge. But this potential can be fully realized only if the data available to researchers is representative of the patient population as a whole. Thus, allowing individual patients to exclude their health information, in keeping with traditional notions of informed consent, may compromise the research enterprise and the medical benefits it produces.

This Article analyzes the tension between realizing societal benefits from medical research and granting individual preferences for privacy. It argues for a shift in the conceptual and regulatory frameworks that govern biomedical research. When studies involve electronic record review rather than human experimentation, the traditional, autonomy-dominated model should give way to one that emphasizes the common good. In record-based studies, the limited benefits of individual informed consent come at too high a cost—difficult administrative burdens, significant expenses, and a tendency to create selection biases that distort study outcomes. Other mechanisms can better protect data subjects' privacy and dignitary interests without compromising research opportunities.

In this Article, we formulate a novel, multi-faceted approach to achieve these ends. This approach recognizes that technical means for achieving identity concealment and information security are necessary but not sufficient to protect patients' medical privacy and to foster public trust while

* Professor of Law and 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 Jessica Berg, Jonathan Entin, Jessie Hill, Jacqueline Lipton, Maxwell Mehlman, Andrew Pollis, and Cassandra Robertson for their very helpful comments on drafts of this paper. The Article was presented at a faculty workshop at Case Western Reserve University School of Law at which the authors received additional helpful input. Professor Hoffman presented "Privacy and e-Health Innovation," from which this paper grew, at the Yale Law School Privacy and Innovation Symposium. We are also grateful for the skilled research assistance of Isaac Figueras.

facilitating research. Hence, we call for supplementing such means with (1) an oversight process that is tailored to record-based research and applies even to de-identified patient records, which are currently exempt from scrutiny, and (2) public notice and education about the nature and potential benefits of such research.

TABLE OF CONTENTS

I. INTRODUCTION

THE shift from hard-copy medical files to electronic health records (EHR) systems is transforming medical research in the United States.[1] One of the great promises of EHR technology is its dramatic potential to expand opportunities for biomedical research.[2] Digitizing medical files opens new frontiers for record-based research because electronic searches and computer analysis permit fast and inexpensive

1. David Blumenthal & Marilyn Tavenner, *The "Meaningful Use" Regulation for Electronic Health Records*, 363 NEW ENG. J. MED. 501, 501 (2010).

2. PRICEWATERHOUSECOOPERS, TRANSFORMING HEALTHCARE THROUGH SECONDARY USE OF HEALTH DATA 3 (2009), *available at* http://www.pwc.com/us/en/healthcare/publications/secondary-health-data.jhtml; Sharona Hoffman & Andy Podgurski, *Finding a Cure: The Case for Regulation and Oversight of Electronic Health Record Systems*, 22 HARV. J.L. & TECH. 103, 117–19 (2008); 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).

data synthesis.[3] EHR systems will enable the creation of sizeable record databases or networks of smaller databases that facilitate large-scale studies.[4] Researchers could pose queries to databases that include very large numbers of patients with diverse demographics who have been treated in different clinical settings over long periods of time.[5] This wealth of information could yield significant discoveries concerning the effectiveness of various treatments.[6] The secondary use of health data[7] could thus promote Comparative Effectiveness Research (CER)[8] and help fill significant gaps in medical knowledge.

But EHR-based research also raises new questions. Traditionally, a paramount principle of biomedical research ethics is human subject autonomy, which is realized through informed consent.[9] The regulatory requirement of informed consent dictates that researchers supply potential participants in biomedical research with information about the anticipated benefits and risks of each research project so that the potential participants can make educated decisions about whether to enroll.[10] Individuals must be free to decline to participate in studies if they so choose, and federal regulations provide detailed guidance concerning the contents of informed consent forms.[11]

This paradigm, however, is a poor fit for research based on EHRs. EHR systems' enormous potential to transform medical research has generated significant debate about the appropriate extent of regulatory pro-

---

3. *See* Abel N. Kho et al., *Electronic Medical Records for Genetic Research: Results of the eMERGE Consortium*, 3 SCI. TRANSLATIONAL MED. 79rel, 5 (2011), *available at* http://stm.sciencemag.org/content/3/79/79rel.abstract; Mark G. Weiner & Peter J. Embi, *Toward Reuse of Clinical Data for Research and Quality Improvement: The End of the Beginning?*, 151 ANNALS INTERNAL MED. 359, 359–60 (2009).

4. JEFFREY BROWN ET AL., AGENCY FOR HEALTHCARE RESEARCH AND QUALITY, DESIGN SPECIFICATIONS FOR NETWORK PROTOTYPES AND COOPERATIVE TO CONDUCT POPULATION-BASED STUDIES AND SAFETY SURVEILLANCE 2–3 (2009), *available at* http://www.effectivehealthcare.ahrq.gov/ehc/products/54/150/2009_0728DE-cIDE_DesignSpecNetCoopPopSafety.pdf; Til Sturmer et al., *Nonexperimental Comparative Effectiveness Research Using Linked Healthcare Databases*, 22 EPIDEMIOLOGY 298, 299 (2011).

5. Douglas Peddicord et al., *A Proposal to Protect Privacy of Health Information While Accelerating Comparative Effectiveness Research*, 29 HEALTH AFF. 2082, 2087 (2010).

6. *Id.*

7. Secondary use can be defined as "non-direct care use of . . . [data] including but not limited to analysis, research, quality/safety measurement, public health, payment, provider certification or accreditation, and marketing and other business including strictly commercial activities." Safran, *supra* note 2, at 4.

8. *See* discussion *infra* Part III.A. (reviewing CER).

9. COMM. ON HEALTH RESEARCH & THE PRIVACY OF HEALTH INFO.: THE HIPAA PRIVACY RULE, IOM, BEYOND THE HIPAA PRIVACY RULE: ENHANCING PRIVACY, IMPROVING HEALTH THROUGH RESEARCH 247 (Sharyl J. Nass et al. eds., 2009) [hereinafter IOM REPORT] (noting that "[t]he principle of autonomy currently dominates the ethical landscape" for clinical research).

10. NAT'L INSTS. OF HEALTH, THE BELMONT REPORT: ETHICAL PRINCIPLES AND GUIDELINES FOR THE PROTECTION OF HUMAN SUBJECTS OF RESEARCH (1979), *available at* http://ohsr.od.nih.gov/guidelines/belmont.html [hereinafter BELMONT REPORT].

11. 45 C.F.R. § 46.116 (2010).

tections for human subjects in studies that are solely record-based.[12] Are consent requirements barriers to conducting effective research? If human subjects will not undergo experimentation in the course of the study, and researchers will examine only their medical records, do the subjects really need the full panoply of regulatory protections? Does informed consent make sense in the context of EHR database research? Is the cost of extensive regulation too high?

At the same time, one might also ask whether EHR research actually requires more, rather than less, regulatory protection. Computerization of health information poses new risks of privacy breaches that did not exist when paper files could simply be locked away.[13] In addition, data subjects whose records are used in research without their consent might arguably suffer other dignitary harms. Harms to dignity include not only privacy violations, but also group stigmatization due to research findings, inability to control whether one's records will be used for objectionable purposes, and a lack of opportunity to share in profits acquired by data users.[14]

This Article employs an interdisciplinary approach, drawing upon the legal, bioethics, and informatics literature to develop a full understanding of the regulatory, ethical, and technical complexities of EHR data use. Part I of the Article provides background information. It describes existing initiatives to create EHR research databases and discusses the regulations that govern EHR-based research. Part II evaluates the benefits and potential harms of EHR research. Part III analyzes the concept of informed consent and argues that a requirement of informed consent is inappropriate for record-based research. For the sake of simplicity, we use the terms EHR and EHR systems to designate electronic health records and the systems in which they operate, though we mean for EHR to be synonymous with what others call the electronic medical record (EMR).[15] It is also important to emphasize that this Article focuses on

12. *See* IOM REPORT, *supra* note 9, at 33–35; Fred H. Cate, *Protecting Privacy in Health Research: The Limits of Individual Choice*, 98 CALIF. L. REV. 1765, 1798–1801 (2010); Henry T. Greely, *Breaking the Stalemate: A Prospective Regulatory Framework for Unforeseen Research Uses of Human Tissue Samples and Health Information*, 34 WAKE FOREST L. REV. 737, 752–58 (1999); Franklin G. Miller, *Research on Medical Records Without Informed Consent*, 36 J.L. MED. & ETHICS 560, 564 (2008).

13. Mark A. Rothstein, *Improve Privacy in Research by Eliminating Informed Consent? IOM Report Misses the Mark*, 37 J.L. MED. & ETHICS 507, 509–11 (2009).

14. Daniel Kim et al., *A Physician's Role Following a Breach of Electronic Health Information*, 21 J. CLINICAL ETHICS 30, 31 (2010) ("dignitary harms . . . may result when a patient's autonomy is undermined"); Mark A. Rothstein, *Is Deidentification Sufficient to Protect Health Privacy in Research?*, 10 AM. J. BIOETHICS 3, 6–7 (2010). *See also* discussion *infra* Part III.B.

15. Peter Garrett & Joshua J. Seidman, *EMR vs EHR—What is the Difference?*, HEALTHITBUZZ (Jan. 4, 2011, 12:07 PM), 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.").

record-based medical research rather than interventional research. Record-based research only involves review of existing patient records, which we assume will be entirely electronic in the future. By contrast, interventional research involves physical or psychological testing, and thus experimentation on human beings.[16]

This Article makes several important contributions. Part IV argues for a change in conceptual framework that rejects the primacy of autonomy and informed consent in the context of non-interventional research. In light of past research abuses, such as the Nazi concentration camp experiments and the infamous Tuskegee syphilis trial,[17] it is not surprising that the major research ethics codes emphasize human subject autonomy, the avoidance of individual harm, and consent.[18] But electronic-database queries were not part of those prior atrocities, and they present a drastically diminished risk of gross abuses that will inflict acute physical or mental pain; thus, in the context of record-based research, autonomy should be secondary to the common good.[19] This Article explains that advancing the common good will require concessions from both data subjects and the health care industry.

The many obstacles that hinder the attainment of informed consent in large-scale, record-based studies further justify a change in conceptual framework. The consent process can be extremely burdensome and costly and can distort research results by introducing selection bias.[20] This Article thoroughly explains and illustrates how different forms of selection bias can impact a variety of study types.[21] Furthermore, while informed consent provides subjects with a choice, it does not provide them with any added protection against privacy breaches, which are the focus of most commentators' concern.[22]

In Part V, this Article formulates several practical recommendations that seek to balance the individual interests of those whose records will be used in research with societal needs to maximize the potential for medical discoveries and achieve improvements in human health. To address apprehension about privacy, this Article analyzes two identity concealment techniques. One option is to create large databases exclusively

16. *See* IOM REPORT, *supra* note 9, at 19 (differentiating between clinical trials and information-based research).

17. Sharona Hoffman, *Continued Concern: Human Subject Protection, the Institutional Review Board, and Continuing Review*, 68 TENN. L. REV. 725, 730–31 (2001).

18. *See* discussion *infra* Part IV.A.

19. Many articles describe the tension between these values. *See, e.g.*, Don E. Detmer, *Your Privacy or Your Health—Will Medical Privacy Legislation Stop Quality Health Care?*, 12 INT'L J. QUALITY HEALTH CARE 1, 1 (2000); Miller, *supra* note 12, at 562; Khadija Robin Pierce, *Comparative Architecture of Genetic Privacy*, 19 IND. INT'L & COMP. L. REV. 89, 92 (2009); Charity Scott, *Is Too Much Privacy Bad for Your Health? An Introduction to the Law, Ethics, and HIPAA Rule on Medical Privacy*, 17 GA. ST. U. L. REV. 481, 482 (2000).

20. *See* discussion *infra* Part IV.D.

21. *See* discussion *infra* Part IV.D.1.

22. *See* Peddicord et al., *supra* note 5, at 2,087 ("Consent is, at best, a rough proxy for protection from privacy harm.").

for research that would include only EHRs that have been de-identified.[23] A second approach is secure statistical analysis of distributed databases, which allows researchers to query the EHR databases of medical facilities or trusted aggregators, but enables them to receive only summary statistics in response.[24] Although not all studies can utilize these techniques, the two identity concealment mechanisms will work well in many cases.

Second, data subjects should be protected through additional oversight. We make the novel recommendation that all research protocols, including those involving only de-identified data, which are currently exempt from scrutiny,[25] be reviewed and monitored by an ethics board with expertise in record-based research. The degree of oversight should depend on the extent to which records contain identifiers that can be linked to specific patients. Research using identifiable records should be subject to a thorough approval process, and protocols in which patients' identities will be concealed should undergo a streamlined registration process. All studies should be subject to continuing review and potential unannounced audits. In addition, security safeguards for electronic databases should be bolstered.[26]

Third, this Article emphasizes the need for notification and education in lieu of consent for EHR-based research. Notification and education, like consent, can demonstrate researchers' respect for human subjects and promote a sense of autonomy.[27] Transparency and accountability on the part of researchers should prevent data subjects from suffering serious research abuses and should inspire enthusiasm about biomedical research. Furthermore, armed with knowledge and a political voice, informed members of the public can seek to influence elected officials to reverse objectionable policies through the legislative process.

II. BACKGROUND

EHR research databases are not a futuristic idea—they are fast becoming a reality. This Part provides background information concerning contemporary efforts to build EHR resources for research purposes. It also discusses the federal oversight structure for record-based research.

A. Existing Initiatives to Create EHR Research Databases

A variety of initiatives are already underway to create large databases of EHRs or networks of smaller databases, called federated networks,[28]

23. *See* discussion *infra* Part VI.A.1.
24. *See* discussion *infra* Part VI.A.3.
25. *See* discussion *infra* Part II.B.
26. *See* discussion *infra* Part VI.B.
27. IOM Report, *supra* note 9, at 266 (stating that one of the primary aims of consent is to "provide respect for the person").
28. 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." Wilson D.

that can be used for research purposes. We describe below a sample of projects that the federal government, states, and private industry have undertaken.

For many years, Department of Veterans Affairs (VA) researchers have used records collected from particular VA facilities or consolidated at a regional level.[29] The VA is now working to create a nationwide centralized data repository of de-identified patient charts.[30] In 2009, another major health care system, Kaiser Permanente, received a multi-million-dollar federal grant to establish a national electronic research database that will include health information from 30 million current and past patients in eight geographic regions.[31]

The Centers for Medicare & Medicaid Services created a research database called the Chronic Condition Data Warehouse (CCW) pursuant to Section 723 of the Medicare Modernization Act of 2003.[32] CCW provides researchers with information about Medicare and Medicaid beneficiaries, claims for services, and assessment data.[33] Researchers must submit requests through the Research Assistance Data Center and can ask for either identifiable data files or limited data sets.[34] Requests for identifiable data are scrutinized to ensure that disclosure will not violate privacy requirements.[35]

The Food and Drug Administration (FDA) Amendments Act of 2007 authorized the creation of the Sentinel health data network encompassing records from 100 million individuals.[36] The FDA does not plan to establish its own database.[37] Rather, it intends to send queries concerning potential product safety problems to various participating data holders, such as health care facilities and insurers who would have their own EHR or

---

PACE ET AL., AGENCY FOR HEALTHCARE RESEARCH AND 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.

29. U.S. Dep't of Veterans Affairs, *Study Will Boost Role of Electronic Records in Care, Research*, VA RESEARCH CURRENTS, Aug. 2009, at 3, 7, *available at* http://www.research.va.gov/news/research_highlights/records-080309.cfm.

30. *Id.* (noting that, significantly, the records in the database will include free text, such as doctors' notes, that is changed into structured data).

31. Press Release, Kaiser Permanente, National Institutes of Health Awards More Than $54 Million to Kaiser Permanente to Conduct Health Research (Oct. 12, 2009), *available at* http://www.dor.kaiser.org/external/dorexternal/news/press_releases/press_release.aspx?id=3361.

32. *About Chronic Condition Data Warehouse*, CHRONIC CONDITION DATA WAREHOUSE, http://www.ccwdata.org/about/index.htm (last visited Oct. 27, 2011).

33. *Id.*

34. *Requesting CMS Data*, RESEARCH DATA ASSISTANCE CENTER, http://www.resdac.org/Medicare/requesting_data.asp (last modified Oct. 11, 2011).

35. *Id.*

36. Barbara J. Evans, *Authority of the Food and Drug Administration to Require Data Access and Control Use Rights in the Sentinel Data Network*, 65 FOOD & DRUG L.J. 67, 67 (2010); *FDA's Sentinel Initiative-Transforming How We Monitor The Safety of FDA-Regulated Products*, U.S. FOOD & DRUG ADMIN., http://www.fda.gov/Safety/FDAsSentinelInitiative/ucm2007250.htm (last updated Oct. 5, 2011) [hereinafter *FDA's Sentinel Initiative*].

37. *FDA's Sentinel Initiative*, *supra* note 36.

claims databases.[38] Using automated mechanisms, the data holders would assess their records and send summary responses to the FDA.[39]

At the state level, the California Office of Statewide Health Planning & Development established a database of inpatient hospital discharge data.[40] Within thirty days of discharge, hospitals must report a large number of details, including diagnoses, treatments, and drug intake.[41] Selected datasets that do not directly identify patients are available for purchase by the public and thus could be used for research.[42]

In the private sector, Geisinger Health Systems established a company called MedMining that extracts EHR data, de-identifies it, and offers it to researchers.[43] MedMining asserts on its website that its customers include numerous major pharmaceutical, medical device, and biotech customers.[44] The data sets it delivers to customers feature "lab results, vital signs, medications, procedures, diagnoses, lifestyle data, and detailed costs" from both inpatient and outpatient settings.[45]

Yet another initiative is the Distributed Ambulatory Research in Therapeutics Network (DARTNet), a federated network of EHR data from eight large organizations serving over 400,000 patients.[46] DARTNet is funded by the Agency for Healthcare Research and Quality (AHRQ).[47] For each DARTNet member organization, relevant clinical information is captured in a standardized database and then transferred to another database that presents de-identified data for query access through a secure web-portal.[48] DARTNet researchers query the de-identified federated databases, consisting of data from EHRs, laboratories, imaging centers, pharmacies, and billing systems, though the patient EHRs themselves never leave the clinical sites at which they are stored.[49]

Other agencies and organizations are creating electronic registries and databases to focus on specific disease categories and to support research through data sharing. These include the Cancer Biomedical Informatics Grid,[50] the Interagency Registry for Mechanically Assisted Circulatory

---

38. *Id.*
39. *Id.*
40. *Inpatient Hospital Discharge Data*, CAL. DIABETES PROGRAM, http://www.caldiabetes.org/content_display.cfm?contentID=487&CategoriesID=31 (last updated Nov. 18, 2010).
41. Marc A. Rodwin, *Patient Data: Property, Privacy & the Public Interest*, 36 AM. J.L. & MED. 586, 615 (2010).
42. *Id.*
43. PRICEWATERHOUSECOOPERS, *supra* note 2, at 13; *Welcome to MedMining*, MEDMINING, http://www.medmining.com/index.html (last visited Oct. 27, 2011).
44. *Welcome to MedMining*, *supra* note 43.
45. *Id.*
46. PACE ET AL., *supra* note 28, at 1.
47. *Id.*
48. *Id.* at 2.
49. *Id.*
50. *About caBIG*, NAT'L CANCER INST., https://cabig.nci.nih.gov/overview/ (last modified July 2, 2011) (stating that the initiative's goal is to "[b]uild or adapt tools for collecting, analyzing, integrating, and disseminating information associated with cancer research and care.").

Support,[51] the Extracorporeal Life Support Organization,[52] and the United Network for Organ Sharing.[53]

These few examples illustrate the increasing use and importance of EHR databases for research purposes. The law does not ignore the use of medical records in research and addresses its permissibility in key federal regulations.

B. EHR-Based Research and the Law

Ordinarily, biomedical research protocols require institutional review board (IRB) approval,[54] and patients must authorize the release of identifiable information to researchers under the Health Insurance Portability and Accountability Act (HIPPA) Privacy Rule.[55] By contrast, research using de-identified EHRs can be conducted with few regulatory burdens; research involving solely de-identified records need not be approved by an IRB[56] and is not subject to coverage by the HIPAA Privacy Rule.[57] Consequently, health care providers, including clinicians and medical facilities, can disclose de-identified data to researchers without obtaining patient consent or applying HIPAA's privacy safeguards to the de-identified data. This section reviews provisions of the federal research regulations and the HIPAA Privacy Rule that apply to EHR-based research.

1. The Federal Research Regulations

The federal regulations that require IRB review and participant consent, known as the Common Rule,[58] cover only research on human subjects, and define a human subject as "a living individual about whom an investigator . . . obtains (1) [d]ata through intervention or interaction with the individual, or (2) [i]dentifiable private information."[59] Because of the very minimal risk of harm to participants, the regulations specifically exempt research "involving the collection or study of existing data, documents, [or] records . . . if the information is recorded by the investigator in such a manner that subjects cannot be identified, directly or through

51. *INTERMACS Description*, INTERMACS, http://www.uab.edu/ctsresearch/intermacs/description.htm (last visited Oct. 27, 2011) (explaining that analysis of the collected data is expected to improve patient care and to "influence future research").

52. *ELSO Registry Information Data Policy*, 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).

53. *Data*, United Network for Organ Sharing, http://www.unos.org/donation/index.php?topic=data (last visited Oct. 27, 2011) (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").

54. 45 C.F.R. § 46.109 (2010).

55. *Id.* § 164.508(b)(3)(i).

56. *Id.* § 46.101(b)(4).

57. See *id.* § 160.103 for the definition of "Protected Health Information."

58. *See Federal Policy for the Protection of Human Subjects ('Common Rule')*, U.S. Dep't Health & Hum. Servs., http://www.hhs.gov/ohrp/humansubjects/commonrule/index.html (last visited Oct. 27, 2011).

59. 45 C.F.R. § 46.102.

identifiers linked to the subjects."[60] The regulations provide no details as to which identifiers need to be removed to render data de-identified.[61]

The Common Rule provides IRBs with flexibility and allows them to exercise discretion in appropriate circumstances. IRBs may waive the requirement of informed consent if they find that a study involves no more than minimal risk for subjects and entails no procedure for which consent would be required in the treatment setting.[62] Accordingly, even record-based research involving personally identifiable health information may be exempted from the informed consent mandate.

Research utilizing a database of EHRs that have been previously de-identified would not be covered by the research regulations.[63] Furthermore, the applicability of the federal regulations to research involving medical records rather than interaction with patients depends on the method by which data is *recorded* by the investigator.[64] Consequently, even research through a service that queries EHRs with identifiable patient data but presents results to researchers in summary, non-identifiable form would most likely be exempt from IRB review. At most, however, such project proposals would be sent to IRBs, would be deemed to pose only minimal risk to subjects, and would consequently require no informed consent process.

2. *The HIPAA Privacy Rule*

The HIPAA Privacy Rule generally prohibits disclosure of individually identifiable health information without patient consent, unless the information is transmitted for purposes of treatment, payment, or health care operations.[65] The HIPPA Privacy Rule's application to research activities is analyzed in this section.

a. De-Identified Information

Like the Common Rule, the HIPAA Privacy Rule covers only "individually identifiable health information."[66] Thus, the Rule does not prohibit covered entities[67] from disclosing de-identified data to third parties, including researchers. The regulations provide that information can be considered de-identified: (1) if an appropriate expert determines that

60. *Id.* § 46.101(b)(4).
61. *Id.*
62. *Id.* § 46.117(c).
63. *Human Research Protections Frequently Asked Questions: Can I Analyze Data That Are Not Individually Identifiable, Such as Medication Databases Stripped of Individual Patient Identifiers, for Research Purposes Without Having to Apply the HHS Protection of Human Subjects Regulations?*, U.S. DEP'T OF HEALTH & HUM. SERVS., http://answers.hhs.gov/ohrp/questions/7284 (last updated Dec. 30, 2010).
64. 45 C.F.R. § 46.101(b)(4).
65. *Id.* § 164.506.
66. *Id.* § 160.103 (defining "protected health information").
67. The HIPAA Privacy Rule applies to health plans, health care clearinghouses, health care providers who transmit health information electronically for particular purposes (generally claims or benefits activities), and their business associates. *Id.* §§ 160.102–160.103; 42 U.S.C. § 17931 (2006).

there is only a "very small" risk that the information could be re-identified, and (2) the expert documents his or her analysis.[68] This criterion is known as the HIPAA "statistical standard."[69] In the alternative, information is deemed de-identified according to the HIPAA Privacy Rule's "safe harbor" provision[70] if the following eighteen identifiers are removed:

(A) Names;

(B) All geographic subdivisions smaller than a State, including street address, city, county, precinct, zip code, and their equivalent geocodes, except for the initial three digits of a zip code if, according to the current publicly available data from the Bureau of the Census:

(1) The geographic unit formed by combining all zip codes with the same three initial digits contains more than 20,000 people; and

(2) The initial three digits of a zip code for all such geographic units containing 20,000 or fewer people is changed to 000;

(C) All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older;

(D) Telephone numbers;

(E) Fax numbers;

(F) Electronic mail addresses;

(G) Social security numbers;

(H) Medical record numbers;

(I) Health plan beneficiary numbers;

(J) Account numbers;

(K) Certificate/license numbers;

(L) Vehicle identifiers and serial numbers, including license plate numbers;

(M) Device identifiers and serial numbers;

(N) Web Universal Resource Locators (URLs);

(O) Internet Protocol (IP) address numbers;

(P) Biometric identifiers, including finger and voice prints;

(Q) Full face photographic images and any comparable images; and

(R) Any other unique identifying number, characteristic, or code.[71]

The requirements for de-identification under this provision are far more specific than those of the Common Rule. It is, therefore, possible

68. 45 C.F.R. § 164.514(b)(1).

69. Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 1701, 1737 (2010).

70. Standard for Privacy of Individually Identifiable Health Information, 67 Fed. Reg. 53,182, 53,232 (Aug. 14, 2002) (codified at 45 C.F.R. pts. 160 & 164).

71. 45 C.F.R. § 164.514(b)(2)(i). In addition, information will not be considered de-identified if an entity has "actual knowledge that the information could be used alone or in combination with other information to identify an individual who is a subject of the information." *Id.* § 164.514(b)(2)(ii).

that a protocol would be exempt from the Common Rule's consent mandate because some identifiers will be removed, but would still require patient authorization under the HIPAA Privacy Rule because not all eighteen safe harbor identifiers are redacted.[72]

b. Other HIPAA Exemptions

The HIPAA Privacy Rule contains several other exceptions that apply to research use of health data. Covered entities[73] may disclose "limited data sets" without patient consent if the recipient signs a data use agreement that prohibits re-identification of the data.[74] Limited data sets allow somewhat more liberal disclosures than the safe harbor provision because they make three modifications to the eighteen-factor list: disclosure of all elements of dates, including exact birth dates, is permitted, and while specific addresses must be withheld, patients' towns or cities and zip codes can be revealed.[75] The limited data set provision also eliminates the catch-all item of "any other unique" identifier.[76]

In addition, the HIPAA Privacy Rule does not protect records of decedents that are used for research purposes.[77] Researchers can obtain further exemptions with approval of an IRB or privacy board in accordance with regulatory guidance.[78]

III. THE BENEFITS AND RISKS OF RESEARCH USING EHR DATA

The advent of EHRs has the potential to transform medical research by enabling investigators to conduct computerized searches that will yield an unprecedented wealth of information about patient care and treatment efficacy. By the same token, the prospect of electronic research raises serious concerns that cannot be ignored. The possible benefits and harms of EHR-based research are thoroughly analyzed in this part.

A. THE CONTRIBUTIONS OF EHR-BASED RESEARCH

EHR technology could make an invaluable contribution to medical research because it can facilitate large-scale observational studies that will fill existing knowledge gaps. Contemporary medical practice involves a startling amount of guesswork.[79] According to some estimates, as few as

72. IOM REPORT, *supra* note 9, at 173.

73. *See supra* note 67 and accompanying text.

74. 45 C.F.R. § 164.514(e)(1); *see also* § 164.514(e)(4) (containing details concerning data use agreements).

75. 45 C.F.R. § 164.514(e)(2).

76. *Id.*

77. *Id.* § 164.512(i)(1)(iii).

78. *Id.* § 164.512(i)(1)(i). Identifiable medical records may also be used without patient consent to prepare (but not carry out) research protocols as long as the records do not leave the facility in which they are stored. *Id.* § 164.512(i)(1)(ii).

79. 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'

20 to 25% of treatments have been definitively proven effective.[80] In many instances, physicians initially try particular treatment plans, medications, or dosages knowing that these will likely need to be changed or adjusted before the patient receives optimal treatment.[81]

Both the Obama Administration and the Institute of Medicine (IOM) have recognized the importance of CER.[82] 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."[83] CER's aim is to generate improved patient outcomes while maximizing the benefit of health care expenditures.[84] A 2009 IOM Report similarly emphasized the need for CER and proposed initial CER priorities.[85] Such research could lead to a significant reduction in human suffering, disease-related death rates, and health care costs.

CER is to be conducted through a wide variety of means, including both clinical trials and observational studies.[86] Randomized, controlled clinical trials are considered to be the gold standard of medical studies.[87] 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."[88] In a randomized experiment, subjects are randomly assigned to receive one of the interventions under study (possibly including no intervention). For example, investigators might design a clinical trial to include two groups to which eligible patients are randomly assigned: one

---

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").

80. John Casey, *Medical Guesswork*, BUSINESSWEEK, May 29, 2006, at 72 (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%").

81. *See id.*

82. *See* 42 U.S.C. § 1320(e) (2010); Laxmaiah Manchikanti et al., *Facts, Fallacies, and Politics of Comparative Effectiveness Research: Part I Basic Consideration*, 13 PAIN PHYSICIAN, E23, E38–39 (2010).

83. 42 U.S.C. § 1320e(a)(2)(A).

84. *Id.* § 1320e(d)(2)(A); 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); Manchikanti et al., *supra* note 82, at E39.

85. INST. OF MED., INITIAL NATIONAL PRIORITIES FOR COMPARATIVE EFFECTIVENESS RESEARCH 1-3 (2009), *available at* http://www.iom.edu/Reports/2009/ComparativeEffectivenessResearchPriorities.aspx.

86. 42 U.S.C. § 1320e(d)(2)(A).

87. Friedrich K. Port, *Role of Observational Studies Versus Clinical Trials in ESRD Research*, 57 KIDNEY INT'L S-3, S-3 (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 Research or Unethical Practice?*, 33 CONN. L. REV. 449, 452–54 (2001) (describing different designs of clinical trials).

88. BRYAN F. J. MANLY, THE DESIGN AND ANALYSIS OF RESEARCH STUDIES 1 (1992).

group receives Angiotensin-Converting Enzyme (ACE) inhibitors for heart failure, and the second group receives ACE inhibitors in combination with a different drug for the same condition.[89] The goal of this experimental study would be to determine which treatment is more effective as reflected by one or more outcome measures.

By contrast, research can also be accomplished through observational studies.[90] One source defines an "observational study" as "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."[91] Thus, rather than conducting a controlled experiment, investigators might review the charts or electronic files of patients receiving different medications or different types of surgery to treat a particular condition to determine the efficacy of each approach.[92] For example, in exploring the utility of EHRs for genetic research, a recent study found that data captured from EHRs could identify disease characteristics with sufficient accuracy to be used in genome-wide association studies.[93] Observational studies are often conducted when the FDA requires post-marketing studies to verify the safety of drugs.[94]

Observational studies, such as reviews of EHR data, are vulnerable to several criticisms.[95] These studies are not randomized, and the absence of randomization may introduce biases that skew results.[96] For example, if investigators review only records that come from a particular wealthy, suburban medical practice, the results derived may not apply to low-income populations with higher levels of stress, poorer diets, and inferior

89. Sharona Hoffman, *"Racially-Tailored" Medicine Unraveled*, 55 AM. U. L. REV. 395, 400–02 (2005) (describing a clinical trial for heart failure medication).

90. *See* MANLY, *supra* note 88, at 1 (explaining that observational studies involve the collection of data "by observing some process which may not be well-understood"); *see also* CHARLES P. FRIEDMAN & JEREMY C. WYATT, EVALUATION METHODS IN BIOMEDICAL INFORMATICS 369 (2d ed. 2006) (defining observational studies as involving an "[a]pproach to study design that entails no experimental manipulation" in which "[i]nvestigators typically draw conclusions by carefully observing . . . [subjects] with or without an information resource").

91. PAUL R. ROSENBAUM, OBSERVATIONAL STUDIES vii (2d ed. 2002).

92. Kjell Benson & Arthur J. Hartz, *A Comparison of Observational Studies and Randomized, Controlled Trials*, 342 NEW ENG. J. MED. 1878, 1879–83 (2000).

93. Kho et al., *supra* note 3, at 4–5.

94. U.S. DEP'T OF HEALTH & HUMAN SERVS. FOOD & DRUG ADMIN., GUIDANCE FOR INDUSTRY POSTMARKETING STUDIES AND CLINICAL TRIALS—IMPLEMENTATION OF SECTION 505(O)(3) OF THE FEDERAL FOOD, DRUG, AND COSMETIC ACT, 2–7 (2011), *available at* http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM172001.pdf.

95. *See* Gary Taubes, *Do We Really Know What Makes Us Healthy?*, N.Y. TIMES, Sept. 16, 2007, at 52 (describing the limitations of observational studies and stating that they "can only provide what researchers call hypothesis-generating evidence—what a defense attorney would call circumstantial evidence").

96. *See* Benson & Hartz, *supra* note 92, at 1878 (stating that "[c]oncern about inherent bias" in observational studies "has limited their use in comparing treatments"); *see also* MANLY, *supra* note 88, at 4–5.

access to medical care.[97] Similar skewing, however, may occur in interventional research if the population from which subjects are recruited is not sufficiently diverse.[98] However, there are ways of controlling for the bias problem in creating (or extracting data from) an EHR database, such as ensuring that the database is both large enough and drawn randomly from the EHRs of a diverse patient population.[99]

A second concern is that observational study results could be confounded by uncontrolled variables because the assignment of different treatments, including placebos, to patients is not randomized.[100] Thus, any changes that are observed might be caused not by the intervention of interest but by factors, such as age or sex, that influence both the treatment patients receive and the outcomes they have.[101] If researchers do not carefully monitor and adjust for these factors, any conclusion concerning the efficacy of the drug at issue is likely to be questionable.

Third, EHR database studies may also be affected by data quality problems.[102] Researchers cannot assume that EHR data is completely accurate. The data in EHRs may be incomplete or erroneous because, among other reasons, clinicians make typing mistakes, do not have enough time to create comprehensive and error-free records, or have difficulty navigating the EHR system.[103] To estimate error rates and magnitudes, researchers may need to validate the EHRs of a sample of patients, which would entail contacting them or their physicians.[104]

Other complications may compromise the quality of EHR data as well. Medical terminology lacks standardization, and physicians can use the same abbreviations to mean very different things.[105] For example, "MS" can mean "mitral stenosis," "multiple sclerosis," "morphine sulfate," and "magnesium sulfate."[106] In addition, patients who see doctors at different medical facilities whose EHR systems are not interoperable may have fragmented records and pieces of their medical histories in different EHRs.[107]

Problems with the completeness and accuracy of EHR data can be mitigated in part through increased use of electronic means for collecting

97. *See* discussion *infra* Part IV.D.1 (explaining how informed consent can lead to selection bias).

98. *See* MANLY, *supra* note 88, at 4–5.

99. *See id.* at 16.

100. *See id.* at 4–5.

101. *See id.* at 9.

102. *See* Sharona Hoffman & Andy Podgurski, *E-Health Hazards: Provider Liability and Electronic Health Record Systems*, 24 BERKELEY TECH. L.J. 1523, 1537–45 (2009).

103. *See id.*

104. *See id.* at 1565–69, 1577.

105. Christopher G. Chute, *Medical Concept Representation*, *in* MEDICAL INFORMATICS: KNOWLEDGE MANAGEMENT AND DATA MINING IN BIOMEDICINE 163, 170–71 (Hsinchin Chen et al. eds., 2005).

106. *Id.* at 170, tbl. 6-1.

107. Barbara Evans, *Much Ado About Data Ownership*, 25 HARV. J.L. & TECH. 70, 88, 93–94 (2011).

patient data, such as remote patient monitoring.[108] It must also be recognized that data integrity problems are not unique to observational studies. Clinical trials are often criticized for design flaws and other deficiencies.[109] Researchers must be aware of the limitations of their research tools and techniques and strive continuously to improve them.

In fact, observational studies have several advantages over clinical trials.[110] EHR databases could allow researchers to access vast amounts of information about patients with diverse demographics collected over a much longer period of time than that encompassed by clinical trials, which typically last only a few years.[111] The data used in observational studies, consequently, may be far more comprehensive than the data generated by clinical trials, which often include fewer than 3,000 patients.[112] Observational studies can also be considerably less costly and time-consuming than experimental research because the data used already exist.[113]

In some cases, it is impossible to conduct clinical trials.[114] This may be because it is too difficult to recruit a large enough subject population to yield statistically significant results, such as when the condition is very rare.[115] Clinical studies may also be unrealistic because it would be unethical to conduct them.[116] For example, investigators could not examine the outcomes of patients who receive the wrong treatment by deliberately

---

108. Kevin D. Blanchet, *Remote Patient Monitoring*, 14 TELEMEDICINE & E-HEALTH 127, 128–30 (2008).

109. *See, e.g.*, Lorena Baccaglini et al., *Design and Statistical Analysis of Oral Medicine Studies: Common Pitfalls*, 16 ORAL DISEASES 233, 233–40 (2010); Ron Dagan & George H. McCracken, *Flaws in Design and Conduct of Clinical Trials in Acute Otitis Media*, 21 PEDIATRIC INFECTIOUS DISEASES J. 894, 894–901 (2002); 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, 439–50 (2010); Martha Clare Morris & Christine C. Tangney, *A Potential Design Flaw of Randomized Trials of Vitamin Supplements*, 305 JAMA 1348, 1348–49 (2011); Stephen D. Simon, *Is the Randomized Clinical Trial the Gold Standard of Research?*, 22 J. ANDROLOGY 938, 938–42 (2001).

110. Benson & Hartz, *supra* note 92, at 1878 (citing the advantages of diminished cost, timeliness, and a broader spectrum of patients).

111. *See, e.g.*, 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; Evans, *supra* note 109, at 446 ("Phase III trials typically last one to four years and may include 1000 to 10,000 patients of whom only a few hundred patients typically receive the new drug for more than three to six months."); 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"); 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).

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

113. Benson & Hartz, *supra* note 92, at 1878 (mentioning "greater timeliness" as an advantage of observational studies); Port, *supra* note 87, at S-3, S-4.

114. Benson & Hartz, *supra* note 92, at 1878.

115. *See* Etheredge, *supra* note 111, at w107.

116. Benson & Hartz, *supra* note 92, at 1878.

giving some individuals incorrect medications.[117] By contrast, review of EHR databases could allow for a broader range of research.[118] Investigators could gain access to patient records all over the country, including those of individuals with very rare illnesses.[119] In addition, researchers could study data relating to actual patients who are treated in a clinical setting, rather than in the controlled environment of a research trial, and could analyze care that is of varying quality, including substandard care.[120]

It is not anticipated that EHR-based observational studies will replace randomized clinical trials.[121] However, observational studies are an indispensable addition to the research tool kit.[122] In the words of one commentator, EHRs "will offer the capacity for real-time learning from the experience of tens of millions of people and will greatly increase the ability to generate and test hypotheses."[123]

B. Potential Harms Associated with EHR-Based Research

While the anticipated benefits of EHR-based research are significant, such research is not devoid of risks. Data subjects may risk privacy violations as well as other dignitary harms, all of which are addressed in this part.

1. Privacy

The terms "privacy" and "confidentiality" are at times used interchangeably or inconsistently, but the IOM offers illuminating definitions of these words.[124] According to the IOM, privacy focuses on the "collection, storage, and use of personal information" and thus on questions of access to data.[125] Confidentiality concerns the duty to avoid improper disclosure of information that is conveyed in an intimate relationship.[126] Inappropriate disclosures of EHR data may involve violations of both privacy and confidentiality. However, for purposes of simplicity, we use the word "privacy" to encompass all aspects of the concern about data disclosure.

---

117. *See* Manly, *supra* note 88, at 13–14.
118. *See* Etheredge, *supra* note 111, at w107.
119. *See id.* at w109.
120. *See id.* at w109–w116.
121. *See id.* at w108.
122. *See* Benson & Hartz, *supra* note 92, at 1878, 1884 (concluding, based on a literature review, that "observational studies and randomized, controlled trials usually produce similar results"); Port, *supra* note 87, at S-5 (arguing that both observational studies and clinical studies have their place and complement each other). *But see* Gordon H. Guyatt et al., *Randomized Trials Versus Observational Studies in Adolescent Pregnancy Prevention*, 53 J. Clinical Epidemiology 167, 173 (2000) (cautioning researchers about the risks of observational studies and stating that recommendations should be based on randomized trials whenever possible).
123. Etheredge, *supra* note 111, at w108.
124. *See* IOM Report, *supra* note 9, at 16–17.
125. *Id.* at 16–17, 76.
126. *Id.* at 76.

a. Privacy Breach Harms

Once information is digitized, it is vulnerable to privacy breaches resulting from hacking; stolen or misplaced laptops and storage devices; accidental disclosures, such as e-mails inadvertently sent to the wrong recipient; or even intentional misconduct.[127] The news media and other organizations have provided accounts of many such violations during the last several years.[128] The Department of Health and Human Services (HHS) website lists almost 300 health care providers and insurers that have reported significant breaches since September of 2009.[129]

The personal and sensitive information contained in medical records might be of interest to a large number of parties.[130] Employers wish to hire healthy workers who will not have productivity and absenteeism problems or submit costly medical claims for reimbursement.[131] Various types of insurers (e.g., life, disability, long-term care) want to find clients who are low-risk and whose premium payments will exceed claims.[132] Lenders are interested in borrowers who can work and earn salaries that will enable them to pay off their loans.[133]

Advertisers and marketers hope to influence doctors' prescribing decisions and patients' medical purchasing choices; political operatives may hope to use health information to disqualify or embarrass candidates; and blackmailers or other criminals may seek financial gain through the possession and use of such data.[134]

If health information contained in research databases can be linked to the names of data subjects, those with access to the data could theoretically sell or distribute it to interested third parties. Comprehensive EHRs will include psychiatric records, reproductive and sexual histories, HIV status, serious illnesses such as cancer, and much more.[135] Thus, patients whose information falls into inappropriate hands could face employment or insurance discrimination;[136] lose financial and other opportunities; become victims of criminal conduct; or suffer public embarrassment, though some of these harms may be mitigated by existing

127. *See* Sharona Hoffman & Andy Podgurski, *In Sickness, Health, and Cyberspace: Protecting the Security of Electronic Private Health Information*, 48 B.C. L. REV. 331, 332–34 (2007).

128. *See id.* at 332–33; *see also* IOM REPORT, *supra* note 9, at 95–96 tbl. 2-2; Milt Freudenheim, *Breaches Lead to Push to Protect Medical Data*, N.Y. TIMES, May 30, 2011, at B1.

129. *Health Information Privacy: Breaches Affecting 500 or More Individuals*, U.S. DEP'T OF HEALTH & HUM. SERVS., http://www.hhs.gov/ocr/privacy/hipaa/administrative/breachnotificationrule/breachtool.html (last visited Oct. 27, 2011) [hereinafter *Breaches Affecting*).

130. Hoffman & Podgurski, *supra* note 127, at 334–35.

131. *Id.*

132. *Id.*

133. *Id.*

134. *Id.*

135. Sharona Hoffman, *Employing E-Health: The Impact of Electronic Health Records on the Workplace*, 19 KAN. J.L. & PUB. POL'Y 409, 421–24 (2010).

136. *See id.*

anti-discrimination laws.[137]

It is important to note, however, that the danger of electronic privacy breaches arises as soon as providers convert patients' medical files from paper format to EHRs, and clinicians do not consult patients about whether to undertake this transition. To date, data breaches have in fact generally occurred in the clinical rather than research setting.[138] Furthermore, patients routinely face privacy risks not only because of security vulnerabilities in EHR systems, but also because of vulnerabilities in their own computers or other electronic devices, data mining of data sources such as purchase records, and elicitation of sensitive information directly from patients by websites such as social networking services.[139] Thus, patients should not perceive research activities involving EHRs as generating privacy risks that would otherwise be entirely nonexistent.

b. Privacy and De-Identification

One technique that could reduce privacy risks is de-identification of records.[140] Nevertheless, commentators worry that de-identification does not provide sufficient protection to data subjects.[141] The potential shortcomings of de-identification are analyzed below.

i. De-Identification Procedures

Some experts question the reliability of contemporary de-identification techniques.[142] The quality of de-identification may vary among different EHR systems; de-identification capacity often is not designed into EHR systems, and, thus, it must be added after data is exported from an EHR system.[143] Different parts of the EHR, such as patient demographics, clinicians' free-text notes, laboratory and imaging reports, and hospitalization records, may have to be de-identified separately, and, thus, the process might be very labor-intensive and time-consuming.[144] Furthermore, a fragmented and complex process could result in many instances in which identifiers are overlooked and retained in the record.[145] Thus, if de-identification is not automated, it would need to be assigned to trusted

137. *Id.*; *see also* discussion *infra* Part VI.C.2.

138. *See Breaches Affecting*, *supra* note 129.

139. DANIEL J. SOLOVE, THE DIGITAL PERSON 41–55 (Jack M. Balkin & Beth Simone Noveck eds., 2004); Balachander Krishnamurthy & Craig E. Wills, *On the Leakage of Personally Identifiable Information Via Online Social Networks*, 40 COMPUTER COMM. REV. 112, 112 (2010).

140. *See supra* Part II.B. (discussing the application of federal regulations to de-identified records); *see also* discussion *infra* Part VI.A.1 (discussing de-identification techniques).

141. Rothstein, *supra* note 14, at 5–6.

142. *Id.* at 5.

143. *Id.*; *see also* Ben Wellner et al., *Rapidly Re-targetable Approaches to Deidentification in Medical Records*, 14 J. AM. MED. INFORMATICS ASS'N 564, 572 (2007).

144. Rothstein, *supra* note 14, at 5; *see also* Ishna Neamatullah et al., *Automated De-identification of Free-Text Medical Records*, 8 BMC MED. INFORMATICS & DECISION MAKING 32, 33 (2008), *available at* http://www.biomedcentral.com/content/pdf/1472-6947-8-32.pdf.

145. *See supra* Part II.B.2.a (listing identifiers discussed in the HIPAA Privacy Rule).

professionals. In addition, it is possible that a cryptographic key will have to be retained in case researchers need to conduct follow-up studies that require re-identification so that data can be linked to specific individuals.[146] Such a key would need to be carefully safeguarded so that it does not fall into the hands of potential wrongdoers.

ii. The Possibility of Re-Identification

Experts have found that de-identified information can be re-identified using publicly available resources, such as voter registration records.[147] The risk may be small, but it exists.

In general, de-identification is based on assumptions that third parties do not have certain information about data subjects that may facilitate re-identification; however, adversaries may legally or illegally obtain such information from a variety of sources and then correlate it to de-identified records to achieve re-identification.[148] For example, information about patients' medication purchases or evidence of the web links on which an individual clicks can be useful for this purpose.[149]

It is estimated that between 63% and 87% of the U.S. population could be accurately identified based on the three factors of gender, zip code, and date of birth, without any need for details such as name, social security number, or a precise address.[150] Latanya Sweeney, a leading authority, asserts that 0.04% of records that comply with the de-identification requirements of the HIPAA Privacy Rule[151] could be re-identified.[152] Dr. Sweeney is famous for having identified the health records of Massachusetts Governor William Weld when she was a graduate student in 1996 based on anonymized hospital discharge data that was released to the public and voter registration information that was also publicly available.[153]

---

146. Patricia Kosseim & Megan Brady, *Policy by Procrastination: Secondary Use of Electronic Health Records for Health Research Purposes*, 2 McGILL J.L. & HEALTH 5, 28 (2008).

147. *Id.* at 28–29; Ohm, *supra* note 69, at 1703 ("Clever adversaries can often *reidentify* or *deanonymize* the people hidden in an anonymized database.") (emphasis added).

148. GEORGE T. DUNCAN ET AL., STATISTICAL CONFIDENTIALITY: PRINCIPLES & PRACTICE 37 (2011).

149. *Id.* at 29–31 (discussing use of microdata).

150. Philippe Golle, *Revisiting the Uniqueness of Simple Demographics in the U.S. Population*, *in* ASS'N FOR COMPUTIVE MACHINERY WORKSHOP ON PRIVACY IN THE ELECTRONIC SOC'Y 77, 77 (2006), *available at* http://crypto.stanford.edu/~pgolle/papers/census.pdf; Latanya Sweeney, *Simple Demographics Often Identify People Uniquely* 2 (Carnegie Mellon Univ., Working Paper No. 3, 2000), *available at* dataprivacy.org/projects/identifiability/paper1.pdf.

151. *See supra* Part II.B.2.a (discussing de-identification standards under the HIPAA Privacy Rule).

152. NAT'L COMM. ON VITAL & HEALTH STATISTICS, REPORT TO THE SECRETARY OF HEALTH AND HUMAN SERVICES ON ENHANCED PROTECTIONS FOR USES OF HEALTH DATA: A STEWARDSHIP FRAMEWORK FOR "SECONDARY USES" OF ELECTRONICALLY COLLECTED AND TRANSMITTED HEALTH DATA 36 n.16 (2007), *available at* www.ncvhs.hhs.gov/071221lt.pdf.

153. Kathleen Benitez & Bradley Malin, *Evaluating Re-identification Risks with Respect to the HIPAA Privacy Rule*, 17 J. AM. MED. INFORMATICS ASS'N 169, 169 (2010).

A study published in 2010 by Kathleen Benitez and Bradley Malin[154] found that even records that have been de-identified in accordance with HIPAA Privacy Rule specifications[155] are potentially vulnerable to re-identification. The degree of risk varies from state to state and depends on what demographic information is available to the public through voter registration records.[156] When all eighteen HIPAA safe harbor provision identifiers are removed, the percentage of a state's population vulnerable to unique re-identification was estimated to range from 0.01% to 0.25%.[157] When the identifiers permitted by HIPAA for limited data sets were added in, the risk percentage rose to between 10% and 60%, depending on the state.[158] In 2011, the same authors published a second paper in which they assessed their own method of de-identification—consistent with HIPAA's statistical standard.[159] They quantified the risk of re-identification in this case as ranging "from 0.01% to 0.19%."[160]

Both of the Benitez and Malin studies make particular assumptions about the re-identification scheme and the external data used to implement it.[161] They focus on a "marketer attack" using demographic data about patients, such as that found in voter registration records.[162] In a "marketer attack," the adversary simply tries to identify as many records as possible and does not focus on a particular record or subset of records.[163] The authors also assume that adversaries will use publicly available data and not engage in illegal activity, such as hacking.[164] In addition, attackers are assumed to be private individuals rather than business entities that might have more information about targeted data subjects.[165] Needless to say, these assumptions may not apply in actual attempts at re-identification, and, thus, the risk figures supplied by Benitez and Malin may be misleading.

A recent paper by the Technology Policy Institute, a nonprofit, asserted that "there is no evidence that re-identification by a true adversary (somebody other than a researcher or journalist interested in the efficacy

154. *Id.*

155. *See supra* Part II.B.2.a.

156. Benitez & Malin, *supra* note 153, at 176.

157. *Id.* at 169.

158. *Id.*; *see also supra* notes 74–76 and accompanying text (discussing limited data sets and the HIPAA Privacy Rule).

159. Bradley Malin et al., *Never Too Old for Anonymity: A Statistical Standard for Demographic Data Sharing Via the HIPAA Privacy Rule*, 18 J. AM. MED. INFORMATICS ASS'N 3, 3 (2011). The statistical standard is articulated in 45 C.F.R. § 164.514(b)(1) (2010) (stating that information can be considered de-identified if an appropriate expert determines that there is only a "very small" risk that the information could be re-identified and documents her analysis).

160. Malin et al., *supra* note 159, at 7.

161. *Id.* at 4–5; *see also* Benitez & Malin, *supra* note 153, at 170.

162. Benitez & Malin, *supra* note 153, at 170; Malin et al., *supra* note 159, at 4.

163. Benitez & Malin, *supra* note 153, at 170; Malin et al., *supra* note 159, at 4.

164. Benitz & Malin, *supra* note 153, at 170.

165. Benitez & Malin, *supra* note 153, at 170; Malin et al., *supra* note 159, at 4.

of privacy protections) has actually happened."[166] The authors asserted that because re-identification is very difficult to achieve, it may be possible "only for small populations under unusual conditions."[167] Still, even a fraction of a percent of re-identification risk could mean that hundreds of thousands of Americans' de-identified records would be vulnerable.[168]

2. *Harms Not Related to Privacy*

While the potential for privacy breaches has received significant attention in the literature, other possible harms to the dignity or autonomy of patients have raised concerns as well.[169] If patients are not asked to consent to research that involves their EHRs, they will have no opportunity to determine whether they are willing to accept the risks of dignitary harms. As Professor Mark Rothstein has argued, these harms include group stigmatization, inadvertently supporting medical developments that one finds morally objectionable, and enabling commercial enterprises to garner large profits in which data subjects do not share.[170]

a. Group Stigmatization

Group stigmatization may occur if researchers find that individuals with particular ancestry are more vulnerable to a specific illness than other groups or have better outcomes with treatment that is different from standard therapy.[171] For example, the genetic abnormalities BRCA1 and BRCA2 are associated with an increased risk of breast and ovarian cancer and are found more commonly in Ashkenazi Jews.[172] When genetic testing was developed to identify the BRCA1 and BRCA2 mutations, some members of the Jewish community became anxious that Jews would be perceived as having a flawed genetic makeup or as being unusually diseased.[173] Likewise, the FDA's 2005 approval of the drug BiDil only for African-Americans generated significant concern about the implications of ethnopharmacology.[174] Would race-based prescriptions lead some to assume that African-Americans were biologically different from and measurably inferior to others?[175] Data subjects whose de-identified information is used in research without their consent will

166. JANE YAKOWITZ & DANIEL BARTH-JONES, TECH. POLICY INST., THE ILLUSORY PRIVACY PROBLEM IN *Sorrell v. IMS Health* 7 (2011), *available at* www.techpolicyinstitute.org/files/the%20illusory%20privacy%20problem%20in%20sorrell.pdf.

167. *Id.*

168. *See* discussion *infra* Part IV.D.2.b.

169. *See, e.g.*, Golle, *supra* note 150, at 77.

170. Rothstein, *supra* note 14, at 6–7.

171. *Id.*; *see also* Sharona Hoffman, *"Racially-Tailored" Medicine Unraveled*, 55 AM. U. L. REV. 395, 423–27 (2005).

172. Roxana Moslehi, *BRCA1 and BRCA2 Mutation Analysis of 208 Ashkenazi Jewish Women with Ovarian Cancer*, 66 AM. J. HUM. GENETICS 1259, 1264 (2000).

173. Hoffman, *supra* note 171, at 423.

174. *Id.* at 396–97.

175. *Id.* at 424.

likely not have opportunities to opt out of studies that could conceivably lead to stigmatization of groups with which they strongly identify.

b. Moral Objections

Biomedical research could also lead to outcomes that some data subjects find unacceptable.[176] For example, research may reveal that particular fetal abnormalities can be discovered in-utero, and testing for the abnormality may ultimately induce parents to abort fetuses that they would have otherwise kept.[177] A patient who opposes abortion may find it abhorrent to have her medical file play a role in such research, even if it is merely subject to an automated query as part of a large database of de-identified files. Yet, without an informed consent process, she will be given no choice in the matter.

c. No Share in Commercial Profits

Biomedical research, at its most successful, can enable pharmaceutical and device manufacturers to enjoy significant monetary rewards. However, manufacturers achieve commercial success only after the investment of considerable time and money in product development and then only in a minority of instances. The cost of bringing a drug from initial clinical testing to FDA approval has been estimated at $802 million, and the process takes an average of 90.3 months.[178] Furthermore, according to a study of clinical trial data from 2003 to 2010, only 10% of drugs actually progress from phase one trials to FDA approval.[179] However, when medical products are marketed, they can be very lucrative, generating billions of dollars of revenue,[180] and these profits are not shared with the research subjects who participated in the relevant studies.[181]

Informed consent forms often include language that explains the possibility that the research sponsor or another party will benefit financially from the research.[182] A 2008 Canadian study found that research participants were particularly concerned about their ability to consent if others

176. *See* Miller, *supra* note 12, at 561 ("[S]ome individuals whose data are used might object to the purpose of the research.").

177. *See* Greely, *supra* note 12, at 760–61 (providing the examples of research concerning "genetic associations with intelligence, violence, or sexual orientation or research into human evolution," all of which might be offensive to some individuals); Rothstein, *supra* note 14, at 7.

178. Joseph A. DiMasi et al., *The Price of Innovation: New Estimates of Drug Development Costs*, 22 J. HEALTH ECON. 151, 164, 166 (2003).

179. David Thomas, *Release of BIO/Biomedtracker Drug Approval Rates Study*, BIOTECH NOW (Feb. 15, 2011), www.biotech-now.org/events12011/021/release-of-biobiomedtracker-drug-approval-rates-study/.

180. *See* PFIZER INC., 2010 FINANCIAL REPORT 25 (2010), *available at* www.pfizer.com/files/annualreport/2010/financial/financial2010.pdf (indicating that, in 2010, Pfizer earned $10.733 billion from Lipitor, $1.928 billion from Viagra, and $1.718 billion from Effexor).

181. Rothstein, *supra* note 14, at 7.

182. *Id.*

might gain financial benefits from use of their data.[183] If patients are not asked to consent, they cannot opt out no matter how strongly they object to this possibility. It should be noted, however, that it is extremely unlikely that lucrative medical products will be developed entirely based on observational studies using EHRs. Randomized, controlled clinical trials remain the gold standard for drug and device approval.[184] Thus, manufacturers seeking to make large profits will still conduct studies for which they will need to gain the consent of participants who will in turn have the opportunity to decline enrollment.

IV. INFORMED CONSENT

Because there is some possibility that record-based research will result in harm to patients, some would argue that data subjects should be given an opportunity to withhold consent to release their files for EHR studies. This Part will address the origins of the informed consent doctrine and the appropriateness of applying it to EHR database studies. It makes the case that obtaining informed consent is sensible with respect to clinical trials that involve human experimentation but is generally unnecessary for research projects that are restricted to accessing EHR databases. As we will argue in Part V of the Article, other safeguards that protect data subjects and are better suited to EHR-based research should replace the informed consent framework.

A. HUMAN EXPERIMENTATION VS. RECORD-BASED STUDIES

Informed consent undoubtedly has taken root as a normative component of medical research. But, examining the origins of the doctrine reveals that, historically, the underlying concern was largely protecting subjects against abusive experimental interventions rather than against unwanted observational studies.

A commitment to informed consent in research emerged from the ruins of World War II, during which Nazi doctors conducted brutal experiments on prisoners.[185] The importance of informed consent was initially recognized in the Nuremberg Code, the first major international document to provide guidelines on research ethics.[186] The Nuremberg Code opens by stating that "[t]he voluntary consent of the human subject is absolutely essential."[187] The provision goes on to discuss the need to inform each subject of "the nature, duration, and purpose of the experiment" and of "the effects upon his health or person which may possibly

183. Donald J. Willison et al., *Alternatives to Project-Specific Consent for Access to Personal Information for Health Research: Insights from a Public Dialogue*, 9 BMC MED. ETHICS 18, 27 (2008).

184. Hoffman & Podgurski, *supra* note 2, at 118; Port, *supra* note 87, at S-5.

185. Sharona Hoffman, *supra* note 87, at 471.

186. *Id.*

187. NAT'L INSTS. OF HEALTH, NUREMBERG CODE ¶ 1 (1949), *available at* http://ohsr.od.nih.gov/guidelines/nuremberg.html.

come from his participation in the experiment."[188] The studies contemplated by the Nuremberg Code, therefore, involve physical interventions that affect the body, such as the testing perpetrated by the Nazis, rather than the database queries at issue in this Article.[189]

A second international document that embodies research ethics guidance, the Declaration of Helsinki, was adopted in 1964 and has been revised multiple times since.[190] Several provisions of the Declaration detail informed consent requirements,[191] though the consent mandate applies only to personally identifiable medical data or biological material.[192] Furthermore, the Declaration of Helsinki recognizes that "[t]here may be situations where consent would be impossible or impractical to obtain for such research or would pose a threat to the validity of the research. In such situations the research may be done only after consideration and approval of a research ethics committee."[193] Under the Declaration of Helsinki, research utilizing de-identified data would not require consent, and further exceptions could be made for use of individually identifiable data in appropriate circumstances.[194]

In the United States, the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research issued the *Belmont Report* in 1979.[195] This project was undertaken in the wake of the infamous Tuskegee syphilis trial. The trial took place from 1932 until 1972 and involved 600 African-American men, 399 of whom had syphilis.[196] In the course of the study, researchers withheld penicillin from the subjects after it was proven to be effective in treating syphilis because they wanted to learn about the natural course of the disease.[197] The *Belmont Report* identified "respect for persons" as one of three foundational principles for ethical research and demands that investigators obtain informed and voluntary consent from all human subjects.[198] Specifically, the *Belmont Report* states: "Respect for persons requires that subjects, to the degree that they are capable, be given the opportunity to choose what shall or shall not happen to them."[199] This wording and the historical backdrop of the *Belmont Report* suggest that its primary concern is clinical experimentation rather than the collection of data from

188. *Id.*
189. *See id.* ¶ 2 (providing for human consent in an "experiment," not data collection).
190. Hoffman, *supra* note 87, at 474.
191. WORLD MED. ASS'N DECLARATION OF HELSINKI: ETHICAL PRINCIPLES FOR MED. RES. INVOLVING HUMAN SUBJECTS §§ 24–29, *available at* http://www.wma.net/en/30publications/10policies/b3/17c.pdf, [hereinafter DECLARATION OF HELSINKI] (amended 2008).
192. *Id.* § 25.
193. *Id.*
194. *See id.* § 1.
195. BELMONT REPORT, *supra* note 10; Hoffman, *supra* note 87, at 472–73.
196. CTRS. FOR DISEASE CONTROL & PREVENTION, U.S. PUBLIC HEALTH SERVICE SYPHILIS STUDY AT TUSKEGEE (2011), *available at* http://www.cdc.gov/tuskegee/timeline.htm.
197. *Id.*
198. BELMONT REPORT, *supra* note 10, at Parts B.1, C.1.
199. *Id.* at C.1.

existing records for observational studies.[200]

B. The Absence of a Constitutional Right to Control Medical Records

Federal regulations that allow record-based research without consent would likely not violate any constitutional rights.[201] The Supreme Court has not found that patients have either a property right or a privacy right associated with their medical records.[202]

The question of health data ownership is complicated and lacks a clear answer. Medical records are generally considered to be the property of the physicians and hospitals that create them rather than the property of patients.[203] Several state statutes and judicial decisions acknowledge that healthcare providers own their records.[204] However, the property status of a patient's health data, as opposed to any physical or electronic records containing such data, is far more ambiguous.[205]

Recently, several scholars have posited that patients should not enjoy an absolute ownership right to their health information. For example, Professor Marc Rodwin argued against "treating patient data as private property [because it] precludes forming comprehensive databases required for many of . . . [the] most important public health and safety uses."[206] He proposed that clinicians, hospitals, and insurers be required by federal law to report de-identified patient data to public authorities who would create aggregate databases that researchers could utilize.[207] Rodwin believes that patient data should be treated as public property rather than private property.[208] Similarly, Professor Barbara Evans calls

---

200. The second principle articulated in the *Belmont Report* is beneficence, which encompasses the mandates to "do no harm" and to "maximize potential benefits" while minimizing risks in research. *Id.* at B.2. The third principle is justice, which requires that the benefits and risks of research be distributed fairly and that selection procedures for human subjects be sound and impartial. *Id.* at B.3. *See* discussion *infra* Part V.B (discussing further the concepts of beneficence and justice).

201. Rodwin, *supra* note 41, at 609.

202. *See* Evans, *supra* note 107, at 72–73 (noting ownership is left to state law and state courts have issued inconsistent holdings); *see also* Rodwin, *supra* note 41, at 588–89.

203. *See* Rodwin, *supra* note 41, at 587–88; Marc A. Rodwin, *The Case for Public Ownership of Patient Data*, 302 JAMA 86, 87 (2009); *see also* Mark A. Hall, *Property, Privacy, and the Pursuit of Interconnected Electronic Medical Records*, 95 Iowa L. Rev. 631, 642 (2010).

204. *See* Fla. Stat. Ann. § 456.057(1) (West 2007); Miss. Code. Ann. § 41-9-65 (West 2007); S.C. Code Ann. § 44-115-20 (2002); Tenn. Code Ann. § 68-11-304(a)(1) (2011); Va. Code Ann. § 32.1-127.1:03A (West 2011); Young v. Murphy, 90 F.3d 1225, 1236 (7th Cir. 1996); Holtkamp Trucking Co. v. Fletcher, 932 N.E.2d 34, 44 (Ill. App. Ct. 2010); Estate of Finkle, 395 N.Y.S.2d 343, 344, 552 (N.Y. Surr. Ct. 1977). *But see* Person v. Farmers Ins. Grp. of Cos., 61 Cal. Rptr. 2d 30, 31 (Cal. Dist. Ct. App. 1997) (finding that health records belong to the patient).

205. Hall, *supra* note 203, at 642; Rodwin, *supra* note 41, at 588; Evans, *supra* note 107, at 72–74. *But see* N.H. Rev. Stat. Ann. § 151:21 (LexisNexis 2005) ("Medical information contained in the medical records at any facility licensed under this chapter shall be deemed to be the property of the patient.").

206. Rodwin, *supra* note 41, at 589.

207. *Id.*

208. *Id.* at 590.

for a debate about "appropriate public uses of private data and how best to facilitate these uses while adequately protecting individuals' interests."[209]

Furthermore, the Supreme Court has never recognized a constitutional right to informational privacy.[210] In a 2011 case, *NASA v. Nelson*, the Supreme Court noted that the lower courts have issued inconsistent rulings concerning this purported right.[211] The Court explicitly declined to determine whether a right to informational privacy exists[212] and determined that if it did, the government's inquiries during employment background checks would not violate that right.[213]

C. Patients' Preferences Regarding Consent

According to the IOM, public opinion polls show that "a significant portion of the public would prefer to control all access to their medical records via informed consent."[214] At the same time, empirical data suggests that a majority of Americans are supportive of medical research and recognize its benefits.[215]

Several empirical studies sought to determine patient preferences as to whether they should be asked to consent to research studies that will involve only an examination of their medical files.[216] Although the results are inconclusive, a review of a few of them can be illuminating.

Two studies, one from the United States, and one from Canada, found that patients prefer to be asked for consent and often do not distinguish between identifiable and de-identified data for purposes of their responses.[217] The U.S. study, conducted through telephone interviews of 1,193 patients, focused on research using samples of genetic material.[218] It found that 81% of respondents wanted to know about research if their samples would be identifiable, and 72% wished to be informed if the samples would be anonymous.[219] Of those wanting to know about research involving either identifiable or anonymized samples, 57% would require that their permission be sought, and 43% would be content with notifica-

209. Evans, *supra* note 107, at 77.
210. NASA v. Nelson, 131 S. Ct. 746, 748 (2011).
211. *Id.* at 756–57.
212. *Id.* at 757.
213. *Id.* at 763–64.
214. IOM Report, *supra* note 9, at 268.
215. *Id.* at 119.
216. *Id.* at 81–86.
217. Sara Chandros Hull et al., *Patients' Views on Identifiability of Samples and Informed Consent for Genetic Research*, 8 Am. J. Bioethics 62, 69 (2008) (finding that most patients surveyed did not differentiate between identifiable and non-identifiable genetic samples and questioning whether the regulatory distinction is useful); Donald J. Willison et al., *Patient Consent Preferences for Research Uses of Information in Electronic Medical Records: Interview and Survey Data*, 326 BMJ 373, 375 (2003) (noting a "lack of distinction between identifiable and anonymised information" in the minds of survey participants). The Canadian study involved 123 patients, seventeen of whom were interviewed, while the remainder completed surveys. Willison et al., *supra* at 373–74.
218. Hull et al., *supra* note 217, at 64.
219. *Id.* at 65.

tion alone.[220]

The Canadian study asked 1,230 adults for their reaction to (1) use of data directly from their medical files, and (2) automated abstraction of data from their EHRs with assurances that direct identifiers would not be collected.[221] With respect to use of data directly from medical records, 60% of respondents felt that their permission should be obtained, though only half of those wished for project-by-project consent rather than general consent.[222] Twenty-four percent indicated they would be satisfied with notification alone, and 12% believed that neither notification nor permission was needed.[223] With respect to automated abstraction, 27%, as opposed to 12%, were comfortable with use of information without permission or notification.[224] The study concluded that the majority of patients "wished to maintain some level of control over the use of their information."[225] It is noteworthy, however, that 68% agreed to some degree with the statement: "Research that could be beneficial to people's health is more important than protecting people's privacy."[226]

By contrast, a British study concluded that a majority of patients were willing to share their data without being asked for consent when no identifiers would be disclosed to parties other than their treating physicians.[227] This study examined responses from 166 patients who recently had been discharged from a hospital.[228] The questionnaire clearly stated that doctors, rather than other parties, would access the data in patient records and would use it in anonymous form.[229] It also specified the purposes for which the information would be used, including clinical audits, research, training, comparison of treatment outcomes in different hospitals, and publications about diseases in medical journals.[230] Only 13% of patients questioned indicated that they would definitely want to be asked for permission to use their medical records.[231] Assurances about anonymity, restriction of access to doctors alone, and the constructive purposes for which the data would be used may account for the high degree of patient willingness to share information without burdening physicians with consent requirements.

The disparate results make it difficult to draw definitive conclusions from studies concerning patient preferences and attitudes. The discrep-

220. *Id.* at 66.
221. Donald J. Willison et al., *Alternatives to Project-Specific Consent for Access to Personal Information for Health Research: What Is the Opinion of the Canadian Public?*, 14 J. AM. MED. INFORMATICS ASS'N 706, 707 (2007).
222. *Id.* at 708.
223. *Id.*
224. *Id.* at 709.
225. *Id.* at 710.
226. *Id.* at 708.
227. Bruce Campbell, *Extracting Information from Hospital Records: What Patients Think About Consent*, 16 QUALITY & SAFETY HEALTH CARE 404, 407 (2007).
228. *Id.* at 404
229. *Id.* at 405.
230. *Id.* at 405–06.
231. *Id.* at 406.

ancies may stem from the phrasing of questions in the different studies[232] and from variation between the populations of participants.[233] The studies also reveal some degree of confusion and ambivalence on the part of patients.[234] However, the studies' outcomes suggest that, with further education about the benefits of comprehensive data collection for research and about the safeguards implemented to protect privacy, patients may become increasingly willing to prioritize medical advances (from which they too can benefit) over concerns about risks in the record-based research context.[235]

D. The Trouble with Consent

While consent requirements promote patient autonomy and may be favored by patients, they can also interfere with the scientific integrity of the research enterprise. Consent requirements can result in selection bias that can actually invalidate research outcomes.[236] In addition, contacting thousands or millions of patients who are included in a database can be a very expensive and time-consuming undertaking for researchers and might make it impossible for many studies to proceed.[237]

1. Informed Consent Can Lead to Selection Bias

One major difficulty with informed consent is that it leads to selection bias, which can skew research results.[238] This section argues against routinely granting data subjects a choice concerning inclusion of their records in research because of the unacceptable risk of selection bias.

a. Selection Bias vs. Confounding

Selection biases result from procedures used to select subjects and from other factors that affect study participation.[239] The term "selection bias" is used to describe subtly different kinds of study biases.[240] By one definition, selection bias occurs when those who decide to consent to participate in research constitute a subset of individuals who are not representative of the patient population of interest.[241] This could happen if a disproportionate number of people of one ancestry or economic class opt out of a study. It can likewise happen if individuals with certain behavior

232. *See* IOM Report, *supra* note 9, at 79 ("[H]ow the questions and responses are worded and framed can significantly influence the results and their interpretation.").
233. *Id.* at 70.
234. *Id.* at 201.
235. *See* Miller *supra* note 12, at 565 ("[P]ublic education is important to explain the rationale for access to medical records for research without consent and the safeguards in place to protect private information from being misused.").
236. IOM Report, *supra* note 9, at 201; Miller, *supra* note 12, at 560.
237. Cate, *supra* note 12, at 1789–93.
238. IOM Report, *supra* note 9, at 201; Miller, *supra* note 12, at 560.
239. Kenneth J. Rothman et al., Modern Epidemiology 136 (3d ed. 2008).
240. Miguel A. Hernán et al., *A Structural Approach to Selection Bias*, 15 Epidemiology 615, 615 (2004).
241. IOM Report, *supra* note 9, at 209; Miller, *supra* note 12, at 560.

traits that might be pertinent to a study—such as diet, smoking habits, alcohol or drug consumption, and exercise—disproportionately opt out.

If the process of obtaining patients' informed consent to participate in a research study is subject to this kind of selection bias, then the consenting patients will not comprise a representative sample of the population targeted for study.[242] Consequently, using results from the study population to estimate measures of interest, such as disease prevalence or average treatment effect, will tend to yield estimates that differ systematically from the true values of these measures for the target population.[243] That is, the estimates will not generalize from the set of consenters to the target population.

However, in one type of medical research, known as causal effect studies, accurately estimating population statistics is often not the primary concern.[244] These studies typically assess whether a certain treatment has a beneficial causal effect on patients with a particular condition or whether a certain exposure has a harmful causal effect on individuals.[245] In such a study, use of a representative sample of subjects from a broad population may actually threaten the study's internal validity, due to variations in factors other than the treatment or exposure and the outcome (e.g., genetic abnormalities).[246] Thus, researchers may seek a group of study subjects that is relatively homogeneous, except that some are treated or exposed and others are not.[247] Once the nature and magnitude of a causal effect is established using such a group, researchers may seek to generalize the results to a more diverse population either by reasoning from existing knowledge and theory or by conducting an empirical study with a sample of subjects that is representative of the population.[248] For example, although the causal link between smoking and lung cancer was established mainly through studies of men, the link was assumed by experts to exist in women also, based on the physiological similarity between the lungs of women and men.[249]

In causal effect studies, researchers may consider confounding bias (confounding) to be a greater threat than selection bias to the validity of causal effect estimates.[250] "Classical" confounding occurs when the values of certain variables, called confounders, influence both whether individuals receive a treatment or exposure under study and whether they exhibit the outcomes of interest.[251] For example, doctors' concerns about

242. IOM REPORT, *supra* note 9, at 209; Miller, *supra* note 12, at 560.
243. IOM REPORT, *supra* note 9, at 209–11.
244. Miguel A. Hernán, *A Definition of Causal Effect for Epidemiological Research*, 58 J. EPIDEMIOLOGY & CMTY. HEALTH 265, 265 (2004).
245. *Id.*
246. ROTHMAN ET AL., *supra* note 239, at 146–47.
247. *Id.*
248. *Id.*
249. *Id.* at 147.
250. Sander Greenland, *Quantifying Biases in Casual Models: Classical Confounding vs Collider-Stratification Bias*, 14 EPIDEMIOLOGY 300, 306 (2003).
251. *See id.*

side effects of a new treatment may influence them to favor it for younger, more robust patients who are likely to have better outcomes than older, more frail patients.[252] Such a practice would result in confounding because it would make the new treatment appear far more effective on average than it really is.[253] Elderly, feeble patients who may not do well with any therapy, including the one at issue, are unlikely to receive the treatment in question.[254] If they did take the study drug and their poor outcomes were to be considered, the study drug would likely appear less successful.[255]

In an observational study, if all potential confounding variables are known and are accurately measured, adjustments can be made during statistical analysis of the results that reduce or eliminate confounding bias.[256] Randomized treatment assignment, when feasible, tends to prevent confounding because randomization helps to ensure that the subjects in the treatment and control groups are similar with respect to the values of potential confounding variables, even unknown ones.[257] On the other hand, lack of generalizability to actual patient populations is a recognized limitation of many randomized trials, which EHR-based observational research is meant to address.[258] Moreover, noncompliance and loss to follow-up may cause substantial confounding and selection bias even in randomized trials.[259]

Informed consent itself cannot be a confounding variable in a causal effect study[260] because only patients who consent to participate will be included in the study. That is, consent status is fixed and not a variable at all among the participants. Therefore, one might think that seeking informed consent from subjects and allowing them to decline to participate is not problematic for causal effect studies. However, while informed consent will not cause confounding, it can still produce a type of selection bias that makes it difficult to determine whether a certain treatment or exposure has a causal effect on patients.[261]

The selection bias at issue, also called "collider bias,"[262] is one that involves selection based on a common causal effect of two factors.[263] Like confounding, this bias can cause a group of subjects who received a treatment or exposure to differ from the control group, which did not

252. *See* ROTHMAN ET AL., *supra* note 239, at 488.
253. *See id.*
254. *See id.*
255. *See id.*
256. *Id.* at 58.
257. *Id.* at 88–89 (discussing randomization).
258. Stewart et al., *supra* note 79, at w181. For additional discussion of observational trials, see *supra* Part III.
259. ROTHMAN ET AL., *supra* note 239, at 202.
260. *See supra* notes 244–50, 260–64 and accompanying text (discussing causal effect studies).
261. ROTHMAN ET AL., *supra* note 239, at 136.
262. *Id.* at 185.
263. *Id.* at 136–37 (distinguishing confounding from selection bias); Hernán, *supra* note 244, at 267.

receive it, in ways that seriously distort causal effect estimation.[264] As we will illustrate, this kind of selection bias could arise in an EHR-based study if patients' decisions about permitting research use of their EHRs are influenced by two factors, one of which also influences the treatment or exposure variable later studied and the other of which influences the outcome variable.

Consider, for example, a retrospective EHR-based cohort study undertaken to determine if taking a certain heavily advertised diet medication increases a person's risk of heart attack. Suppose that, among the public, both the probability of individuals using the medication and the probability of them consenting to research uses of their EHRs increase with television viewing, due to advertising and other favorable publicity. Suppose also that chronic stress, though it is not considered in the study, increases individuals' risk of heart attack but decreases the likelihood that they will consent to research uses of their EHRs. Assume that, for these reasons, use of the diet medication among the public is positively correlated with television viewing and with consent, but it is negatively correlated with chronic stress. Thus, non-use of the medication is negatively correlated with television viewing and with consent, but it is positively correlated with chronic stress. Note that these are statistical associations, not causal relationships; neither using the diet medication nor avoiding it should be assumed to cause or prevent television viewing, consent, or chronic stress. Finally, assume there is no one factor that is a common cause of both using, or not using, the diet medication and of having, or not having, a heart attack. The causal influences in this hypothetical scenario are illustrated by the causal diagram in Figure 1.

All subjects in the study cohort must have consented to use of their EHRs in research. Due to the aforementioned correlations, consenters who took the diet medication were more likely to suffer chronic stress than consenters who did not take the medication. The two causes of consent are television viewing and absence of chronic stress. Consenters who did not take the medication were less likely to watch television and hence more likely to be free of chronic stress. Assume that the diet medication does *not* increase the risk of heart attacks. The investigators may erroneously come to the opposite conclusion when they compare the outcomes of the subjects who used the medication to the outcomes of the subjects who did not use it, because, unknown to the researchers, the users suffered more heart attacks due to chronic stress.

Observe that in this hypothetical scenario consent status is causally influenced (positively) by television viewing, which is also a cause of using the diet medication, and is causally influenced (negatively) by chronic stress, which is also a cause of heart attacks. This led to selection bias that falsely indicated that the medication caused heart attacks. This hypothetical scenario illustrates how subject selection influenced by in-

264. ROTHMAN ET AL., *supra* note 239, at 186.

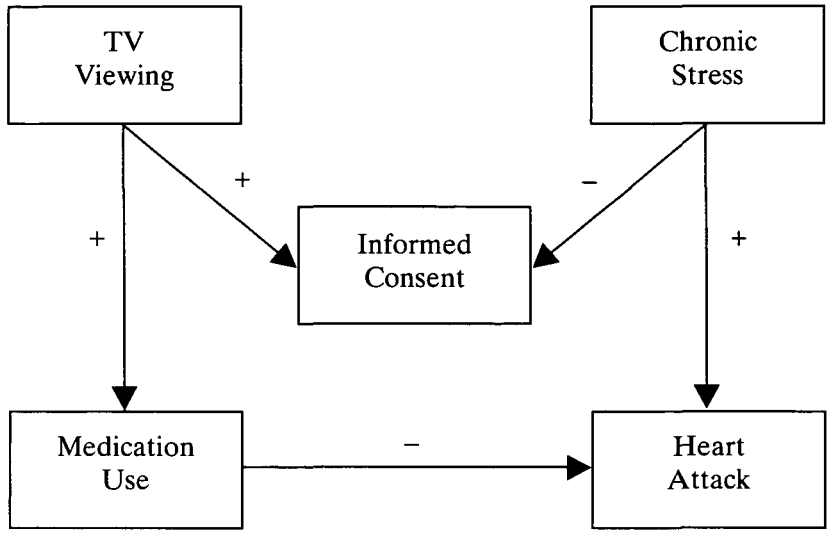

Figure 1: Diagram indicating causal influences among variables in example scenario illustrating selection bias due to informed consent. Plus sign indicates positive influence; minus sign indicates negative influence.

formed consent can distort a causal effect estimate because of collider bias.

b. Selection Bias Is Confirmed by Empirical Evidence

Several studies confirm that selection bias is not merely a theoretical problem. For instance, one study focused on the Registry of the Canadian Stroke Network, which includes twenty Canadian hospitals.[265] Nurse coordinators obtained consent from approximately 3,100 patients, and the reasons for non-consent were most often inability to contact the patient rather than explicit refusal.[266] The authors found major selection biases because of the consent requirement. Specifically, "the in-hospital mortality rate among the enrolled patients was only 6.9%, which is much lower than the true mortality rate among all patients with stroke in Canada."[267] This skewing occurred because nurse coordinators had difficulty obtaining consent from grieving or very distressed families of patients who had died or were critically ill.[268] In addition, many patients could not provide consent because of impairments resulting from their strokes, and no surrogates were available.[269] Thus, usually, only the healthiest patients with the best prognosis provided consent.

In a different research project, 876 Irish patients with ischaemic heart disease returned questionnaires that included a request for consent to

265. Jack Tu et al., *Impracticability of Informed Consent in the Registry of the Canadian Stroke Network*, 350 NEW ENG. J. MED. 1414, 1415 (2004).
266. *Id.* at 1416–17.
267. *Id.* at 1419.
268. *Id.*
269. *Id.*

participate in further research.[270] Of these, 574, or 65.5%, signed the consent form and agreed to participate in the future.[271] Analysis of these patients' records revealed that their willingness to be involved in further research correlated with four distinctive predictors: (1) a prior surgical cardiac intervention, (2) lower blood pressure measurements, (3) lower cholesterol levels, and (4) being an ex-smoker.[272] The investigators found clear indications of selection bias and concluded that if consent is required, study populations may consist disproportionately of individuals "who have made healthy lifestyle decisions, who have previously benefited from healthcare or those whose clinical risk factors are already well managed."[273]

A review of literature about selection bias, however, concluded that no clear factors, such as age, sex, socio-economic status, or medical history, emerged as consistently predictive of which patients would agree or decline to participate in studies.[274] Therefore, future studies cannot easily control for specific factors to combat the problem of selection bias.

An IOM Report, *Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research*, discusses several additional studies of selection bias.[275] The IOM concluded that the HIPAA Privacy Rule's requirement of patient authorization for use of identifiable health information generates biased study samples and jeopardizes the validity of research outcomes.[276]

2. *Obtaining Informed Consent Can Be Costly and Burdensome*

In addition to generating selection bias, consent requirements can be very expensive and work-intensive for investigators. Therefore, they can significantly hinder research projects or even make them impossible to pursue.

a. Consent Options

Consent for research drawing upon EHR databases could be sought in a variety of ways. Each mechanism, however, has its own shortcomings and risks.

First, data subjects could be asked to consent generally to use of their records in observational studies. Thus, subjects would be asked to pro-

---

270. Brian Buckley et al., *Selection Bias Resulting from the Requirement for Prior Consent in Observational Research: A Community Cohort of People with Ischaemic Heart Disease*, 93 HEART 1116, 1117–18 (2007).

271. *Id.* at 1118.

272. *Id.*

273. *Id.* at 1119.

274. Michelle Kho et al., *Written Informed Consent and Selection Bias in Observational Studies Using Medical Records: Systematic Review*, 338 BMJ b866, b873 (2009) (reviewing seventeen published studies and finding "authorisation bias in studies requiring informed consent for use of data from medical records").

275. IOM REPORT, *supra* note 9, at 209–12.

276. *Id.* at 216.

vide broad consent for all future, unspecified studies.[277] This would be the least burdensome option for investigators but could nevertheless introduce significant selection bias.[278] For example, research concerning psychiatric conditions or HIV might be obstructed because individuals with these conditions are particularly worried about privacy leaks and potential stigmatization, and, thus, disproportionately refuse to allow their records to be included in databases. In addition, because the nature of future research projects is unknown, it is arguable that subjects who are asked for consent on a one-time basis cannot realistically make an informed, meaningful choice.[279]

In the alternative, to maximize data subject autonomy, investigators could obtain consent for each separate research project from all data subjects.[280] Such a requirement, however, would be unworkable. The databases of de-identified EHRs or federated systems that we envision would include millions of records.[281] If investigators had to re-contact every data subject for permission before conducting each study, many research projects could be too costly or time-consuming to pursue.[282] Research staff might spend more time seeking permission from patients than actually conducting research to improve health outcomes.

To save time and money, investigators might consider automating consent so that subjects would receive electronic messages about proposed studies and would be asked to respond electronically to indicate their agreement or refusal to have their records included. The response rate to e-mail solicitations, however, is likely to be unsatisfactory, and many may feel that such an impersonal approach deprives patients of the opportunity to provide truly meaningful consent.[283] Numerous commentators argue that even more formal and extensive informed consent procedures are deeply flawed and that subjects often make decisions about participation without sufficient information or comprehension of the data they are given.[284] If research enrollment requests come as one of dozens of e-mails that individuals receive each day and if patients are not alerted to the importance of their decision through more personal contact, they may well default to ignoring such messages or clicking on a box without giving the matter significant thought.

However, other methods of contacting subjects, such as by mail or telephone, may endanger privacy and exacerbate selection bias. Patients

277. *See* Kosseim & Brady, *supra* note 146, at 22–26; Willison et al., *supra* note 183, at 19.

278. *See supra* Part IV.D.1.a.

279. E. Vermeulen, *A Trial of Consent Procedures for Future Research with Clinically Derived Biological Samples*, 101 BRIT. J. CANCER 1505, 1505 (2009) ("Still others argue that informing patients about future research with tissue is impossible, even in basic terms, and that consent cannot be truly 'informed'").

280. Kosseim & Brady, *supra* note 146, at 20–22; Willison et al., *supra* note 183, at 19.

281. Kosseim & Brady, *supra* note 146, at 9 n.11.

282. *Id.* at 25.

283. *Id.* at 20, 24.

284. *See* Hoffman, *supra* note 87, at 484–87.

would need to be re-identified each time consent is sought for individual studies, and their identities would be linked to their records, which would be included or excluded according to their preferences.[285] Whoever handles the consent communication would therefore be able to scrutinize identifiable medical data, and the data could be subject to eavesdropping by hackers or other intruders.[286] In addition, patients who know the precise nature of each project for which they are asked to allow use of their records may selectively deny permission based on their feelings about the study or how relevant they believe it is to their own health problems.[287] Moreover, a process by which patients are frequently contacted by investigators, asked for consent, and reminded of the risks of inclusion, may make patients needlessly anxious about research participation and encourage them to refuse to allow inclusion of their records.[288]

Several middle-ground options exist as well. For example, data subjects might be permitted to describe particular categories of studies from which they want their data to be excluded, and their choices would be included in their EHRs.[289] To illustrate, they could indicate that they do not want their data used in studies concerning genetic abnormalities or psychological illnesses. Combing through all data subjects' records to determine their preferences, however, would be a very work-intensive task for researchers unless the function could be fully automated.[290] This option may also create significant selection bias.[291] Large numbers of individuals may decline to participate because they have the condition at issue and fear being identified, but these are precisely the individuals whose records might be most valuable. Similarly, individuals may disproportionately withhold their records because of specific political, cultural, or other beliefs, and their absence from the study population may skew research results.[292]

Alternatively, patients could describe outcomes they wish to avoid by stating that they wish to be excluded from studies that might promote abortion or result in commercial profits for pharmaceutical companies.[293] However, it will likely be impossible for researchers to predict which studies will ultimately lead to particular outcomes that are objectionable to specific individuals.[294] For example, it may be difficult to determine in advance whether a research project will ultimately lead to a genetic dis-

285. IOM REPORT, *supra* note 9, at 252.

286. *Id.* at 103.

287. *Id.* at 251–52.

288. *Id.*

289. *Id.* at 102; Mark A. Rothstein, *Debate of Patient Privacy Control in Electronic Health Records*, BIOETHICS FORUM (Feb. 17, 2011, 10:09 AM), www.thehastingscenter.org/bioethicsforum/post.aspx?id=5139.

290. Rothstein, *supra* note 289.

291. *See* IOM REPORT, *supra* note 9, at 209.

292. *See supra* Part IV.D.1 (discussing selection bias).

293. *See* Willison et al., *supra* note 221, at 707.

294. *Id.* at 711.

covery that could cause some women to abort fetuses because of genetic abnormalities that became detectable.

Yet another approach would be to allow subjects to refuse disclosure of certain categories of information.[295] These might include sensitive information such as psychiatric conditions, HIV status, or sexual history.[296] Thus, data subjects would agree to have all but the designated parts of their records accessible to researchers.[297] But sequestering such data would surely compromise the integrity of studies in some instances.[298] Details of medical history, such as HIV status, psychiatric conditions, and reproductive problems, may well be relevant to the outcomes of various biomedical studies or to deciding whether a subject's records should be included in the first place.[299] Without these details, the other data contained in an EHR may at times be essentially meaningless.

Arguably, the least damaging alternative would be presumed consent with an opt-out opportunity. Records would be available to researchers as a default unless the data subject specifically requested that her record be excluded.[300] In addition, opting out could be made difficult so that only those who are truly committed to having their records excluded pursue it. To illustrate, rather than checking a box, individuals may be required to write out a request and may be asked to renew their opt-out indication annually so that they revisit their decisions. Although this approach is more appealing than those described above, it still raises concerns about data integrity. Supplying an opt-out choice could be quite burdensome for researchers if data subjects were to be given the option for each separate study. Even a general opt-out choice that covered all EHR studies could lead to selection bias.[301] When Iceland adopted a presumed consent and opt-out approach for inclusion of records in the country's Health Sector Database, at least 7% of the population, or 20,200 individuals, opted out.[302] In the United States, if CER is enthusiastically promoted by government authorities,[303] it is possible that political opponents, media personalities with political agendas, and others who are suspicious of government initiatives will encourage large numbers of followers to opt out, thus diminishing the quality of research databases.

295. Rothstein, *supra* note 289.

296. *Id.*; *see also* Nicolas P. Terry & Leslie P. Francis, *Ensuring the Privacy and Confidentiality of Electronic Health Records*, 2007 U. ILL. L. REV. 681, 702 (2007). These articles discuss the option of allowing patients to sequester data in their EHRs in the clinical setting so that it will be hidden from other clinicians; the papers do not address research questions specifically. *See generally* Terry & Francis, *supra*; Rothstein, *supra* note 289.

297. Terry & Francis, *supra* note 296, at 702; Rothstein, *supra* note 289.

298. *See* Rothstein, *supra* note 289.

299. *Id.*

300. Willison et al., *supra* note 183, at 19.

301. *Id.* at 23.

302. Jamaica Potts, *At Least Give the Natives Glass Beads: An Examination of the Bargain Made Between Iceland and DeCODE Genetics with Implications for Global Bioprospecting*, 7 VA. J.L. & TECH. 8, 51–52 (2002); David E.Winickoff, *Genome and Nation: Iceland's Health Sector Database and Its Legacy*, 1 INNOVATION 80, 90 (2006).

303. *See supra* note 83 and accompanying text (discussing the Obama Administration's support for CER).

b. Empirical Evidence Concerning the Cost of Consent Mandates

Empirical data supports the contention that consent requirements are associated with significant costs. A 2007 survey of 1,527 epidemiologists found that the HIPAA Privacy Rule's authorization requirements had significantly hindered research.[304] Respondents expressed frustration with the cost and delays associated with regulatory compliance.[305] Other studies reveal similar objections and even suggest that some health care providers are opting out of conducting research altogether.[306]

Several studies have attempted to quantify the cost and time demands of consent processes. The study of the Registry of the Canadian Stroke Network, discussed above, concluded that nurse coordinators spent a median of forty minutes with each patient or surrogate for consent purposes, including the time spent arranging interviews.[307] In addition, of the 2 million Canadian dollars spent on the registry during the first two years, $500,000 was spent on consent activities alone.[308] A British study estimated that the cost of obtaining consent through a combination of e-mail, mail, and telephone calls for review of records of prostate cancer patients was $248 for each man who consented.[309] In a U.S. study, 2,228 mothers who were likely to deliver preterm infants were approached in person for consent to a study of neonatal care. Consent was found to take between 1,735 and 2,790 hours and to cost between $65,945 and $106,029, depending on staff salaries.[310] Yet another study focused on parental consent to the participation of 2,496 middle-school-aged children in a survey.[311] Consent, involving three mailings and follow-up telephone calls to non-responders, was estimated to cost at least $50,000.[312]

It is difficult to determine how these figures would translate into a cost estimate for obtaining consent, on either a one-time basis or a case-by-case basis, from potential EHR research subjects. It is clear, however, that an effort to contact and obtain consent from all or most U.S. patients for purposes of creating and using national databases would be extremely expensive.

---

304. Roberta B. Ness, *Influence of the HIPAA Privacy Rule on Health Research*, 298 JAMA 2164, 2167 (2007).

305. *Id.*

306. IOM REPORT, *supra* note 9, at 199–209.

307. Tu et al., *supra* note 265, at 1418.

308. *Id.*

309. Sian Noble et al., *Feasibility and Cost of Obtaining Informed Consent for Essential Review of Medical Records in Large-Scale Health Services Research*, 14 J. HEALTH SERVICES RES. & POL'Y 77, 79–80 (2009). Of the 230 individuals who were sent consent forms, 179 consented. *Id.*

310. Wade D. Rich et al., *Antenatal Consent in the SUPPORT Trial: Challenges, Costs, and Representative Enrollment*, 126 PEDIATRICS e215, e217–18 (2010).

311. Finn-Aage Esbensen et al., *Differential Attrition Rates and Active Parental Consent*, 23 EVALUATION REV. 316, 320, 322, & 329 (1999).

312. *Id.*

V. RECONSTRUCTING THE CONCEPTUAL FRAMEWORK

In the words of the IOM, "[t]he principle of autonomy currently dominates the ethical landscape for both medical care and clinical research in the United States."[313] The current, consent-centered ethical framework is based on the assumption that research will involve human experimentation, and is firmly rooted in a history of shocking research abuses.[314] As recently as 2011, Wellesley College Professor Susan Reverby discovered evidence of previously unknown human subject exploitation.[315] From 1946 to 1948, American researchers deliberately infected Guatemalan prison inmates, mental patients, and soldiers with venereal diseases to test the efficacy of penicillin.[316] All instances of serious research abuse, however, have occurred in the context of interventional studies.[317] With respect to record-based research, and in light of the great promise of EHR research databases, it is appropriate to shift the discussion from autonomy to a new focus on the goal of promoting the common good.

We wish to emphasize that we address only research that involves record review without clinical testing. One might be concerned that focusing on the common good will lead to a slippery slope and, eventually, to rationalizing away informed consent altogether. In the research context, however, it is easy to draw a bright-line distinction between interventional and record-based studies. In the case of interventional studies, concerns about harm to human subjects should not be subordinated to the goal of promoting social benefits, and consent should not be abandoned. The same is not true, however, for noninterventional research so long as all studies are subject to stringent oversight and privacy protections.[318]

A. THE IMPORTANCE OF THE COMMON GOOD

The traditional concepts of informed consent center upon the individual rights of research subjects because the research contemplated is generally physically or psychologically invasive. With the advent of large EHR databases and the proliferation of research studies that involve only record review, it is appropriate to turn to the value of the common good as a counterweight to concern about individual risk. When human beings are not subject to any physical or psychological testing in research, and only their records are scrutinized, the value of the common good should prevail over individual interests. Society's interests in achieving medical advances should outweigh the individual risks of privacy breaches and

313. IOM REPORT, *supra* note 9, at 247.
314. *See supra* Parts III.A, IV.A.
315. Donald G. McNeil Jr., *U.S. Infected Guatemalans with Syphilis in '40s*, N.Y. TIMES, Oct. 2, 2010, at A1.
316. *Id.*
317. *See* David J. Rothman, *Ethics and Human Experimentation: Henry Beecher Revisited*, 317 NEW ENG. J. MED. 1195, 1195–96 (1987).
318. *See* discussion *infra* Part VI.B.

non-privacy-related dignitary injuries[319] when all reasonable efforts are made to prevent such harms.

All patients benefit from medical care improvements that have been made possible by past research studies. Thus, it is arguably irresponsible or inequitable for some patients to prohibit researchers from accessing their data and decline to make their own contribution to the research endeavor.[320] Refusal to participate in research can be characterized as "free riding" because there is no practical way to prevent those who do not contribute their records to research from enjoying the benefits of improved treatment resulting from biomedical studies.[321]

Subordinating individual freedom to the common good because individuals profit from societal initiatives is consistent with the philosopher Jean-Jacques Rousseau's theory of social consent. Rousseau spoke of a social contract by which individuals willingly give up freedom and autonomy to enjoy the advantages of living in society.[322] Individuals who are residents of a political state necessarily accept its benefits, and by doing so, citizens tacitly consent to the laws that enable governmental authority to function.[323] The concept of social consent may be applied to the medical arena as well. Because essentially all individuals will at some time in their lives receive medical care, they may be deemed to tacitly consent to having their EHRs available for research that makes treatment possible.

A few bioethicists have gone as far as to argue that individuals have a moral duty to participate in biomedical research, which extends even to interventional studies.[324] However, one need not take a position regarding whether participation in research rises to the level of a moral duty to argue that it is ethically sound to prohibit patients from withholding their information from EHR databases.

B. The Common Good as Embodied in Beneficence and Justice

The value of the common good has already been incorporated into biomedical ethics through the second and third concepts articulated in the

---

319. *See supra* Part III.B.

320. Miller, *supra* note 12, at 564.

321. *Id.*; *see also* Sarah Chan & John Harris, *Free Riders and Pious Sons—Why Science Research Remains Obligatory*, 23 Bioethics 161, 162–64 (2009); G. Owen Schaefer et al., *The Obligation to Participate in Biomedical Research*, 302 JAMA 67, 68 (2009).

322. Jean-Jacques Rousseau, On the Social Contract 52–53 (Roger D. Masters ed., Judith R. Masters trans., St. Martin's Press 1978).

323. *Id.*; Edward A. Harris, Note, *From Social Contract to Hypothetical Agreement: Consent and the Obligation to Obey the Law*, 92 Colum. L. Rev. 651, 676 (1992).

324. *See, e.g.*, John Harris, *Scientific Research Is a Moral Duty*, 31 J. Med. Ethics 242, 247 (2005); Rosamond Rhodes, *In Defense of the Duty to Participate in Biomedical Research*, 8 Am. J. Bioethics 37, 38 (2008); Rosamond Rhodes, *Rethinking Research Ethics*, 5 Am. J. Bioethics 7, 15 (2005) ("reasonable people should endorse policies that make research participation a social duty"); Schaefer et al., *supra* note 321, at 67. *But see* Stuart Rennie, *Viewing Research Participation as a Moral Obligation: In Whose Interests?*, 41 Hastings Center Rep. 40, 46 (2011) (arguing that the moral status of research participation should not be changed).

Belmont Report.[325] Thus, it is not foreign to the field of research ethics.

Beneficence mandates that researchers do no harm and maximize potential benefits while minimizing research risks.[326] Beneficence most clearly dictates that investigators eschew harming individual research participants.[327] However, the *Belmont Report* also recognizes the importance of societal interests and instructs that the benefits "that may result from the improvement of knowledge and from the development of novel medical, psychotherapeutic, and social procedures" must be considered in determining whether to proceed with research studies.[328] The *Belmont Report*'s explanation of the principle's application states that at times the potential to gain significant societal benefits from research will justify risks to individual human subjects and that the loss of such potential benefits is of serious concern.[329]

The principle of justice requires that the benefits and risks of research be distributed fairly and that selection procedures for human subjects be sound and impartial.[330] This principle prohibits exploitation of vulnerable groups for the benefit of those who are more advantaged.[331] The vulnerable must not bear a disproportionate burden in research initiatives, and those who will benefit must make a fair contribution.[332] EHR-based research is likely to encompass many, if not most medical conditions, and it is impossible to predict in advance what knowledge it will yield over the years and who will benefit from it. Consequently, the principle of justice supports inclusion of all Americans in EHR databases to promote the common good.

C. The Common Good as Applied to the Health Care Industry

The common good principle supports the imposition of certain burdens on patients, namely, depriving them of choice as to whether their EHRs are accessible to researchers. At the same time, the common good requires concessions from health care providers who create EHRs, including physicians, clinics, hospitals, and others. Despite having ownership claims to medical files,[333] health care entities must make their records available to researchers to facilitate treatment improvements.[334] In addition, providers should not be able to charge excessive fees for access to the records they control. As discussed above, to be useful, a research

325. See *supra* notes 195–99 and accompanying text for prior discussion of the *Belmont Report*.
326. *See* Belmont Report, *supra* note 10, at B.2.
327. *See id.*
328. *Id.*
329. *Id.* at C.2.
330. *Id.* at B.3.
331. *Id.*
332. *Id.*
333. *See supra* notes 202–04 and accompanying text.
334. *See* IOM Report, *supra* note 9, at 230 (stating that researchers report that they "have difficulty obtaining deidentified information" from health care entities).

database must be sufficiently large and contain records that are representative of all segments of the patient population so that it generates reliable and generalizable research outcomes.[335] Providers, like patients, will benefit from medical advances because they will enjoy professional success, enhanced reputations, career satisfaction, and perhaps larger incomes resulting from improved knowledge and more effective treatment protocols. It would be entirely unfair to deprive patients of the right to control their health information and the opportunity to consent to their research use but leave providers at liberty to refuse to contribute their files. Achieving social benefits will thus require cooperation and concessions on the part of both patients and the health care industry.

D. Public Health Precedents

Public policy already places the common good ahead of concerns about privacy and autonomy in establishing a large number of reporting requirements. Physicians are required by law to report to authorities cases of particular infectious diseases, including tuberculosis, sexually transmitted diseases, infection by bioterrorism agents, and new epidemic illnesses.[336] State legislatures have also imposed reporting requirements with respect to conditions that affect a patient's ability to drive safely and injuries resulting from child abuse, elder abuse, or violence against an intimate partner or dependent adult.[337] To comply with these mandates, physicians must supply personally identifiable information without asking patients for consent or deferring to patients' privacy concerns.[338]

The public health reporting requirements produce information that is conveyed only to the government and that addresses particular medical problems.[339] The research initiatives contemplated in this Article are distinguishable in that they would open EHR databases to private sector researchers and would yield less certain and less predictable public benefits. Nevertheless, the reporting mandates constitute precedent for an approach that assigns primacy to public welfare over individual privacy and other dignitary concerns.

VI. PRACTICAL SOLUTIONS: PROTECTING DATA SUBJECTS WHILE PROMOTING RECORD-BASED RESEARCH

Even with greater focus on the principle of promoting the public good, concerns about the privacy vulnerabilities of EHR-based research and the risk of harm to data subjects cannot be taken lightly. The opportunity to consent, however, does not protect data subjects from harm if they

335. *See supra* Part IV.D.1.

336. John C. Moskop et al., *From Hippocrates to HIPAA: Privacy and Confidentiality in Emergency Medicine—Part I: Conceptual, Moral, and Legal Foundations*, 45 Annals Emergency Med. 53, 57 (2005).

337. *Id.*

338. *Id.*

339. *See* 45 C.F.R. § 164.512 (2010).

choose to participate in research studies.[340] Consent merely allows individuals to assume the risks knowingly or to opt out completely.

As noted earlier, the federal research regulations and the HIPAA Privacy Rule do not prohibit record-based research in the absence of consent.[341] Neither covers de-identified EHRs; limited data sets can be employed without patient authorization; and even research using clearly identifiable information can proceed without informed consent if authorized by an IRB or privacy board.[342] This Article, therefore, does not propose a radical departure from the current regulatory regime. It argues only that, as a norm, consent need not be sought for non-interventional research.

Nevertheless, record-based research without consent will be ethically justified only if a number of important safeguards are implemented. This Part first analyzes identity concealment techniques and urges that they be used as often as possible. Second, it recommends additional oversight mechanisms that help compensate for the limitations of identity concealment techniques and are tailored to EHR-based research. Finally, it proposes that notice be provided to all individuals whose records might be included in research projects (even in de-identified form), and emphasizes the need for public education about the nature and benefits of EHR-based research.

A. Identity Concealment Techniques

One mechanism to address concerns about privacy is identity concealment. A large body of work exists concerning a variety of identity concealment techniques, including k-anonymity, l-diversity, and others.[343] A comprehensive treatment of the topic is beyond the scope of this Article. Here, we detail recommendations for only two options: 1) building large databases of de-identified records to which researchers can have direct access; and 2) establishing federated systems through which researchers can conduct statistical analyses of distributed databases and receive summary information without direct identifiers.

1. Large Databases of De-Identified Data

Patient privacy may be protected through de-identification. All eighteen safe harbor provision identifiers would need to be removed to mini-

340. Peddicord, *supra* note 5, at 2087.

341. *See supra* Part II.B.

342. *See supra* Part II.B.

343. Ashwin Machanavajjhala et al., *L-Diversity: Privacy Beyond K-Anonymity*, 1 ACM Transactions on Knowledge Discovery from Data 1, 1, 2 (2007); *see generally*, Charu C. Aggarwal & Philip S. Yu, Privacy-Preserving Data Mining: Models and Algorithms (1st ed. 2008); Arvind Narayanan & Vitaly Shmatikov, *Privacy and Security Myths and Fallacies of "Personally Identifiable Information"*, 53 Comm. of the ACM 24, 25 (2010); Latanya Sweeney, *K-Anonymity: A Model for Protecting Privacy*, 10 Int'l J. Uncertainty, Fuzziness & Knowledge-Based Systems 557, 557–70 (2002).

mize the possibility of re-identification.[344] To ensure appropriate de-identification, information technology experts would de-identify patient records and copy them to a separate database that would be available to researchers.[345]

If it is to yield reliable and widely applicable research results, the database should be as comprehensive as possible.[346] As we suggested in previous work, ideally, a national research database would include all Americans' de-identified EHRs.[347] An important question is whether, as Professor Marc Rodwin recommended, health care providers should be required by law to submit their EHRs to a research database.[348]

A strict legal mandate, while scientifically justifiable, may generate an outcry from the medical, and perhaps patient communities, fueled by politicians and news media who wish to generate distrust and resentment of "big government" initiatives.[349] Such an outcry could hinder compliance and foster public resentment of the entire biomedical research endeavor.

Consequently, policy makers may consider alternatives to mandated participation in a comprehensive national research database. A system of incentives and disincentives would be needed to encourage providers to contribute their EHRs.[350] For example, access to the database should be available only to investigators whose institutions contribute their records to it.[351] Another incentive may be access to commercial services that serve as electronic resources for clinicians.[352] For example, we have proposed the development of services for conducting personalized comparisons of treatment effectiveness (PCTE).[353] For a given patient seeking the most appropriate treatment for her condition, a PCTE service would characterize the relative effectiveness of the available treatments by analyzing EHRs for a cohort of treated patients who, when treated, were similar to the given patient with respect to clinically relevant factors.[354] The creation of such services would give institutions and individual providers that do not conduct research a stake in the success of the database, thereby increasing the likelihood that they will contribute the EHRs under their control.[355] Use of the service could be denied to those who

344. *See supra* Part III.B.1.b.ii (discussing re-identification risks).

345. *See* IOM REPORT, *supra* note 9, at 173.

346. *See id.* at 146–47.

347. *See* Hoffman & Podgurski, *supra* note 2, at 162–64.

348. Rodwin, *supra* note 41, at 615–16 (recommending that federal law mandate the creation of a national database of de-identified data and require health care providers to submit their records to the government for this purpose).

349. *See id.* at 589–90.

350. *See id.* at 599–600.

351. One complication may be that some research institutions will not have their own EHR systems.

352. *See* PRICEWATERHOUSECOOPERS, *supra* note 2, at 5.

353. Sharona Hoffman & Andy Podgurski, *Improving Health Care Outcomes Through Personalized Comparisons of Treatment Effectiveness Based on Electronic Health Records*, 39 J.L. MED. & ETHICS 425, 425 (2011).

354. *Id.*

355. *See* PRICEWATERHOUSECOOPERS, *supra* note 2, at 3.

fail to do so.

The principal drawback of relying on incentives to encourage contributions to the database rather than establishing an enforceable mandate is that incentives may not be strong enough to induce full participation, especially if there are commercial advantages to nonparticipation.[356] There would be nothing to stop providers from opting out based on their own cost–benefit calculations.[357]

2. *Does De-Identification Compromise Data Quality?*

To be deemed de-identified under the HIPAA Privacy Rule, records must have eighteen types of identifiers removed.[358] Researchers may be legitimately concerned that removing so many identifiers will compromise the quality or usefulness of research data. One author asserts that removal of the eighteen HIPAA identifiers would render data "useless for most medical research."[359] Another paper concluded that elimination of the HIPAA data elements reduced data by 31% and precluded access to information that is vitally important for research purposes.[360] Of particular importance may be the elimination of all elements of dates other than year, which could prevent researchers from determining the time that elapsed between episodes of care.[361] Similar objections were voiced in comments submitted to HHS concerning the proposed HIPAA Privacy Rule in 2000 and 2002.[362]

In response, HHS explained that it very carefully researched the data elements to be included in the HIPAA safe harbor provision[363] and strove to "balance the need to protect individuals' identities with the need to allow de-identified databases to be useful."[364] The safe harbor provision allows retention of some information about geographic location, including the relevant state and, in most cases, the first three zip code digits.[365] It also allows disclosure of dates, including age, by year,

---

356. Hoffman & Podgurski, *supra* note 2, at 126–28.

357. *See id.*

358. 45 C.F.R. § 164.514(b)(2)(i) (2010).

359. Cate, *supra* note 12, at 1789.

360. Infectious Diseases Soc'y of Am., *Grinding to a Halt: The Effects of the Increasing Regulatory Burden on Research and Quality Improvement Efforts*, 49 CLINICAL INFECTIOUS DISEASES 328, 330 (2009).

361. IOM REPORT, *supra* note 9, at 175.

362. Standards for Privacy of Individually Identifiable Health Information, 65 Fed. Reg. 82,462, 82,710 (Dec. 28, 2000) (codified at 45 C.F.R. pts. 160 & 164); Standards for Privacy of Individually Identifiable Health Information, 67 Fed. Reg. 53,182, 53,232 (Aug. 14, 2002) (codified at 45 C.F.R. pts. 160 & 164); *see also* IOM REPORT, *supra* note 9, at 232–33.

363. Standards for Privacy of Individually Identifiable Health Information, 65 Fed. Reg. at 82,710–12; Standards for Privacy of Individually Identifiable Health Information, 67 Fed. Reg. at 53,232–34.

364. Standard of Privacy of Individual Identifiable Health Information, 67 Fed. Reg. at 53,232.

365. 45 C.F.R. § 164.514(b)(2)(i)(B) (2010). The initial three digits of a zip code cannot be disclosed if 20,000 or fewer people live in that zip code, but, according to HHS, as of 2002, only seventeen zip codes were excluded for this reason. Standards for Privacy of Individually Identifiable Health Information, 67 Fed. Reg. at 53,234.

though not by more specific units.[366] Finally, important details such as race, sex, religion, and income need not be redacted from records in order to render them de-identified.[367] HIPAA-qualified, de-identified data should thus be sufficient for some studies.

3. Secure Statistical Analysis of Distributed Databases

Nevertheless, for other studies, it may be crucial to include identifiers beyond those permitted by the limited data set provision. In these cases, a possible alternative is a technique known as secure statistical analysis of distributed databases.[368]

Secure statistical analysis of distributed databases involves querying databases that participate in a federated system, using special algorithms intended to prevent disclosure of sensitive information.[369] In a federated system, such as the FDA's Sentinel Initiative and DARTNet,[370] each institution manages and maintains control of its own database, but distributed queries are possible through a standard web service.[371] Ideally, all health care providers in the country would participate in a comprehensive federated system, but smaller federated systems may be created at least as a first step. Researchers with approved research projects[372] would submit statistical queries via the Internet using software that interfaces with the federated system's distributed query service. The query service would interact with all relevant databases[373] to initiate operations, communicate intermediate results, and return the final results to researchers. Individual databases in the federation would cooperate to compute summary statistics, but they would not share individual records or sensitive statistics that would identify particular organizations.[374] The query service

366. 45 C.F.R. § 164.514(b)(2)(i)(C). Ages over 89 must be aggregated into a single category of 90 or older, presumably because relatively few people reach that age range. *See id.*

367. *See* § 164.514(b)(2)(i).

368. Alan F. Karr et al., *Secure Regression on Distributed Databases*, 14 J. COMPUTATIONAL & GRAPHICAL STAT. 263, 263–64 (2005).

369. *Id.*; *see also* Oren E. Livne et al., *Federated Querying Architecture for Clinical and Translational Health IT*, *in* PROCEEDINGS OF THE 1ST ACM INTERNATIONAL HEALTH INFORMATICS SYMPOSIUM 250, 251–54 (2010); Wilson D. Pace et al., *An Electronic Practice-Based Network for Observational Comparative Effectiveness Research*, 151 ANNALS INTERNAL MED. 338, 338–39 (2009).

370. *See supra* notes 36–37, 46–49 and accompanying text.

371. Griffin M. Weber et al., *The Shared Health Research Information Network (SHRINE): A Prototype Federated Query Tool for Clinical Data Repositiones*, J. AM. MED. INFORMATICS ASS'N 624, 624 (2009).

372. *See* discussion *infra* Part VI.B.1.b (discussing approval and oversight mechanisms).

373. Queries may be qualified in various ways. For example, a researcher may limit the query to a particular state or geographic region.

374. Alan F. Karr, *Secure Statistical Analysis of Distributed Databases, Emphasizing What We Don't Know*, 1 J. PRIVACY & CONFIDENTIALITY 197, 197, 199 (2009). The approach could support analyses using a number of standard statistical techniques, but it would not permit investigators to employ whatever techniques they choose. For example, entities could fit a linear regression model $Y = \hat{a}X + \hat{a}$ to their global data, consisting of values for the predictor variable(s) X and the outcome variable Y, and share the coefficient(s) $\hat{a}$ of the fitted model, without disclosing to each other either individual-level or entity-level data.

would provide researchers with a somewhat restricted choice of standard statistical query types, enabling them to compute, for example, estimates of population or subpopulation means, proportions, and ratios and estimates of regression coefficients.[375] A requirement that users of the statistical query service be authorized researchers would limit access by illicit users, and improper use of the service by authorized users could be detected after the fact by analyzing logs recording their interactions with the service.[376]

The following are two illustrations to elucidate the use of secure statistical analysis of distributed databases. First, a researcher might submit a query asking for the prevalence of a particular disease among the population represented by the combined records contained in the federated system. After statistical analysis, the researcher would receive an estimate of the proportion of the population diagnosed with that disease. Second, an investigator could conduct more complex CER[377] using the service. The investigator would indicate the treatments at issue, the outcome measures of interest, and any known confounders, and would select the desired analytical approach. Given these parameters, the query service would conduct the statistical analysis and provide results to the researcher. For example, the query service might select and compare two treatment groups, each of which received a different treatment, ensuring that the groups are balanced with respect to values of known confounding variables. Ultimately, the investigator would receive a numerical estimate of the difference in the average treatment effects for the two groups.

Statistical databases in general are not invulnerable to attack, and a number of technical issues must be resolved for secure statistical analysis of distributed databases to become widely applicable.[378] Ideally, the approach would yield useful research data by allowing original, non-redacted medical records to be queried at their facilities of origin while protecting patient privacy because investigators would only see information summarizing aggregate data. The participating organizations would each need to support the same data schema, communication protocol, querying interface, and security policy.

Given the health information technology resources needed to support the requirements for a participating database within a federated system, it is likely that small or resource-poor health care providers would have to use trusted third-party aggregators[379] to provide the query service. These

375. *See* Hoffman & Podgurski, *supra* note 2, at 118–19.

376. *See id.* at 154–55 (discussing audit trails).

377. *See supra* notes 82–86 and accompanying text (discussing comparative effectiveness).

378. *See* Cynthia Dwork, *A Firm Foundation for Private Data Analysis*, 54 COMMS. ACM 86, 86–95 (2011); Karr, *supra* note 374, at 202–95.

379. *See* David Kitte, *Report of the Data Sharing and Aggregation Workshop*, AGENCY FOR HEALTHCARE RES. & QUALITY, http://ahrq.gov/qual/performance3/perfm3c.htm (last visited Nov. 2, 2011).

providers would need to upload new and updated EHRs regularly to the aggregator but would not have to support statistical queries themselves.

Other providers may also choose to use trusted aggregators in order to avoid conflicts of interest among competing health care entities conducting commercially oriented research. If researchers from such entities had to submit queries directly to competitors, the queries themselves might reveal the exact nature of the research. Such disclosure might cause the querying entity to lose its competitive advantage and diminish potential profits from research discoveries. Trusted aggregators can serve as intermediaries who hold copies of medical records and process queries from researchers without revealing them to other parties. For this reason, the aggregators should be government contractors subject to rigorous oversight. Given that there is a risk of disclosures from any database, the maximum number of EHRs under the control of any one aggregator should be limited.

B. Strengthening Research Oversight

Identity concealment is a useful safeguard against research abuses, but it cannot fully shield data subjects from harm and must be supplemented by other protections. This Article has argued that informed consent requirements should be suspended for all record-based studies.[380] However, in lieu of having an opportunity to consent, data subjects should enjoy the benefits of rigorous oversight and feel as confident as possible about its efficacy. Because of the risk of re-identification, even studies using de-identified records should undergo an approval procedure, though it can be streamlined. This section outlines a tiered review process that would apply some degree of scrutiny to all research projects. It also emphasizes the importance of continuing review and offers recommendations for enhanced security measures to protect EHR databases.

1. Ethics Board Review

Regulations that require approval and monitoring of all studies by an ethics board could go far to protect data subjects from the risks of record-based research. The IOM developed a relevant proposal, which is described and critiqued. We then offer an alternative framework that would provide data subjects with more comprehensive protections.

a. IOM Proposal

In *Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health through Research*, the IOM detailed a proposal to remove barriers to record-based research.[381] It recommended that waivers of informed consent be granted so long as consent is replaced by other protection

380. *See supra* Part III.
381. IOM Report, *supra* note 9.

mechanisms.[382] Accordingly, researchers who believe direct identifiers are necessary for their studies and who do not wish to obtain consent, would seek approval from an ethics oversight board with expertise in reviewing records-based research.[383] The board could grant waivers for studies using identifiable health information after considering the following: (1) measures that will be taken to safeguard data security, (2) possible harms to which inappropriate disclosure would expose subjects, and (3) the study's potential benefits.[384] The IOM did not recommend ethics board oversight for studies in which no direct identifiers would be available to investigators.[385]

The IOM proposal has several strengths. The IOM specifies that ethics boards would need to have special expertise with respect to record-based research, unlike traditional IRBs that often focus largely on studies involving clinical testing.[386] In addition, boards would be specifically directed to scrutinize the privacy safeguards that investigators plan to implement.[387]

However, the IOM's proposal does not go far enough. First, it does not define the term "direct identifiers" and, thus, does not clarify which data elements would trigger ethics board review.[388] Second, the IOM does not support subjecting studies that do not involve "direct identifiers" to any oversight.[389] Third, the IOM relies excessively on pre-approval of research protocols.[390] The ethics board is envisioned as scrutinizing only security measures that researchers *plan to* implement without following up to ensure that they have been employed and are effective.[391] The IOM recommendations do not take into account the multiplication of risk that occurs when de-identified health information is promulgated to more and more research groups, each of which is a point of potential vulnerability to security and privacy violations.[392] The IOM recommendations also do not take into account the highly changeable nature of security threats.

b. Proposed Regulatory Approach

As detailed above, data de-identification and identity concealment in general do not entirely eliminate the risk of privacy violations. With some effort, adversaries could re-identify at least a small percentage of records, and this risk cannot be ignored.[393]

382. *Id.* at 34.
383. *See id.*
384. *Id.*
385. *Id.*
386. *Id.* at 265.
387. *Id.*
388. *Id.*
389. *Id.*
390. *Id.* at 264.
391. *Id.* at 265.
392. *Id.*
393. *See supra* Part III.B.1.b.ii.

Consequently, this Article proposes that all record-based studies undergo approval by an ethics board with expertise in non-interventional research and in information security. This task could be assigned to existing IRBs, but because these bodies are already overworked and may not have the requisite expertise,[394] separate reviewing entities could be established exclusively for record-based studies. We use the term "ethics boards" in this section but do not mean to suggest that these must necessarily be different from IRBs.

The degree of scrutiny that ethics boards apply to studies should depend on the extent to which researchers or others may be able to identify patient data and on the severity of the potential harm. Studies using any identifiers that are excluded by the HIPAA safe harbor provision, including limited data sets,[395] should undergo a thorough approval process. Limited data sets should be subject to careful scrutiny because they can include birthdates and zip codes, which significantly increase the possibility of re-identification.[396] The ethics board should pay particular attention to the security measures that will be implemented. It should also verify the credentials of applicants to ensure that they are bona fide researchers who have a genuine research project in mind.

If, for some reason, researchers must obtain directly identifiable data such as names or social security numbers, ethics boards should remain free, at their discretion, to require patient consent. For example, informed consent may be appropriate for a small study that allows investigators to obtain patient names and view sensitive medical information including psychiatric or gynecological records.

Studies in which researchers will view only data that are de-identified in accordance with the HIPAA safe harbor provision should undergo a streamlined process through which investigators register their projects and their identities are confirmed. The researchers should also promise in writing that they will not attempt to re-identify data, will not convey the records they obtain to individuals who are not members of the research team, and will refrain from using data for purposes outside the scope of the study. In addition, researchers should commit to disposing of any records they have obtained in identifiable or de-identified form,

394. Joseph A. Catania et al., *Survey of U.S. Boards That Review Mental Health-Related Research*, 3 J. EMPIR. RES. HUM. RES. ETHICS 71, 71 (2009) (noting concern about IRB workloads); Ezekiel J. Emanuel et al., *Oversight of Human Participants Research: Identifying Problems to Evaluate Reform Proposals*, 141 ANNALS INTERNAL MED. 282, 284 (2004) ("IRBs may review research for which they lack expertise and rely on information given by investigators without corroboration"); David A. Hyman, *Institutional Review Boards: Is This the Least Worst We Can Do?*, 101 NW. U. L. REV. 750, 761 (2007) (discussing IRBs' heavy workload).

395. *See supra* notes 74–76 and accompanying text.

396. One study found that for limited data sets, the risk of re-identification ranges from 10% to 60%, depending on the information that different states make publicly available. *See* Benitez & Malin, *supra* note 153, at 169. Currently, the Common Rule does not make clear whether research using limited data sets would require IRB approval and consent, and the HIPAA Privacy Rule requires data use agreements but no patient authorization for such studies. *See supra* notes 60–61, 78 and accompanying text.

using approved means, at the end of a designated period.[397]

Furthermore, it would be essential for ethics boards to conduct continuing reviews of all research studies.[398] Researchers should be required to submit annual reports and to inform the board immediately of any adverse events, such as hacking or inappropriate disclosure of data to third parties. If data are not appropriately safeguarded, ethics boards may not approve future studies by the same investigators, may require corrective action, or may withdraw approval of the study and mandate that it be stopped.[399] Monitoring by the boards should be supplemented with oversight by HHS, which is charged with HIPAA Security Rule enforcement.[400] HHS should be authorized to conduct unannounced audits of all research projects, including those using de-identified data, to ensure that investigators are safeguarding privacy with appropriate security measures and are engaging in valid research activities rather than misusing data.[401] HHS should also ensure that the ethics boards are responsibly fulfilling their duties.

Ethics board oversight for all record-based studies is a novel recommendation that departs from the IOM's more modest proposal and proposals made by other analysts. Professor Rodwin has suggested that after records are fully de-identified, they be made available to the public, perhaps for a fee.[402] We believe it would be irresponsible to allow any member of the public to access de-identified EHRs without any oversight because, over time, such a policy would likely lead to abuses.

If a large number of de-identified EHRs were to be publicly available, data miners would gain many targets for re-identification, and they could expend as much time and computational power as they have available. Even with a low success rate, they may be able to re-identify a large number of EHRs.[403] For example, if data miners had access to a database of de-identified EHRs for every person in the United States (over 311 million people)[404] and they de-identified records with a 0.10% success rate,[405] they would be able to de-identify EHRs of over 311,000 people. Moreover, some data miners, such as certain commercial enterprises, may

397. *See* Ohm, *supra* note 69, at 1767.

398. 45 C.F.R. § 46.103 (2010) (discussing continuing review by IRBs); Hoffman, *supra* note 17, at 738–43.

399. *See* 45 C.F.R. § 46.113 (2010) (empowering IRBs to suspend or terminate studies that they approved).

400. *Id.* § 160.308.

401. The HIPAA Security Rule already authorizes HHS to engage in enforcement activities, including conducting compliance reviews; this authority should be expanded to all research activities, including those involving databases of de-identified records. *See id.*; *see also* discussion *infra* Part VI.B.1.c (discussing security safeguards).

402. Rodwin, *supra* note 41, at 615.

403. *See* PETER WINKELSTEIN, MEDICAL INFORMATICS KNOWLEDGE AND DATA MINING IN BIOMEDICINE 153–54 (HsinChin Chen et al. eds., 2005).

404. U.S. Census Bureau, *U.S. & World Population Clocks*, CENSUS.GOV, http://www.census.gov/main/www/popclock.html (last visited Nov. 10, 2011).

405. Benitez & Malin, *supra* note 153, at 169 (finding that between 0.01% and 0.25% of a state's population is vulnerable to re-identification if data is de-identified in accordance with the HIPAA safe harbor provision).

have access to external data, gleaned from various sources, that greatly facilitates re-identification.[406]

Regulators should also consider whether research proposals should be subject to further limitations. For example, should ethics boards limit the number of records to which investigators have access?[407] Should an upper limit be set for the number of queries a research team submits to a statistical database on the theory that an unreasonable number of queries might indicate that the data are being used for inappropriate purposes? These questions require further study by security experts who would need to balance the needs of researchers against the need to optimize privacy protection.

Ethics board review and supervision of all record-based projects will surely entail costs, though the streamlined approach for de-identified records is designed to curb expenses. To finance ethics board operations, federal regulations could require applicants who seek project approval to pay a fee to HHS. A precedent for such an approach is set by the Prescription Drug User Fee Act[408] and the Medical Device User Fee and Modernization Act of 2002,[409] which require drug and device manufacturers seeking FDA approval to pay certain fees.

c. Security Safeguards

As emphasized throughout this Article, a major concern relating to record-based research is the risk of privacy breaches. HHS addressed security concerns relating to electronically stored health information by promulgating the HIPAA Security Rule in 2005.[410] These regulations require implementation of a variety of administrative, physical, and technical safeguards.[411] The Security Rule, however, applies only to health plans, health care clearinghouses, health care providers who transmit health information in electronic form for particular purposes, and their business associates.[412] Other entities are not required to employ any of the Rule's security measures no matter how much health data they may store or process.[413]

We have suggested improvements to the HIPAA Security Rule in prior work.[414] A critical modification would be expanding the definition of "covered entities" to ensure that all researchers, as well as entities or individuals who operate research databases, are subject to regulatory requirements. The term "covered entity" should apply to "any person who

406. MEHMED KANTARDZIC, DATA MINING: CONCEPTS, MODELS, AND ALGORITHMS 380 (2d ed. 2011).
407. *See* Ohm, *supra* note 69, at 1767.
408. 21 U.S.C. § 379h (Supp. IV 2010) (detailing fees that must be paid by those submitting human drug applications).
409. *Id.* § 379j(a).
410. 45 C.F.R. §§ 164.302–.318 (2010).
411. *Id.*
412. *Id.* § 160.103; 42 U.S.C. § 17931 (2006).
413. Hoffman & Podgurski, *supra* note 127, at 344–45.
414. *Id.* at 359–84.

knowingly stores or transmits individually identifiable health information in electronic form for any business or research purpose related to the substance of such information."[415] No distinctions should be made based on the source of the health information.

The HIPAA Security Rule, as currently written, exempts databases of de-identified records that meet the safe harbor standard from complying with the specified security measures.[416] Because determined adversaries may even be able to re-identify records that are de-identified in accordance with safe harbor guidelines,[417] there is reason to question whether this exemption is sound, and the matter merits further examination by security experts.

To protect patient privacy, HHS will also need to enforce the HIPAA Security Rule aggressively. Responsibility for enforcement is delegated to the agency's Office of Civil Rights.[418] The regulations empower HHS to investigate complaints of violations and to conduct self-initiated compliance reviews of covered entities.[419] HHS states on its website that in 2008 and 2009 it conducted ten compliance reviews.[420] With the proliferation of EHR databases and EHR-based research projects, HHS will need to augment its monitoring activities to prevent privacy abuses and may require additional funding to do so. HHS will need to be ever-vigilant in overseeing the security of large databases or federated systems because security threats evolve rapidly over time and often cannot be anticipated.

C. Notice and Education

Effective protection of patient privacy through identity concealment, robust oversight for all protocols, and enhanced security should considerably alleviate anxiety about the potential risks of EHR-based research. But, even these measures would not address concerns about the autonomy rights of data subjects. Some advocates may still favor consent as a matter of principle or because they are concerned about the potential for group stigmatization, objectionable outcomes, and commercial exploitation.[421]

This part proposes that notice and education replace consent in record-based research studies. Notice and education, admittedly, will not enable data subjects to make a choice about inclusion of their records in databases. But they can empower data subjects in other ways. In a dem-

415. *Id.* at 360.

416. *See* 45 C.F.R. § 160.103 (defining "protected health information" as "individually identifiable health information"); 45 C.F.R. § 164.302 (establishing that the Security Rule applies to "electronic protected health information").

417. *See supra* Part III.B.1.b.ii.

418. U.S. Dep't of Health & Hum. Servs., *Security Rule Enforcement*, HHS.GOV, http://www.hhs.gov/ocr/privacy/hipaa/enforcement/cmsenforcemain.html (last visited Nov. 2, 2011) [hereinafter *Security Rule Enforcement*].

419. 45 C.F.R. §§ 160.306, 160.308.

420. *Security Rule Enforcement*, *supra* note 418.

421. *See supra* Part III.B.2.

ocratic society, an educated public can effectively dismantle policies that are objectionable. The democratic political system can bring change by fostering communication with elected representatives, permitting referenda, or ultimately using elections to replace government officials. Therefore, notice and education, like consent, can promote autonomy and respect for persons.

1. Notice

The HIPAA Privacy Rule requires that health care providers notify patients that their health data might be used in some instances without their consent.[422] Permissible uses include disclosure of information for treatment, payment, health care operations, public health initiatives, law enforcement, and other purposes.[423] But de-identified information is not covered by the privacy regulations,[424] so the Privacy Rule does not entitle patients to any notice regarding research uses of data without identifiers.[425]

We propose expanding the HIPAA Privacy Rule notice requirement to apply to all research uses. For research using identifiable records, notice should replace requests for patient authorization, and notices should also explain that authorized investigators might access patient information in de-identified form. Health care providers whose EHRs may be available to researchers through any venue and in any format should be obligated to furnish patients with a notice that describes in general terms how and under what circumstances their data might be used.

The notice should be provided in written form and also be discussed verbally with patients by either a physician or a knowledgeable nurse. Notice should be supplied to patients at least once by each health care provider, such as the doctor, hospital, laboratory, etc., whose EHR will be used for research purposes.

The notice should briefly explain the benefits of observational studies and the potential to determine the comparative effectiveness of treatments and achieve medical advances that will improve health care outcomes for all. If applicable, it should also explain that no individually identifiable data will be disclosed to researchers. These explanations should be written in simple language that is accessible to an average reader.[426] In addition, the notice could acknowledge that research can ultimately lead to commercial profits that are not shared with data

422. 45 C.F.R. § 164.520.

423. *Id.* §§ 164.506(c), 164.512.

424. *Id.* § 164.514.

425. If individually identifiable information will be disclosed to researchers, the covered entity must currently obtain patient authorization in advance and cannot merely provide notice. *Id.* § 164.508(b)(3)(i).

426. According to experts, the average reading comprehension level in the United States is at most an eighth grade level. *See What Is Health Literacy?*, PARTNERSHIP FOR CLEAR HEALTH COMM., http://www.npsf.org/pchc/health-literacy.php (last visited Nov. 2, 2011); *Comprehension and Reading Level*, INFORMATICS REV., http://www.informatics-review.com/FAQ/reading.html (last visited Nov. 2, 2011) ("Research tells us that to commu-

subjects.[427]

2. Public Education Initiatives

It would be naïve to assume that the public will automatically embrace a policy allowing researchers to access de-identified EHR information without patient consent. Those generally resistant to federal initiatives as manifestations of "big government" and as infringing upon individual autonomy may be very vocal in their opposition and gain support from significant segments of the population. The media may also be complicit in fueling public discontent.[428] In the 1990s, President Clinton's efforts to achieve health care reform were derailed by opposition from conservatives, libertarians, and the health care industry with the help of a highly effective "Harry and Louise" television advertisement.[429] In 2009, as the country debated the merits of President Obama's health care reform initiative, the idea that the government would utilize "death panels" to ration care gained surprising traction.[430]

To gain public trust, promoters of comparative effectiveness and other EHR-based research initiatives should launch their own public education campaign.[431] This responsibility should be shared by HHS, private research institutions, and highly respected professional organizations, such as the American Medical Association. Educational messages can take the form of public service announcements and news stories through media outlets such as television, radio, medical websites, and e-mail. Researchers should also distribute updates concerning ongoing research projects and their outcomes to the media so that the public can remain apprised of the uses to which EHRs are put and the new knowledge that is acquired as a result. The costs of educational initiatives can be covered, at least in part, through user fees charged to commercial research organizations that apply for access to EHR databases or federated systems.[432]

In addition, local researchers could conduct community meetings to educate the public and address concerns about EHR-based research. A similar approach is used when investigators seek a waiver of informed consent for research regarding emergency care in circumstances in which obtaining consent will be impossible.[433] The federal regulations require

---

nicate effectively with a general audience in the U.S., we need to write at a 6th-8th grade reading level.").

427. *See supra* Part III.B.2.c.

428. Victor R. Fuchs & Arnold Milstein, *The $640 Billion Question—Why Does Cost-Effective Care Diffuse So Slowly?*, 364 NEW ENG. J. MED. 1985, 1986 (2011) (stating that the media "is the principal source" of misunderstandings about "who really pays for health care" and "the *relative* benefit of clinical interventions").

429. Natasha Singer, *Harry and Louise Return, with a New Message*, N.Y. TIMES, July 17, 2009, at B3, *available at* http://www.nytimes.com/2009/07/17/business/media/17adco.html.

430. Earl Blumenauer, *My Near Death Panel Experience*, N.Y. TIMES, Nov. 15, 2009, at WK12.

431. IOM REPORT, *supra* note 9, at 144.

432. *See supra* notes 401–05 and accompanying text.

433. *See* 21 C.F.R. § 50.23(a) (2011).

public disclosure of the project and consultation with community representatives in the area from which the subjects will be drawn.[434] Similar meetings could be conducted in public libraries, community centers, and other easily accessible locations to discuss EHR-based observational studies for which consent will not be sought. These meetings would not be consultations that seek input from community members but would be an opportunity for the public to interact in person with researchers and gain an in-depth understanding of record-based research.

3. The Benefits of Notice and Education

A notice mandate and public education initiatives would go far beyond the current regulatory mandates with respect to de-identified data. So long as health care providers do not disclose personally identifiable information to researchers, neither the Common Rule nor the HIPAA Privacy Rule requires that data subjects receive any information at all about studies.[435] Health care providers can thus submit data that meet the HIPAA safe harbor provision's requirements to a research database without any regulatory oversight, and they are free to leave data subjects in complete ignorance of such research activities.[436]

It is also noteworthy that the law does not require physicians to seek patients' permission to create medical files in the first place. In addition, providers do not ask patients to consent to the transition from paper records to EHRs, even though computerized records can expose patients to privacy breach risks that do not exist when records are limited to hard-copy files that can be locked away in cabinets.[437] In fact, it is more likely that privacy breaches will occur in clinical settings than in the generally more focused and controlled setting of a research project that involves a limited number of professionals who are dedicated to achieving accurate study outcomes.[438]

Notice and education will not empower data subjects to make choices about inclusion of their records in observational studies. However, a mandate that researchers share comprehensive and truthful information with the public, together with intensified oversight, should prevent abuses and exploitation of data subjects. Historically, such abuses occurred when data subjects were vulnerable because of ignorance, poverty, or imprisonment.[439] In addition, notice and education will enable the public to voice its concerns and influence research policies through the democratic process.

---

434. *Id.* § 50.24(a)(7).

435. *See supra* Part II.B.

436. *See supra* Part II.B.

437. *See supra* Part III.B.1.a.

438. *Breaches Affecting*, *supra* note 129 (reporting large privacy breaches at various health care entities).

439. *See supra* Part IV.A.

D. Additional Safeguards to Protect Data Subject Interests

In this final section we briefly review several laws that protect patients from misuse of their data by third parties. In addition, we suggest a few further interventions that could be implemented to minimize the risk of harm to data subjects.

Federal and state laws address discrimination based on biological and health-related factors. The Americans with Disabilities Act prohibits employment discrimination against qualified individuals with disabilities.[440] The Genetic Information Nondiscrimination Act prohibits employers and health insurers from engaging in discrimination based on genetic information.[441] Many state legislatures have passed other anti-discrimination laws.[442] The statutory restrictions on how employers and insurers can use health data may reduce the frequency or impact of privacy breaches because it could make health information less attractive to hackers or those to whom they seek to sell their bounty.[443] The efficacy of these laws depends on a variety of factors, among which are judicial interpretation of statutory provisions and robust administrative enforcement.[444] However, the statutes' existence sends an important message to those who would be inclined to subject the vulnerable to discrimination, deters at least some misconduct, and provides potential remedies for aggrieved individuals.[445]

There are additional steps that could protect data subjects against certain dignitary harms associated with record-based studies. Researchers should be scrupulous and conscientious in reporting research results to avoid group stigmatization.[446] As stated in Section 30 of the Helsinki Declaration, authors are responsible for the accuracy and completeness of their research outcome reports.[447] To illustrate, assume that researchers determine that a genetic abnormality exists across populations but is somewhat more prevalent among individuals of a particular ancestry. In interviews and publications, investigators must clearly communicate that the genetic abnormality is not unique to a particular minority group and accurately describe its prevalence variations. Neither the media nor researchers should be tempted to generate sensationalist, misleading headlines that might receive public attention but be inflammatory and damaging to a minority group.

---

440. 42 U.S.C. § 12112 (2006).

441. Genetic Information Nondiscrimination Act (GINA) of 2008, Pub. L. No. 110-223, 122 Stat. 881 (2008).

442. Sharona Hoffman, *The Importance of Immutability in Employment Discrimination Law*, 52 Wm. & Mary L. Rev. 1483, 1488 (2011).

443. Sharona Hoffman, *Settling the Matter: Does Title I of the ADA Work?*, 59 Ala. L. Rev. 305, 307 (discussing plaintiffs' low win rates in court and Equal Employment Opportunity Commission enforcement of the Americans with Disabilities Act).

444. *Id.* at 308–11, 314–16.

445. *Id.* at 307 (discussing the benefits of the Americans with Disabilities Act).

446. Hoffman, *supra* note 87, at 450–54.

447. Declaration of Helsinki, *supra* note 191, § 30.

The problem of offensive outcomes[448] could be partially addressed through regulatory intervention. Regulations could prohibit investigators from undertaking designated types of studies that are particularly controversial without data subject consent even if researchers will see only de-identified information. For example, consent could be required for studies that focus directly on facilitating abortion. This approach would have to be limited to a very small number of study categories because it will be costly and burdensome and could threaten the integrity of research projects through selection bias.[449] However, it may need to be considered to avoid public outcries and resistance to the EHR database research enterprise.

VII. CONCLUSION

Individual interests in privacy and autonomy may conflict with society's need for the best possible research outcomes. This Article has sought to balance competing goals and values, and it has proposed a multi-faceted approach to maximize research opportunities and to protect the valid interests of data subjects.

The traditional autonomy-focused model is inappropriate for large-scale, record-based research enabled by EHR technology. Instead, the research ethics framework must shift to emphasize the common good in noninterventional research. This conceptual change must be combined with a variety of measures that will replace informed consent and effectively safeguard patient privacy and other dignitary interests. These include: (1) the development of research techniques that yield adequate data but conceal patient identifiers from researchers; (2) ethics board oversight for all record-based studies, including those using de-identified data; (3) scrupulous attention to the security of databases and revision of the HIPAA Security Rule; and (4) notice and educational initiatives.

In July 2011 HHS issued an Advance Notice of Proposed Rulemaking (ANPRM) titled "Human Subjects Research Protections: Enhancing Protections for Research Subjects and Reducing Burden, Delay, and Ambiguity for Investigators."[450] The ANPRM represents the first effort to modernize the research regulations in over two decades, and it generated nearly 1,100 comments during the first comment period, which ended on October 26, 2011.[451] It is unclear when the new rule will be finalized or what its contents will be, but the process is likely to be lengthy.[452] In the meantime, we hope that regulators will consider the concerns we raise.

---

448. *See supra* Part III.B.2.b.

449. *See supra* Part IV.D.

450. 76 Fed. Reg. 44512 (proposed July 26, 2011) (to be codified at 45 C.F.R. pts. 46, 160, 164, and 21 C.F.R. pts. 50, 56).

451. Jeannie Baumann, *Human Subject Protection: OHRP Commended for Proposed Changes to Common Rule; Some Areas Questioned*, 10 MED. RES. L. POL'Y REP. (BNA) 723, 723 (Nov. 2, 2011).

452. *Id.*

Observational research involving EHR databases is not without some risk of privacy violations or other dignitary harms, and these should not be ignored. But with appropriate interventions, the risks can be minimized. Health information technology creates the potential for unprecedented scientific discoveries and dramatic improvements in human health. Society must not squander this opportunity.