2014

Citizen Science: The Law and Ethics of Public Access to Medical Big Data

Sharona Hoffman
Case Western Reserve University School of Law, sharona.hoffman@case.edu

CITIZEN SCIENCE: THE LAW AND ETHICS OF PUBLIC ACCESS TO MEDICAL BIG DATA

Sharona Hoffman[†]

ABSTRACT

Patient-related medical information is becoming increasingly available on the Internet, spurred by government open data policies and private sector data sharing initiatives. Websites such as HealthData.gov, GenBank, and PatientsLikeMe allow members of the public to access a wealth of health information. As the medical information terrain quickly changes, the legal system must not lag behind. This Article provides a base on which to build a coherent health data policy. It canvasses emergent data troves and wrestles with their legal and ethical ramifications.

Publicly accessible medical data have the potential to yield numerous benefits, including scientific discoveries, cost savings, new patient support tools, improved healthcare quality, greater government transparency, and public education. At the same time, the availability of electronic personal health information that can be mined by any Internet user raises concerns related to privacy, discrimination, erroneous research findings, and litigation. This Article analyzes the benefits and risks of health data sharing and proposes balanced legislative, regulatory, and policy modifications to guide data disclosure and use.

DOI: http://dx.doi.org/10.15779/Z385Z78

† Edgar A. Hahn Professor of Law and Professor of Bioethics, Co-Director of Law-Medicine Center, Case Western Reserve University School of Law; B.A., Wellesley College; J.D., Harvard Law School; LL.M. in Health Law, University of Houston. Professor Hoffman was a Distinguished Scholar in Residence at the Centers for Disease Control and Prevention's (CDC) Center for Surveillance, Epidemiology and Laboratory Services during the spring semester of 2014. This Article grew out of the author's work with the CDC, and she wishes to thank the many colleagues who discussed these important issues with her. The author also thanks Jaime Bouvier, Jessie Hill, Tony Moulton, Andy Podgurski, Andrew Pollis, and Timothy Webster for their thoughtful comments on prior drafts. Tracy (Yeheng) Li provided invaluable research assistance throughout this project.

TABLE OF CONTENTS

I. INTRODUCTION 1744

II. PUBLICLY AVAILABLE BIG DATA SOURCES 1748

A. FEDERAL AND STATE DATABASES 1748
1. Federal Government Data at HealthData.gov 1748
a) CDC Wonder 1749
b) Chronic Condition Data Warehouse 1749
2. State Government Health Data Websites 1750
3. Healthcare Cost and Utilization Project 1750
4. GenBank 1751
5. All-Payer Claims Databases 1752
B. PRIVATE SECTOR DATABASES 1753
1. Dryad Digital Repository 1753
2. PatientsLikeMe 1753
3. The Personal Genome Project 1754

III. THE BENEFITS OF PUBLIC ACCESS TO HEALTH INFORMATION 1755

A. SCIENTIFIC DISCOVERY 1755
B. RESEARCH COST REDUCTIONS 1757
C. TOOLS TO HELP PATIENTS NAVIGATE THE HEALTHCARE SYSTEM 1761
D. GOVERNMENT TRANSPARENCY AND PUBLIC EDUCATION 1761
E. IMPROVEMENTS IN HEALTHCARE QUALITY AND PUBLIC HEALTH POLICY 1762

IV. RISKS OF PUBLIC ACCESS TO HEALTH DATA 1763

A. PRIVACY THREATS 1764
1. Privacy Law 1764
a) The HIPAA Privacy Rule 1765
b) The Privacy Act 1765
c) State Laws 1765
2. De-identification 1766
3. Does Public-Use Medical Data Pose a Real Privacy Threat? 1768
a) Data Holders Not Covered by the HIPAA Privacy Rule 1768
b) Re-identification of Fully De-identified Health Records 1770
c) The Peculiarities of Genetic Information 1771
B. DISCRIMINATION AND SPECIAL TARGETING 1772

1. Employers 1773
a) Using Identifiable or Re-Identifiable Data 1774
b) De-identified Information as a Basis for Multi-Factor Discrimination and Discrimination by Proxy 1776
2. Financial Institutions and Marketers 1778
C. PROPAGATION OF INCORRECT AND HARMFUL RESEARCH CONCLUSIONS 1780
1. Error Sources 1782
2. Potential Harms 1783
D. LITIGATION 1785
1. Defamation 1786
2. Other Causes of Action 1787
3. Anti-SLAPP Legislation 1788
V. RECOMMENDATIONS 1789
A. PRIVACY AND DATA STEWARDSHIP 1790
1. HIPAA Privacy Rule Modifications 1790
a) Expanding the Definition of "Covered Entity" and Creating National Data Release and De-identification Standards 1790
b) Prohibiting Re-identification 1792
2. Data Release Review Boards 1793
3. Data Use Agreements, Privacy Training, Registries, and Consent Procedures 1793
B. ANTI-DISCRIMINATION PROTECTIONS 1796
1. Detecting, Deterring, and Prosecuting Multi-Factor Discrimination 1797
2. Requiring Disclosure of Data Mining for Disability Proxies and Predictors 1798
3. Addressing Data Mining in the ADA's Definition of Disability 1799
C. CITIZEN SCIENTIST CHAPERONING 1800
D. TORT CLAIM LITIGATION STRATEGIES 1803
VI. CONCLUSION 1804

I. INTRODUCTION

On May 9, 2013, President Barack Obama issued an executive order entitled "Making Open and Machine Readable the New Default for Government Information."[1] The Order directed that, to the extent permitted by law, the government must release its data to the public in forms that are easy to find, access, and use.

Health information drawn from patient records is among the most useful but sensitive types of data that are becoming commonly available to the public pursuant to President Obama's policy and other public and private initiatives that will be discussed in this Article. This is the first article to canvass these emergent data troves and to wrestle with their legal and ethical ramifications. As federal agencies gear up to post increasing amounts of information on the Internet in order to comply with Executive Order 13,642,[2] it is time to carefully consider the benefits and the risks of public access to medical data. The Article also formulates guidelines for data use in order to protect privacy, deter discrimination, and prevent other harms.

Ordinary citizens can now easily find and access patient-related medical data on the Internet.[3] This is the era of "Citizen Science" and "Do-It-Yourself Biology."[4] Citizen Science is "the practice of public participation and collaboration in scientific research" through data collection, monitoring, and analysis for purposes of scientific discovery,

1. Exec. Order No. 13,642, Making Open and Machine Readable the New Default for Government Information, 78 Fed. Reg. 28111 (May 14, 2013), http://www.gpo.gov/fdsys/pkg/FR-2013-05-14/pdf/2013-11533.pdf. The Order states, in relevant part:

> To promote continued job growth, Government efficiency, and the social good that can be gained from opening Government data to the public, the default state of new and modernized Government information resources shall be open and machine readable. Government information shall be managed as an asset throughout its life cycle to promote interoperability and openness, and, wherever possible and legally permissible, to ensure that data are released to the public in ways that make the data easy to find, accessible, and usable.

2. *Id.*; *see also* OFFICE OF MGMT. & BUDGET, EXEC. OFFICE OF THE PRESIDENT, OMB MEMORANDUM M-13-13, OPEN DATA POLICY—MANAGING INFORMATION AS AN ASSET (2013), http://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf.

3. *See infra* Part II.

4. Heidi Ledford, *Garage Biotech: Life Hackers*, 467 SCIENCE 650, 650–52 (2010); Amy Dockser Marcus, *Citizen Scientists*, WALL STREET. J., Dec. 3, 2011.

usually without compensation.[5] Do-It-Yourself Biology (DIYbio) is an international movement "spreading the use of biotechnology beyond traditional academics and industrial institutions and into the lay public."[6]

Large data resources are often called "big data," which is characterized by its sizeable volume, variety, and velocity, that is, the speed with which it is produced.[7] Increasingly, government and private sector sources furnish data collections to the public, and this supply stream will expand considerably in the future.[8] In this Article, publicly available resources will be called "public-use data" or "open data."

The potential benefits of public access to health information are considerable. At a time of diminishing government funding for research,[9] the widespread availability of high-quality datasets at little to no cost may be very important to continued scientific advancement. Professional researchers as well as talented and dedicated students and amateurs could make important discoveries and answer pressing medical questions,[10] and they could do so without undertaking the expense, time, and work involved in recruiting patients and developing original datasets.[11] Open data has also enabled entrepreneurs to create tools that assist patients in navigating the complexities of the contemporary healthcare system by facilitating searches about symptoms and treatments, listing medical providers by location, and furnishing physician ratings and price information.[12] In addition, federal and state data sharing initiatives promote government transparency and educational initiatives about health and medicine.[13] Finally, data sharing may promote improvements in government-provided services. Easily accessible and navigable public-use

5. *Citizen Science*, NAT'L GEOGRAPHIC, http://education.nationalgeographic.com/education/encyclopedia/citizen-science (last visited Sept. 17, 2015).

6. DANIEL GRUSHKIN ET AL., SYNTHETIC BIOLOGY PROJECT, SEVEN MYTHS & REALITIES ABOUT DO-IT-YOURSELF BIOLOGY 4 (2013), http://www.synbioproject.org/process/assets/files/6676/7_myths_final.pdf.

7. PHILIP RUSSOM, TDWI RESEARCH, BIG DATA ANALYTICS 6 (2011), https://tdwi.org/research/2011/09/~/media/TDWI/TDWI/Research/BPR/2011/TDWI_BPReport_Q411_Big_Data_Analytics_Web/TDWI_BPReport_Q411_Big%20Data_ExecSummary.pdf.

8. *See infra* Part II.

9. Nora Macaluso, *Decade-Long Funding Decline at NIH May Be Poised for Reversal, Collins Says*, 13 BLOOMBERG BNA MED. RES. L. & POL'Y REP. 311 (2014) (indicating that "the chances of a project's getting a grant from NIH have fallen to about 16 percent from 25 percent to 30 percent before 2003").

10. *See infra* Section III.A.

11. *See infra* note 111 and accompanying text.

12. *See infra* Section III.C.

13. *See infra* Section III.D.

data may help administrators determine how to allocate resources more effectively and engage in quality enhancement activities. Furthermore, media attention focused on healthcare inequities and inefficiencies may catalyze positive policy changes.

At the same time, public access policies are not devoid of risks. First, the possibility of privacy breaches can never be fully eliminated.[14] No matter how carefully data custodians de-identify patient information, at least a small risk of re-identification will always remain. If data holders do not thoroughly anonymize data, the risk of re-identification grows exponentially.[15] Furthermore, the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule does not cover most entities that operate public-use databases, and, therefore, those entities are not subject to detailed privacy regulations.[16] Second, open data may enable discrimination by employers, financial institutions, and anyone with a stake in people's health.[17] These entities may attempt to re-identify publicly available health records that belong to applicants or to employees. In the alternative, they may mine medical data to find statistical associations between particular attributes, habits, or behaviors (for example, obesity or smoking) and health risks. Then, based on their findings, entities could formulate discriminatory policies that exclude from employment, financial, or other opportunities individuals they perceive as high-risk.[18] Third, amateurs may reach incorrect conclusions and foster misconceptions among the public about human health or the healthcare industry. Amateurs could disseminate their findings broadly through the Internet without the filter mechanism of having articles reviewed and accepted by peer-reviewed journals.[19] While some errors will be innocent, others might be intentional, with data manipulated to promote personal agendas, such as maligning certain ethnic groups, hurting business competitors, or supporting particular political viewpoints. In turn, parties who believe that they have been hurt by adverse research findings may initiate litigation, asserting claims such as defamation or interference with

14. *See infra* Section IV.A.

15. *See* Sharona Hoffman & Andy Podgurski, *Balancing Privacy, Autonomy, and Scientific Needs in Electronic Health Records Research*, 65 SMU L. REV. 85, 105–107 (2012) (discussing re-identification). For further discussion, see *infra* Sections IV.A.2, IV.A.3.b and IV.A.3.c.

16. *See infra* Section IV.A.3.a.

17. *See* Sharona Hoffman & Andy Podgurski, *In Sickness, Health, and Cyberspace*, 48 B.C. L. REV. 331, 334–35 (2007) (discussing the many parties who might be interested in obtaining medical information about individuals).

18. *See infra* Section IV.B.

19. *See infra* Section IV.C.

economic advantage.[20] In some cases, parties will bring lawsuits merely to intimidate and harass citizen scientists and will needlessly burden the courts.[21]

It is too early to tell whether the benefits of open data will outweigh the risks. However, it is noteworthy that the research projects contemplated in this Article will not be subject to the federal research regulations. The regulations exempt studies based on records or data that are publicly available, and they apply only to studies funded or conducted by federal agencies or submitted to the Food and Drug Administration (FDA) in support of applications for marketing approval.[22] Citizen scientists will therefore operate in a regulatory vacuum with no governing standards or processes for approval and monitoring. This Article argues for the implementation of moderate safeguards and oversight mechanisms that will balance the needs of all stakeholders: patients, researchers, clinicians, industry, federal and state governmental entities, and the public at large.[23]

The Article will proceed as follows. Part II will sample some of the many data collections that various government and private entities have already made publicly available, examining their content and any requirements for data use. Part III will analyze the benefits of public access to medical data, and Part IV will assess its risks. Part V will formulate a detailed proposal for legal and policy interventions designed to promote responsible health data stewardship and to protect those impacted by open data. The first set of recommendations addresses privacy concerns and includes changes to the HIPAA Privacy Rule; establishment of data release review boards; and requirements for data use agreements, privacy training, registries, and consent procedures. Other recommendations in Part V call for clarification and modest expansion of anti-discrimination protections; suggest the development of research guidance, peer review, and publication opportunities for citizen scientists; and address litigation and liability avoidance strategies pertaining to public-use data.

---

20. *See infra* Section IV.D.

21. *See id.*

22. *See* 45 C.F.R. § 46.101(a) (2013) (stating that the regulations apply to "all research involving human subjects conducted, supported or otherwise subject to regulation by any federal department or agency"); 21 C.F.R. § 50.1 (describing the FDA regulations' scope of coverage); 45 C.F.R. § 46.101(b)(4) (2013) (exempting "[r]esearch involving the collection or study of existing data, documents, records, pathological specimens, or diagnostic specimens, if these sources are publicly available or if the information is recorded by the investigator in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects").

23. *See infra* Part V.

II. PUBLICLY AVAILABLE BIG DATA SOURCES

Many large databases offer public access to patient-related health information. Federal and state governments as well as private sector enterprises have established these databases. No comprehensive catalogue of these sources exists. This Part lists a representative sample of databases that feature public-use medical data.

A. FEDERAL AND STATE DATABASES

1. Federal Government Data at HealthData.gov

HealthData.gov, launched in 2011, is a Department of Health and Human Services website that makes over 1000 data sets available to researchers, entrepreneurs, and the public free of charge.[24] It predates Executive Order 13,642 by two years and establishes a home for the federal government's open data. Several states and federal government agencies such as the Centers for Disease Control and Prevention (CDC), the Centers for Medicare & Medicaid Services (CMS), the National Institutes of Health (NIH), and the Administration for Children and Families provide the data sets.[25]

All users can search for information by key words, agency type, and subject area.[26] As just one example, users can access a table entitled "Vaccination coverage among children 19–35 months of age for selected diseases, by race, Hispanic origin, poverty level, and location of residence in metropolitan statistical area."[27] HealthData.gov offers many interactive analysis tools and will continue to grow and be refined over the coming years.[28] Users can access a number of separate federal agency databases

24. *About*, HEALTHDATA.GOV, www.healthdata.gov/content/about (last visited Nov. 23, 2015).

25. *Id.*

26. HEALTHDATA.GOV, http://healthdata.gov (last visited Nov. 23, 2015). The subject areas listed are administrative, biomedical research, children's health, epidemiology, healthcare cost, healthcare providers, Medicaid, Medicare, population statistics, quality measurement, safety, treatments, and other.

27. Ctrs. for Disease Control & Prevention, *Vaccination Coverage Among Children 19–35 Months of Age for Selected Diseases*, HEALTHDATA.GOV (Oct. 14, 2015), http://www.healthdata.gov/dataset/selected-trend-table-health-united-states-2011-vaccination-coverage-among-children-19–35.

28. *See, e.g.*, Harnam Singh, *The National Practitioner Data Bank (NPDB) Introduces Interactive Data Analysis Applications*, HEALTHDATA.GOV (May 29, 2014), http://healthdata.gov/blog/national-practitioner-data-bank-npdb-introduces-interactive-data-analysis-applications; Damon Davis et al., *Health Data Initiative Strategy & Execution Plan Released and Ready for Feedback*, HEALTHDATA.GOV (Oct. 23, 2013), http://www.healthdata.gov/blog/health-data-initiative-strategy-execution-plan-released-and-ready-feedback.

through Healthdata.gov. The CDC database, CDC Wonder,[29] and the CMS database, Chronic Condition Data Warehouse,[30] are discussed below.

a) CDC Wonder

CDC Wonder enables researchers and the public at large to access a wide variety of public health information.[31] This includes data sets about deaths, births, cancer, HIV and AIDS, tuberculosis, vaccinations, census data, and more.[32] The website features statistical research data, reference material, reports, and guidelines related to public health.[33] Users conduct queries by selecting items from drop-down menus and completing fill-in-the-blank forms.[34] Prior to receiving data, users must read a short "data use restrictions" screen and click "I agree," thereby promising to comply with instructions concerning data use and disclosure that are designed to protect the privacy of data subjects.[35]

b) Chronic Condition Data Warehouse

The CMS established the Chronic Condition Data Warehouse (CCW) to allow users to purchase data about Medicare and Medicaid beneficiaries and claims.[36] Researchers can apply for access to identifiable or partially identifiable data, and CCW administrators scrutinize all requests.[37] CCW also offers public-use files that contain aggregated summary level health information for which no data use agreement or

29. *See* Ctrs. for Disease Control & Prevention, *CDC Wonder: Births*, HEALTHDATA.GOV (Oct. 30, 2015), http://healthdata.gov/dataset/cdc-wonder-births-0.

30. Dep't of Health & Human Servs., *Chronic Condition Data Warehouse*, HEALTHDATA.GOV (Oct. 30, 2015), http://www.healthdata.gov/dataset/chronic-condition-data-warehouse.

31. *What Is CDC Wonder?*, CDC WONDER, http://wonder.cdc.gov/wonder/help/main.html#What is WONDER (last updated Jan. 25, 2016).

32. *Id.*

33. *Id.*

34. *Id.*

35. *See, e.g.*, *About Natality, 2007–2013*, CDC WONDER, http://wonder.cdc.gov/natality-current.html (last visited Nov. 23, 2015). See *infra* note 309 and accompanying text for further discussion of data use agreements.

36. CENTERS FOR MEDICARE & MEDICAID SERVICES, CHRONIC CONDITIONS DATA WAREHOUSE, https://www.ccwdata.org/web/guest/home (last visited Nov. 23, 2015).

37. *CMS Data Request Center*, RESEARCH DATA ASSISTANCE CENTER, http://www.resdac.org/cms-data/request/cms-data-request-center (last visited Nov. 23, 2015).

privacy board review is required.[38] For example, the Medicaid State Drug Utilization File contains information about outpatient drugs for which state Medicaid agencies have paid.[39]

2. State Government Health Data Websites

Like the federal government, many states offer publicly available health data on government websites. Examples are Health Data New York,[40] New Jersey State Health Assessment Data,[41] North Carolina State Center for Health Statistics,[42] FloridaHealthFinder.gov,[43] and Minnesota Center for Health Statistics.[44] All these websites provide a wealth of information free of charge to the public and offer a variety of interactive tools and query options.

3. Healthcare Cost and Utilization Project

The Healthcare Cost and Utilization Project (HCUP) is sponsored by the Agency for Healthcare Research and Quality[45] and offers a variety of databases for purchase. These include the following:

- Nationwide Inpatient Sample
- Kids' Inpatient Database
- Nationwide Emergency Department Sample
- State Inpatient Databases
- State Ambulatory Surgery Databases
- State Emergency Department Databases[46]

38. *Public Use Files (PUF)/Non-Identifiable Data Requests*, RESEARCH DATA ASSISTANCE CENTER, http://www.resdac.org/cms-data/request/public-use-files (last visited Nov. 23, 2015).

39. *Medicaid State Drug Utilization File*, RESEARCH DATA ASSISTANCE CENTER, http://resdac.advantagelabs.com/cms-data/files/medicaid-state-drug-utilization (last visited Nov. 23, 2015).

40. *Health Data NY*, N.Y. DEP'T OF HEALTH, https://health.data.ny.gov (last visited Nov. 23, 2015).

41. *NJSHAD: New Jersey's Public Health Data Resource*, N.J. DEP'T OF HEALTH, https://www26.state.nj.us/doh-shad/home/Welcome.html (last updated Jan. 5, 2016).

42. *Statistics and Reports*, N.C. STATE CEN. FOR HEALTH STATISTICS, http://www.schs.state.nc.us/data/minority.cfm (last updated Jan. 5, 2016).

43. *State Health Data Directory*, FLA. AGENCY FOR HEALTH CARE ADMIN., http://www.floridahealthfinder.gov/StateHealthDataDirectory (last visited Nov. 23, 2015).

44. *Selected Public Health Data Websites*, MINN. CENTER FOR HEALTH STAT., http://www.health.state.mn.us/divs/chs/countytables/resources.htm (last updated Jan. 21, 2016).

45. *Overview of HCUP*, HEALTHCARE COST & UTILIZATION PROJECT, http://www.hcup-us.ahrq.gov/overview.jsp (last updated Jan. 28, 2016).

46. *Id.*

HCUP databases offer "a core set of clinical and nonclinical information found in a typical [hospital] discharge abstract including all-listed diagnoses and procedures, discharge status, patient demographics, and charges for all patients, regardless of payer (e.g., Medicare, Medicaid, private insurance, uninsured)."[47] Patient demographics may include sex, age, and—for some states—race, but no other attributes that more directly identify patients.[48] The databases are available for purchase, and purchasers must complete a training course and sign a data use agreement prior to receiving data.[49] Users must agree to use the data solely for research and statistical purposes and not to attempt to identify any individual.[50] Those wishing to purchase information from state databases must also explain how they intend to use the data.[51] Prices may vary significantly, depending on the type of data sought and the type of entity with which the applicant is affiliated (for example, for-profit or non-profit organization), with significant discounts available to students.[52]

4. GenBank

GenBank is the National Institutes of Health's genetic sequence database, which includes all DNA sequences that are publicly available.[53] The data are free, and GenBank places no restriction on their use.[54] According to scientists at the National Center for Biotechnology Information, GenBank contains "over 900 complete genomes, including the draft human genome, and some 95,000 species."[55] Leading journals

47. *Databases and Related Tools from HCUP: Fact Sheet*, AGENCY FOR HEALTHCARE RESEARCH & QUALITY, http://archive.ahrq.gov/research/findings/factsheets/tools/hcupdata/datahcup.html (last updated Mar. 2011).

48. *Overview of the State Inpatient Databases*, HEALTHCARE COST & UTILIZATION PROJECT, http://www.hcup-us.ahrq.gov/sidoverview.jsp (last updated Jan. 20, 2016).

49. *Purchase HCUP Data*, HEALTHCARE COST & UTILIZATION PROJECT, http://www.hcup-us.ahrq.gov/tech_assist/centdist.jsp (last updated Nov. 18, 2015).

50. HEALTHCARE COST & UTILIZATION PROJECT, HCUP NATIONWIDE INPATIENT SAMPLE APPLICATION (2015), http://www.hcup-us.ahrq.gov/db/nation/nis/NISApp_Final.pdf.

51. *Purchase HCUP Data*, *supra* note 49.

52. HEALTHCARE COST & UTILIZATION PROJECT, SID/SASD/SEDD APPLICATION KIT (2015), http://www.hcup-us.ahrq.gov/db/state/SIDSASDSEDD_Final.pdf (listing prices that range from $35 to over $1600).

53. *GenBank Overview*, NAT'L CEN. FOR BIOTECH. INFO. (NCBI), http://www.ncbi.nlm.nih.gov/genbank (last visited Nov. 23, 2015).

54. *Id.*

55. Jo McEntyre & David J. Lipman, *GenBank—A Model Community Resource?*, NATURE (Apr. 5, 2001), http://www.nature.com/nature/debates/e-access/Articles/lipman.html.

now require authors to deposit their sequences in GenBank, and all publicly funded laboratories also do so as a matter of policy.[56]

GenBank provides a variety of data search and retrieval tools, such as the Basic Local Alignment Search Tool (BLAST), which finds similarities between sequences.[57] Public-use data available on GenBank have enabled scientists and commercial enterprises to conduct research and generate new products, including assemblies of the human genome produced by Celera Genomics and the University of California at Santa Cruz.[58]

5. All-Payer Claims Databases

A large number of states have launched all-payer claims databases that collect information about private and public insurance related to medical, dental, and pharmacy services.[59] Typically, the collected data include information regarding patient demographics; insurance contracts; healthcare providers; payments made by insurers and patients; dates on which medical services were received; and codes for diagnoses, procedures, and drugs.[60] Consumers, employers, and other stakeholders can access data to learn about healthcare costs, compare prices, and make more informed decisions about insurance plans and healthcare providers.[61]

Similarly, CMS has released Medicare provider utilization and payment data that is available free of charge.[62] The website offers information pertaining to the 100 most commonly performed inpatient services, thirty frequently provided outpatient services, and more.[63] Thus, for instance, users may obtain hospital-specific charges for particular services and compare prices.[64]

---

56. *Id.*

57. *Id.*; *Genbank Overview*, *supra* note 53.

58. McEntyre & Lipman, *supra* note 55.

59. JO PORTER ET AL., THE BASICS OF ALL-PAYER CLAIMS DATABASES: A PRIMER FOR STATES 1 (2014), http://www.apcdcouncil.org/sites/apcdcouncil.org/files/The%20Basics%20of%20All-Payer%20Claims%20Databases.pdf.

60. *Id.* at 2.

61. *Id.* at 3; *Colorado Medical Price Compare*, CTR. FOR IMPROVING VALUE IN HEALTH CARE, https://www.cohealthdata.org (last visited Nov. 23, 2015); *CHIA Data*, CTR. FOR HEALTH INFO. & ANALYSIS, http://www.chiamass.gov/chia-data (last visited Nov. 23, 2015) (requiring applications for Massachusetts data).

62. *Medicare Provider Utilization and Payment Data*, CTRS. FOR MEDICARE & MEDICAID SERVS., http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data (last updated Apr. 30, 2015).

63. *Id.*

64. *Medicare Provider Utilization and Payment Data: Inpatient*, CTRS. FOR MEDICARE & MEDICAID SERVS., http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Inpatient.html

B. PRIVATE SECTOR DATABASES

1. Dryad Digital Repository

Dryad is an international repository containing data files associated with peer-reviewed scientific articles and other "reputable sources (such as dissertations)."[65] It is a nonprofit organization supported by fees from its members and data submitters.[66] Researchers submit data underlying their publications directly to Dryad, and any member of the public can access the collection at no cost.[67] The website provides a search tool that allows users to enter key words or other search criteria and takes them to data associated with particular publications.[68]

2. PatientsLikeMe

PatientsLikeMe is a for-profit website that enables patients who sign up for membership to share their health data and disease experiences.[69] Users can report and obtain information about treatments and connect with others who have the same condition.[70] PatientsLikeMe acknowledges that it sells de-identified information submitted by users to its "partners," which it describes as "companies that can use that data to improve or understand products or the disease market."[71] Members may choose different privacy settings and may determine whether non-members will

(last updated June 1, 2015). *But see* Patrick T. O'Gara, *Caution Advised: Medicare's Physician-Payment Data Release*, 371 NEW ENG. J. MED. 101 (2014) (discussing the limitations of payment data released by CMS); Dawn Fallik, *For Big Data, Big Questions Remain*, 33 HEALTH AFF. 1111, 1111 (2014) (stating that "Medicare's release of practitioner payments highlights the strengths and weaknesses of digging into big data").

65. *The Organization: Overview*, DRYAD, http://datadryad.org/pages/organization (last updated Oct. 22, 2015); *Frequently Asked Questions*, DRYAD, http://datadryad.org/pages/faq#depositing (last updated Jan. 5, 2016).

66. *Pricing Plans and Data Publishing Prices*, DRYAD, http://datadryad.org/pages/pricing (last updated Jan. 5, 2016).

67. *Frequently Asked Questions*, DRYAD, http://datadryad.org/pages/faq#using (last updated Jan. 5, 2016).

68. *The Repository: Key Features*, DRYAD, http://datadryad.org/pages/repository (last updated Feb. 15, 2015).

69. PATIENTSLIKEME, http://www.patientslikeme.com (last visited Nov. 23, 2015).

70. *What Is PatientsLikeMe?*, PATIENTSLIKEME, https://support.patientslikeme.com/hc/en-us/articles/201186434-What-is-PatientsLikeMe- (last visited Nov. 23, 2015).

71. *Does PatientsLikeMe Sell My Data?*, PATIENTSLIKEME, https://support.patientslikeme.com/hc/en-us/articles/201245770-Does-PatientsLikeMe-sell-my-information- (last visited Nov. 23, 2015).

be able to view any of their data.[72] PatientsLikeMe releases reports of aggregated data concerning symptoms and treatments to the public.[73] In addition, members may opt into a public registry that will make their profiles and shared data available to anyone with access to the Internet.[74] PatientsLikeMe makes public-use information available on its website at no cost and does not require applications or data use agreements.[75]

3. The Personal Genome Project

George Church launched the Personal Genome Project in 2005 at Harvard University, and it is now an international enterprise involving thousands of patients.[76] It aims to promote research and offers genomic, environmental, and human trait information from volunteer participants to any interested party.[77] Users can easily access a wealth of information directly from the website, including genome data, genome reports, trait and survey data, participant profiles, and microbiome data.[78] Data files list the date of birth, gender, zip code, height, weight, and race of individual participants, though names are not displayed.[79] The Personal Genome Project states explicitly that its participants must be "willing to waive expectations of privacy" in order to make "a valuable and lasting contribution to science."[80]

72. *Privacy Policy*, PATIENTSLIKEME, http://www.patientslikeme.com/about/privacy (last updated Mar. 5, 2012).

73. *See, e.g.*, *Treatments*, PATIENTSLIKEME, http://www.patientslikeme.com/treatments (last updated Feb. 2, 2016).

74. *See, e.g.*, *Welcome to the PatientsLikeMe Public ALS Registry*, PATIENTSLIKEME, http://www.patientslikeme.com/registry (last visited Nov. 23, 2015); *What Information is Visible on Public Profiles?*, PATIENTSLIKEME, https://support.patientslikeme.com/hc/en-us/articles/201245830-What-information-is-visible-on-public-profiles- (last visited Nov. 23, 2015).

75. *Conditions at PatientsLikeMe*, PATIENTSLIKEME, http://www.patientslikeme.com/conditions (last updated Feb. 5, 2016).

76. *About PGP Harvard*, PERSONAL GENOME PROJECT, HARV. MED. SCH., http://www.personalgenomes.org/harvard/about-pgp (last visited Nov. 23, 2015).

77. *Id.*

78. *Id.; Data & Samples*, PERSONAL GENOME PROJECT, HARV. MED. SCH., http://www.personalgenomes.org/harvard/data (last visited Nov. 23, 2015) (Microbiome data focuses on "the types of bacteria in and on a participant's body.").

79. *See, e.g.*, *Public Genetic Data*, PERSONAL GENOME PROJECT, HARV. MED. SCH., https://my.pgp-hms.org/public_genetic_data (last visited Nov. 23, 2015).

80. *About PGP Harvard*, *supra* note 76.

III. THE BENEFITS OF PUBLIC ACCESS TO HEALTH INFORMATION

Public-use data potentially offer many valuable benefits. These include new scientific discoveries, research cost savings, new tools to help patients navigate the healthcare system, greater government transparency, public education about science and medicine, improved healthcare quality, and positive healthcare policy changes.

A. SCIENTIFIC DISCOVERY

One of the great hopes of health data sharing is that it will promote scientific discovery and medical advances. Citizen scientists may be extremely motivated and dedicated researchers, perhaps especially if they are focusing on diseases that afflict them or their loved ones. Citizen scientists who would not otherwise have access to health data and lack the means to collect original data for studies may nevertheless have the skills, talent, and creativity to make significant contributions given the appropriate data tools.[81]

In his May 2013 executive order, President Obama stated that public information resources have enabled entrepreneurs and innovators "to develop a vast range of useful new products and businesses."[82] Similarly, proponents of DIYbio enthuse that it "can inspire a generation of bioengineers to discover new medicines, customize crops to feed the world's exploding population, harness microbes to sequester carbon, solve the energy crisis, or even grow our next building materials."[83]

Citizen scientists have proven themselves to be capable inventors whose contributions aid many people. For example, three Dutch DIY biologists created Amplino, an inexpensive diagnostic system that can be used in developing countries to detect malaria with a single drop of blood in less than forty minutes.[84] Likewise, Katherine Aull, a graduate of the Massachusetts Institute of Technology whose father suffered from

81. Huseyin Naci & John P. A. Ioannidis, *Evaluation of Wellness Determinants and Interventions by Citizen Scientists*, 314 JAMA 121, 122 (2015), http://jama.jamanetwork.com/article.aspx?articleid=2330497.

82. Exec. Order No. 13,642, *supra* note 1.

83. Grushkin et al., *supra* note 6, at 4.

84. Thomas Landrain et al., *Do-It-Yourself Biology: Challenges and Promises for an Open Science and Technology Movement*, 7 SYST. SYNTHETIC BIOLOGY 115, 121 (2013); Linda Nordling, *DIY Biotech: How to Build Yourself a Low-Cost Malaria Detector*, GUARDIAN (Apr. 25, 2014), http://www.theguardian.com/global-development-professionals-network/2014/apr/25/diy-detector-malaria-eradication-amplino (reporting that Amplino "is almost ready for field-testing in rural Zambia").

hemochromatosis, a condition that causes the body to absorb excessive amounts of iron and can permanently damage vital organs, developed a homemade genetic test to determine whether she was vulnerable to this inherited disease.[85] She built a lab in her closet and used equipment purchased from eBay or found in her kitchen.[86]

New troves of publicly available data promise to facilitate and accelerate the work of professional researchers and citizen scientists. Public data sources have already led to important discoveries. For example, Project Tycho is a University of Pittsburgh initiative designed to promote the availability and use of public health data by facilitating its analysis and redistribution.[87] Tycho researchers have digitized disease surveillance data from the years 1888 to 2011 published in the CDC's *Morbidity and Mortality Weekly Report* and estimate that since 1924, 103 million incidents of childhood diseases were prevented because of immunizations.[88] This finding will be useful for public health authorities, who at times meet resistance to vaccination efforts.

Among the more creative initiatives was a crowdsourcing contest called the Dialogue on Reverse Engineering Assessment and Methods (DREAM7) focused on breast cancer prognosis.[89] Crowdsourcing can be defined as "a participative online activity in which an individual, an institution, a non-profit organization, or company proposes to a group of individuals . . . via a flexible open call, the voluntary undertaking of a task."[90] DREAM7 provided participants with access to genetic and clinical data from Sage's Synapse, an informatics platform that allows users to

85. Ana Delgado, *DIYbio: Making Things and Making Futures*, 48 FUTURES 65, 70 (2013); *Biopunks Tinker with the Building Blocks of Life*, NPR (May 19, 2011), http://www.npr.org/2011/05/22/136464041/biopunks-tinker-with-the-building-blocks-of-life.

86. Delgado, *supra* note 85, at 70.

87. *About Project Tycho Data*, U. OF PITTSBURGH, https://www.tycho.pitt.edu/about.php (last visited Nov. 23, 2015).

88. Willem G. van Panhuis et al., *Contagious Diseases in the United States from 1888 to the Present*, 369 NEW ENG. J. MED. 2152, 2156 (2013).

89. Michael Eisenstein, *Crowdsourced Contest Identifies Best-In-Class Breast Cancer Prognostic*, 7 NATURE BIOTECH. 578, 578 (2013).

90. Enrique Estellés-Arolas & Fernando González-Ladrón-de-Guevara, *Towards an Integrated Crowdsourcing Definition*, 38 J. INFO. SCI. 189, 197 (2012); *see also* Thea C Norman et al., *Leveraging Crowdsourcing to Facilitate the Discovery of New Medicines*, 3 SCI. TRANSLATIONAL MED. mr1, 2 (2011) (defining crowdsourcing as "the act of outsourcing tasks traditionally performed by an employee to an undefined, large group of people or community (a 'crowd')").

share data and access programming codes and analytical tools.[91] The contest challenged the crowd to "provide an unbiased assessment of models and methodologies for the prediction of breast cancer survival."[92] A winner was selected from among 1400 entries, and results were published in a scientific journal.[93]

Crowdsourcing is an increasingly popular phenomenon.[94] It has been used for projects ranging from locating over 1400 automated external defibrillators in public places in Philadelphia to developing a predictive algorithm for regions of local similarity between genetic sequences that is superior to the NIH's standard algorithm, BLAST.[95] The availability of vast amounts of publicly accessible data may make crowdsourcing all the more prevalent. Researchers will likely continue to harness the talents and expertise of citizen scientists to make important contributions to medical science.[96]

B. RESEARCH COST REDUCTIONS

Open data resources will be of particular value in an era of diminished research funding. NIH appropriations peaked at $36.4 billion in fiscal year 2010 thanks to funding from the American Recovery and Reinvestment Act, but they declined to $29.9 billion by fiscal year 2014. In 2014, the NIH funded 18.1% of grant proposals compared to 31.5% in 2000.[97]

At the same time, despite the abundance of information and medical technology available in the twenty-first century, "more than half of medical treatments are used without sufficient proof of their

91. SYNAPSE, ABOUT SYNAPSE (2013), https://s3.amazonaws.com/static.synapse.org/About_Synapse.pdf.

92. GUSTAVO STOLOVITZSKY & ANDREA CALIFANO, *DREAM CHALLENGE* (2013), http://www.slideshare.net/tulipnandu/dream-challenge.

93. Eisenstein, *supra* note 89, at 578.

94. Benjamin M. Good & Andrew I. Su, *Crowdsourcing for Bioinformatics*, 29 BIOINFORMATICS 1925, 1925 (2013).

95. *The Accelerating World of Drug Discovery and Commercialization*, TRENDS MAG., Oct. 2013, at 30 (2013); *Basic Local Assignment Search Tool (BLAST)*, NAT'L CENTER FOR BIOTECHNOLOGY INFORMATION, http://blast.ncbi.nlm.nih.gov/Blast.cgi (last visited Nov. 23, 2015).

96. Benjamin L. Raynard et al., *Crowdsourcing—Harnessing the Masses to Advance Health and Medicine, A Systematic Review*, 29 J. GEN. INTERNAL MED. 187, 187 (2014) (concluding that "[u]tilizing crowdsourcing can improve the quality, cost, and speed of a research project while engaging large segments of the public and creating novel science").

97. *Research Project Success Rates by NIH Institute for 2014*, U.S. DEP'T OF HEALTH & HUMAN SERVICES, http://www.report.nih.gov/success_rates/Success_ByIC.cfm (last updated Mar. 22, 2012).

effectiveness."[98] For example, experts have recently raised new questions about the efficacy of mammography, a well-established practice that was long considered life-saving and a key element of preventive medicine.[99] Likewise, although physicians have prescribed and studied hormone replacement therapy for post-menopausal women for decades, experts are still unsure as to whether it is advisable or whether its risks outweigh its benefits, at least for some subgroups of patients.[100] A third illustration is a debate over the risks of a particular class of antidepressants called selective serotonin reuptake inhibitors (SSRIs) in light of evidence that they may induce suicidal thoughts and behavior in adolescent patients.[101] No consensus has formed regarding this side effect, and further study is necessary.[102]

Professional researchers and citizen scientists will be able to use open data to reduce the expense of clinical trials and to conduct low-cost records-based research. While many will focus on well-known and widespread health problems, open data may also stimulate the study of subjects for which little to no public funding is available. For example, because of vigorous lobbying by the National Rifle Association, the CDC was prohibited for many years from analyzing the impact of firearms on public health.[103] Similarly, there is often limited interest in or funding for research relating to rare diseases.[104] Citizen scientists, however, may be highly motivated, for personal rather than profit-seeking reasons, to research those diseases.

98. Eric B. Larson, *Building Trust in the Power of "Big Data" Research to Serve the Public Good*, 309 JAMA 2443, 2444 (2013).

99. Nikola Biller-Andorno & Peter Jüni, *Abolishing Mammography Screening Programs? A View from the Swiss Medical Board*, 370 NEW ENG. J. MED. 1965, 1965–67 (2014).

100. HERBERT I. WEISBERG, BIAS AND CAUSATION: MODELS AND JUDGMENT FOR VALID COMPARISONS 18–21 (2010) (noting that the risks may include slight elevations in the incidence of coronary heart disease and breast cancer).

101. *Id.* at 21–23.

102. *Id.*

103. Michael Luo, *Sway of N.R.A. Blocks Studies, Scientists Say*, N.Y. TIMES (Jan. 26, 2011), http://www.nytimes.com/2011/01/26/us/26guns.html. The moratorium was lifted by a memorandum issued by President Obama in January of 2013. Memorandum, Engaging in Public Health Research on the Causes and Prevention of Gun Violence, 78 Fed. Reg. 4295 (Jan. 16, 2013).

104. NAT'L ORG. FOR RARE DISORDERS, RESEARCH GRANT POLICY (2015), https://rarediseases.org/wp-content/uploads/2015/05/NORD-Research-Grant-Policy.pdf.

The gold standard of medical research has traditionally been randomized, controlled clinical trials.[105] Phase 3 clinical trials, conducted as the final step before approval of a drug, cost an average of $20 million, involve 300 to 3000 people, and last one to four years.[106] These experimental studies are conducted through "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."[107] Thus, investigators might design a clinical trial to compare two drugs for a particular ailment or to compare a drug to a placebo. If researchers share data from prior clinical trials, they may be able to improve study quality and efficiency by honing in on patient sub-groups that are most likely to be responsive to the drug in question.[108] For example, a bladder cancer study determined that one participant who responded unusually well to the drug everolimus had a particular genetic mutation, and thus future testing of the drug could focus on subjects with that mutation to determine whether it enhances responsiveness to the drug.[109]

In the alternative, researchers can undertake observational studies by reviewing existing records and data sets rather than conducting experiments.[110] Professional researchers and citizen scientists will be able to use the large quantities of open data that are now becoming available to

105. Friedrich K. Port, *Role of Observational Studies Versus Clinical Trials in ESRD Research*, 57 KIDNEY INT'L (SUPPLEMENT 74) S3, S3 (2000), http://www.kidney-international.org/article/S0085-2538(15)47033-4/pdf (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 clinical trial designs).

106. U.S. DEP'T OF HEALTH & HUMAN SERVICES, OFFICE OF THE ASSISTANT SECRETARY FOR PLANNING AND EVALUATION, EXAMINATION OF CLINICAL TRIAL COSTS AND BARRIERS FOR DRUG DEVELOPMENT (2014), http://aspe.hhs.gov/report/examination-clinical-trial-costs-and-barriers-drug-development; *Step 3: Clinical Research*, U.S. FDA, (2015), http://www.fda.gov/ForPatients/Approvals/Drugs/ucm405622.htm. These sources also discuss the earlier stages of clinical trials, Phase 1 and Phase 2.

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

108. Eisenstein, *supra* note 89, at 580.

109. *Id.*; Gopa Iyer et al., *Genome Sequencing Identifies a Basis for Everolimus Sensitivity*, 338 SCIENCE 221, 221 (2012).

110. Observational studies involve the review of existing records or data. *See* CHARLES P. FRIEDMAN & JEREMY C. WYATT, EVALUATION METHODS IN BIOMEDICAL INFORMATICS 369 (Kathryn J. Hannah & Marion J. Ball eds., 2nd ed. 2006) (defining observational studies as involving an "[a]pproach to study design that entails no experimental manipulation").

minimize research expenses. Researchers may find that existing data collections contain all the raw data that they need and be spared the work and cost of recruiting human subjects for original data. Public-use data can thus prevent costly duplication of effort.[111]

Furthermore, an emerging trend called crowdfunding can fund relatively inexpensive big data projects.[112] Crowdfunding is an Internet-based method of fundraising by which one can solicit money from numerous donors, who usually contribute small amounts.[113] Typically, crowdfunding for scientific projects raises less than $10,000,[114] but enterprising fund-raisers have frequently surpassed that sum.[115] Public-use data may enable a growing number of projects to have very limited costs that researchers can cover creatively rather than through the traditional channels of government-allocated grant awards.

---

111. CDC, CDC-GA-2005-14, CDC/ATSDR POLICY ON RELEASING AND SHARING DATA 5–6 (2005), http://www.cdc.gov/maso/Policy/ReleasingData.pdf.

112. Vural Özdemir et al., *Crowd-Funded Micro-Grants for Genomics and "Big Data": An Actionable Idea Connecting Small (Artisan) Science, Infrastructure Science, and Citizen Philanthropy*, 17 OMICS 161, 162 (2013).

113. Stuart R. Cohn, *New Crowdfunding Registration Exemption: Good Idea, Bad Execution*, 64 FLA. L. REV. 1433, 1434 (2012).

114. Rachel E. Wheat et al., *Raising Money for Scientific Research Through Crowdfunding*, 28 TRENDS ECOLOGY & EVOLUTION 71, 72 (2013), http://jarrettbyrnes.info/pdfs/Wheat_et_al_2012.pdf.

115. Ethan O. Perlstein, *Anatomy of the Crowd4Discovery Crowdfunding Campaign*, 2 SPRINGERPLUS 560, 561 (2013), http://www.springerplus.com/content/pdf/2193-1801-2-560.pdf (reporting that the authors raised $25,460 from 390 donors in 15 countries for a pharmacological research project); Joe Palca, *Scientists Get Research Donations from Crowd Funding*, NPR (Mar. 15, 2013), http://www.npr.org/2013/02/14/171975368/scientist-gets-research-donations-from-crowdfunding (reporting that UBiome and American Gut together raised over $600,000 for projects designed to discover how microbiomes (tiny organisms that reside in the human body) influence health when donors were promised an analysis of the bacteria in their own digestive tracts). The Internet offers a large number of platforms for crowdfunding, including the aptly named Kickstarter, Experiment, and Indiegogo, among others. *See* KICKSTARTER, https://www.kickstarter.com; EXPERIMENT, https://experiment.com; INDIEGOGO, https://www.indiegogo.com. Crowdfunding has become so popular that it is being used not only by enterprising individuals and companies but also by several universities, such as the University of Virginia and Tulane, that are seeking to compensate for the dearth of funding from traditional sources. Morgan Estabrook, *New Crowdfunding Site Allows Public to Advance U. Va. Research Projects Through Targeted Donations*, UVA TODAY (May 15, 2013), http://news.virginia.edu/content/new-crowdfunding-site-allows-public-advance-uva-research-projects-through-targeted-donations; Keith Brannon, *Tulane University Launches Crowdfunding Partnership for Medical Research*, TULANE U. (Dec. 10, 2013), http://tulane.edu/news/releases/pr_12102013.cfm. To enhance their likelihood of success and attract donors, those pursuing crowdfunding are well-advised to post convincing videos on funding websites and to follow up with blog entries and media coverage of their projects, to the extent possible. Perlstein, *supra*, at 561.

C. TOOLS TO HELP PATIENTS NAVIGATE THE HEALTHCARE SYSTEM

Open health data can promote not only research but also services that are helpful for patients. Several enterprises are developing tools to help patients obtain suitable and affordable medical care. Aidin is a small startup that uses CMS data on health facilities and nursing homes to provide hospitals and patients with information about options for care after discharge from the hospital.[116] Aidin offers its clients listings of available providers, quality of care ratings, and reviews. It also helps hospitals track patient experiences and outcomes so that they can determine which providers are the best fit for patients with specific health conditions.[117]

Similarly, iTriage is a free mobile app and website that allows patients to look up their symptoms and learn about possible causes and treatments.[118] In addition, it assists patients in locating and selecting appropriate care options by providing a variety of information, including hospital wait times and physician ratings.[119] iTriage uses publicly available data from HHS, the FDA, and other sources.[120]

Other examples are the state all-payer claims databases, Medicare's Provider Utilization and Payment Data, and Medicare's Hospital Compare.[121] These educate patients about healthcare costs and quality and allow patients to compare prices for various inpatient and outpatient services.[122]

D. GOVERNMENT TRANSPARENCY AND PUBLIC EDUCATION

Proponents of government transparency will be pleased by the proliferation of open data. Databases such as HealthData.gov, Genbank, and others[123] allow viewers to gain significant insight into the information that the government has collected about individuals and the healthcare

116. *Former Sec. Sebelius Celebrates Aidin in Annual "Health Datapalooza" Speech*, AIDIN (June 12, 2014), http://www.myaidin.com/articles/june_2014/002.html; *Our Story*, AIDIN, http://www.myaidin.com/ourstory.html (last visited Nov. 23, 2015).

117. *Our Story*, AIDIN, *supra* note 116.

118. *What is iTriage?*, ITRIAGE, https://about.itriagehealth.com/itriage-what-is (last visited Nov. 23, 2015).

119. *Id.*

120. *About Our Medical Content*, ITRIAGE, https://about.itriagehealth.com/company-info/medical-content (last visited Nov. 23, 2015).

121. *See supra* Section II.A.5; *Hospital Compare*, MEDICARE.GOV, http://www.medicare.gov/hospitalcompare/search.html (last visited Nov. 23, 2015).

122. *See Hospital Compare*, *supra* note 121.

123. *See supra* Part II.

industry. In some cases, such insight may generate public debate and critique of government investigative policies that could lead to positive policy changes.[124]

In addition, public-use data can function as an important educational tool.[125] Patients can research their own conditions, find doctors with special expertise, better prepare for their medical appointments, and assess different treatment options that they are given.[126] Furthermore, the general public can learn about the healthcare system, healthcare costs, disease trends, genetics, research and public health initiatives, and much more.[127] Ordinary citizens and students will be able to access raw data themselves and engage in research exercises, either within the framework of academic programs or on their own. For example, the New York University School of Medicine is leveraging open data resources to enhance its curriculum. It is creating patient snapshots from New York hospital discharge data and developing sophisticated training tools based on these cases.[128] Active learning and engagement with health data might also inspire greater public enthusiasm about medical research and more vocal support for government funding of this vital activity.

E. IMPROVEMENTS IN HEALTHCARE QUALITY AND PUBLIC HEALTH POLICY

Open data can fuel improvements in healthcare quality and health policies. A report from New York State provided a number of compelling illustrations.[129] In 2011, in preparation for Hurricane Irene, nursing home administrators used publicly available weekly bed census reports to identify facilities to which they could evacuate residents.[130] Likewise, annual

124. CDC/ATSDR POLICY ON RELEASING AND SHARING DATA, *supra* note 111, at 4 (stating that data sharing can "build trust with outside partners and the public by allowing open critique of CDC investigations").

125. GRUSHKIN, *supra* note 6, at 4 (stating that "wider access to the tools of biotechnology, particularly those related to the reading and writing of DNA, has the potential to spur global innovation and promote biology education and literacy").

126. Internet searches, however, should not replace consultation with medical experts, and often have pitfalls. Patients should not panic based on their independent research and become convinced that they suffer from a dreaded disease or have a poor prognosis before being examined by a physician. Patients also should not go to the doctor with a closed mind, unwilling to accept the expert's own assessment and treatment recommendations.

127. *See supra* Part II.

128. Erika G. Martin et al., *Liberating Data to Transform Healthcare: New York's Open Data Experience*, 311 JAMA 2481, 2481 (2014).

129. *Id.*

130. *Id.*

reports of cardiac surgery mortality rates, linked to the hospitals and surgeons who provide care, induced low-scoring facilities to undertake quality improvement initiatives and several physicians who performed poorly to leave practice.[131] A different study, published in 2015 in *Health Affairs*, concluded that Medicare's Hospital Compare "slowed the rate of price increases in a majority of states that had not previously been exposed to comparable information through their own public reporting systems."[132]

Once data are released, they are available not only to the general public, but also to the media. Media stories about health-related inequities can be particularly potent tools to effect policy changes. After officials released New York childhood obesity statistics that were organized by school district, news outlets highlighted the disparities in 2013 and some school administrators decided to improve their policies despite cost concerns.[133] A 2014 report in *Crain's New York Business* that publicized hospital cost disparities (for example, hip replacements that cost $103,725 at New York University Hospitals Center but only $15,436 at Bellevue Hospital Center) is likewise expected to catalyze pricing and reimbursement changes.[134]

IV. RISKS OF PUBLIC ACCESS TO HEALTH DATA

Although the benefits of opening health data resources to the public are considerable, the risks are not inconsequential. The federal research regulations cover only studies that are funded or conducted by federal government agencies or that do not use publicly available data.[135] Therefore, studies without federal funding and ones that use publicly available data are not subject to any formal oversight. Furthermore, the HIPAA Privacy Rule and state privacy laws most likely will not govern open databases.[136] This Part analyzes several potential risks associated with open access to patient-related health information: 1) privacy breaches; 2) discrimination and special targeting by employers, financial institutions, and marketers, among others; 3) propagation of incorrect and harmful research conclusions; and 4) litigation.

131. *Id.*

132. Avi Dor et al., *Medicare's Hospital Compare Quality Reports Appear to Have Slowed Price Increases for Two Major Procedures*, 34 HEALTH AFF. 71, 75 (2015) (focusing on coronary artery bypass grafts and percutaneous coronary interventions).

133. Martin et al., *supra* note 128, at 2481.

134. *Id.*

135. *See* 45 C.F.R. §§ 46.101(a), 46.101(b)(4) (2013).

136. *See infra* Section IV.A.1.

A. PRIVACY THREATS

I recently logged onto the Personal Genome Project and looked at the Participant Profiles section.[137] To my surprise, several profiles disclosed the names of patients along with their date of birth, sex, weight, height, blood type, race, health conditions, medications, allergies, procedures, and more.[138] I wondered if these patients understood that anyone with a computer could view all of this information. Other profiles excluded the name of the participant but provided all of the other details, which could potentially allow a clever and motivated viewer to identify the patient.

Privacy threats are the first risk that may come to mind with respect to public use of patient-related medical big data. The HIPAA Privacy Rule,[139] the Privacy Act,[140] and numerous state privacy laws govern the disclosure of medical records.[141] However, the laws and regulations do not cover all data holders who make medical information publicly available.[142] In addition, public-use data is most often presented in de-identified form[143] and thus is exempt from the disclosure restrictions established in these laws and regulations.[144] Moreover, even with thorough de-identification, at least a small risk of re-identification remains. Privacy concerns thus deserve thorough analysis.

1. Privacy Law

Many federal and state laws address medical privacy. None of the laws, however, provide patients with comprehensive protection, and even in the aggregate, they leave many gaps. The following discussion describes laws and regulations relevant to the disclosure of patient-related data for public use.

---

137. *Participant Profiles*, PERSONAL GENOME PROJECT, HARV. MED. SCH., https://my.pgp-hms.org/users (last visited Nov. 23, 2015).

138. *Public Profile -- hu43860C*, PERSONAL GENOME PROJECT, HARV. MED. SCH., https://my.pgp-hms.org/profile/hu43860C (last updated Sept. 4, 2015).

139. 45 C.F.R. §§ 160.101–.534 (2013).

140. 5 U.S.C. § 552a (2010).

141. *See* AMERICANS HEALTH LAWYERS ASSOCIATION, STATE HEALTHCARE PRIVACY LAW SURVEY (2013); Sarah Hexem, *Public Health Departments and State Patient Confidentiality Laws Map*, LAWATLAS, http://lawatlas.org/preview?dataset=public-health-departments-and-state-patient-confidentiality-laws (last visited Nov. 23, 2015).

142. *See infra* Sections IV.A.1 and IV.A.3.a.

143. *See supra* Part II.

144. *See infra* Section IV.A.1.

a) The HIPAA Privacy Rule

The HIPAA Privacy Rule establishes that, with some exceptions, entities covered by the regulations must obtain patients' permission before disclosing their medical information to third parties.[145] The Rule, however, covers only health plans, healthcare clearinghouses, healthcare providers who transmit health information electronically for purposes of HIPAA-relevant transactions, and their business associates.[146] It does not apply to government agencies or private enterprises that are not acting in these capacities. Thus, HIPAA does not regulate many of the websites discussed in Part II of this Article, such as those operated by state governments, CDC, Dryad or PatientsLikeMe.

Moreover, the HIPAA Privacy Rule protects only "individually identifiable health information" that is electronically or otherwise transmitted or maintained.[147] Consequently, the federal regulations do not govern data that custodians de-identify[148] and open to the public.

b) The Privacy Act

The Privacy Act is a federal law that governs the collection, storage, use, and disclosure of information by the federal government.[149] The law provides that the federal government may not disclose records without the data subject's permission, unless specific exceptions apply. However, the Privacy Act defines the term "record" as an item that contains a person's "name, or the identifying number, symbol, or other identifying particular assigned to the individual."[150] Consequently, the Privacy Act exempts the government's dissemination of de-identified information on HealthData.gov or other websites.

c) State Laws

All states have recognized a common law or statutory right to privacy[151] and have statutes that address privacy concerns.[152] A thorough

145. 45 C.F.R. §§ 164.508–.510 (2013).
146. 45 C.F.R. §§ 160.102–.103 (2013); 42 U.S.C. § 17934 (2010).
147. 45 C.F.R. § 160.103 (2013).
148. *See infra* Section IV.A.2 (discussing HIPAA's requirements for de-identification).
149. 5 U.S.C. § 552a (2010).
150. *Id.* at § 552a(a)(4).
151. *See* Corrine Parver, *Patient-Tailored Medicine, Part Two: Personalized Medicine and the Legal Landscape*, 2 J. HEALTH & LIFE SCI. L. 1, 32 (2009).
152. *See* AMERICANS HEALTH LAWYERS ASSOCIATION, *supra* note 141; LAWATLAS, *supra* note 141.

analysis of state law is beyond the scope of this Article.[153] In general, state laws are varied and inconsistent, often providing piecemeal protection for some types of data but not others.[154] Moreover, like the HIPAA Privacy Rule and the Privacy Act, states typically allow disclosure of de-identified health information without patient authorization.[155] Therefore, most of the public-use data resources contemplated in this Article would not be governed by state law.

2. De-identification

The foregoing discussion raises the following critical question: what does "de-identified" mean, and how can data holders achieve de-identification? The HIPAA Privacy Rule provides a detailed answer. It states that health information is de-identified if (1) a qualified expert determines that there is only a "very small" risk that the data can be re-identified, and (2) the expert documents his or her analysis.[156] The Department of Health and Human Services issued guidance that endorsed several de-identification techniques:

- *Suppression*, which involves redaction of particular data features prior to disclosure (e.g., removing zip codes, birthdates, income);
- *Generalization*, which involves transforming particular information into less specific representations (e.g., indicating a 10-year age range instead of exact age); and
- *Perturbation*, which involves exchanging certain data values for equally specific but different values (e.g., changing patients' ages).[157]

153. For detailed information about state privacy and confidentiality laws, s*ee* AMERICANS HEALTH LAWYERS ASSOCIATION, *supra* note 141; LAWATLAS, *supra* note 141.

154. *See* Deven McGraw et al., *Privacy as an Enabler, Not an Impediment: Building Trust into Health Information Exchange*, 28 HEALTH AFF. 416, 420 (2009) (noting that "[a]lthough the states have an important role to play in privacy policy, state privacy laws are fragmentary and inconsistent, providing neither developers nor consumers with the assurances they deserve, especially for services of nationwide reach").

155. Scott Burris et al., *The Role of State Law in Protecting Human Subjects of Public Health Research and Practice*, 31 J.L. MED. & ETHICS 654, 656 (2003).

156. 45 C.F.R. § 164.514(b)(1) (2013).

157. *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule*, U.S. DEP'T OF HEALTH & HUMAN SERVS. (Nov. 26, 2012), http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html

In the alternative, according to the HIPAA Privacy Rule, de-identification is achieved 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

#guidancedetermination (noting that techniques such as suppression and generalization are often used in combination).

> (R) Any other unique identifying number, characteristic, or code . . . .[158]

Health information that has all eighteen identifiers removed in accordance with the HIPAA "safe harbor" provision is considered per se de-identified and exempted from HIPAA coverage unless a covered entity knows that the data can be used on its own or together with other information to identify a data subject.[159] For example, if researchers request only data pertaining to a very small geographic area in which most people know each other, it may be impossible to truly de-identify the information.[160] In such a case, experts may need to aggregate data from several locations or to combine suppression with other techniques.

3. Does Public-Use Medical Data Pose a Real Privacy Threat?

Data custodians offering public-use data may try hard to de-identify patient records or to ask for patients' consent to disclosure.[161] Nevertheless, many are not required to do so because they are not covered by the HIPAA Privacy Rule and its data disclosure and de-identification guidelines. Consequently, the patient authorization and de-identification practices that data custodians choose to implement may deviate from HIPAA standards and leave data more vulnerable to attack by hackers or other wrongdoers.

Moreover, even with careful de-identification, sophisticated adversaries may be able to re-identify at least a small number of records. Successful de-identification of genetic information may be particularly challenging. With voluminous de-identified medical data available to the public, re-identification attempts are likely to occur. Perpetrators may have malevolent intent, such as identity theft, or may simply be interested in determining whether they can meet the challenge of re-identification.

a) Data Holders Not Covered by the HIPAA Privacy Rule

The HIPAA Privacy Rule's health information disclosure and de-identification requirements do not apply to most suppliers of publicly available health data, because they are either government agencies or non-

158. 45 C.F.R. § 164.514(b)(2)(i) (2013). Removal of the eighteen identifiers is a comprehensive form of suppression.

159. 45 C.F.R. § 164.514(b)(2)(ii) (2013).

160. Khaled El Emam et al., *Evaluating Predictors of Geographic Area Population Size Cut-offs to Manage Re-identification Risk*, 16 J. AM. MED. INFORMATICS ASS'N 256, 256–57 (2009); Felix T. Wu, *Defining Privacy and Utility in Data Sets*, 84 U. COLO. L. REV. 1117, 1156 (2013).

161. *See supra* Part II.

covered private entities.[162] Consequently, these data holders may not be diligent about obtaining meaningful patient authorization for disclosure of identifiable information. In addition, if they de-identify records, they may choose to do so in ways that provide far less privacy protection to their subjects than does the HIPAA safe harbor provision. Stripping medical records of names alone does little to conceal patients' identities, and even leaving just a few specific details may make it easy to ascertain who the individual is. One startling study found that almost 98% of Montreal residents could be identified based on their full postal code, date of birth, and gender.[163]

Data holders' de-identification practices vary. A 2013 survey found that thirty-three states released patient hospital discharge data to the public, but only seven de-identified them in a manner that would conform to the HIPAA Privacy Rule's standard.[164] Many states released the month or quarter of hospital admission and/or discharge and patients' five-digit zip codes.[165] Datasets with these details are more vulnerable to re-identification than those that are de-identified in accordance with HIPAA guidance. The more personal details a publicly available health record contains, the more likely it is to be matched to other open datasets that include names, such as voter registration lists, purchasing records,[166] or news reports.[167] Thus, the more overlapping information fields there are between the medical records and other datasets, such as zip codes, ages, and details of illness, the more likely an adversary will be able to link names to the purportedly anonymized health information.

Scholars confirm that concern about re-identification is well-grounded, as demonstrated by a variety of re-identification successes. In a particularly infamous case, Latanya Sweeney, now a computer scientist at Harvard University, identified the health records of Massachusetts' Governor William Weld when she was a graduate student at the

162. *See supra* note 146 and accompanying text.

163. Khaled El Emam, *The Re-identification Risk of Canadians from Longitudinal Demographics*, 11 BMC MED. INFORMATICS & DECISION MAKING 46, 51 (2011).

164. SEAN HOOLEY & LATANYA SWEENEY, SURVEY OF PUBLICLY AVAILABLE STATE HEALTH DATABASES 4 (2013), http://dataprivacylab.org/projects/50states/1075-1.pdf.

165. *Id.* at 4–7.

166. *See infra* note 196 and accompanying text (discussing information that third parties can purchase about individuals).

167. Arvind Narayanan & Vitaly Shmatikov, *Privacy and Security: Myths and Fallacies of "Personally Identifiable Information,"* COMM. ACM, June 2010, at 24, 26; *Re-identification*, ELECTRONIC PRIVACY INFORMATION CENTER, http://epic.org/privacy/reidentification (last visited Nov. 23, 2015).

Massachusetts Institute of Technology in 1996.[168] She compared birth date, gender, and zip code information that was retained in publicly released hospital discharge records to the same identifiers in publicly available voter registration lists and could match voter names to hospital records.[169]

In a more recent effort, Dr. Sweeney and colleagues worked to re-identify publicly available profiles in the Personal Genome Project[170] that contained medical and genomic information as well as date of birth, gender, and zip code.[171] They linked the demographic data to voter lists or other public records that featured names and were able to identify eighty-four to ninety-seven percent of Personal Genome Project profiles.[172]

In a third project, Dr. Sweeney focused on Washington State hospital discharge data, which contained many demographic details other than names and addresses and could be purchased for fifty dollars. She attempted to match hospitalization records to eighty-one newspaper stories about accidents and injuries in 2011 and was able to determine the name of the patient to whom the records belonged in thirty-five (or forty-three percent) of the cases, based on the news accounts.[173]

b) Re-identification of Fully De-identified Health Records

Theoretically, de-identification in accordance with the HIPAA Privacy Rule's guidelines should make it impossible for anyone to determine the identity of data subjects. Nevertheless, experts have concluded that there remains a small risk that highly skilled and motivated attackers will be able to re-identify records that have been de-identified in

168. Jonathan Shaw, *Exposed: The Erosion of Privacy in the Internet Era*, HARV. MAG., Sept.–Oct. 2009, at 38, http://harvardmagazine.com/2009/09/privacy-erosion-in-internet-era.

169. *Id.*; 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).

170. *See supra* Section II.B.3.

171. Latanya Sweeney et al., *Identifying Participants in the Personal Genome Project by Name* (Harv. U. Data Privacy Lab, White Paper 1021-1, Apr. 24, 2013), http://dataprivacylab.org/projects/pgp/1021-1.pdf.

172. *Id.* at 1. The researchers found that some Personal Genome Project profiles contained the data subject's name, and in other instances, when the downloadable DNA files were uncompressed, they had a file name that included the data subjects' first and last names. *Id.* at 3.

173. Latanya Sweeney, *Matching Known Patients to Health Records in Washington State Data* (Harv. U. Data Privacy Lab, White Paper 1089-1, July 4, 2013), http://dataprivacylab.org/projects/wa/1089-1.pdf; Jordan Robertson, *States' Hospital Data for Sale Puts Privacy in Jeopardy*, BLOOMBERG (Jun 4, 2013), http://www.bloomberg.com/news/2013-06-05/states-hospital-data-for-sale-puts-privacy-in-jeopardy.html.

compliance with HIPAA guidelines.[174] Re-identification may occur when perpetrators have access to non-medical open data, such as voter registration records, that they can link to anonymized health information. Studies have estimated that the risk of re-identification of HIPAA de-identified records falls in the range of 0.01% to 0.25%.[175] Although this percentage seems tiny, it translates into a risk of tens of thousands or even hundreds of thousands of records being re-identified if one thinks in terms of the 323 million individuals in the American population.[176]

Furthermore, the HIPAA Privacy Rule's safe harbor provision does not ban the disclosure of certain details whose presence could make it easier to identify individuals. For example, according to Dr. Khaled El Emam, if hospital discharge data includes length of stay and time since last visit, which are not among the eighteen prohibited identifiers, as many as 16.57% of the records could have a high likelihood of re-identification.[177]

c) The Peculiarities of Genetic Information

The HIPAA Privacy Rule does not provide explicit guidance concerning the de-identification of genetic information,[178] such as the genetic sequences available through GenBank.[179] Many commentators have expressed concern that adversaries could re-identify anonymized genetic information using a variety of techniques.[180] Researchers believe that people can be uniquely identified through a sequence of only thirty to eighty out of thirty million single-nucleotide polymorphisms (SNPs).[181] In

174. Hoffman & Podgurski, *supra* note 15, at 105–07.

175. Khaled El Emam et al., *A Systematic Review of Re-Identification Attacks on Health Data*, 6 PLOS ONE e28071 (2011), http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0028071 (finding a re-identification rate of 0.013%); NAT'L COMM. ON VITAL & HEALTH STATISTICS, ENHANCED PROTECTIONS FOR USES OF HEALTH DATA 36 n.16 (2007), http://www.ncvhs.hhs.gov/wp-content/uploads/2014/05/071221lt.pdf.

176. *See U.S. and World Population Clock*, U.S. CENSUS BUREAU, http://www.census.gov/popclock (last visited Feb. 5, 2016).

177. Khaled El Emam, *Methods for the De-identification of Electronic Health Records for Genomic Research*, 3 GENOME MED. 25, 27 (2011).

178. 45 C.F.R. § 164.514(b)(2)(i) (2013); El Emam, *supra* note 177, at 27.

179. *See supra* Section II.A.4; Melissa Gymrek et al., *Identifying Personal Genomes by Surname Inference*, 339 SCIENCE 321, 321 (2013) (noting that "[s]haring sequencing data sets without identifiers has become a common practice in genomics").

180. El Emam, *supra* note 177, at 27; Dina N. Paltoo et al., *Data Use Under the NIH GWAS Data Sharing Policy and Future Directions*, 46 NATURE GENETICS 934, 937 (2014).

181. El Emam, *supra* note 177, at 27; Liina Kamm et al., *A New Way to Protect Privacy in Large-Scale Genome-Wide Association Studies*, 29 BIOINFORMATICS 886, 886

one study, researchers identified family names by matching short sequences of DNA bases on an individual's Y chromosome to entries in recreational genetic genealogy databases.[182] These short sequences are repeated different numbers of times in different individuals, and hence they are called short tandem repeats or Y-STRs. Even providing only summary-level genetic information cannot always fully protect the identities of data subjects.[183] Given genotype frequencies for a study cohort, it is possible to determine if a particular individual is in the cohort if one knows the individual's genotype and has a reference set of allele frequencies for the underlying population.[184] Thus, genetic information may be more difficult to de-identify effectively than other types of data.

B. DISCRIMINATION AND SPECIAL TARGETING

Medical big data can serve as a treasure trove of information for parties who will use it to further their own economic interests.[185] The release of patient data for public use, alongside advances in re-identification capabilities, raises significant concern regarding potential discrimination or targeting by parties with a stake in individuals' health and economic welfare.[186] This Section will focus on three examples: employers, financial institutions, and marketers. Employers have a strong incentive to identify

(2013). A single-nucleotide polymorphism is a "variation at a single position in a DNA sequence among individuals." *Single Nucleotide Polymorphism*, NATURE, http://www.nature.com/scitable/definition/single-nucleotide-polymorphism-snp-295 (last visited Nov. 23, 2015).

182. Melissa Gymrek et al., *Identifying Personal Genomes by Surname Inference*, 339 SCIENCE 321, 321 (2013).

183. David W. Craig et al., *Assessing and Managing Risk When Sharing Aggregate Genetic Variant Data*, 12 NATURE REVS. GENETICS 730, 730 (2012).

184. *Id.* at 734–35. An allele is one of several variations of a gene. *Allele*, GENETICS HOME REFERENCE, U.S. NAT'L LIBR. OF MED. (Feb. 1, 2016), http://ghr.nlm.nih.gov/glossary=allele.

185. *See* Narayanan & Shmatikov, *supra* note 167, at 26 (noting "increasing economic incentives for potential attackers"); Kate Crawford & Jason Schultz, *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms*, 55 B.C. L. REV. 93, 96–99 (2014) (discussing business use of big data to obtain personal health information about consumers); Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1, 3 (2014) (stating that in today's world "[p]redictive algorithms mine personal information to make guesses about individuals' likely actions and risks[]" and "[p]rivate and public entities rely on predictive algorithmic assessments to make important decisions about individuals").

186. EXEC. OFFICE OF THE PRESIDENT, BIG DATA: SEIZING OPPORTUNITIES, PRESERVING VALUES 51 (2014), http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf (stating that "[a]n important conclusion of this study is that big data technologies can cause societal harms beyond damages to privacy, such as discrimination against individuals and groups").

and select the healthiest workers in order to avoid attendance and productivity problems and high health insurance costs. Likewise, lenders are interested in borrowers who will have income and be able to pay off their loans. For their part, advertisers and marketers wish to tailor their marketing campaigns to reach the most lucrative markets, and thus, they might target particular individuals based on known health conditions[187] or offer special promotions to some consumers but not others.[188]

1. Employers

Employers go to great lengths to select employees carefully in order to maximize business productivity and profitability. Sick or disabled employees can be very expensive for employers because of absenteeism, performance shortcomings, high insurance costs, loss of customers who are uncomfortable interacting with the individual, erosion of workforce morale if other workers feel overburdened while the employer accommodates the ill or impaired employee, and other problems.[189] Employers may have good economic reasons to strive for the healthiest possible workforce, but they are constrained by federal and state laws prohibiting discrimination based on a variety of protected classifications, including disability and genetic information.[190] Moreover, if employers make assumptions about people's health and apply rigid, generalized rules to determine which employees are undesirable, they will deprive many qualified individuals of job opportunities.

The advent of publicly available data may enable employers to discriminate against individuals who are perceived to be at high risk of poor health in ways that are subtle and difficult to detect. Some employers are already embracing advanced technologies such as smart badges that enable them to monitor employee conduct and analyze workplace

187. Lori Andrews, *Facebook is Using You*, N.Y. TIMES (Feb. 4, 2012), http://www.nytimes.com/2012/02/05/opinion/sunday/facebook-is-using-you.html.

188. EXEC. OFFICE OF THE PRESIDENT, BIG DATA, *supra* note 186, at 47.

189. *See* Bruce Japsen, *U.S. Workforce Illness Costs $576B Annually From Sick Days To Workers Compensation*, FORBES (Sept. 12, 2012), http://www.forbes.com/sites/brucejapsen/2012/09/12/u-s-workforce-illness-costs-576b-annually-from-sick-days-to-workers-compensation; Jessica L. Roberts, *Healthism and the Law of Employment Discrimination*, 99 IOWA L. REV. 571, 580–89 (2014) (analyzing the rationales for health-driven employment policies).

190. *See* Sharona Hoffman, *The Importance of Immutability in Employment Discrimination Law*, 52 WM. & MARY L. REV. 1483, 1489–94 (2011) (discussing the forms of discrimination prohibited by anti-discrimination legislation).

interactions as never before.[191] They may well pursue opportunities to use identifiable, re-identifiable, and even non-identifiable medical data to develop new screening tools and hiring policies.

a) Using Identifiable or Re-Identifiable Data

Individuals who agree to share identifiable or easily re-identifiable medical data with the public on websites such as PatientsLikeMe or the Personal Genome Project[192] should understand that it will be accessible to everyone. This includes not only fellow patients or others with benign interests, but also employers who may take adverse action based on health concerns.

Many employers reportedly access public profiles that applicants post on social media sites as part of their investigation of candidates' credentials.[193] They also ask applicants for permission to obtain their credit reports.[194] It is therefore not far-fetched to assume that employers will search publicly available health profiles as well. It is also possible that employers will hire data miners to re-identify medical information when doing so is not excessively difficult. Employers or their agents may be able to re-identify health records that feature certain items such as postal codes, birthdates, and gender, with the aid of demographic information and names contained in voter registration lists, credit reports, or job applications.[195]

Employers may also be able to hire experts who can re-identify information that is thoroughly de-identified in compliance with the HIPAA safe harbor standard if they have a sufficient amount of related, identifiable data about applicants and employees to which they can match the de-identified records. For example, data miners may be able to obtain individuals' detailed purchasing histories or web-browsing histories from

191. Steve Lohr, *Unblinking Eyes Track Employees: Workplace Surveillance Sees Good and Bad*, N.Y. TIMES (June 21, 2014), http://www.nytimes.com/2014/06/22/technology/workplace-surveillance-sees-good-and-bad.html.

192. *See supra* Sections II.B.2 and II.B.3.

193. Greg Fish & Timothy B. Lee, *Employer Get Outta My Facebook*, BLOOMBERG BUSINESSWEEK (Mar. 20, 2008), http://www.businessweek.com/debateroom/archives/2010/12/employers_get_outta_my_facebook.html; Phyllis Korkki, *Is Your Online Identity Spoiling Your Chances?*, N.Y. TIMES (Oct. 9, 2010), http://www.nytimes.com/2010/10/10/jobs/10search.html.

194. Gary Rivlin, *The Long Shadow of Bad Credit*, N.Y. TIMES (May 12, 2013), http://www.nytimes.com/2013/05/12/business/employers-pull-applicants-credit-reports.html.

195. *See supra* Section IV.A.3.a.

database marketers such as Acxiom,[196] and by some estimates, approximately 4000 data brokers already exist.[197] If these lists suggest that particular workers have certain health conditions, data miners may be able to link anonymized health records to names on the lists and thereby identify patients and obtain their medical details.

Experienced data miners, aided by contemporary technology, often have no difficulty achieving re-identification. Interested buyers can purchase lists of patients with depression, erectile dysfunction, diabetes, Alzheimer's disease, and Parkinson's disease.[198] In a 2010 article, two computer scientists, Arvind Narayanan and Vitaly Shmatikov, went as far as to say that "advances in the art and science of re-identification, increasing economic incentives for potential attackers, and ready availability of personal information about millions of people (for example, in online social networks) are rapidly rendering [de-identification] obsolete."[199]

The Americans with Disabilities Act (ADA) prohibits employers from engaging in disability-based discrimination.[200] The law allows employers to conduct medical inquiries and examinations within certain limits to determine fitness for duty,[201] but workers who feel that an employer denied them a job opportunity because of information it discovered may sue the employer.[202] Unlike medical exams, publicly shared medical data would enable employers to view workers' health information without the individuals' knowledge and, consequently, with little concern about being

196. *See* Alice E. Marwick, *How Your Data Are Being Deeply Mined*, N.Y. REV. BOOKS, Jan. 9, 2014, http://www.nybooks.com/articles/archives/2014/jan/09/how-your-data-are-being-deeply-mined (discussing the development of "database marketing," an industry that collects, aggregates, and brokers personal data from sources such as "home valuation and vehicle ownership, information about online behavior tracked through cookies, browser advertising, and the like, data from customer surveys, and 'offline' buying behavior"); *see also* ACXIOM, http://www.acxiom.com (last visited Nov. 23, 2015) (describing a company that gives "clients the power to successfully manage audiences, personalize customer experiences and create profitable customer relationships" using big data analytics).

197. Frank Pasquale, *The Dark Market for Personal Data*, N.Y. TIMES (Oct. 17, 2014), http://www.nytimes.com/2014/10/17/opinion/the-dark-market-for-personal-data.html.

198. Shannon Pettypiece & Jordan Robertson, *For Sale: Your Name and Medical Condition*, BLOOMBERG BUSINESS (Sept. 18, 2014), http://www.bloomberg.com/bw/articles/2014-09-18/for-sale-your-name-and-medical-condition.

199. Narayanan & Shmatikov, *supra* note 167, at 26; *see also* ELECTRONIC PRIVACY INFORMATION CENTER, *supra* note 167 (stating that "'anonymized' data can easily be re-identified").

200. 42 U.S.C. § 12112(a) (2010).

201. 42 U.S.C. § 12112(d) (2010).

202. 42 U.S.C. § 12117(a) (2010).

accused of disability discrimination in case of adverse employment decisions.

b) De-identified Information as a Basis for Multi-Factor Discrimination and Discrimination by Proxy

Employers may use publicly available medical data for purposes of screening workers even without attempting to re-identify records. Some websites feature information concerning disease trends that might induce employers to try to exclude certain classes of employees. For instance, CDC Wonder allows users to search for cancer incidence by age, sex, race, ethnicity, and region.[203] As a hypothetical example, the results of a search might lead an employer to conclude that Hispanic women over fifty are more prone to several cancers than other individuals, and consequently, to decline to hire Hispanic women over fifty.[204]

Some researchers have in fact focused on particular ethnic sub-groups and concluded that they have more health problems than others. A prime example is the PINE Study, for which investigators interviewed 3,018 Chinese adults aged 60 to 105 who lived in the Chicago area between 2011 and 2013.[205] The study concluded that "Chinese older adults experience disproportionate health disparities," suffering from significant physical, psychological, financial, and social challenges.[206] Though this was far from the study's intention, readers of the report may think twice about hiring people of Chinese ancestry who are sixty or older. While investigators used interviews for this study, they could also undertake record reviews in the future if sufficient information is available. The study's findings could encourage employers to pursue similar research using open medical data, because it will yield clear categories of individuals who should be excluded as likely to become problematic employees.

The civil rights laws prohibit discrimination by race, color sex, and age, among other categories,[207] but discrimination based on a combination of two or more factors would be very difficult to detect and prove. If accused of discrimination, the employer would be able to show that it has

203. *United States Cancer Statistics, 1999–2010 Incidence Archive Request*, CDC WONDER, http://wonder.cdc.gov/cancer-v2010.html (last visited Nov. 23, 2015).

204. *See* Jourdan Day, *Closing the Loophole—Why Intersectional Claims are Needed to Address Discrimination Against Older Women*, 75 OHIO ST. L.J. 447, 448 (2014).

205. XINQI DONG ET AL., THE PINE REPORT, at v (2013), http://chinesehealthyaging.org/files/PINE_Final_Reports/All.pdf.

206. *Id.* at v, 40.

207. *See* Title VII of the Civil Rights Act of 1964, 42 U.S.C. § 2000e-2(a) (2006); Age Discrimination in Employment Act, 29 U.S.C. §§ 623(a), 631(a) (2006).

Hispanic, female, and older employees in its workforce. A plaintiff would need to be clever enough to discern that the employer is excluding only a subgroup that falls at the intersection of several protected categories and then somehow decipher the employer's motivation for doing so. Furthermore, many courts disallow multi-factor claims involving age.[208] These courts perceive "age plus" cases as prohibited by a Supreme Court decision, *Gross v. FBL Financial Services, Inc.*, that held that a plaintiff claiming age discrimination must prove that age was the "but for" reason for the adverse action at issue.[209]

Anonymized data can provide other opportunities for discrimination as well.[210] Employers, who are highly motivated to develop means to screen out workers at high risk of health problems, may undertake their own citizen science projects or hire experts to do so. Employers or their agents may mine medical data using sophisticated algorithms to detect associations between individual characteristics or behaviors and poor physical or mental health.[211] Then, through job applications, interviews, and reference or background checks, employers could try to determine whether applicants have those attributes or behaviors.

Concern that employers would attempt to find reliable predictors of applicants' future health status is not fanciful. In the words of two prominent scholars, "predictive algorithms . . . are increasingly rating people in countless aspects of their lives."[212] Several websites, such as "Lifespan Calculator" and "How Long Will I Live?," invite users to calculate their longevity based on a series of questions. These websites' calculations may or may not be trustworthy or illuminating, but they reflect deep interest in creating health-related predictive tools.[213] The websites ask users about their height, weight, education, income, marital status, exercise habits, smoking, drinking, driving, seat belt use, work history, eating, sleeping, and more.[214] They also ask a small number of

208. Day, *supra* note 204, at 449.

209. *Id.* at 466–67; Gross v. FBL Fin. Servs., Inc., 557 U.S. 167, 177–78 (2009).

210. Michael Schrage, *Big Data's Dangerous New Era of Discrimination*, HARV. BUS. REV. (Jan. 29, 2014), http://blogs.hbr.org/2014/01/big-datas-dangerous-new-era-of-discrimination.

211. *See* EXEC. OFFICE OF THE PRESIDENT, BIG DATA, *supra* note 186, at 45–47 (discussing algorithms).

212. Citron & Pasquale, *supra* note 185, at 2.

213. *See Lifespan Calculator*, NORTHWESTERN MUTUAL, http://media.nmfn.com/tnetwork/lifespan (last visited Nov. 23, 2015); Dean P. Foster, Choong Tze Chua, & Lyle Y. Ungar, *How Long Will I Live?*, U. PENN., http://gosset.wharton.upenn.edu/mortality/perl/CalcForm.html (last visited Nov. 23, 2015).

214. *See Lifespan Calculator*, *supra* note 213; *How Long Will I Live?*, *supra* note 213.

questions about family and personal medical history. If employers asked such questions directly, they could be found liable for violations of federal anti-discrimination law.[215] However, as data mining science continues to develop and demand for its products grows, experts will likely develop dependable tools that do not require such explicit questions. While employers may not care about whether employees will live to be eighty or ninety, they will be interested in determining whether they will remain healthy and productive during their working lives.

Already, some employers are known to reject candidates who are obese or smoke because of anticipated health problems.[216] In the future, they might disqualify applicants for many more forms of conduct or characteristics. Applicants could routinely be questioned during interviews about their eating, exercise, travel, and other habits. Employers may then base employment decisions on proxies for disease or predictions of later illness without violating state and federal anti-discrimination laws. As Professor Jessica Roberts explains, those statutes prohibit discrimination based on attributes (for example, race or disability) rather than on behavior (for example, consumption of fatty food or a sedentary lifestyle).[217] Furthermore, the laws focus only on *current* disabilities and genetic information and do not govern any assumptions employers might make about individuals' future ailments that do not relate to off-limits genetic information.[218]

2. Financial Institutions and Marketers

Like employers, financial institutions collect information about individuals. Banks routinely maintain databases with data about customers who previously overdrew their accounts or bounced checks.[219] Nothing will

215. *See* Genetic Information Nondiscrimination Act, 42 U.S.C. § 2000ff(4) (2008) (including "the manifestation of a disease or disorder in family members" in the definition of "genetic information" that employers are forbidden to seek); Americans with Disabilities Act, 42 U.S.C. § 12112(d)(2) (prohibiting employers from conducting most medical inquiries and tests prior to extending a job offer to the applicant).

216. Roberts, *supra* note 189, at 577–79.

217. *Id.* at 604–07.

218. *See* Hoffman, *supra* note 190, at 1489–94 (2011) (discussing the forms of discrimination prohibited by anti-discrimination legislation). The Genetic Information Nondiscrimination Act prohibits employers from discriminating based on genetic information, and therefore, employers should refrain from mining data collections for genetic information, even if it is abundantly available. Genetic Information Nondiscrimination Act, 42 U.S.C. §§ 2000ff(4), 2000ff-1(a) (2008).

219. Jessica Silver-Greenberg & Michael Corkery, *Bank Account Screening Tool is Scrutinized as Excessive*, N.Y. TIMES (June 15, 2014), http://dealbook.nytimes.com/2014/06/15/bank-account-screening-tool-is-scrutinized-as-excessive.

prevent them from adding health information to their databases in order to hone their ability to screen out applicants with a high risk of defaulting on loans if such data is attainable at low cost. As suggested above, financial institutions may utilize identifiable and easily re-identifiable information and may mine databases to discern associations between health risks and various attributes or behaviors.[220]

The ADA prohibits disability-based discrimination by places of public accommodation, that is, establishments that provide services to the public, including banks and other financial institutions.[221] However, customers are unlikely to suspect or discover that banks viewed their health information while assessing their loan applications and thus, such acts of discrimination will probably go unchallenged.

Marketers and advertisers also have an interest in individuals' health data. The more they know about potential customers, the more they can tailor their materials to appeal to those individuals.[222] For example, individuals who are known to have diabetes might receive advertisements about sugar-free products, which some may perceive as a troubling invasion of privacy. Consumers may be particularly resentful when the health condition at issue is sensitive, as noted in a 2012 Forbes magazine article entitled "How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did."[223]

Marketers may also engage in discriminatory practices, offering promotions and discounts to some customers but not others, or advertising selectively so that they reach only certain consumers. They may mine health records for clues regarding individuals' purchasing potential and aggressively pursue the most likely or wealthiest customers. A 2014 presidential report provided the following account:

> [S]ome . . . retailers were found to be using an algorithm that generated different discounts for the same product to people based on where they believed the customer was located. While it may be that the price differences were driven by the lack of competition in certain neighborhoods, in practice, people in

220. *See supra* Section IV.B.1 (discussing potential discrimination by employers).

221. 42 U.S.C. §§ 12181(7)(F), 12182(a) (2010).

222. Andrews, *supra* note 187.

223. Kashmir Hill, *How Target Figured Out a Teen Girl Was Pregnant Before Her Father Did*, FORBES (Feb. 16, 2012), http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did (discussing Target's practice of data-mining its customers' purchasing records in order "to figure out what you like, what you need, and which coupons are most likely to make you happy").

> higher-income areas received higher discounts than people in lower-income areas.[224]

While this practice already exists, access to open medical data may enable industry to refine marketing campaigns even further, to the dismay of some customers. Moreover, courts are unlikely to find that selective advertising or promotional offers and discounts violate anti-discrimination laws.[225] Marketers will generally be able to argue convincingly that their decisions were based on economic factors rather than on race, disability, or other protected categories.[226]

C. PROPAGATION OF INCORRECT AND HARMFUL RESEARCH CONCLUSIONS

Citizen science can lead to valuable and illuminating discoveries.[227] At the same time, however, amateurs may reach incorrect conclusions.[228] Furthermore, anyone can widely publicize information on the Internet, whether it be correct or erroneous. Advice as to how to gain broad exposure is abundantly available on the Internet and can be found in webpages such as "12 Ways to Promote Your Blog"[229] and "How to Promote Your Article Online."[230] In some cases, the media, celebrities, and politicians highlight the work of ordinary citizens,[231] and they may well do so with respect to scientific discoveries that they find intriguing or that support their own agendas. In other cases, individuals can gain

224. EXEC. OFFICE OF THE PRESIDENT, BIG DATA, *supra* note 186, at 46–47.

225. Schrage, *supra* note 210 (stating that it is unclear "where value-added personalization and segmentation end and harmful discrimination begins").

226. Crawford & Schultz, *supra* note 185, at 101 (stating that "housing providers could design an algorithm to predict the relevant PII [personally identifiable information] of potential buyers or renters and advertise the properties only to those who fit these profiles" and do so without violating fair housing laws).

227. *See supra* Section III.A.

228. INSTITUTE OF MEDICINE, DISCUSSION FRAMEWORK FOR CLINICAL TRIAL DATA SHARING 13 (2014), https://globalhealthtrials.tghn.org/site_media/media/medialibrary/2014/01/IOM_data_sharing_Report.pdf (stating that "shared clinical trial data might be analyzed in a manner that leads to biased effect estimates or invalid conclusions").

229. Sally Kane, *12 Ways to Promote Your Blog: Blog Promotion Tips for Lawyers and Legal Professionals*, ABOUT.COM, http://legalcareers.about.com/od/practicetips/tp/10-Ways-To-Promote-Your-Blog.htm.

230. Daniel Vahab & Lisa Chau, *How to Promote Your Article Online*, SOCIAL MEDIA MONTHLY (Nov. 30, 2012), http://thesocialmediamonthly.com/how-to-promote-your-article-online.

231. Well-known examples are singing sensation Susan Boyle and conservative activist "Joe the Plumber."

attention through word of mouth and social media, as happens when a YouTube video or blog post "goes viral."[232]

While professional researchers most often seek publication in peer-reviewed journals that carefully scrutinize submissions, nothing will stop citizen scientists from posting their study results on blogs, personal web pages, and other electronic publications, making them instantaneously available to a worldwide audience.[233] Some commentators describe this phenomenon in terms of a shift from "intermediation" to "apomediation."[234] Traditionally, peer reviewed journals served as necessary intermediaries between researchers and readers and thus gatekeepers for scientific knowledge. The Internet has now triggered disintermediation and increased use of apomediaries, agents or tools that guide readers to information without any middlemen required.[235] Many reports published on websites appear highly professional and credible to general readers, who are not always sophisticated about distinguishing between reliable and questionable sources of information.[236]

---

232. *See* Seth Mnookin, *One of a Kind: What Do You Do if Your Child Has a Condition That is New to Science?*, NEW YORKER (July 21, 2014), http://www.newyorker.com/magazine/2014/07/21/one-of-a-kind-2 (describing how a father posted a blog entry about his disabled son's extremely rare genetic abnormality in order to identify other patients with the condition, and the blog went viral, yielding contact with several other families).

233. R.J.W. Cline & K.M. Haynes, *Consumer Health Information Seeking on the Internet: The State of the Art*, 16 HEALTH EDUC. RES. 671, 679 (2001) (stating that "the Internet is characterized by uncontrolled and unmonitored publishing with little peer review").

234. Dan O'Connor, *The Apomediated World: Regulating Research When Social Media Has Changed Research*, 41 J.L. MED. & ETHICS 470, 471 (2013); Gunther Eysenbach, *Medicine 2.0: Social Networking, Collaboration, Participation, Apomediation, and Openness*, 10 J. MED. INTERNET RES. e22 (2008) (coining the term "apomediation").

235. Eysenbach, *supra* note 234, at 5.

236. *See, e.g.*, Geraldine Peterson et al., *How Do Consumers Search For and Appraise Information on Medicines on the Internet? A Qualitative Study Using Focus Groups*, 5 J. MED. INTERNET RES. e33 (2003) (concluding "that there was a range of search and appraisal skills among [study] participants, with many reporting a limited awareness of how they found and evaluated Internet-based information on medicines"); Cline & Haynes, *supra* note 233, at 680 (cautioning that many consumers have weak information-evaluation skills); Miriam J. Metzger, *Making Sense of Credibility on the Web: Models for Evaluating Online Information and Recommendations for Future Research*, 58 J. AM. SOC'Y INFO. SCI. & TECH. 2078, 2079 (2007) (noting that "studies have found that users are seldom diligent in checking the accuracy of the information they obtain online"). *But see* S. Mo Jang, *Seeking Congruency or Incongruency Online? Examining Selective Exposure to Four Controversial Science Issues*, 36 SCI. COMM. 143, 159 (2014) (finding that "online users may not be as susceptible to confirmation bias [a tendency to favor information that confirms one's views] as some scholars . . . have argued," although "[t]hose who were more religious tended to avoid science news articles that challenged their existing views").

Incorrect findings are unlikely to be a rarity. They will stem from a variety of failings and potentially lead to a number of different harms.

1. Error Sources

Erroneous findings could be caused by poor data quality in the original dataset or flawed study design.[237] Data quality deficiencies may result from clinicians' data entry errors in electronic health records, fragmented or incomplete electronic health records, data coding inaccuracies, or problems with software that processes or analyzes data.[238] Highly skilled analysts should be able to recognize data quality problems, adjust for them, and estimate error rates, but amateurs may not know how.[239]

Furthermore, scientific studies can be flawed due to a variety of biases. Selection bias arises when the group of subjects studied is not representative of the population as a whole, and thus, researchers cannot generalize study results.[240] For example, researchers using information from PatientsLikeMe or the Personal Genome Project should assume that individuals who choose to make their medical information public on such websites are a self-selected group (perhaps more educated and more interested in research) that is not typical of average patients. Confounding bias occurs when there are relevant variables that researchers neglect to consider that affect treatment choices and outcomes, and thus, the study's results are skewed.[241] For example, low income may be a confounder because it may cause individuals to select inferior, inexpensive treatments and may also separately lead to poor health because of stress or inadequate nutrition.[242] Measurement bias is a concern when measurements are inaccurate because equipment has failed, patients have reported facts incorrectly, or other problems have occurred in the process of collecting and measuring values.[243] Consequently, researchers face many hurdles and must conduct their studies very skillfully in order to derive valid results.

237. Sharona Hoffman & Andy Podgurski, *The Use and Misuse of Biomedical Data: Is Bigger Really Better?*, 39 AM. J.L. & MED. 497, 515–27 (2013).
238. *Id.* at 515–21.
239. *Id.* at 530–32.
240. *Id.* at 521–23.
241. *Id.* at 523–25.
242. Sharona Hoffman & Andy Podgurski, *Big Bad Data: Law, Public Health, and Biomedical Databases*, 41 J.L. MED. & ETHICS (SUPPLEMENT ON 2012 PUB. HEALTH L. CONF.) 56, 58 (2013).
243. *Id.*

Researchers must be particularly sensitive to the difference between *association* and *causation*.[244] They may identify associations between certain behaviors, exposures, or treatments and particular outcomes but wrongly assume that there is a causal relationship between the two.[245] To illustrate, suppose that a citizen scientist concludes that people who eat acai berries live longer than those who do not eat this fruit. Does this mean that acai berry consumption prolongs life? Probably not. The explanation for this finding may well be that individuals who purchase this exotic fruit are generally well-off and have the means to make careful food choices, to exercise, to limit their stress, and to obtain top-notch medical care. Thus, it may be true that eating acai berries is *associated* with a longer life on average; but it does not follow that acai berries have some property that actually *causes* people to live longer.

Crowdfunding[246] may add another element of uncertainty to research quality. Crowdfunding does not depend on peer review of carefully written grant proposals by professional experts.[247] Rather, researchers aim to appeal to a large number of donors through videos and social media campaigns.[248] Some commentators have accused crowdfunding of turning "science into a popularity contest."[249] It is certainly possible that the "crowd" will ignore the most meritorious proposals and opt to fund projects that are less deserving but more media-friendly and tantalizing.[250] Consequently, studies that are funded in this manner may not be of the highest quality.

2. Potential Harms

While many mistaken conclusions will be benign, some could be harmful. Patients reading incorrect information about their diseases may become unnecessarily anxious or, in the opposite case, overly sanguine about their symptoms and fail to seek needed medical care.

244. *See, e.g.*, Austin Bradford Hill, *The Environment and Disease: Association or Causation?*, 58 PROC. ROYAL SOC'Y MED. 295, 295–300 (1965); Arvid Sjölander, *The Language of Potential Outcomes*, *in* CAUSALITY: STATISTICAL PERSPECTIVES AND APPLICATIONS 6, 9 (Carlo Berzuini et al. eds., 2012).

245. *See* Stephen Choi et al., *The Power of Proxy Advisors: Myth or Reality?* 59 EMORY L.J. 869, 879–85 (2010) (discussing the difference between correlation and causation).

246. *See supra* notes 113–115 and accompanying text.

247. Karen Kaplan, *Crowd-Funding: Cash on Demand*, 497 NATURE 147, 148 (2013).

248. *Id.*

249. Palca, *supra* note 115.

250. Kaplan, *supra* note 247, at 148.

Worse yet, individuals with personal agendas may undertake scientific studies with malevolent intent. They may use findings to inflame passion and prejudice against particular minority groups. Some may attempt to further political agendas by "proving" that their opponents' policies have adverse effects on human health or the healthcare system. Others with selfish economic interests may aim to hurt competitors by claiming that their products cause particular ailments.[251]

Even peer-reviewed journals have published articles whose conclusions are false. A notorious example is a 1998 study published in the prestigious journal *Lancet*, that suggested a link between autism and the measles, mumps, rubella (MMR) vaccination.[252] While the study was later retracted,[253] the belief that vaccinations can lead to autism gained a considerable foothold and still needs to be explicitly repudiated on the CDC's website.[254]

Researchers who are media-savvy or web-savvy and do not submit their findings to peer-reviewed journals for review by experts may be all the more likely to propagate incorrect and potentially harmful views. Manuscripts that are not submitted to journals will not be scrutinized by experts before their authors post them on the Internet, and no filtering mechanism exists to indicate to readers whether the material is valid or trustworthy.[255] The Internet provides publishing opportunities without any need for intermediaries and oversight. Therefore, potentially, millions of readers could view and believe even nonsensical conclusions, especially when authors assert that they based their research on data that the government furnished.

Many myths have in fact gained considerable traction despite the existence of abundant evidence to negate them. Two examples are climate change denial[256] and the outcry that the Patient Protection and Affordable

251. Michelle Mello et al., *Preparing for Responsible Sharing of Clinical Trial Data*, 369 NEW. ENG. J. MED. 1651, 1653 (2013) (cautioning that public access to clinical trial data "could . . . lead unskilled analysts, market competitors, or others with strong private agendas to publicize poorly conducted analyses").

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

253. Simon H. Murch et al., *Retraction of an Interpretation*, 363 LANCET 750, 750 (2004).

254. *Measles, Mumps, and Rubella (MMR) Vaccine Safety Studies*, CDC, http://www.cdc.gov/vaccinesafety/Vaccines/MMR/MMR.html (last updated Aug. 28, 2015).

255. *See supra* notes 233–235 and accompanying text.

256. Aaron M. McCright & Riley E. Dunlap, *Cool Dudes: The Denial of Climate Change Among Conservative White Males in the United States*, 21 GLOBAL ENVTL. CHANGE 1163, 1163 (2011).

Care Act (aka Obamacare) would authorize "death panels" to decide which patients should live and which should die.[257] In both cases, the arguments gained popularity because high-profile public figures embraced them to further their own political agendas, which may occur in many other instances as well.

A particularly pernicious argument was made by Michael Levin in a 1997 book called *Why Race Matters*.[258] The author argued that African-Americans are typically less intelligent and more aggressive, assertive, and impulsive than Whites.[259] In addition, according to the author, African-Americans are more likely to commit crimes because they suffer from "an absence of conscience," lack the ability to engage in self-monitoring, and have less free will and a different moral orientation from Whites.[260] In an era in which anyone in the world can access Internet material without leaving home or paying any money for a publication, these types of purportedly research-backed arguments can be more dangerous than ever before.

D. LITIGATION

Open health data may lead to a proliferation of litigation or threats of litigation in several circumstances. First, parties who feel they were injured by published invalid research outcomes may assert claims such as defamation or interference with economic advantage. Second, business entities may threaten to sue or file frivolous cases against citizen scientists who have acted in good faith and posted legitimate findings because the companies fear that the research outcomes will harm them. Thus, parties could use litigation to intimidate citizen scientists and pressure them to retract and remove purportedly offending materials. Third, data subjects who feel that they are victims of unauthorized disclosure of identifiable medical data may assert common law privacy breach claims. This Section analyzes several potential causes of action and the protection provided in some states by legislation that prohibits strategic lawsuits against public participation (SLAPPs).

257. Brian Beutler, *Republicans' "Death Panel" Smear Was Appallingly Effective*, NEW REPUBLIC (June 23, 2014), http://www.newrepublic.com/article/118313/gop-obamacare-death-panel-smear-putting-peoples-lives-risk.

258. MICHAEL LEVIN, WHY RACE MATTERS: RACE DIFFERENCES AND WHAT THEY MEAN (1997).

259. *Id.* at 213.

260. *Id.* at 213, 322.

1. Defamation

Defamation claims generally require proof of the following elements:

> (1) publication (to a third party) (2) of a defamatory statement (3) "of and concerning" the plaintiff (4) that is false, (5) published with requisite degree of fault (negligence or actual malice), and (6) damages the plaintiff's reputation (which, in some instances, can be presumed).[261]

Establishing a successful defamation claim is no easy task, and plaintiffs must meet a high standard of proof.[262] Electronic speech is entitled to the same stringent First Amendment protections as print communication.[263]

Nevertheless, both individuals and entities may bring defamation claims.[264] For example, a manufacturer may file a defamation suit relating to the publication of intentionally false statements asserting that its product causes health problems. However, as a rule, defamatory statements against groups are not actionable.[265] Thus, if an author published or posted a piece asserting that Jews or African-Americans are biologically inferior in some way, Jewish or African-American plaintiffs could not bring a defamation claim, no matter how baseless and offensive the publication was.

An increasing number of defamation cases involve material posted on the Internet, which is the most likely venue for citizen science publications.[266] For example, businesses have filed defamation suits in response to negative reviews on the website Yelp.[267]

---

261. Matthew E. Kelley & Steven D. Zansberg, *A Little Birdie Told Me, "You're A Crook": Libel in the Twittersphere and Beyond*, 30 COMM. LAW. 34 (2014); RESTATEMENT (SECOND) OF TORTS § 558 (1977).

262. K.J. Greene, *Intellectual Property Expansion: the Good, the Bad, and the Right of Publicity*, 11 CHAP. L. REV. 521, 534 (2008) (stating that "defamation law sets very high standards of proof and injury to prevent conflict with First Amendment principles").

263. Reno v. ACLU, 521 U.S. 844, 870 (1997) (asserting that "our cases provide no basis for qualifying the level of First Amendment scrutiny that should be applied to this medium [the Internet]").

264. Wendy Gerwick Couture, *The Collision Between the First Amendment and Securities Fraud*, 65 ALA. L. REV. 903, 918–20 (2014) (discussing defamation suits brought by entities and individuals).

265. RESTATEMENT (SECOND) OF TORTS § 564A (1977); Ellyn Tracy Marcus, *Group Defamation and Individual Actions: A New Look at an Old Rule*, 71 CALIF. L. REV. 1532, 1533 (1983).

266. Amy Kristin Sanders & Natalie Christine Olsen, *Re-defining Defamation: Psychological Sense of Community in the Age of the Internet*, 17 COMM. L. & POL'Y 355, 365

A particularly memorable defamation case brought by industry involved a discussion on Oprah Winfrey's television show.[268] After scientists linked the consumption of beef from cattle infected by Mad Cow Disease with a new variant of the deadly Creutzfeldt-Jakob Disease, the *Oprah Winfrey Show*, like many other media outlets, covered the story in a segment entitled "Dangerous Foods."[269] At one point in the show Ms. Winfrey stated that she was "stopped cold from eating another burger."[270] Subsequently, several Texas cattlemen sued Ms. Winfrey and other defendants, asserting numerous causes of action, including defamation, and claiming that the beef market suffered significant losses because of the broadcast.[271] Fortunately for Oprah, the defendants prevailed on all claims.[272]

In some cases, plaintiffs may well have legitimate claims against individuals who maliciously publicize damaging information that they know to be false. In fact, the prospect of facing defamation claims may be an important deterrent to such misconduct. However, it is not difficult to imagine that in other instances, the chilling effect of litigation will thwart the dissemination of non-defamatory information. Industry may file lawsuits primarily to intimidate citizen scientists and force them to comply with demands for removal or retraction of material that they researched and posted in good faith. Citizen scientists who are far less powerful and prosperous than Oprah Winfrey may be unable to mount a full defense and simply capitulate.[273]

2. Other Causes of Action

Plaintiffs may file a myriad of other claims, only a few of which will be discussed as examples below. The cattle ranchers who sued Oprah Winfrey alleged not only defamation but also the closely related tort of business disparagement as well as negligence and negligence per se.[274] In addition, companies that feel their products have been inappropriately

(2012) (noting that "[w]ith the increasing number of speakers and messages has come a flurry of litigation as courts struggle to regulate the medium of the masses").

267. Yelp, Inc. v. Hadeed Carpet Cleaning, Inc. 752 S.E.2d 554 (Va. Ct. App. 2014); Bently Reserve L.P. v. Papaliolios, 160 Cal. Rptr. 3d 423 (2013).

268. Texas Beef Group v. Winfrey, 201 F.3d 680 (5th Cir. 2000).

269. *Id.* at 682–84.

270. *Id.* at 688.

271. *Id.* at 682.

272. *Id.* at 680.

273. *But see infra* Section IV.D.3 (discussing anti-SLAPP statutes).

274. Texas Beef Group v. Winfrey, 201 F.3d 680, 682 (5th Cir. 2000); *see id.* at 685 for a discussion of the elements of a business disparagement claim.

denigrated may bring a claim of interference with economic advantage. This theory of liability typically involves proof of the following elements: (1) plaintiff had an economic relationship with a third party that would have likely been economically beneficial for the plaintiff, (2) the defendant knew of the relationship, (3) the defendant engaged in intentional or negligent acts designed to disrupt the relationship, (4) the relationship was in fact disrupted, and (5) the defendant's conduct proximately caused plaintiff to suffer economic harm.[275] Individuals and entities subjected to published criticism or negative commentary often assert allegations of tortious interference with economic advantage alongside defamation claims.[276]

Patients whose data were used for research purposes may also initiate litigation. A patient who believes she did not consent to the posting of her identifiable medical records may assert a claim of public disclosure of private facts, a tort with the following elements: "(1) public disclosure (2) of a private fact (3) which would be offensive and objectionable to the reasonable person and (4) which is not of legitimate public concern."[277] There is no precedent for applying this theory of liability to re-identified data, but in the future, parties may attempt to invoke it in such circumstances. If re-identified medical information were posted on the Internet or otherwise publicized, the affected individuals may well find the conduct objectionable, and courts are likely to agree that the health records are not of public concern, thus ruling for plaintiffs.

3. Anti-SLAPP Legislation

Citizen scientists can take a degree of comfort in the existence of anti-SLAPP legislation in some states.[278] Strategic lawsuits against public participation (SLAPPs) have been defined as "civil complaints or counterclaims (against either an individual or an organization) in which the alleged injury was the result of petitioning or free speech activities protected by the First Amendment of the U.S. Constitution."[279] For

275. Crown Imports, LLC v. Superior Court, 223 Cal. App. 4th 1395, 1404 (2014) (discussing the tort under California law).

276. *Responding to Strategic Lawsuits Against Public Participation (SLAPPs)*, DIGITAL MEDIA LAW PROJECT, http://www.dmlp.org/legal-guide/responding-strategic-lawsuits-against-public-participation-slapps (last visited Nov. 23, 2015).

277. *See* Diaz v. Oakland Tribune, Inc., 139 Cal. Rptr. 762, 768 (Cal. Ct. App. 1983) (listing the elements of the public disclosure tort under California law).

278. DIGITAL MEDIA LAW PROJECT, *supra* note 276.

279. Robert D. Richards, *A SLAPP in the Facebook: Assessing the Impact of Strategic Lawsuits against Public Participation on Social Networks, Blogs, and Consumer Gripe Sites*, 21 DEPAUL J. ART, TECH. & INTELL. PROP. L. 221, 222 (2011).

example, SLAPPs have been filed by businesses as a form of retaliation against consumers who posted negative comments about them on social networking sites.[280] There is thus reason to worry that some companies will file SLAPPs against citizen scientists who claim that their products are inferior to others or cause health-related harms.

Anti-SLAPP statutes have been enacted in twenty-eight states, the District of Columbia, and Guam.[281] These laws enable defendants subject to certain frivolous allegations to have SLAPPs dismissed quickly and to recover costs and attorneys' fees.[282] The statutes can vary significantly.[283] Pennsylvania's law is very narrow, granting immunity to defendants who make "an oral or written communication to a government agency relating to enforcement or implementation of an environmental law or regulation . . . ."[284] By contrast, in California the law is much broader and covers "written or oral statement[s] or writing made in a place open to the public or a public forum in connection with an issue of public interest."[285] The Pennsylvania law allows defendants to request hearings at which the court will determine whether they are entitled to immunity.[286] The California law establishes a somewhat different procedure, allowing a covered defendant to file a special motion to strike, after which the court will require the plaintiff to produce evidence that it is likely to prevail on its claim. In the absence of such evidence, the claim will be dismissed and defendant will recover attorney's fees and costs.[287] Protection is inconsistent across jurisdictions but may be very helpful to some victims of frivolous litigation initiated for purposes of harassment and intimidation.

V. RECOMMENDATIONS

The growing trend of opening patient-related data held by the government and private entities to the public raises hopes for considerable benefits. At the same time, it provokes significant concerns. How should

280. *Id.* at 222–23; Rex Hall, Jr., *Firm Sues WMU Student Over Facebook Page; Towing Company Seeks $750,000 in Damages for Online Criticism*, GRAND RAPIDS PRESS, Apr. 14, 2010, at A6 (discussing litigation that followed the student's posting of an entry on his Facebook page that criticized T & J Towing for wrongly towing his car from a legal parking space and damaging it).

281. DIGITAL MEDIA LAW PROJECT, *supra* note 276.

282. *Id.*

283. Richards, *supra* note 279, at 232.

284. 27 PA. CONS. STAT. ANN. § 8302(a) (2001).

285. CAL. CODE CIV. P. §§ 425.16(e), 425.17 (West 2011).

286. 27 PA. CONS. STAT. ANN. § 8303 (2001).

287. CAL. CODE CIV. P. § 425.16(b)–(c) (2011).

legislators and regulators respond to this emerging phenomenon? The law must balance the interests of a variety of stakeholders: patients, professional researchers, citizen scientists, government, industry, and the public at large. An excessively heavy-handed approach to regulation might discourage citizen scientists from pursuing projects and making important contributions and may deter data custodians from releasing records. However, a regulatory approach that is too timid may result in privacy breaches, discrimination, and other societal harms. This Part formulates recommendations for regulatory and policy modifications to address open data concerns.

A. PRIVACY AND DATA STEWARDSHIP

The risk that anonymized health information will be re-identified and used inappropriately can never be fully eliminated,[288] but it can be minimized. Several legal and policy interventions could enhance privacy protections. First, the HIPAA Privacy Rule should be amended to expand the definition of "covered entity" and to add a provision that prohibits re-identification. Second, any party releasing patient-related data to the public should establish a data release review board that will scrutinize all disclosed data sets to ensure that they are de-identified as effectively as possible. The review board should also oversee other privacy protections, including privacy training for data recipients, data use agreements, user registries, and consent procedures for data subjects opting to share identifiable information.

1. HIPAA Privacy Rule Modifications

Two HIPAA Privacy Rule changes should be made to enhance data subject privacy. The HIPAA statute and regulations should be amended to expand their reach and efficacy through a broader definition of "covered entity" and an explicit prohibition of any attempt to re-identify data.

a) Expanding the Definition of "Covered Entity" and Creating National Data Release and De-identification Standards

The HIPAA Privacy Rule currently governs only healthcare providers, health plans, healthcare clearinghouses, and their business associates.[289] It therefore does not apply to numerous parties that store and disclose health information, including government entities and database operators. Expansion of the definition of "covered entity" in the HIPAA Privacy

288. *See infra* Section IV.A.3.
289. 45 C.F.R. §§ 160.102–.103 (2013); 42 U.S.C. § 17934 (2010).

Rule and its enabling legislation[290] could improve privacy protection for data subjects. Regulators could turn to a Texas privacy statute as a model for more comprehensive coverage. The law defines "covered entity" in relevant part as any party who,

> for commercial, financial, or professional gain, monetary fees, or dues, or on a cooperative, nonprofit, or pro bono basis, engages, in whole or in part, and with real or constructive knowledge, in the practice of assembling, collecting, analyzing, using, evaluating, storing, or transmitting protected health information. The term includes a business associate, health care payer, governmental unit, information or computer management entity, school, health researcher, health care facility, clinic, health care provider, or person who maintains an Internet site.[291]

The HIPAA Privacy Rule's scope of coverage should be similarly broadened, with one modification. The regulations should explicitly reach employers, financial institutions, and amateur researchers, along with the parties listed in the definition above.

The proposed regulatory expansion should not inhibit the release of data to the public. Rather, it would provide all data holders with clear instructions regarding privacy safeguards and create uniform, national standards for data disclosure and de-identification.[292] Those releasing identifiable information, such as PatientsLikeMe or the Personal Genome Project would need to obtain meaningful patient consent,[293] as discussed in greater detail below.[294] Those who wish to be exempt from HIPAA coverage would need to de-identify disclosed data in accordance with the Privacy Rule's de-identification provision.[295]

In some cases, data holders will want to release information that is largely anonymized but contains a few identifiers that are particularly useful for research purposes. In these instances, database operators would follow the Privacy Rule's "limited data set" provision.[296] In limited data

290. 45 C.F.R. §160.103 (2013); 42 U.S.C. §1320d-1(a) (2010).

291. TEX. HEALTH & SAFETY CODE ANN. 181.001(b)(2)(A) (West 2012).

292. Note that the definition of "health information" would also need to be revised because it is currently limited to information that is "created or received by a health care provider, health plan, public health authority, employer, life insurer, school or university, or health care clearinghouse." 45 C.F.R. § 160.103 (2013). It thus fails to include data handled by website operators and others.

293. 45 C.F.R. §164.508 (2013).

294. *See infra* notes 318–321 and accompanying text.

295. 45 C.F.R. §164.514(b)(2013); *See supra* Section IV.A.2 for detailed discussion of de-identification.

296. 45 C.F.R. §164.514(e)(1)–(4) (2013).

sets, custodians redact most of the safe harbor provision's eighteen identifiers but retain dates and geographic locales, including city or town, state, and postal codes.[297] Database operators may release limited data sets without patient authorization so long as data recipients sign data use agreements containing specified restrictions and privacy protections.[298] These agreements are required because the added identifiers, while valuable to analysts, make re-identification considerably easier for skilled attackers.[299]

The proposed change would modify only the definition of "covered entity." It would not impact the exceptions to the HIPAA Privacy Rule that the regulations establish elsewhere.[300] Thus, the proposal would not create hurdles for health care treatment, payment, administration, or the activities of law enforcement and public health officials.[301]

b) Prohibiting Re-identification

The HIPAA Privacy Rule should also be amended to include a general prohibition of any attempt to re-identify information that would apply to any user of de-identified data.[302] This restriction is already an element of data use agreements, which require the recipients of limited data sets to promise that they will not "identify the information or contact the individuals."[303] The proposed change would extend this regulatory proscription to anyone using de-identified information, including employers, financial institutions, and all other parties. The provision could specify exceptions, such as permitting re-identification necessary to respond to medical or public health emergencies. Violators should be subject to HIPAA's enforcement provisions, which incorporate civil and criminal penalties.[304]

297. 45 C.F.R. §164.514(e)(2) (2013).

298. 45 C.F.R. §164.514(e)(4) (2013).

299. 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) (estimating that the risk of re-identification is between 10% and 60%, depending on the state).

300. 45 C.F.R. §§ 164.502, .506, .512 (2013).

301. *Id.*

302. If the HIPAA Privacy Rule's scope of coverage is expanded as suggested above, the prohibition would apply to all covered entities and individuals. If not, the HIPAA statute itself should be amended to include a re-identification prohibition that applies broadly to all de-identified health data users.

303. 45 C.F.R. §164.514(e)(4)(ii)(C)(5) (2013).

304. 45 C.F.R. §§ 160.300–.552 (2013).

2. Data Release Review Boards

In the absence of HIPAA Privacy Rule amendments, data custodians not currently covered by the Rule should implement their own privacy safeguards. Database operators who release patient-related information to the public should institute a thoughtful and thorough process for reviewing the information at issue and establishing strong privacy safeguards.

The CDC's Policy on Releasing and Sharing Data recommends the establishment of data-release review boards, and data custodians would be wise to implement this suggestion.[305] The boards, composed of data mining and privacy experts, would review any data that are to be released to ascertain that they are as effectively de-identified as possible. For example, the board would assess whether the disclosed sample size is so small that data subjects are likely to be identified no matter what variables are stripped away, as may be the case when data is collected about very rare diseases.[306] The board would also determine what statistical methods should be used to achieve de-identification of various data sets, including suppression, perturbation, and generalization.[307] In addition, the board could analyze data quality to ensure that the released information is sufficiently reliable that it will be of value to users.[308] Finally, the data-release review board should oversee all other privacy safeguards that data holders implement.

3. Data Use Agreements, Privacy Training, Registries, and Consent Procedures

Data custodians who release medical information to the public should implement several privacy protection measures beyond board review, and the extent of these procedures should depend on the type of data at issue. Users who access any database of medical information, including aggregate, summary-level data, should be alerted that the information is sensitive and raises privacy concerns. For example, the CDC Wonder website asks viewers who are seeking mortality information from its

305. CDC/ATSDR POLICY ON RELEASING AND SHARING DATA, *supra* note 111, at 9.

306. *See supra* note 160 and accompanying text.

307. CDC/ATSDR POLICY ON RELEASING AND SHARING DATA, *supra* note 111, at 9; *supra* note 157 and accompanying text (discussing the various statistical methods).

308. CDC/ATSDR POLICY ON RELEASING AND SHARING DATA, *supra* note 111, at 9; *see supra* Section IV.C.1 (discussing data quality shortcomings); Hoffman & Podgurski, *supra* note 237, at 530–32 (discussing data quality assessment).

database to agree to a short list of data use restrictions by clicking an "I agree" icon.[309]

For patient-level data that is not aggregated, more elaborate procedures are needed. The Healthcare Cost and Utilization Project's National (Nationwide) Inpatient Sample (NIS) furnishes a useful model. The NIS contains information concerning millions of hospital stays.[310] It provides detailed information about patients and hospitals but is careful to remove identifiers and most likely meets the HIPAA safe harbor standard.[311] Nevertheless, the NIS requires purchasers of the data to take a 15-minute training course that addresses privacy concerns.[312] It also requires purchasers to sign a detailed data use agreement that specifies a variety of use restrictions designed to protect individual and institutional data subjects from privacy violations and other abuses, such as attempts to

309. *About Underlying Cause of Death, 1999–2013*, CDC WONDER, http://wonder.cdc.gov/ucd-icd10.html. Users agree that they will:

- Use these data for health statistical reporting and analysis only.
- For sub-national geography, do not present or publish death counts of 9 or fewer or death rates based on counts of nine or fewer (in figures, graphs, maps, tables, etc.).
- Make no attempt to learn the identity of any person or establishment included in these data.
- Make no disclosure or other use of the identity of any person or establishment discovered inadvertently and advise the NCHS Confidentiality Officer of any such discovery.

Id.

310. *Overview of the National (Nationwide) Inpatient Sample (NIS)*, HEALTHCARE COST & UTILIZATION PROJECT, http://www.hcup-us.ahrq.gov/nisoverview.jsp (last updated Nov. 17, 2015).

311. *Id.* The data elements provided in the overview are:

- Primary and secondary diagnoses and procedures;
- Patient demographic characteristics (e.g., sex, age, race, median household income for ZIP Code);
- Hospital characteristics (e.g., ownership);
- Expected payment source;
- Total charges;
- Discharge status;
- Length of stay;
- Severity and comorbidity measures.

Id.

312. *Welcome to the HCUP Data Use Agreement (DUA) Training!*, HEALTHCARE COST & UTILIZATION PROJECT, http://www.hcup-us.ahrq.gov/tech_assist/dua.jsp (last updated Sept. 23, 2015).

gain commercial or competitive advantage through analysis of released NIS data.[313]

If data users violate the agreement, the NIS would presumably challenge them in court.[314] A useful supplement to the NIS's requirements would be an online test in which examinees would have to demonstrate that they read and understood the training materials and data use agreement.

Admittedly, training courses and data use agreements will not prevent all privacy violations, and data custodians are not likely to dedicate significant resources to their enforcement. However, these measures will alert the public to the importance of privacy and responsible data handling and may avert innocent breaches by citizen scientists who wish to do no harm.

Equally important, the data use agreement requirement will create a record of those accessing data, and data custodians should maintain functional registries of users. Data custodians can require signatories to provide their name, affiliation, and contact information.[315] If the dataset at issue consists of lower-risk, aggregated or summary data and users do no more than click on an "I agree" icon, only their network addresses will be recorded. Nevertheless, if the individuals used their own computers, authorities could link the network addresses to their identities if need be.[316] Data custodians could then preclude those who violate data use agreements by re-identifying data or engaging in other misconduct from downloading information in the future, and the government could subject such violators to other penalties.[317]

In some cases, privacy requirements should apply not only to data users, but also to data subjects. Specifically, individuals choosing to allow

313. HEALTHCARE COST & UTILIZATION PROJECT, DATA USE AGREEMENT FOR THE NATIONWIDE DATABASES (2014), http://www.hcup-us.ahrq.gov/team/HCUP_Nationwide_DUA_051614.pdf.

314. *See id.* (explaining that violation of the data use agreement can lead to fines or imprisonment under federal and state law); *About Underlying Cause of Death*, *supra* note 309 (describing sanctions for violations).

315. *Data Use Agreement*, *supra* note 313.

316. J.D. Sartain, *Can Your IP Address Give Away Your Identity to Hackers, Stalkers and Cybercrooks?*, NETWORKWORLD (Jul. 16, 2013), http://www.networkworld.com/article/2168144/malware-cybercrime/can-your-ip-address-give-away-your-identity-to-hackers--stalkers-and-cybercrooks-.html. Devious persons may, however, use a spoofed Internet address.

317. *About Underlying Cause of Death*, *supra* note 309 (describing sanctions for violations and stating that "[r]esearchers who violate the terms of the data use restrictions will lose access to WONDER and their sponsors and institutions will be notified").

public access to identifiable or easily identifiable data, such as datasets that include birth date, sex, and zip code,[318] should undergo a comprehensive informed consent process.[319] Such data subjects should understand that their personal health information will be viewable not only by researchers with good intentions, but also by employers, marketers, financial institutions, and others who may not have their best interest in mind.[320] To this end, the Harvard Personal Genome Project requires participants to read and sign a lengthy consent document. They also must pass an examination demonstrating their understanding of the material contained in the consent form.[321] Testing data subjects' comprehension of the privacy risks they accept would be an important component of any informed consent process pertaining to sharing individually identifiable data.

B. ANTI-DISCRIMINATION PROTECTIONS

Ironically, while open data policies promote transparency on the government's part,[322] they may provide new opportunities for employers and others to discriminate in non-transparent ways.[323] Based on data about various health risks, entities might discriminate against discrete population sub-groups such as African-American women older than fifty.[324] These multi-factor discrimination cases are much more difficult to detect and prosecute than cases involving traditional protected classes.[325] In addition, entities may retain experts to mine data and develop new applicant

318. *See* El Emam, *supra* note 163, and accompanying text (explaining that it is relatively easy to re-identify such data).

319. Arguably, anyone whose data is released to the public in any form, including as fully de-identified information, should be asked for consent. A full exploration of this issue is beyond the scope of this Article. However, it is unrealistic to expect that government authorities who receive data relating to millions of patients from a variety of sources will have the resources to track down, contact, and obtain consent from all data subjects. Moreover, allowing individuals to opt out of data sharing could lead to selection bias, whereby the people who choose to be included in databases are not representative of the population as a whole. If that is the case, research results based on study of database participants could not be generalized to others, and therefore, would be of very limited scientific use. Therefore, this Article recommends extensive consent procedures only for data subjects who opt to disclose identifiable or easily re-identifiable information. *See* Hoffman & Podgurski, *supra* note 15, at 114–123 (discussing the problems with consent).

320. *See supra* Section IV.B (discussing discrimination concerns).

321. *Participation Documents*, PERSONAL GENOME PROJECT, HARV. MED. SCH., http://www.personalgenomes.org/harvard/sign-up#documents (last visited Nov. 23, 2015).

322. *See supra* notes 123–124 and accompanying text.

323. *See supra* Section IV.B.

324. *See supra* notes 203–204 and accompanying text.

325. *See supra* notes 207–208 and accompanying text.

screening tools that focus on proxies for disability or predictors of bad health that employers can consider without violating any explicit legal prohibition.[326] As open data and data mining proliferate, novel forms of health-based discrimination may become increasingly common and require several changes to anti-discrimination law and practice.

1. Detecting, Deterring, and Prosecuting Multi-Factor Discrimination

As difficult as multi-factor discrimination may be to detect, enforcement agencies and plaintiffs' attorneys will need to recognize the real possibility that it is occurring.[327] An uptick in litigation and enforcement actions relating to multi-factor cases may encourage victims to bring this type of discrimination to light and discourage employers and businesses from engaging in it.

In multi-factor cases, plaintiffs claiming employment discrimination who believe that one of the improperly considered attributes was their age may face particular hurdles because of the Supreme Court's decision in *Gross v. FBL Financial Services, Inc.* This decision barred mixed-motive claims and required "but for" proof of age discrimination.[328] However, in *Gross* the employer allegedly considered a mixture of proper (performance-related) and improper (the plaintiff's age of fifty-four) factors rather than a combination of prohibited categories (for example, age, race, sex).[329] In a future case, the Supreme Court may revisit the question of whether plaintiffs can sue employers for discriminating based on age and one or more other protected classifications and hold that such claims are allowable. In the alternative, Congress could amend the Age Discrimination in Employment Act to add a provision that explicitly permits multi-factor claims.[330]

326. *See supra* notes 211–218 and accompanying text.

327. *See supra* note 208 and accompanying text; Cathy Scarborough, *Conceptualizing Black Women's Employment Experiences*, 98 YALE L.J. 1457, 1476–78 (1989) (discussing Title VII multi-factor claims).

328. Day, *supra* note 204, at 466–67; Gross v. FBL Fin. Servs., Inc., 557 U.S. 167, 177–78 (2009).

329. FBL's defense was that "Gross' reassignment was part of a corporate restructuring and that Gross' new position was better suited to his skills" and no protected classification other than age was at issue. *Gross*, 557 U.S. at 167.

330. *See* Day, *supra* note 204, at 466–67 (proposing legislative action to approve age-plus-sex claims).

2. Requiring Disclosure of Data Mining for Disability Proxies and Predictors

Instances in which employers, financial institutions, or others engage in data mining and exclude individuals based on perceived or anticipated health conditions will also be difficult to detect. Consequently, anti-discrimination laws should include a requirement that businesses disclose their data mining practices to workers, consumers, and other parties that are affected by such practices.

Several other commentators have called for transparency with respect to data mining and predictive modeling activities. Professors Danielle Citron and Frank Pasquale argue that "we need to switch the default in situations like this away from an assumption of secrecy, and toward the expectation that people deserve to know how they are rated and ranked."[331] Similarly, commentators Kate Crawford and Jason Schultz would require parties to provide notice, "disclosing not only the type of predictions they attempt, but also the general sources of data that they draw upon as inputs, including a means whereby those whose personal data is included can learn of that fact."[332]

A disclosure requirement would be a valuable addition to anti-discrimination protections. It would constitute a compromise between prohibiting data mining practices altogether and ignoring them. A tweak of the ADA's medical inquiry and exam provision[333] could add a requirement that employers disclose in writing to applicants and employees any medical data mining activities that they intend to use for purposes of making employment decisions. This information would then be available to plaintiffs' attorneys and government enforcement agencies such as the Equal Employment Opportunity Commission (EEOC),[334] which could investigate whether these activities resulted in unlawful discrimination. Likewise, the ADA's public accommodation title could feature the same provision to cover financial institutions and other businesses.[335] Employment or loan application forms could include disclosure statements so long as the statements were in sufficiently large and readable print or on separate sheets given to applicants.

---

331. Citron & Pasquale, *supra* note 185, at 21.

332. Crawford & Schultz, *supra* note 185, at 125.

333. 42 U.S.C. § 12112(d) (2010).

334. The Equal Employment Opportunity Commission is the federal agency tasked with enforcing the federal anti-discrimination laws. *See Overview*, EQUAL EMPLOYMENT OPPORTUNITY COMMISSION, http://www.eeoc.gov/eeoc (last visited Nov. 23, 2015).

335. *See* 42 U.S.C. § 12182 (2010).

Some may object that such a requirement will open the floodgates of litigation, especially in employment discrimination cases, because any applicant who receives notice of an employer's data mining activities and who is not hired or promoted could claim discrimination. Employment discrimination claimants, however, must exhaust their administrative remedies prior to filing suit.[336] While the EEOC and state administrative agencies would likely be able to hire experts to investigate and interpret employers' data mining activities in selected instances, they pursue litigation in only a handful of cases each year because of limited resources.[337] The vast majority of claimants, whose cases the government will not pursue, will need to find an attorney interested in investing the time and money in delving into the technicalities of data mining activities, which may be no easy task.[338] Furthermore, plaintiffs would have legitimate claims only if they were subjected to discrimination based on legally protected characteristics such as race or disability. Still, the existence of a disclosure requirement may deter at least some employers from engaging in unlawful discrimination and depriving qualified employees of job opportunities.

3. Addressing Data Mining in the ADA's Definition of Disability

The ADA defines "disability" very broadly[339] and prohibits employers, financial institutions, and others from discriminating against individuals based on a belief that they currently have physical or mental impairments. The ADA's "regarded as" provision explicitly states that an individual is protected by the statute if "he or she has been subjected to an action prohibited under this chapter because of an actual or perceived physical or mental impairment whether or not the impairment limits or is perceived to limit a major life activity."[340]

336. 42 U.S.C. § 2000e-5(e)–(f), 42 U.S.C. § 12117 (addressing EEOC enforcement responsibilities). Title III of the ADA, which covers public accommodations such as financial institutions does not include a similar requirement that plaintiffs exhaust administrative remedies. *See* Hill v. Park, No. 03-4677, 2004 WL 180044 (E.D. Pa. Jan. 27, 2004).

337. *See EEOC Litigation Statistics*, EQUAL EMPLOYMENT OPPORTUNITY COMMISSION, http://www.eeoc.gov/eeoc/statistics/enforcement/litigation.cfm (last visited Nov. 23, 2015) (indicating that in fiscal year 2013, the EEOC filed only 148 lawsuits nationwide).

338. *See* Theodore J. St. Antoine, *Mandatory Arbitration: Why It's Better than It Looks*, 41 U. MICH. J.L. REFORM 783, 790 (2008) (estimating that only 5% of individuals with employment discrimination claims who turn to private attorneys for help are actually able to retain counsel).

339. *See* 42 U.S.C. § 12102 (2010).

340. 42 U.S.C. § 12102(3)(A) (2010).

However, the ADA does not ban discrimination against individuals who are neither currently impaired nor perceived as impaired but are deemed to be at risk of being unhealthy in the future because of their eating habits, exposure to toxins, or a myriad of other concerns.[341] Thus, for example, so long as employers do not consider genetic factors,[342] they can exclude such workers without being challenged.

If open data enables discrimination against high health-risk individuals and such discrimination becomes increasingly common, legislators would be wise to respond to it. An easy fix would be to add language to the "regarded as" provision of the ADA indicating that individuals are also regarded as disabled if they have been subjected to an adverse action because they are perceived as likely to develop physical or mental impairments in the future.

C. CITIZEN SCIENTIST CHAPERONING

Several mechanisms should be developed to assist citizen scientists in conducting, validating, and publishing their research. "Chaperoning" citizen scientists by means of research support and filtering tools could reduce the potential for widespread dissemination of erroneous and harmful research conclusions.[343]

First, government agencies, academic institutions, and other research experts should develop educational resources and best practices guidelines to assist citizen scientists in conducting research.[344] These documents or videos could be posted on database websites, and users could be required or encouraged to review them, along with privacy training materials, before signing data use agreements.[345] Data custodians could also test users

341. *See id.*

342. *See* Hoffman, *supra* note 218 and accompanying text (discussing the Genetic Information Nondiscrimination Act).

343. *See supra* Section IV.C.

344. CDC/ATSDR POLICY ON RELEASING AND SHARING DATA, *supra* note 111, at 7 (urging CDC staff to develop "[i]nstructions for non-CDC users on the appropriate use of the data"); JOHN P. HOLDREN, OFFICE SCI. & TECH. POL'Y, EXEC. OFFICE OF THE PRESIDENT, MEMORANDUM, INCREASING ACCESS TO THE RESULTS OF FEDERALLY FUNDED SCIENTIFIC RESEARCH 6 (2013) https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf (urging federal agencies, in coordination with the private sector, to "support training, education, and workforce development related to scientific data management, analysis, storage, preservation, and stewardship").

345. *See supra* Section V.A.3.

on these materials in order to ensure that they have read and understood them prior to signing the agreement.[346]

Second, citizen scientists should have opportunities to have their work vetted, validated, and published in platforms that are recognized as reliable. Without such mechanisms, readers will be unable to discern whether citizen scientists' findings are trustworthy.

One option is to follow the Wikipedia paradigm. Wikipedia allows the public to post and edit articles, but the site provides some degree of oversight and quality control.[347] Authors can request reviews of their entries from peers, and Wikipedia administrators have authority to delete and undelete pages, protect pages from editing, and take other actions.[348] In extreme cases, administrators, of whom there are over 1,400, can temporarily or permanently bar authors from contributing to Wikipedia because of intentional and persistent misconduct.[349] In addition, Wikipedia has an extensive dispute resolution system for disagreements about the contents of Wikipedia pages.[350] Wikipedia encourages readers who find passages that are biased or erroneous to improve them and discuss the problem with the original author. Parties may also ask for a third opinion or for a moderated discussion through the Dispute Resolution Noticeboard, or they can initiate open requests for comments from the community at large or requests for mediation with help from the Mediation Committee.[351]

A similar venue could be established for the publication of citizen scientists' reports and findings that are not submitted to traditional journals. Opportunities for editing by other professional and amateur scientists, dispute resolution mechanisms, and other forms of oversight would significantly enhance the reliability of posted materials. The venue's policy should also require authors to disclose any computer programs that

346. *See Participation Documents*, *supra* note 321 and accompanying text.

347. *Policies and Guidelines*, WIKIPEDIA (Feb. 5, 2016, 11:49 AM), https://en.wikipedia.org/w/index.php?title=Wikipedia:Policies_and_guidelines&oldid=703418205.

348. *Editor Review*, WIKIPEDIA (Dec. 9, 2015, 2:24 PM) https://en.wikipedia.org/w/index.php?title=Wikipedia:Editor_review&oldid=694475551; *Administrators*, WIKIPEDIA (Jan. 28, 2016, 2:56 PM), https://en.wikipedia.org/w/index.php?title=Wikipedia:Administrators&oldid=70211392.

349. *Id.*; *Policies and Guidelines*, *supra* note 347.

350. *Dispute Resolution*, WIKIPEDIA (Oct. 26, 2015, 4:16 PM), https://en.wikipedia.org/w/index.php?title=Wikipedia:Dispute_resolution&oldid=687601794.

351. *Id.*

they used to analyze the data so that others can replicate and verify their research.[352]

Opportunities for peer review of citizen science research outcomes would provide significant benefits. The contemporary scientific community is open to innovation and several hybrid peer review models are emerging. For example, F1000Research is a pioneering open access journal for life scientists.[353] F1000Research reviews submitted articles internally and, if it initially deems them meritorious, it publishes them within a week of submission, together with their underlying datasets, making all materials publicly available. The service only then sends articles for peer review. Another novelty is that F1000Research discloses its reviewers' identities and enables authors to communicate with the reviewers to address their concerns. Authors may publish revised manuscripts,[354] and articles that peer reviewers approve are indexed in external databases such as PubMed.[355]

Peerage of Science offers a second non-traditional approach.[356] Authors submit manuscripts to the service rather than directly to journals. Authors set their own deadlines for reviews, and any qualified reviewer with a prior peer-reviewed publication can submit a review. A second stage of the process reviews the initial reviewers' assessments.[357] Authors can accept offers from participating journals or export reviews outside of Peerage of Science to journals of their choice.[358]

F1000Research and Peerage of Science demonstrate the contemporary spirit of innovation in the academic community. They are not suggested as venues for amateur citizen scientists, because they are designed for professional scientists producing conventional scholarship. The future, however, may herald different models to chaperone citizen scientists. Whether these follow the Wikipedia paradigm or another path, they would assist not only researchers in improving and publicizing their work,

352. Ari B. Friedman, Letter to the Editor, 370 NEW. ENG. J. MED. 484, 484 (2014) (reviewing Michelle M. Mello et al., *Preparing for Responsible Sharing of Clinical Trial Data*, 369 NEW. ENG. J. MED. 1651 (2013)).

353. *How It Works*, F1000RESEARCH, http://f1000research.com/about (last visited Nov. 23, 2015).

354. *Id.*

355. *FAQs*, F1000RESEARCH, http://f1000research.com/faqs (last visited Nov. 23, 2015).

356. *How It Works*, PEERAGE OF SCIENCE, http://www.peerageofscience.org/how-it-works (last visited Nov. 23, 2015).

357. *Process Flow*, PEERAGE OF SCIENCE, http://www.peerageofscience.org/how-it-works/process-flow (last visited Nov. 23, 2015).

358. *Id.*

but also the reading public in filtering out research findings that have no reliable basis.[359]

D. TORT CLAIM LITIGATION STRATEGIES

Parties who are hurt by citizen scientists' wrongdoing will have a variety of avenues to seek redress. Plaintiffs may allege defamation, interference with economic advantage, public disclosure of private facts, and other claims.[360] Database operators who require data recipients to sign data use agreements may also sue for breach of contract if (1) recipients attempt to re-identify information, use data for commercial or competitive purposes, or violate other agreement provisions, and (2) the breaches damage the database's reputation or economic interests.[361]

Of greater concern are instances in which parties may file suit against citizen scientists who act in good faith but publicize information critical of the plaintiffs' products or conduct. Businesses may hope to intimidate and deter citizen scientists and to force them to disavow and remove any offending material.[362] Citizen scientists who publish their data outside of traditional academic journals will not have a defense based on scrutiny and approval by highly qualified peer reviewers. Such citizen scientists will have no academic institution committed to their vigorous defense.

In some states, defendants will be able to utilize anti-SLAPP legislation and have cases quickly dismissed.[363] If amateur researchers make valuable contributions to science but are routinely harassed through frivolous litigation, additional states may respond with anti-SLAPP statutes that cover such cases.

In the meantime, citizen scientist advocacy organizations can develop educational materials that address strategies to minimize the risk of liability. To this end, the Harvard-affiliated Digital Media Law Project offers "Practical Tips for Avoiding Liability Associated with Harms to

359. Admittedly, even experienced scientists often cannot reach consensus about the validity of research findings and disagree about the accuracy of study outcomes. *See supra* notes 98–102 and accompanying text. However, a filtering mechanism could at least screen out material that no educated reviewer would consider reliable.

360. *See supra* Sections IV.D.1 and IV.D.2.

361. *See supra* notes 309–313 and accompanying text.

362. *See supra* Section IV.D.3.

363. *Id.*

Reputation."[364] The long list of detailed suggestions includes, among others:

- Strive to be as accurate as possible;
- Use reliable sources;
- Seek comment from the subjects of your statements, when appropriate;
- Document your research;
- Keep an eye out for "Red Flag" statements [e.g., explicitly accusing someone of criminal or immoral conduct];
- Be cautious when publishing negative information about businesses;
- Where possible, get consent from the people you cover;
- Be willing to correct or retract your mistakes.[365]

Lawsuits can be expensive and traumatic even if they come to a quick end. Precautions will not prevent litigation in every case, but citizen scientists would be wise to heed experts' advice in order to minimize the likelihood of being sued and facing liability.

VI. CONCLUSION

The medical and scientific communities are rapidly adopting a culture of data sharing, and the expansion of open data practices is widely perceived as inevitable.[366] Many stakeholders are grappling with the legal and ethical implications of public access to patient-related data. For example, the prestigious Institute of Medicine is in the process of crafting a document entitled "Strategies for Responsible Sharing of Clinical Trial Data."[367]

Open medical data have the potential to yield numerous benefits, including scientific discoveries, cost savings, new patient support tools, improved healthcare quality, greater government transparency, and public education.[368] At the same time, open data raise several complex legal and

364. *Practical Tips for Avoiding Liability Associated with Harms to Reputation*, DIGITAL MEDIA LAW PROJECT, http://www.dmlp.org/legal-guide/practical-tips-avoiding-liability-associated-harms-reputation (last updated July 22, 2008).

365. *Id.*

366. *See* Exec. Order No. 13,642, *supra* note 1.

367. *Activity: Strategies for Responsible Sharing of Clinical Trial Data*, INSTITUTE OF MEDICINE, http://www.iom.edu/Activities/Research/SharingClinicalTrialData.aspx (last visited Nov. 23, 2015) (describing the project and its timeline); *see supra* note 228 for interim report.

368. *See supra* Part III.

ethical concerns related to privacy, discrimination, erroneous research findings, and litigation.[369]

Scientists and policy-makers must carefully consider the varied implications of making patient-related big data available to the public. In the future, they may devise a detailed regulatory framework for citizen science.[370] Until then, government, industry, data custodians, and others should implement the more modest interventions proposed in this Article to protect all stakeholders: patients, researchers, businesses, and the public at large.

In his May 2013 executive order, President Obama asserted that "making information resources easy to find, accessible, and usable can fuel entrepreneurship, innovation, and scientific discovery that improves Americans' lives . . . ."[371] Unfortunately, without well-considered responses to the legal and ethical implications of open data, the new trend may generate more harm than good. However, with careful data stewardship, society may well enjoy the new policy's promised bounty.

369. *See supra* Part IV.
370. O'Connor, *supra* note 234, at 481.
371. Exec. Order No. 13,642, *supra* note 1.