SCHOOL OF LAW

CASE WESTERN RESERVE UNIVERSITY

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

Sharona Hoffman

Case Research Paper Series in Legal Studies
Working Paper 2014-21
August 2014

*Draft – article forthcoming in Berkeley Technology Law Journal*

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

**Sharona Hoffman***

*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 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, the development of patient support tools, healthcare quality improvement, greater government transparency, public education,*

* 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 her 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.

*and positive changes in healthcare policy. 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.*

## Table of Contents

## 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 make it easy to find, access, and use.

[1] Exec. Order No. 13642, 78 Fed. Reg. 28111, May 14, 2013. 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.

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 paper. 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 the executive order,[2] it is time to consider carefully the benefits and risks of public access to medical data. The Article's further contribution is that it formulates careful guidelines for data use in order to protect privacy, deter discrimination, and prevent other harms.

Patient-related medical data can now easily be found on the Internet.[3] With its help, ordinary citizens interested in scientific research are taking matters into their own hands. 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, usually without compensation.[5] Do-It-

---

[2] Sylvia M. Burwell et al., *Memorandum for the Heads of Executive Departments and Agencies re: Open Data Policy – Managing Information as an Asset*, May 9, 2013 at http://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf.

[3] *See infra* Part I.

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

[5] National Geographic Education, *Citizen Science*, at http://education.nationalgeographic.com/education/encyclopedia/citizen-science/?ar_a=1.

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, data collections are being furnished to the public by government and private sector sources, and this supply stream will expand considerably in the future.[8] Publicly available resources will be called "public-use data" or "open data" in this Article.

The potential benefits of public access to health information are considerable. In an era of diminishing government funding for research,[9] the widespread availability of high-quality datasets at little or no cost may be very important to continued scientific advancement. Professional researchers as well as talented and dedicated

---

[6] Daniel Grushkin et al., *Seven Myths and Realities about Do-It-Yourself Biology*, SYNTHETIC BIOLOGY PROJECT, Nov. 2013, p. 4, at http://www.synbioproject.org/process/assets/files/6676/7_myths_final.pdf.

[7] Philip Russom, *Big Data Analytics*, TDWI BEST PRACTICES REPORT (4th Quarter 2011), p. 6, *available at* http://public.dhe.ibm.com/common/ssi/ecm/en/iml14293usen/IML14293USEN.PDF.

[8] *See infra* Part I.

[9] Nora Macaluso, *Decade-Long Funding Decline at NIH May Be Poised for Reversal, Collins Says*, 13 BLOOMBERG BNA MED. RES. L & POL'Y REPORT 311 (May 15, 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").

students and amateurs could make important discoveries and answer pressing medical questions,[10] and they can 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 the services the government itself provides. Easily accessible and navigable public-use 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 patient information is de-identified, at least a small risk of re-identification will always remain, and if data is not thoroughly anonymized, the risk of re-identification is

[10] *See infra* Part II.A.
[11] *See infra* note 109 and accompanying text.
[12] *See infra* Part II.C.
[13] *See infra* Part II.D.
[14] *See infra* Part III.A.

substantial.[15] Furthermore, the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule does not cover most of the entities that operate public use databases, and, therefore, they are not subject to detailed privacy regulations.[16] Second, open data may enable discrimination by employers, financial institutions, and anyone else with a stake in people's health.[17] These entities may attempt to re-identify publicly available health records that belong to applicants or employees. In the alternative, they may mine medical data to find statistical associations between particular attributes, habits, or behaviors (e.g. obesity or smoking) and health risks and then, based on their findings, formulate discriminatory policies that exclude individuals who are perceived as high-risk from employment, financial, or other opportunities.[18] Third, amateurs may reach incorrect conclusions and foster misconceptions among the public about human health or the health care industry. They may be able to 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 for purposes of promoting personal

---

[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 *supra* Parts III.A.2 and III.A.3.b & c.

[16] *See infra* Part III.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* Part III.B.

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

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 economic advantage.[20] In some cases lawsuits will be brought 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 their risks. However, it is noteworthy that the research projects contemplated in this Article will not be subject to the federal research regulations, which exempt studies based on records or data that are publicly available and 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

---

[20] *See infra* Part III.D.

[21] *Id.*

[22] 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").

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 I will sample some of the many data collections that are already made publicly available by various government and private entities, examining their contents and any requirements for data use. Part II will analyze the benefits of public access to medical data, and Part III will assess its risks. Part IV will formulate a detailed proposal for legal and policy interventions designed to promote responsible data stewardship and 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 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.

## I. PUBLICLY AVAILABLE BIG DATA SOURCES

Many large databases offer public access to patient-related health information. These databases have been established by federal and state governments as well as by private sector enterprises. No comprehensive catalogue of

[23] *See infra* Part IV.

these sources exists. This Part creates a list consisting of a representative sample of databases that feature public-use medical data.

### A. FEDERAL AND STATE DATABASES

#### *1. HealthData.gov*

HealthData.gov, launched in 2011, is a Department of Health and Human Services website that makes over 1,000 data sets available to researchers, entrepreneurs, and the public free of charge.[24] It thus predates the executive order by two years and establishes a home for the federal government's open data. The data sets are provided by government agencies such as the Centers for Disease Control and Prevention, the Centers for Medicare & Medicaid Services, the National Institutes of Health, the Administration for Children and Families, and several states.[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

---

[24] Kathleen Sebelius, *One Thousand Data Sets and Counting*, HEALTHDATA.GOV, Feb. 26, 2014.

[25] *Id.*

[26] *HealthData.gov*, at http://healthdata.gov/. 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] HealthData.gov, *Vaccination Coverage Among Children 19–35 Months of Age for Selected Diseases, by Race, Hispanic Origin, Poverty Level, and Location*

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 through Healthdata.gov. Two examples, the Centers for Disease Control and Prevention's (CDC) database, CDC Wonder[29] and the Centers for Medicare & Medicaid Services' 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 incidence, 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-

---

*of Residence in Metropolitan Statistical Area*, at http://www.healthdata.gov/data/dataset/selected-trend-table-health-united-states-2011-vaccination-coverage-among-children-19.

[28] *See e.g.* Harnam Singh, *The National Practitioner Data Bank (NPDB) Introduces Interactive Data Analysis Applications*, HEALTHDATA.GOV, May 29, 2014, at http://healthdata.gov/blog/national-practitioner-data-bank-npdb-introduces-interactive-data-analysis-applications; Damon Davis, HHS 2013 Year in Health Data Highlights, HEALTHDATA.GOV, Jan. 16, 2014, at https://www.healthdata.gov/blog/hhs-2013-year-health-data-highlights.

[29] *See* HealthData.gov, *CDC Wonder: Births*, at http://healthdata.gov/data/dataset/cdc-wonder-births-0.

[30] HealthData.gov, *Chronic Condition Data Warehouse*, at http://www.healthdata.gov/data/dataset/chronic-condition-data-warehouse.

[31] CDC, *What Is CDC Wonder?* at http://wonder.cdc.gov/wonder/help/main.html#What is WONDER.

[32] *Id.*

[33] *Id.*

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 Centers for Medicare & Medicaid Services has established the Chronic Condition Data Warehouse (CCW), from which users can 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 privacy board review is necessary.[38] For example, the Medicaid State Drug Utilization File contains information about outpatient drugs for which state Medicaid agencies have paid.[39]

---

[34] *Id.*

[35] *See e.g.* CDC Wonder, *About Natality, 2007-2012*, at http://wonder.cdc.gov/natality-current.html. *See infra* note 302 and accompanying text for further discussion of data use agreements.

[36] Centers for Medicare and Medicaid Services, *Chronic Condition Data Warehouse*, at https://www.ccwdata.org/web/guest/home.

[37] Research Data Assistance Center, *CMS Data Request Center*, at http://www.resdac.org/cms-data/request/cms-data-request-center.

[38] Research Data Assistance Center, *Public-Use Files (PUF)/Non-Identifiable Data Requests*, at http://www.resdac.org/cms-data/request/public-use-files.

[39] Research Data Assistance Center, *Medicaid State Drug Utilization File*, at http://resdac.advantagelabs.com/cms-data/files/medicaid-state-drug-utilization.

### 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]the New Jersey State Assessment Health Data,[41] the North Carolina State Center for Health Statistics,[42] FloridaHealthFinder.gov,[43] and the Minnesota Center for Health Statistics.[44] All of 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;

---

[40] New York State Department of Health, *Health Data New York*, at https://health.data.ny.gov/.

[41] New Jersey State Department of Health, *Welcome to NJSHAD: New Jersey's Public Health Data Resource*, at http://www4.state.nj.us/dhss-shad/home.

[42] North Carolina State Center for Health Statistics, *Statistics and Reports*, at http://www.schs.state.nc.us/data/minority.cfm.

[43] Agency for Healthcare Administration, *State Health Data Directory*, at http://www.floridahealthfinder.gov/StateHealthDataDirectory/.

[44] Minnesota Center for Health Statistics, *Selected Public Health Data Websites*, at http://www.health.state.mn.us/divs/chs/countytables/resources.htm.

[45] Healthcare Cost and Utilization Project, *Overview of HCUP*, at http://www.hcup-us.ahrq.gov/overview.jsp.

- State Ambulatory Surgery Databases; and
- State Emergency Department Databases.[46]

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 attributes that more directly identify patients.[48]

The databases are available for purchase, and purchasers are required to 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

---

[46] *Id.*

[47] Agency for Healthcare Research and Quality, *Databases and Related Tools from the Healthcare Cost and Utilization Project (HCUP) Fact Sheet*, at http://www.ahrq.gov/legacy/data/hcup/datahcup.htm.

[48] Healthcare Cost and Utilization Project, *Overview of the State Inpatient Databases*, at http://www.hcup-us.ahrq.gov/sidoverview.jsp.

[49] Healthcare Cost and Utilization Project, *HCUP Central Distributor*, at http://www.hcup-us.ahrq.gov/tech_assist/centdist.jsp.

[50] Healthcare Cost and Utilization Project, *HCUP Nationwide Inpatient Sample Application*, at http://www.hcup-us.ahrq.gov/db/nation/nis/NISApp_Final.pdf.

[51] Healthcare Cost and Utilization Project, *HCUP Central Distributor*, at http://www.hcup-us.ahrq.gov/tech_assist/centdist.jsp.

(e.g. 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 now require authors to deposit their sequences in GenBank, and all publicly funded laboratories do so as a matter of policy as well.[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 has 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]

---

[52] Healthcare Cost and Utilization Project, *SID/SASD/SEDD Application Kit* (2014), at http://www.hcup-us.ahrq.gov/db/state/SIDSASDSEDD_Final.pdf. Prices range from $35 to over $1600.

[53] GenBank, *GenBank Overview*, at http://www.ncbi.nlm.nih.gov/genbank/.

[54] *Id.*

[55] Jo McEntyre and David J. Lipman, *GenBank – A Model Community Resource?* NATURE WEB DEBATES, at http://www.nature.com/nature/debates/e-access/Articles/lipman.html.

[56] *Id.*

[57] *Id.*, Genbank, *supra* note 53.

[58] McEntyre and Lipman, *supra* note 55.

### 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 in order to learn about healthcare costs, compare prices, and make more informed decisions about insurance plans and healthcare providers. [61]

Similarly, the Centers for Medicare and Medicaid Services (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,

---

[59] Jo Porter et al., *The Basics of All-Payer Claims Databases: A Primer for States* 1 (Jan. 2014), at 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 All Payer Claims Database*, at https://www.cohealthdata.org/; Center for Health Information and Analysis, *All-Payer Claims Database*, at http://www.mass.gov/chia/researcher/hcf-data-resources/apcd/ (requiring applications for Massachusetts data).

[62] Centers for Medicare and Medicaid Services, *Medicare Provider Utilization and Payment Data*, at http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/.

[63] *Id.*

users may obtain hospital-specific charges for particular services and compare prices.[64]

### B. PRIVATE SECTOR DATABASES

#### *1. Dryad Digital Repository*

Dryad is an international repository containing data files that are associated with peer-reviewed scientific articles and other "reputable sources (such as dissertations)."[65] It is a nonprofit organization that is 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]

---

[64] Centers for Medicare and Medicaid Services, *Medicare Provider Utilization and Payment Data: Inpatient*, at http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Inpatient.html. *But see* Patrick T. O'Gara, *Caution Advised: Medicare's Physician-Payment Data Release*, 371 N. ENGL. J. MED. 101 (2014) (discussing the limitations of payment data released by CMS); Dawn Fallik, *For Big Data, Big questions Remain*, 33 HEALTH AFFAIRS 1111, 1111 (2014) (stating that "Medicare's release of practitioner payments highlights the strengths and weaknesses of digging into big data").

[65] Dryad, *The Organization: Overview*, at http://datadryad.org/pages/organization; Dryad, *Frequently Asked Questions*, at http://datadryad.org/pages/faq#depositing.

[66] Dryad, *Pricing Plans and Data Publishing Prices*, at http://datadryad.org/pages/pricing.

[67] Dryad, *Frequently Asked Questions*, at http://datadryad.org/pages/faq#using.

[68] Dryad, *The Repository: Key Features*, at http://datadryad.org/pages/repository.

### 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 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

---

[69] PatientsLikeMe, *Live Better, Together!* at http://www.patientslikeme.com.

[70] PatientsLikeMe, *What Is PatientsLikeMe?* at https://support.patientslikeme.com/hc/en-us/articles/201186434-What-is-PatientsLikeMe-.

[71] PatientsLikeMe, *Does PatientsLikeMe Sell My Data?* at https://support.patientslikeme.com/hc/en-us/articles/201245770-Does-PatientsLikeMe-sell-my-information-.

[72] PatientsLikeMe, *Privacy Policy*, at http://www.patientslikeme.com/about/privacy.

[73] *See e.g.* PatiensLikeMe, *Treatments*, at http://www.patientslikeme.com/treatments.

[74] *See e.g.* PatientsLikeMe, *Welcome to the PatientsLikeMe Public ALS Registry*, at http://www.patientslikeme.com/registry; PatientsLikeMe, *What Information Is Visible on Public Profiles?* at https://support.patientslikeme.com/hc/en-us/articles/201245830-What-information-is-visible-on-public-profiles-.

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*

The Personal Genome Project was launched in 2005 by George Church at Harvard University and 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 relate to individual participants and list date of birth, gender, zip code, height, weight, and race, 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]

## II. THE BENEFITS OF PUBLIC ACCESS TO HEALTH INFORMATION

---

[75] *See e.g.* PatientsLikeMe, *Conditions at PatientsLikeMe*, at http://www.patientslikeme.com/conditions.

[76] Personal Genome Project: Harvard Medical School, *About the PGP*, at http://www.personalgenomes.org/harvard/about-pgp.

[77] *Id.*

[78] Personal Genome Project: Harvard Medical School, *Data & Samples*, http://www.personalgenomes.org/harvard/data. Microbiome data focuses on "the types of bacteria in and on a participant's body." *Id.*

[79] *See e.g.* Personal Genome Project, *Public Genetic Data*, at https://my.pgp-hms.org/public_genetic_data.

[80] Personal Genome Project: Harvard Medical School, *supra* note 76.

Public-use data offers the potential for many valuable benefits. These include new scientific discoveries, research cost savings, the development of tools to help patients navigate the healthcare system, greater government transparency, public education about science and medicine, improvements in 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.

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."[81] 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."[82]

---

[81] Exec. Order No. 13642, *supra* note 1.

[82] Grushkin et al., *supra* note 6, at 4.

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.[83] Likewise, Katherine Aull, a graduate of the Massachusetts Institute of Technology whose father suffered from 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.[84] She built a lab in her closet and used equipment purchased from e-Bay or found in her kitchen.[85]

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

---

[83] Thomas Landrain et al., *Do-It-Yourself Biology: Challenges and Promises for an Open Science and Technology Movement*, 7 SYST. SYNTH. BIOL. 115, 121 (2013); Linda Nordling, *DIY Biotech: How to Build Youself a Low-Cost Malaria Detector*, THE GUARDIAN, April 25, 2014, *available at* http://www.theguardian.com/global-development-professionals-network/2014/apr/25/diy-detector-malaria-eradication-amplino (reporting that Amplino "is virtually ready for area-testing in rural Zambia").

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

[85] Delgado, *supra* note 84, at 70.

analysis and redistribution.[86] Tycho researchers were able to digitize disease surveillance data from the years 1888 to 2011 published in the CDC's *Morbidity and Mortality Weekly Report* and to estimate that since 1924, 103 million incidents of childhood diseases were prevented because of immunizations.[87] This finding will be a useful tool for public health authorities who at times meet resistance to vaccination efforts.

Among the more creative initiatives is a crowdsourcing contest called the Dialogue on Reverse Engineering Assessment and Methods that focused on breast cancer prognosis (DREAM7).[88] 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."[89] DREAM7 provided participants with access to genetic and clinical data from Sage's Synapse, an informatics platform that allows users to share data and access programming codes and analytical

---

[86] University of Pittsburgh, *About Project Tycho™ Data*, at https://www.tycho.pitt.edu/about.php

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

[88] Michael Eisenstein, *Crowdsourced Contest Identifies Best-In-Class Breast Cancer Prognostic*, 7 NATURE BIOTECHNOLOGY 578, 578 (2013).

[89] Enrique Estellés-Arolas and Fernando González-Ladrón-de-Guevara, *Towards an Integrated Crowdsourcing Definition*, 38 J. INFORM. SCIENCE 189, 197 (2012). *See also*, Thea C Norman et al., *Leveraging Crowdsourcing to Facilitate the Discovery of New Medicines*, 3 SCI. TRANSL. MED. 88mr1, 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')").

tools.[90] The contest challenged the crowd to "provide an unbiased assessment of models and methodologies for the prediction of breast cancer survival."[91] A winner was selected from among 1400 entries, and results were published in a scientific journal.[92]

Crowdsourcing is an increasingly popular phenomenon.[93] It has been used for projects ranging from identifying over 1400 automated external defibrillators in public places in Philadelphia to developing a predictive algorithm that is superior to the National Institutes of Health's standard algorithm, the Basic Local Alignment Search Tool (BLAST), which finds regions of local similarity between genetic sequences.[94] The availability of vast amounts of publicly accessible data may make crowdsourcing all the more prevalent. Researchers will likely continue to harness the

---

[90] *About Synapse*, at https://s3.amazonaws.com/static.synapse.org/About_Synapse.pdf.

[91] DREAM, *Sage Bionetworks-DREAM Breast Cancer Prognosis Challenge*, at http://www.the-dream-project.org/challenges/sage-bionetworks-dream-breast-cancer-prognosis-challenge.

[92] Eisenstein, *supra* note 88, at 578.

[93] Benjamin M. Good and Andrew I. Su, *Crowdsourcing for Bioinformatics*, 29 BIOINFORMATICS 1925, 1925 (2013).

[94] *The Accelerating World of Drug Discovery and Commercialization*, 10 TRENDS MAG. Vol. 10, p. 30 (2013), *available at* http://www.cornerstoneadvisors.com/perspectives/october2013/trend-4-the-accelerating-world-of-drug-discovery-and-commercialization/page.aspx?id=1374; BLAST, at http://blast.ncbi.nlm.nih.gov/Blast.cgi.

talents and expertise of willing and able citizen scientists to make important contributions to medical science.[95]

### 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, applicants for NIH grants were estimated to have at most a sixteen percent chance of being funded.[96]

At the same time, despite the abundance of information and medical technology available in the 21st Century, "more than half of medical treatments are used without sufficient proof of their effectiveness."[97] 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

---

[95] Benjamin L. Raynard et al., *Crowdsourcing – Harnessing the Masses to Advance Health and Medicine, a Systematic Review*, 29 J. GEN. INTERN. MED. 187, 201 (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.").

[96] Macaluso, *supra* note 9; Lauren Ingeno, *Crowdfunding Academic Research*, INSIDE HIGHER ED, June 10, 2013, *available at* http://www.insidehighered.com/news/2013/06/10/academic-researchers-using-crowdfunding-platforms#sthash.ziC5DeXs.dpbs (quoting Jonathan Thon, a Harvard Medical School Fellow as stating that the NIH "funds about 10 percent of projects"); Jeannie Baumann, *Neuroscientist Says NIH Funding Squeeze Causing 'Crisis in Biomedical Enterprise*, 13 MED. RES. L. & POL'Y REP. 407, 407 (2014).

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

medicine.[98] Likewise, although physicians have prescribed and studied hormone replacement therapy for post-menopausal women for many 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.[99] 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.[100] No consensus has formed regarding this side effect, and further study is necessary.[101]

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 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 from analyzing the impact of firearms on public health for many years.[102] Likewise, there is often very limited

---

[98] Nikola Biller-Andorno & Peter Jüni, *Abolishing Mammography Screening Programs? A View from the Swiss Medical Board*, 370 N. ENGL. J. MED. 1965, 1965-67 (2014).

[99] 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).

[100] *Id.* at 21-23.

[101] *Id.*

[102] Michael Luo, *N.R.A. Stymies Firearms Research, Scientists Say*, N.Y. TIMES, Jan. 26, 2011, at A1. The moratorium was lifted by an executive order signed by President Obama in January of 2013, entitled "Engaging in Public Health Research on the Causes and Prevention of Gun Violence." 78 Fed. Reg. 4295.

interest in or funding for research relating to rare diseases.[103] Citizen scientists, however, may be highly motivated, for personal rather than profit-seeking reasons, to research those diseases.

The gold standard of medical research has traditionally been randomized, controlled clinical trials.[104] These experimental studies are conducted by means of "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."[105] 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.[106] 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

---

[103] National Organization for Rare Disorders, *Research Grant Policy*, at https://www.rarediseases.org/medical-professionals/research-grants/policy.

[104] Friedrich K. Port, *Role of Observational Studies Versus Clinical Trials in ESRD Research*, 57 KIDNEY INT'L S3, S3 (2000) *available at* http://www.nature.com/ki/journal/v57/n74s/full/4491615a.html (stating that "[r]andomized controlled clinical trials have been considered by many to be the only reliable source for information in health services research."). *See also* Sharona Hoffman, *The Use of Placebos in Clinical Trials: Responsible Research or Unethical Practice?* 33 CONN. L. REV. 449, 452-54 (2001) (describing different clinical trial designs).

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

[106] Eisenstein, *supra* note 88, at 580.

testing of the drug could focus on subjects with that mutation to determine whether it is associated with enhanced responsiveness to the drug.[107]

In the alternative, researchers can undertake observational studies by reviewing existing records and data sets rather than conducting experiments.[108] Professional researchers and citizen scientists will be able to use the large quantities of open data that are now becoming available and minimize research expenses. Researchers may find that existing data collections contain all of the raw data that they need and be spared the work and cost of recruiting human subjects to gather original data. Public-use data can thus prevent costly duplication of effort.[109]

Furthermore, relatively inexpensive big data projects can be funded by an emerging trend called crowdfunding.[110] Crowdfunding is an Internet-based method of fundraising by which one can solicit money from numerous donors, who usually contribute small amounts.[111] Typically,

---

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

[108] 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").

[109] *CDC/ATSDR Policy on Releasing and Sharing Data*, Sept. 7, 2005, at 6, *available at* http://www.cdc.gov/maso/policy/releasingdata.pdf.

[110] 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).

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

crowdfunding for scientific projects raises less than $10,000,[112] but enterprising fund-raisers have frequently surpassed that sum.[113] Public-use data may enable a growing number of projects to have very limited costs that researchers can cover

---

[112] Rachel E. Wheat et al., *Raising Money for Scientific Research Through Crowdfunding*, 28 TRENDS ECOL. & EVOL. 71, 72 (2013), *available at* http://jarrettbyrnes.info/pdfs/Wheat_et_al_2012.pdf.

[113] Ethan O. Perlstein, *Anatomy of the Crowd4Discover Crowdfunding Campaign*, 2 SPRINGERPLUS 560, 561 (2013), *available at* 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 Pass the Hat for Research Funding*, NPR, Feb. 14, 2013, at 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. Kickstarter, at https://www.kickstarter.com/; Experiment, at https://experiment.com/; Indiegogo, at 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, at 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*, Dec. 10, 2013, at 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, at 561.

in creative ways rather than through the traditional channels of government-allocated grant awards.

### 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 in order to provide hospitals and patients with information about options for care after discharge from the hospital.[114] 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.[115]

Similarly, iTriage is a free mobile app and website that allows patients to look up their symptoms and learn about possible causes and treatments.[116] In addition, it assists patients in locating and selecting appropriate care options by providing a variety of information including hospital wait

---

[114] Sebelius, *supra* note 24; Aidin, *Our Story*, at http://www.myaidin.com/ourstory.html.

[115] Aidin, *supra* note 114.

[116] iTriage, *iTriage Helps People Take charge of Their Health*, https://about.itriagehealth.com/for-consumer/what-is-itriage/.

times and physician ratings.[117] iTriage uses publicly available data from HHS, the FDA, and other sources.[118]

Other examples are the state all-payer claims databases and the Medicare providers utilization and payment data.[119] These enable patients to become more educated about healthcare costs and to compare prices for various inpatient and outpatient services.[120]

### 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[121] allow viewers to gain significant insight into the information that the government has collected about individuals and the healthcare industry. In some cases, such insight may generate public debate and critique of government investigative policies that could lead to positive policy changes.[122]

In addition, public-use data can function as an important educational tool.[123] Patients can research their own

---

[117] *Id.*

[118] iTriage, *About Our Medical Content*, at https://about.itriagehealth.com/company-info/medical-content/; Sebelius, *supra* note 24.

[119] *See supra* Part I.A.5.

[120] *Id.*

[121] *See supra* Part I.

[122] *CDC/ATSDR Policy on Releasing and Sharing Data*, *supra* note 109, at 4 (stating that data sharing can "build trust with outside partners and the public by allowing open critique of CDC investigations).

[123] Grushkin, *supra* note 6, at 4 (stating that "wider access to the tools of biotechnology, particularly those related to the reading and writing of

conditions, find doctors with special expertise, be better prepared for their medical appointments, and assess different treatment options that they are given.[124] Furthermore, the general public can learn about the healthcare system, healthcare costs, disease trends, genetics, research and public health initiatives, and much more.[125] Ordinary citizens and students at all levels 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, 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 real cases.[126] 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

DNA, has the potential to spur global innovation and promote biology education and literacy.").

[124] 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.

[125] *See supra* Part I.

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

Open data can fuel improvements in healthcare quality and health policies. A recent report from New York State provided a number of compelling illustrations.[127] 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.[128] Likewise, annual reports of cardiac surgery mortality rates, linked to the hospitals and surgeons who provided care, induced low-scoring facilities to undertake quality improvement initiatives and several physicians who had performed poorly to leave practice.[129]

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, organized by school district, and news outlets highlighted the disparities in 2013, some school administrators decided to improve school policies despite cost concerns.[130] A 2014 report in *Crain's New York Business* that publicized hospital cost disparities (e.g. 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.[131]

---

[127] *Id.*
[128] *Id.*
[129] *Id.*
[130] *Id.*
[131] *Id.*

## III. RISKS OF PUBLIC ACCESS TO HEALTH DATA

Although the benefits of opening health data resources to the public are considerable, the risks of doing so are not inconsequential. The federal research regulations do not cover studies that are not funded or conducted by federal government agencies or that use publicly available data,[132] and therefore, such studies are not subject to any formal oversight. Furthermore, the HIPAA Privacy Rule and state privacy laws most likely will not govern open databases.[133] 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.

### A. PRIVACY BREACHES

The potential for privacy breaches is the first risk that may come to mind with respect to public use of patient-related medical big data. Disclosure of medical records is governed by the HIPAA Privacy Rule,[134] the Privacy Act,[135] and numerous state privacy laws.[136] However, the laws and

---

[132] 45 C.F.R. §§46.101(a) & (b)(4) (2013).

[133] *See infra* Part III.A.1.

[134] 45 C.F.R. §§ 160.101-.534 (2013).

[135] 5 U.S.C. § 552a (2010).

[136] AMERICANS HEALTH LAWYERS ASSOCIATION, STATE HEALTHCARE PRIVACY LAW SURVEY (2013); LawAtlas, *Public Health Departments and State Patient Confidentiality Laws Map*, at http://lawatlas.org/preview?dataset=public-health-departments-and-state-patient-confidentiality-laws.

regulations do not cover all data holders who make medical information publicly available.[137] In addition, public-use data is generally presented in de-identified form[138] and thus is exempt from the disclosure restrictions established in these laws and regulations.[139] 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, however, provides patients with comprehensive protection, and even in the aggregate, they leave many gaps. The following discussion describes laws and regulations that are relevant to the disclosure of patient-related data for public-use.

#### 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.[140] 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.[141] It does not apply to any government agencies or private enterprises that are not acting in these capacities.

---

[137] *See infra* Parts III.A.1 and III.A.3.a.
[138] *See supra* Part I.
[139] *See infra* Part III.A.1.
[140] 45 C.F.R. §§ 164.508 - .510 (2013).
[141] 45 C.F.R. §§ 160.102-160.103 (2013); 42 U.S.C. §17934 (2010).

Thus, HIPAA does not regulate many of the websites discussed in Part I of this Article, such as those operated by the 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.[142] Consequently, the federal regulations do not govern data that custodians de-identify[143] 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.[144] 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."[145] Consequently, the Privacy Act exempts the government's dissemination of de-identified information on HealthData.gov or other websites.

**c. State Laws**

---

[142] 45 C.F.R. § 160.103 (2013).

[143] *See infra* Part III.A.2 (discussing HIPAA's requirements for de-identification).

[144] 5 U.S.C. §552a (2010).

[145] *Id*. at §552a(a)(4).

All states have recognized a common law or statutory right to privacy,[146] and all have statutes that address privacy concerns.[147] A thorough analysis of state law is beyond the scope of this Article.[148] In general, the state laws are varied and inconsistent, often providing piecemeal protection for some types of data but not others.[149] Moreover, like the HIPAA Privacy Rule and the Privacy Act, the states typically allow disclosure of de-identified health information without patient authorization.[150] Therefore, most of the public-use data resources contemplated in this Article would not be governed by state law.

---

[146] Corrine Parver, *Patient-Tailored Medicine, Part Two: Personalized Medicine and the Legal Landscape*, 2 J. HEALTH & LIFE SCI. L. 1, 32 (2009).

[147] AMERICANS HEALTH LAWYERS ASSOCIATION, *supra* note 136; LawAtlas, *supra* note 136.

[148] For detailed information about state privacy and confidentiality laws s*ee* AMERICANS HEALTH LAWYERS ASSOCIATION, *supra* note 136; LawAtlas, *supra* note 136.

[149] Deven McGraw et al., *Privacy As An Enabler, Not An Impediment: Building Trust Into Health Information Exchange*, 28 HEALTH AFFAIRS 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.").

[150] 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).

### 2. *De-identification*

The foregoing discussion raises the following critical questions: 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.[151] The Department of Health and Human Services issued guidance that endorsed several de-identification techniques:

1) *Suppression*, which involves redaction of particular data features prior to disclosure (e.g. removing zip codes, birthdates, income);
2) *Generalization*, which involves transforming particular information into less specific representations (e.g. indicating a 10-year age range instead of exact age); and
3) *Perturbation*, which involves exchanging certain data values for equally specific but different values (e.g. changing patients' ages).[152]

---

[151] 45 C.F.R. § 164.514(b)(1) (2013).

[152] U.S. Department of Health and Human Services, *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule* (2012), at http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html#guidancedetermination (noting that techniques such as suppression and generalization are often used in combination).

The HIPAA Privacy Rule offers detailed guidance concerning suppression, listing 18 identifiers that should be removed in order to render data fully de-identified. They are:

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

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

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

(C) All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older;
(D) Telephone numbers;
(E) Fax numbers;
(F) Electronic mail addresses;
(G) Social security numbers;
(H) Medical record numbers;
(I) Health plan beneficiary numbers;
(J) Account numbers;
(K) Certificate/license numbers;
(L) Vehicle identifiers and serial numbers, including license plate numbers;
(M) Device identifiers and serial numbers;
(N) Web Universal Resource Locators (URLs);
(O) Internet Protocol (IP) address numbers;
(P) Biometric identifiers, including finger and voice prints;
(Q) Full face photographic images and any comparable images; and

(R) Any other unique identifying number, characteristic, or code….[153]

Health information that has all 18 identifiers removed in accordance with the HIPAA "safe harbor" provision is considered per se de-identified unless a covered entity knows that the data can be used on its own or together with other information to identify a data subject.[154] 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.[155] 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.[156] Nevertheless, many are not required to do so because not all data holders are covered by the HIPAA Privacy Rule and its data disclosure and de-identification guidelines. Consequently, the patient authorization and de-identification practices that they choose to implement may deviate from HIPAA standards and leave data more vulnerable to attack.

---

[153] 45 C.F.R. § 164.514(b)(2)(i) (2013).
[154] *Id*. at §164.514(b)(2)(ii).
[155] Khalid El Emam et al., *Evaluating Predictors of Geographic Area Population Size Cut-offs to Manage Re-identification Risk*, 16 J AM MED INFORM ASSOC. 256, 256-57 (2009); Felix T. Wu, *Defining Privacy and Utility in Data Sets*, 84 U. COLO. L. REV. 1117, 1156 (2013).
[156] *See supra* Part I.

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 testing their skills and determining whether they are able to 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-covered private entities.[157] 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 uniquely

[157] *See supra* note 141 and accompanying text.

identified based on their full postal code, date of birth, and gender.[158]

Data holders' de-identification practices do, in fact, vary. A 2013 survey found that thirty-three states released patient hospital discharge data to the public, but only seven de-identify them in a manner that would conform to the HIPAA Privacy Rule's standard.[159] Many states released the month or quarter of hospital admission and/or discharge and 5-digit zip codes.[160] Datasets with these details are more vulnerable to re-identification than those that are de-identified in accordance with HIPAA guidance. That is because 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,[161] or news reports.[162] 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 it is that an adversary will be able to link names to the purportedly anonymized health information.

---

[158] Khaled El Emam, *The Re-identification Risk of Canadians from Longitudinal Demographics*, 11 BMC MED. INFORM. DECIS. MAK. 46, 51 (2011).

[159] Sean Hooley & Latanya Sweeney, *Survey of Publicly Available State Health Databases* (2013), p. 4, at http://dataprivacylab.org/projects/50states/1075-1.pdf

[160] *Id.* at 4-7.

[161] *See infra* note 191 and accompanying text (discussing information that third parties can purchase about individuals).

[162] Arvind Narayanan and Vitaly Shmatikov, *Privacy and Security: Myths and Fallacies of "Personally Identifiable Information,"* 53 COMMUNICATIONS OF THE ACM 24, 26 (June 2010); Electronic Privacy Information Center, *Re-Identification*, at http://epic.org/privacy/reidentification/.

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 Massachusetts Institute of Technology in 1996.[163] 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 was able to match voter names to hospital records.[164]

In a more recent effort, Dr. Sweeney and colleagues worked to re-identify publicly available profiles in the Personal Genome Project[165] that contained medical and genomic information as well as date of birth, gender, and zip code.[166] They linked the demographic data to voter lists or other public records that featured names and were able to identify 84 to 97 percent of Personal Genome Project profiles.[167]

---

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

[164] *Id.*; Kathleen Benitez & Bradley Malin, *Evaluating Re-identification Risks with Respect to the HIPAA Privacy Rule*, 17 J. AM MED. INFORM. ASSOC. 169, 169 (2010).

[165] *See supra* Part I.B.3.

[166] Latanya Sweeney et al., *Identifying Participants in the Personal Genome Project by Name*, Harvard University Data Privacy Lab, White Paper 1021-1, April 24, 2013. http://dataprivacylab.org/projects/pgp/.

[167] *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.

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 $50. She attempted to match hospitalization records to 81 newspaper stories about accidents and injuries in 2011 and was able to determine the name of the patient to whom the records belonged in 35 (or forty-three percent) of the cases, based on the news accounts.[168]

### 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, in some circumstances, will be able to re-identify records that have been de-identified in compliance with HIPAA guidelines.[169] 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-

[168] Latanya Sweeney, *Matching Known Patients to Health Records in Washington State Data*, at http://dataprivacylab.org/projects/wa/1089-1.pdf; Jordan Robertson, *States' Hospital Data for Sale Puts Privacy in Jeopardy*, BLOOMBERG, Jun 5, 2013, at http://www.bloomberg.com/news/2013-06-05/states-hospital-data-for-sale-puts-privacy-in-jeopardy.html.

[169] Hoffman & Podgurski, *supra* note 15, at 105-07.

identified records falls in the range of 0.01 to 0.25%.[170] 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 American population as a whole, consisting of 319 million individuals.[171]

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.[172]

### c. The Peculiarities of Genetic Information

The HIPAA Privacy Rule does not provide explicit guidance concerning the de-identification of genetic

---

[170] Khaled El Emam et al., *A Systematic Review of Re-Identification Attacks on Health Data*, 6 PLoS ONE e28071, *available at* http://www.plosone.org/article/fetchObject.action?uri=info%3Adoi%2F10.1371%2Fjournal.pone.0028071&representation=PDF (finding a re-identification rate of .013%); NAT'L COMM. ON VITAL & HEALTH STATISTICS, REPORT TO THE SECRETARY OF HEALTH AND HUMAN SERCIES ON ENHANCED PROTECTIONS FOR USES OF HEALTH DATA: A STEWARDSHIP FRAMEWORK FOR "SECONDARY USES" OF ELECTRONICALLY COLLECTED AND TRANSMITTED HEALTH DATA 36 n.16 (2007), available at www.ncvhs.hhs.gov/071221lt.pdf.

[171] *See* United States Census Bureau, *U.S. and World Population Clock*, at http://www.census.gov/popclock/.

[172] Khaled El Emam, *Methods for the De-identification of Electronic Health Records for Genomic Research*, 3 Genome Medicine 25, 27 (2011).

information,[173] such as the genetic sequences available through GenBank.[174] Many commentators have expressed concern that adversaries could re-identify anonymized genetic information using a variety of techniques.[175] Researchers believe that people can be uniquely identified through a sequence of only 30 to 80 out of 30 million single-nucleotide polymorphisms (SNPs).[176] In 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.[177] 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.[178] Given genotype frequencies for a study cohort, it is possible to determine if a particular individual is in the cohort if one knows the

---

[173] 45 C.F.R. § 164.514(b)(2)(i) (2013); El Emam, *supra* note 172, at 27.

[174] *See supra* Part I.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.").

[175] El Emam, *supra* note 172, at 27; Dina N. Paltoo et al., *Data Use under the NIH GWAS Data Sharing Policy and Future Directions*, 46 NATURE GEN. 934, 937 (2014).

[176] El Emam, *supra* note 172, at 27; Liina Kamm et al., *A New Way to Protect Privacy in Large-Scale Genome-Wide Association Studies*, 29 BIOINFORMATICS 886, 886 (2013). A single-nucleotide polymorphism is a "variation at a single position in a DNA sequence among individuals." Scitable, *Glossary*, at http://www.nature.com/scitable/definition/single-nucleotide-polymorphism-snp-295.

[177] Melissa Gymrek et al., *Identifying Personal Genomes by Surname Inference*, 339 Science 321, 321 (2013).

[178] David W. Craig et al., *Assessing and Managing Risk when Sharing Aggregate Genetic Variant Data*, 12 NAT. REV. GENET. 730, 730 (2012).

individual's genotype and has a reference set of allele frequencies for the underlying population.[179] 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 many parties who will use it to further their own economic best interests.[180] 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.[181] This part will focus on three examples: employers, financial institutions,

---

[179] *Id.* at 734-35. An allele is one of several variations of a gene. U.S. National Library of Medicine, *Genetics Home Reference*, at http://ghr.nlm.nih.gov/glossary=allele.

[180] *See* Narayanan & Shmatikov, *supra* note 162, 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.").

[181] *Id.*; EXECUTIVE OFFICE OF THE PRESIDENT, BIG DATA: SEIZING OPPORTUNITIES, PRESERVING VALUES 51 (May 2014), at 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 marketers. Employers have a strong incentive to identify 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 an 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[182] or offer special promotions to some consumers but not others.[183]

### *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.[184] Employers may have good economic reasons to strive for the healthiest

---

[182] Lori Andrews, *Facebook Is Using You*, N.Y. TIMES, Feb. 4, 2012, *available at* http://www.nytimes.com/2012/02/05/opinion/sunday/facebook-is-using-you.html?pagewanted=all&_r=0.

[183] EXECUTIVE OFFICE OF THE PRESIDENT, *supra* note 181, at 47.

[184] *See* Bruce Japsen, *U.S. Workforce Illness Costs $576B Annually From Sick Days To Workers Compensation*, FORBES, Sept. 12, 2012, *available at* 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).

possible workforce, but they are constrained by federal and state laws that prohibit discrimination based on a variety of protected classifications, including disability and genetic information.[185] 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 interactions as never before.[186] 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[187] should understand that it will be accessible to anyone and everyone.

---

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

[186] Steve Lohr, *Unblinking Eyes Track Employees: Workplace Surveillance Sees Good and Bad*, N.Y TIMES, June 21, 2014, *available at* http://www.nytimes.com/2014/06/22/technology/workplace-surveillance-sees-good-and-bad.html?emc=eta1.

[187] *See supra* Parts I.B.2 & 3.

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.[188] They also ask applicants for permission to obtain their credit reports.[189] It is therefore not far-fetched to assume that they 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.[190]

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-

---

[188] Greg Fish & Timothy B. Lee, *Employer Get Outta My Facebook*, BLOOMBERG BUSINESSWEEK, at 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, October 9, 2010, *available at* http://www.nytimes.com/2010/10/10/jobs/10search.html.

[189] Gary Rivlin, *The Long Shadow of Bad Credit in a Job Search*, N.Y. TIMES, May 11, 2013, *available at* http://www.nytimes.com/2013/05/12/business/employers-pull-applicants-credit-reports.html?pagewanted=all.

[190] *See supra* Part III.A.3.a.

identified records. For example, data miners may be able to obtain individuals' detailed purchasing histories or web-browsing histories from database marketers such as Acxiom.[191] 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 all of their medical details.

Experienced data miners, aided by contemporary technology, often have no difficulty achieving re-identification. 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."[192]

The Americans with Disabilities Act prohibits employers from engaging in disability-based discrimination.[193] The law allows employers to conduct medical inquiries and

---

[191] *See* Alice E. Marwick, *How Your Data Are Being Deeply Mined*, N.Y. REV. BOOKS, Jan. 9, 2014, *available at* 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"); Acxiom at http://www.acxiom.com/.

[192] Narayanan & Shmatikov, *supra* note 162, at 26. *See also*, Electronic Privacy Information Center, *supra* note 162 (stating that "'anonymized' data can easily be re-identified").

[193] 42 U.S.C. § 12112(a) (2010).

examinations within certain limits to determine fitness for duty,[194] but workers who feel that an employer denied them a job opportunity because of information it discovered, may sue the employer.[195] 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 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.[196] As a hypothetical example, the results of a search might lead an employer to conclude that Hispanic women over 50 are more prone to several cancers than other individuals, and consequently, to decline to hire Hispanic women over 50.[197]

---

[194] 42 U.S.C. § 12112(d) (2010).

[195] 42 U.S.C § 12117(a) (2010).

[196] CDC Wonder, *United States Cancer Statistics, 1999-2010 Incidence Request*, http://wonder.cdc.gov/cancer-v2010.HTML.

[197] *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).

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.[198] The study concluded that "Chinese older adults experience disproportionate health disparities," suffering from significant physical, psychological, financial, and social challenges.[199] Though this was far from the study's intention, readers of the report may think twice about hiring people of Chinese ancestry who are 60 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,[200] 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 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

---

[198] XinQi Dong et al., *The PINE Report*, p. v, *available at* http://chinesehealthyaging.org/files/PINE_Final_Reports/All.pdf.

[199] *Id.* at v & 40.

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

protected categories and then somehow decipher the employer's motivation for doing so. Furthermore, many courts disallow multi-factor claims involving age.[201] These courts perceive "age plus" cases as prohibited by a Supreme court decision, *Gross v. FBL Financial Services, Inc.*, that held that an age discrimination plaintiff must prove that age is the "but for" reason for the adverse action at issue.[202]

Anonymized data can provide other opportunities for discrimination as well.[203] 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 in order to detect associations between individual characteristics or behaviors and poor physical or mental health.[204] Then, they would try to determine from job applications, interviews, and reference or background checks 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."[205] Several websites such as "Lifespan Calculator" and

---

[201] Day, *supra* note 197, at 449.
[202] *Id.* at 466-67; Gross v. Fin. Servs., Inc., 557 U.S. 167, 177-78 (2009).
[203] Michael Schrage, *Big Data's Dangerous New Era of Discrimination*, HARV. BUS. REV., Jan. 29, 2014, at http://blogs.hbr.org/2014/01/big-datas-dangerous-new-era-of-discrimination/.
[204] *See* EXECUTIVE OFFICE OF THE PRESIDENT, *supra* note 181, at 45-47 (discussing algorithms).
[205] Citron & Pasquale, *supra* note 180 at 2.

"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.[206] 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.[207] They also ask a small number of questions about family and personal medical history, which, for employers, could trigger violations of federal anti-discrimination law.[208] 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 means to determine 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.[209] In the future, they might disqualify applicants

---

[206] Northwestern Mutual, *Lifespan Calculator*, at http://media.nmfn.com/tnetwork/lifespan/; *How Long Will I Live?* at http://gosset.wharton.upenn.edu/mortality/perl/CalcForm.html.

[207] *See supra* note 206.

[208] *See* Genetic Information Nondiscrimination Act, 42 U.S.C. §§ 2000ff(4) (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).

[209] Roberts, *supra* note 184, at 577-79.

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 (e.g. race or disability) rather than on behavior (e.g. consumption of fatty food or a sedentary lifestyle).[210] 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.[211]

### 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.[212] Nothing will prevent them

---

[210] *Id.* at 604-07.

[211] *See* Hoffman, supra note 185, 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 Non-Discrimination Act, Pub. L. No. 110-233, 122 Stat. 881 §§ 201(4) & 202(a) (2008); 42 U.S.C. §§ 2000ff(4) & 2000ff-1(a) (Supp. 2010).

[212] Jessica Silver-Greenberg and Michael Corkery, *Bank Account* Screening *Tool Is Scrutinized as Excessive, N.Y. Times*, June 15, 2014, *available at* http://dealbook.nytimes.com/2014/06/15/bank-account-screening-tool-is-scrutinized-as-excessive/?_php=true&_type=blogs&emc=eta1&_r=0.

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.[213]

The Americans with Disabilities Act prohibits disability-based discrimination by places of public accommodation, that is, establishments that provide services to the public, including banks and other financial institutions.[214] 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 most probably go unchallenged.

Marketers and advertisers too 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.[215] 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

[213] *See supra* Part III.B.1 (discussing potential discrimination by employers).
[214] 42 U.S.C. § 12181(7)(F) & 12182(a) (2010).
[215] Andrews, *supra* note 182.

2012 *Forbes* magazine article entitled "How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did."[216]

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 higher-income areas received higher discounts than people in lower-income areas.[217]

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, selective advertising or promotional offers and discounts are unlikely

---

[216] Kashmir Hill, *How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did*, FORBES, Feb. 16, 2012, *available at* 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").

[217] EXECUTIVE OFFICE OF THE PRESIDENT, *supra* note 181, at 46-47.

to be found to violate anti-discrimination laws.[218] 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.[219]

### C. PROPAGATION OF INCORRECT AND HARMFUL RESEARCH CONCLUSIONS

Citizen science can lead to valuable and illuminating discoveries.[220] At the same time, however, amateurs may reach incorrect conclusions.[221] 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

---

[218] Schrage, *supra* note 203 (stating that it is unclear "where value-added personalization and segmentation end and harmful discrimination begins.").

[219] Crawford & Schultz, *supra* note 180, 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).

[220] *See supra* Part II.A.

[221] INSTITUTE OF MEDICINE, DISCUSSION FRAMEWORK FOR CLINICAL TRIAL DATA SHARING: GUIDING PRINCIPLES, ELEMENTS, AND ACTIVITIES 13 (2014) *available at* 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").

Blog"[222] and "How to Promote Your Article Online."[223] In some cases, the media, celebrities, and politicians highlight the work of ordinary citizens,[224] 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 attention through word of mouth and social media as happens when a YouTube video or blog post "goes viral."[225]

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.[226] Some commentators describe this phenomenon in terms of a shift from

---

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

[223] Daniel Vahab & Lisa Chau, *How to Promote Your Article Online*, THE SOCIAL MEDIA MONTHLY, Nov. 30, 2012, at http://thesocialmediamonthly.com/how-to-promote-your-article-online/.

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

[225] *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, *available at* http://www.newyorker.com/reporting/2014/07/21/140721fa_fact_mnookin?currentPage=all&mobify=0 (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).

[226] R. J. 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.").

"intermediation" to "apomediation."[227] Traditionally, peer reviewed journals served as necessary intermediaries between researchers and readers and thus were gatekeepers for scientific knowledge. The Internet has now triggered disintermediation and increased use of apomediaries, agents or tools that guide readers to information without being required middlemen.[228] Many reports published on websites look highly professional and may seem credible to general readers, who are not always sophisticated about distinguishing between reliable and questionable sources of information.[229]

---

[227] 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), p. 5 (coining the term "apomediation").

[228] Eysenbach, *supra* note 227 at 5.

[229] *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 226, 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. INF. SCI. 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 SCIENCE COMMUNICATION 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.[230] 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.[231] Highly skilled analysts should be able to recognize data quality problems, adjust for them, and estimate error rates, but amateurs may not know how to do so.[232]

Furthermore, scientific studies can be flawed because of a variety of biases. Selection bias arises when the group of subjects studied is not representative of the population as a whole, and thus, researches cannot generalize study results.[233] 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

---

[230] Sharona Hoffman & Andy Podgurski, *The Use and Misuse of Biomedical Data: Is Bigger Really Better?*, 39 AM. J. L. & MED. 497, 515-27 (2013).
[231] *Id.* at 515-21.
[232] *Id.* at 530-32.
[233] *Id.* at 521-23.

consider that affect treatment choices and outcomes, and thus, the study's results are skewed.[234] 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.[235] 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.[236] Consequently, researchers face many hurdles and must conduct their studies very skillfully in order to derive valid results.

Researchers must be particularly sensitive to the difference between *association* and *causation*.[237] 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.[238] 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

---

[234] *Id.* at 523-25.

[235] Sharona Hoffman & Andy Podgurski, *Big Bad Data: Law, Public Health, and Biomedical Databases*, 41 J. L. MED. & ETHICS 56, 58 (Supp. 2013).

[236] *Id.*

[237] Austin Bradford Hill, *The Environment and Disease: Association or Causation*, 58 PROC. ROYAL SOC'Y MED. 295, 295-300 (1965); Arvid Sjolander, *The Language of Potential Outcomes*, in CAUSALITY: STATISTICAL PERSPECTIVES AND APPLICATIONS 6, 9 (Carlo Berzuini et al. eds. 2012).

[238] 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).

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 in fact 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[239] may add another element of uncertainty to research quality. Crowdfunding does not depend on peer review of carefully written grant proposals by professional experts.[240] Rather, researchers aim to appeal to a large number of donors through videos and social media campaigns.[241] Some commentators have accused crowdfunding of turning "science into a popularity contest."[242] 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.[243] Consequently, studies that are funded in this manner may not always 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

---

[239] *See supra* notes 111-113 and accompanying text.
[240] Karen Kaplan, *Crowd-Funding: Cash on Demand*, 497 Nature 147, 148 (2013).
[241] *Id.*
[242] Palca, *supra* note 113.
[243] Kaplan, *supra* note 240, at 148.

the opposite case, overly sanguine about their symptoms and fail to seek needed medical care.

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.[244]

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.[245] While the study was later retracted,[246] the belief that vaccinations can lead to autism gained a considerable foothold and still needs to be explicitly repudiated on the CDC's website.[247]

---

[244] Michelle Mello et al., *Preparing for Responsible Sharing of Clinical Trial Data*, 369 N. ENGL. 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.").

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

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

[247] *Measles, Mumps, and Rubella (MMR) Vaccine*, CENTERS FOR DISEASE CONTROL AND PREVENTION,

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.[248] 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[249] and the outcry that the Patient Protection and Affordable Care Act (aka Obamacare) would authorize "death panels" to decide which patients should live and which should die.[250] In both cases, the arguments gained popularity because high-profile public

---

http://www.cdc.gov/vaccinesafety/Vaccines/MMR/MMR.html (last updated Feb. 7, 2011).

[248] *See supra* notes 226-228 and accompanying text.

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

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

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*.[251] The author argued that African-Americans are typically less intelligent and more aggressive, assertive, and impulsive than Whites.[252] In addition, according to the author, African-Americans are more likely to commit crimes because they suffer from "an absence of conscience" and an inability to engage in self-monitoring and have less free will and a different moral orientation from Whites.[253] 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 research outcomes that were invalid 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 in some way. Thus,

---

[251] MICHAEL LEVIN, WHY RACE MATTERS: RACE DIFFERENCES AND WHAT THEY MEAN (Praeger 1997).
[252] *Id.* at 213.
[253] *Id.* at 213, 322.

litigation could be used 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).

### *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).[254]

Establishing a successful defamation claim is no easy task, and plaintiffs must meet a high standard of proof.[255]

---

[254]Matthew E. Kelley & Steven D. Zansberg, *A Little Birdie Told Me, "You're A Crook": Libel in the Twittersphere and Beyond*, 30-Mar COMMUNICATION LAWYER 1 (2014); Restatement (Second) of Torts § 558 (1977).

[255] 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.").

Electronic speech is entitled to the same stringent First Amendment protections as print communication.[256]

Nevertheless, both individuals and entities may bring defamation claims.[257] 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.[258] 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.[259] For example, defamation

---

[256] Reno v. American Civil Liberties Union, 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].").

[257] 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).

[258] Restatement (2d) 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).

[259] 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 (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.").

suits have been filed by businesses in response to negative reviews on the website Yelp.[260]

A particularly memorable defamation case brought by industry involved a discussion on Oprah Winfrey's television show.[261] 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."[262] At one point in the show Ms. Winfrey stated that she was "stopped cold from eating another burger."[263] Subsequently, several Texas cattle men 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.[264] Fortunately for Oprah, the defendants prevailed on all claims.[265]

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

---

[260] Yelp, Inc. v. Hadeed Carpet Cleaning, Inc. 752 S.E.2d 554 (Va.App. 2014); Bently Reserve L.P. v. Papaliolios, 218 Cal.App.4th 418 (1st Dist. 2013).

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

[262] *Id.* at 682-84.

[263] *Id.* at 688.

[264] *Id.* at 682.

[265] *Id.* at 680.

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.[266]

### *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.[267] In addition, companies that feel their products have been inappropriately 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.[268] Individuals and entities that have been subjected to published criticism or negative commentary

---

[266] *But see infra* Part III.D.3 (discussing anti-SLAPP statutes).

[267] Winfrey, *supra* note 261 at 682. *See id.* at 685 for a discussion of the elements of a business disparagement claim.

[268] Crown Imports, LLC v. Superior Court, 223 Cal. App. 4th 1395, 1404 (2d Dist. 2014).

often assert allegations of tortious interference with economic advantage alongside defamation claims.[269]

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."[270] 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.[271] 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

---

[269] Digital Media Law Project, *Responding to Strategic Lawsuits Against Public Participation (SLAPPs)*, at http://www.dmlp.org/legal-guide/responding-strategic-lawsuits-against-public-participation-slapps.
[270] Diaz v. Oakland Tribune, Inc., 139 Cal.Rprt. 762, 768 (1983).
[271] Digital Media Law Project, *supra* note 269.

by the First Amendment of the U.S. Constitution."[272] For example, SLAPPs have been filed by businesses as a form of retaliation against consumers who posted negative comments about them on social networking sites.[273] 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.[274] These laws enable defendants subject to certain frivolous allegations to have SLAPPS dismissed quickly and to recover costs and attorneys' fees.[275] The statutes can vary significantly.[276] Pennsylvania's 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 …."[277] By contrast, in California the law is much broader[278] and covers "written or oral statement[s] or writing made in a place

---

[272] 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).

[273] *Id*. at 222-23; Rex Hall, Jr., *Firm Sues WMU Student Over Facebook Page; Towing Company Seeks $750,000 in Damages for Online Criticism*, THE GRAND RAPIDS PRESS, April 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).

[274] Digital Media Law Project, *supra* note 269.

[275] *Id.*

[276] Richards, *supra* note 272, at 232.

[277] 27 PA. CONS. STAT. ANN. § 8302(a) (2001).

[278] CAL. CODE CIV. PRO. §§ 425.16(e) & 425.17 (West 2011).

open to the public or a public forum in connection with an issue of public interest." The Pennsylvania law allows defendants to request hearings at which the court will determine whether they are entitled to immunity.[279] 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.[280] Protection is inconsistent across jurisdictions but may be very helpful to some victims of frivolous litigation initiated for purposes of harassment and intimidation.

## IV. 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 that it provokes significant concerns. Should legislators and regulators respond in any way to this emerging phenomenon? The law must balance the interests of a variety of stakeholders: patients, professional researchers, citizen scientists, the 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,

[279] 27 PA. CONS. STAT. ANN. § 8303 (2001).
[280] CAL. CODE CIV. PRO. § 425.16(b)-(c) (2011).

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,[281] 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.

[281] *See infra* Part III.A.3.

### a. Expand 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.[282] 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 Rule and its enabling legislation[283] 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:

> (A) 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, healthcare payer, governmental unit, information or computer management entity, school, health researcher, healthcare facility, clinic, healthcare provider, or person who maintains an Internet site.[284]

---

[282] 45 C.F.R. §§ 160.102-160.103 (2013); 42 U.S.C. §17934 (2010).
[283] 45 C.F.R. §160.103 (2013) and 42 U.S.C. §1320d-1(a) (2010).
[284] TEX. HEALTH & SAFETY CODE ANN. 181.001(b)(2) (West 2012).

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

The proposed change 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.[285] Those releasing identifiable information, such as PatientsLikeMe or the Personal Genome Project would need to obtain meaningful patient consent,[286] as discussed in greater detail below.[287] 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.[288]

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.[289] In limited data

---

[285] 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 healthcare provider, health plan, public health authority, employer, life insurer, school or university, or healthcare clearinghouse." 45 C.F.R. § 160.103 (2013). It thus fails to include data handled by website operators and others.

[286] 45 C.F.R. §164.508 (2013).

[287] *See infra* notes 311-314 and accompanying text.

[288] 45 C.F.R. §164.514(b)(2013); *See supra* Part III.A.2 for detailed discussion of de-identification.

[289] 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.[290] 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.[291] These agreements are required because the added identifiers, while valuable to analysts, make re-identification considerably easier for skilled attackers.[292]

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.[293] Thus, the proposal would not create hurdles for health care treatment, payment, administration, or the activities of law enforcement and public health officials.[294]

### 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.[295] This restriction is already an element of data use

---

[290] 45 C.F.R. §164.514(e)(2) (2013).

[291] 45 C.F.R. §164.514(e)(4) (2013).

[292] Kathleen Benitez & Bradley Malin, *Evaluating Re-identification Risks with Respect to the HIPAA Privacy Rule*, 17 J. AM MED. INFORM. ASSOC. 169, 169 (2010) (estimating that the risk of re-identification is between 10% and 60%, depending on the state).

[293] 45 C.F.R. §§ 164.502, .506 & .512 (2013).

[294] *Id.*

[295] 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.

agreements, which require the recipients of limited data sets to promise that they will not "identify the information or contact the individuals."[296] 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 that is necessary in order to respond to medical or public health emergencies. Violators should be subject to HIPAA's enforcement provisions, which incorporate civil and criminal penalties.[297]

### *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 Centers for Disease Control and Prevention's "Policy on Releasing and Sharing Data" recommends the establishment of data-release review boards, and data custodians would be wise to adopt this suggestion.[298] 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

---

[296] 45 C.F.R. §164.514(e)(4)(ii)(C)(5) (2013).
[297] 45 C.F.R. §§ 160.300-.552 (2013).
[298] *CDC/ATSDR Policy on Releasing and Sharing Data*, *supra* note 122 at 9.

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.[299] It would also determine what statistical methods should be used to achieve de-identification of various data sets, including suppression, perturbation, and generalization.[300] 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.[301] 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 to the fact that the information is sensitive and raises privacy concerns. For example, the CDC Wonder website asks viewers who are seeking mortality information from its database to agree to a

---

[299] *See supra* note 155 and accompanying text.

[300] *CDC/ATSDR Policy on Releasing and Sharing Data, supra* note 122, at 9; *supra* note 152 and accompanying text (discussing the various statistical methods).

[301] *CDC/ATSDR Policy on Releasing and Sharing Data, supra* note 122, at 9; *see supra* Part III.C.1 (discussing data quality shortcomings); Hoffman & Podgurski, *supra* note 230, at 530-32 (discussing data quality assessment).

short list of data use restrictions by clicking an "I agree" icon.[302]

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.[303] It provides detailed information about patients and hospitals but is careful to remove identifiers and most likely meets the HIPAA safe harbor standard.[304] Nevertheless, NIS requires purchasers of the data to take a 15-minute training course that addresses

---

[302] CDC Wonder, *About Underlying Cause of Death, 1999-2010*, at http://wonder.cdc.gov/ucd-icd10.html. Users agree to:

- 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.

[303] Healthcare Cost and Utilization Project, *Overview of the National (Nationwide) Inpatient Sample (NIS)*, at http://www.hcup-us.ahrq.gov/nisoverview.jsp#purchase.

[304] *Id.* The data elements that are provided 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

privacy concerns.[305] 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 gain commercial or competitive advantage through analysis of released NIS data.[306]

If data users violate the agreement, the NIS would presumably challenge them in court. 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 contents of the 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.[307] However, these measures will alert the public to the importance of privacy and responsible data handling and may well avert innocent breaches by citizen scientists who wish to do no harm.

Equally important is the fact that the data use agreement requirement will create a record of those accessing data, and

---

[305] Healthcare Cost and Utilization Project, *Welcome to the HCUP Data Use Agreement (DUA) Training!* at http://www.hcup-us.ahrq.gov/tech_assist/dua.jsp.

[306] Healthcare Cost and Utilization Project, *Data Use Agreement for the Nationwide Databases from the Healthcare Cost and Utilization Project Agency for Healthcare Research and Quality*, at http://www.hcup-us.ahrq.gov/team/HCUP_Nationwide_DUA_051614.pdf.

[307] *See id.* (explaining that violation of the data use agreement can lead to fines or imprisonment under federal and state law); CDC Wonder, *supra* note 302 (describing sanctions for violations).

data custodians should maintain functional registries of users. Signatories can be required to provide their name, affiliation, and contact information.[308] 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.[309] Those who are found to violate data use agreements by re-identifying data or engaging in other misconduct, could be precluded from downloading information in the future and be subjected to other penalties.[310]

In some cases, privacy requirements should apply not only to data users, but also to data subjects. Specifically, individuals choosing to allow public access to identifiable or easily identifiable data, such as datasets that include birth date, sex, and zip code,[311] should undergo a comprehensive informed consent process.[312] Such data subjects should

---

[308] Healthcare Cost and Utilization Project, *supra* note 306.

[309] Julie Sartain, *Can Your IP Address Give Away Your Identity to Hackers, Stalkers and Cybercrooks?*, NETWORKWORLD, Jul. 16, 2013, at 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.

[310] CDC Wonder, *supra* note 302 (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.").

[311] *See supra* note 158 and accompanying text (explaining that it is relatively easy to re-identify such data).

[312] 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.

understand that their personal health information will be viewable not only by researchers who have good intentions but also by employers, marketers, financial institutions, and others who may not have their best interest in mind.[313] To this end, the Harvard Personal Genome Project requires participants to read and sign a lengthy consent document. They also must pass an examination and thereby demonstrate their understanding of the material contained in the consent form.[314] Testing data subjects' comprehension of the privacy risks they are accepting 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[315] they may provide new opportunities for employers and others to discriminate in

---

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 consent 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).

[313] *See supra* Part III.B (discussing discrimination concerns).

[314] Personal Genome Project: Harvard Medical School, *Participation Documents*, at http://www.personalgenomes.org/harvard/sign-up#documents.

[315] *See supra* notes 121-122 and accompanying text.

non-transparent ways.[316] Based on data about various health risks, entities might discriminate against discrete population sub-groups such as African-American women older than fifty.[317] These multi-factor discrimination cases are much more difficult to detect and prosecute than cases involving traditional, broad protected classes.[318] In addition, entities may retain experts to mine data and develop new applicant screening tools that focus on proxies for disability or predictors of bad health that can be considered without violating any explicit legal prohibition.[319] As open data and data mining proliferates, 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.[320] 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.

---

[316] *See supra* Part III.B.

[317] *See supra* notes 196-197 and accompanying text.

[318] *See supra* notes 200-201 and accompanying text.

[319] *See supra* notes 204-211 and accompanying text.

[320] *See supra* note 201 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).

In multi-factor cases, employment discrimination plaintiffs who believe that one of the attributes that was improperly considered was their age may face particular hurdles because of the Supreme Courts' decision in *Gross v. FBL Financial Services, Inc.* This decision barred mixed-motive claims and required "but for" proof of age discrimination.[321] However, in *Gross* the employer allegedly considered a mixture of proper (performance-related) and improper (the plaintiff's age of 54) factors rather than a combination of prohibited categories (e.g. age, race, sex).[322] The Supreme Court should revisit the question of whether plaintiffs may sue employers for discriminating based on age and one or more other protected classifications in a future case and clarify its intentions with respect to such claims. In the alternative, Congress could amend the Age Discrimination in Employment Act to add a provision that explicitly permits multi-factor claims.[323]

### *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 specific perceived or anticipated health conditions will also be difficult to detect. Consequently, anti-discrimination laws should include a requirement that businesses disclose

---

[321] Day, *supra* note 197, at 466-67; Gross v. Fin. Servs., Inc., 557 U.S. 167, 177-78 (2009).

[322] 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. *Id.* at 169.

[323] *See* Day, *supra* note 197, at 466-67 (proposing legislative action to approve age-plus-sex claims).

their data mining practices to workers, consumers, and other parties that are affected by them.

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."[324] 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."[325]

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 Americans with Disabilities Act's (ADA) medical inquiry and exam provision[326] 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

[324] Citron & Pasquale, *supra* note 180, at 21.
[325] Crawford & Schultz, *supra* note 180, at 125.
[326] 42 U.S.C. §12112(d) (2010).

Opportunity Commission (EEOC),[327] 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.[328] Disclosure statements could be included on employment or loan application forms so long as they were in sufficiently large and readable print or on separate sheets given to applicants.

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 will claim discrimination. Employment discrimination claimants, however, must exhaust their administrative remedies prior to filing suit.[329] 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.[330] The vast majority of claimants, whose cases the

---

[327] The Equal Employment Opportunity Commission is the federal agency tasked with enforcing the federal anti-discrimination laws. *See* Equal Employment Opportunity Commission, *Overview*, at http://www.eeoc.gov/eeoc/.

[328] 42 U.S.C. § 12182 (2010).

[329] 42 U.S.C. §§ 2000e-5(e)-(f) & 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*, 2004 WL 180044 (E.D. Pa. Jan. 27, 2004).

[330] *See* U.S. Equal Employment Opportunity Commission, *EEOC Litigation Statistics, FY 1997 through FY 2013*, at http://www.eeoc.gov/eeoc/statistics/enforcement/litigation.cfm

government will not pursue, will need to find an attorney who is interested in investing the time and money in delving into the technicalities of data mining activities, which may be no easy task.[331] Furthermore, plaintiffs would have legitimate claims only if they were persons with disabilities, as defined by the statute, and could prove that data mining practices actually led to harm in the form of exclusion from an employment opportunity for health-related reasons. 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[332] and prohibits employers, financial institutions, and others from discriminating against individuals based on a belief that they currently have physical or mental impairments, even if the belief is unfounded. 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

(indicating that in fiscal year 2013, the EEOC filed only 148 lawsuits nationwide).

[331] *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).

[332] *See* 42 U.S.C. § 12102 (2010).

mental impairment whether or not the impairment limits or is perceived to limit a major life activity."[333]

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.[334] Thus, for example, so long as employers do not consider genetic factors,[335] they can exclude such workers without being challenged.

If discrimination against high health-risk individuals is enabled by open data and becomes increasingly common, legislators would be wise to respond to it. An easy fix would be to add language to the "regarded as" provision 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

---

[333] 42 U.S.C. § 12102 (3) (A) (2010).

[334] *Id.*

[335] *See supra* note 211 and accompanying text (discussing the Genetic Information Nondiscrimination Act).

widespread dissemination of erroneous and harmful research conclusions.[336]

First, government agencies, academic institutions, and other research experts should develop educational resources and best practices guidelines to assist citizen scientists in conducting research.[337] These documents or videos could be posted on database websites, and users could be required or encouraged to review them, along with the privacy training materials, before signing data use agreements.[338] Data custodians could also test users on these materials in order to ensure that they have read and understood them prior to allowing them to sign the agreement.[339]

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 have no way to discern whether citizen scientists' findings are trustworthy.

One option is to follow the Wikipedia paradigm. Wikipedia allows any member of the public to post articles

---

[336] *See supra* Part III.C.

[337] *CDC/ATSDR Policy on Releasing and Sharing Data*, *supra* note 122, at 7 (urging CDC staff to develop "[i]nstructions for non-CDC users on the appropriate use of the data"); John P. Holdren, *OSTP Memo to Departments and Agencies: Increasing Access to the Results of Federally Funded Scientific Research*, Feb. 22, 2013, at 6, *available at* http://www.spaceref.com/news/viewsr.html?pid=43429 (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").

[338] *See supra* Part IV.A.3.

[339] *See supra* note 314 and accompanying text.

and anyone to edit those entries but provides some degree of oversight and quality control.[340] 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.[341] In extreme cases, administrators, of whom there are over 1400, can temporarily or permanently bar authors from contributing to Wikipedia because of intentional and persistent misconduct.[342] In addition, Wikipedia has an extensive dispute resolution system for disagreements about the contents of Wikipedia pages.[343] Readers who find passages that are biased or erroneous are encouraged to improve it 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.[344]

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

---

[340] Wikipedia, *Wikipedia: Policies and Guidelines*, at http://en.wikipedia.org/wiki/Wikipedia:Policies_and_guidelines.

[341] Wikipedia, *Wikipedia: Editor Review*, at http://en.wikipedia.org/wiki/Wikipedia:Editor_review; Wikipedia, *Wikipedia: Administrators*, at http://en.wikipedia.org/wiki/Wikipedia:Administrators.

[342] *Id*. Wikipedia, *supra* note 340;

[343] Wikipedia, *Wikipedia: Dispute Resolution*, at http://en.wikipedia.org/wiki/Wikipedia:Dispute_resolution.

[344] *Id.*

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 they used so that their research can be replicated and verified.[345]

Opportunities for peer review of citizen science research outcomes would be of significant benefit. 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.[346] 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 them to address their concerns. Authors may publish revised manuscripts,[347] and articles that peer reviewers approve are indexed in external databases such as PubMed.[348]

---

[345] Ari B. Friedman, *Preparing for Responsible Sharing of Clinical Trial Data (letter to the editor)*, 370 N. ENGL. J. MED. 484, (2014).

[346] F1000Research, *F1000Research, an Open Science Journal*, at http://f1000research.com/about.

[347] *Id.*; F1000Research, *Why publish with F1000Research?* at http://f1000research.com/why-submit;

[348] F1000Research, *The First Open Science Journal for Life Scientists*, at http://f1000research.com/.

Peerage of Science offers a second non-traditional approach.[349] 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.[350] Authors can accept offers from participating journals or export reviews outside of Peerage of Science to journals of their choice.[351]

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, but also the reading public in distinguishing between valid research findings and those that have no reliable basis.[352]

### D. TORT CLAIM LITIGATION STRATEGIES

---

[349] Peerage of Science, *How It Works*, at http://www.peerageofscience.org/how-it-works/.

[350] Peerage of Science, *Process Flow*, at http://www.peerageofscience.org/how-it-works/process-flow/.

[351] *Id.*

[352] 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 97-101 and accompanying text. However, a filtering mechanism could at least screen out material that no educated reviewer would consider reliable.

Parties who are hurt by citizen scientists' wrongdoing will have a variety of avenues by which to seek redress. Plaintiffs may allege defamation, interference with economic advantage, public disclosure of private fact, and other claims.[353] Database operators who have required data recipients to sign data use agreements may also sue for breach of contract if recipients attempt to re-identify information, use data for commercial or competitive purposes, or violate other agreement provisions, and the breaches have damaged the database's reputation or economic interests.[354]

Of greater concern are instances in which parties may file suit against citizen scientists who act in good faith but publicize information that is 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.[355] 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 and 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.[356] If amateur researchers make valuable contributions to science but are routinely harassed through frivolous litigation,

---

[353] *See supra* Parts III.D.1 and III.D.2.

[354] *See supra* notes 302-306 and accompanying text.

[355] *See supra* Part III.D.3.

[356] *Id.*

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 Reputation."[357] 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.[358]

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

[357] Digital Media Law Project, *Practical Tips for Avoiding Liability Associated with Harms to Reputation*, at http://www.dmlp.org/legal-guide/practical-tips-avoiding-liability-associated-harms-reputation.
[358] *Id.*

experts' advice in order to minimize the likelihood of being sued and especially of facing liability.

## V. 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.[359] Many 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."[360]

Open medical data have the potential to yield numerous benefits, including scientific discoveries, cost savings, the development of patient support tools, healthcare quality improvement, greater government transparency, public education, and positive changes in healthcare policy.[361] At the same time, open data raise several complex legal and ethical concerns related to privacy, discrimination, erroneous research findings, and litigation.[362]

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

---

[359] *See* Exec. Order No. 13642, *supra* note 1.

[360] Institute of Medicine, *Activity: Strategies for Responsible Sharing of Clinical Trial Data*, at http://www.iom.edu/Activities/Research/SharingClinicalTrialData.aspx (describing the project and its timeline; *see supra* note 221 for interim report.

[361] *See supra* Part II.

[362] *See supra* Part III.

detailed regulatory framework for citizen science.[363] Until then, the government, industry, data custodians, and others, should implement the more modest interventions proposed in this Article in order 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."[364] 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.

[363] O'Connor, *supra* note 227 at 481.
[364] Exec. Order No. 13642, *supra* note 1.