Journal of Law and the Biosciences, 1–12
https://doi.org/10.1093/jlb/lsab028
Original Article

Protecting research data of publicly revealing participants

Kyle J. McKibbin[1,†], Bradley A. Malin[2,3,4,‡] and Ellen Wright Clayton[1,5,6,7,*,**]

[1]Vanderbilt University Law School, Nashville, TN
[2]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN
[3]Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN
[4]Department of Computer Science, Vanderbilt University, Nashville, TN
[5]Center for Biomedical Ethics and Society, Vanderbilt University Medical Center, Nashville, TN
[6]Department of Pediatrics, Vanderbilt University Medical Center, Nashville, TN
[7]Department of Health Policy, Vanderbilt University Medical Center, Nashville, TN
*Corresponding author. Ellen.clayton@vumc.org

ABSTRACT

Biomedical researchers collect large amounts of personal data about individuals, which are frequently shared with repositories and an array of users. Typically, research data holders implement measures to protect participants' identities and unique attributes from unauthorized disclosure. These measures, however, can be less effective if people disclose their participation in a research study, which they may do for many reasons. Even so, the people who provide these data for research often understandably expect that their privacy will be protected. We discuss the particular challenges posed by self-disclosure and identify various steps that researchers should take to protect data in these cases to protect both the individuals and the research enterprise.

KEYWORDS: re-identification, risk mitigation, self-disclosure, informed consent, privacy

† Kyle J. McKibbin graduated this year from Vanderbilt Law School, where he focused on legal implications of technology for privacy.

‡ Bradley Malin is one of the best-known experts in the world in data privacy, both as a technological innovator and a policy advisor. He has developed numerous methods to assess and mitigate privacy risks when sharing data.

** Ellen Wright Clayton is an internationally known expert on the ethical and legal issues posed by creating and using biorepositories. Malin and Clayton are co-PIs of the transdisciplinary Center for Excellence in ELSI Research, Genetic Privacy and Identity in Community Settings.

The decision to participate in research can be influenced by numerous factors, but ultimately is personal. While some individuals choose to participate without accolade or attribution, there are many reasons why people disclose that they participate in research. For instance, they may be proud of their contribution to the health of others. As one person publicly intimated:

> 'I signed up as [a research] participant because my family is seriously affected by heart disease. I'm hoping that my data will help someone find answers to why some people are at greater risk for specific health conditions and how to prevent them, because I want my children and grandchildren to not have to experience the same health issues that I have.'[1]

At times, research participants may want to inspire others to take part by setting an example. Thus, Walter M. Kimbrough, President of Dillard University, and C. Reynold Verret, President of Xavier University, wrote in an open letter about their participation in a vaccine trial:

> 'We appeal to the students, faculty, staff and alumni of Dillard, Xavier, and our sibling institutions to consider participating in this [COVID-19 vaccine] trial or others being conducted. The people and communities we serve look to us as an example.'[2]

Others, including study sponsors, sometimes encourage participants to tell their stories publicly. For instance, the Million Veteran Program of the U.S. Department of Veteran Affairs (VA) supports a webpage dedicated to testimonials, where named participants explain their rationale for participating.[3] As one participant stated:

> 'I thought this would be a good way to help the Veterans of tomorrow. Improving healthcare for Veterans requires assistance from many parties, including Veterans. Being part of MVP provides me an opportunity to enhance the care I receive, as well as my fellow warriors.'[4]

As the previous examples illustrate that the decision to self-disclose can also be beneficial to researchers and study sponsors, primarily as a recruitment tool. Investigators may need to recruit a broad and diverse group of participants to ensure the success of a project. Alternatively, they may be interested in recruiting individuals from underserved or underrepresented groups, where participant self-disclosure might

1 *Dorothy Farrar Edwards Leads USA in All of Us Research Program*, 20 Quarterly (University of Wisconsin Medical Alumni Association) 34 (2018).

2 William M. Kimbrough & C. Reynold Verret, *A message from the presidents of Dillard and Xavier* (2020), https://dillard.edu/communications/news/xavier-dillard-vaccine-trials.php.

3 Official Blog of the U.S. Department of Veteran Affairs, *VA's Million Veteran Program seeking female Veterans for genetic-based studies*(2020), https://www.blogs.va.gov/VAntage/75603/va-million-veterans-program-seeking-female-veterans-genetic-based-studies/; see also *All of Us* Participant Partners, https://allofus.nih.gov/about/who-we-are/all-us-participant-partners.

4 https://www.blogs.va.gov/VAntage/75603/va-million-veterans-program-seeking-female-veterans-genetic-based-studies/. VA Million Veterans Program, *Testimonials* (2020), https://www.mvp.va.gov/webapp/mvp-web-participant/#/public/testimonials.

help generate trust. While self-disclosure is not new,[5] modern technology has made it dramatically easier for people to reveal information about themselves on a wide, and essentially permanent, scale.

Yet, people who disclose the fact that they are taking part in research may not necessarily intend to reveal all the data about themselves held by the project. Even the person who publicly discloses taking part in a study of a rare disease, thereby revealing some aspect of her health, may not be willing to have her entire research record accessed by third parties, which re-identification of the research record would permit. Indeed, most people who participate in research typically expect that some, if not all, data about them obtained over the course of a study will be protected.[6] Researchers typically promise to provide protection to the extent possible as part of the consent process.[7] Yet, the degree to which people who publicly disclose their participation in a research project can or should be able to agree to the possible revelation of other research data about them as well the reciprocal issue of the extent to which their expectations of privacy can be met are open questions. The answers to these questions, which we discuss below, require a more in depth understanding of the ecology of data, the likelihood of privacy intrusions (e.g., re-identification of seemingly anonymous records), as well as the current state of law and regulation in the USA.

I. THE DATA ENVIRONMENT—DEFINING THE RISK

I.A. Publicly accessible data

People are social beings. They often like to tell their stories, to connect with others, and to exchange knowledge and support. Many post information about themselves,[8] their opinions,[9] their friends and family,[10] and even their research participation on social media.[11] Even the wary tend to discount future risks of harms from unwanted use of in favor of the immediate benefits they receive from information sharing.[12] Additionally, people are not the only sources of accessible information about themselves as large amounts of data on topics ranging from internet searches, voting records, purchasing

5 Natalya N. Bazarova & Yoon Hyung Choi, *Self-Disclosure in Social Media: Extending the Functional Approach to Disclosure Motivations and Characteristics on Social Network Sites*, 64 J. Commun. 635 (2014); Mina Tsay-Vogel et al., *Social Media Cultivating Perceptions of Privacy: A 5-Year Analysis of Privacy Attitudes and Self-Disclosure Behaviors among Facebook Users*, 20 New Media Soc. 141 (2018).

6 Ellen W Clayton et al., *A systematic literature review of individuals' perspectives on privacy and genetic information in the United States*, 13 PLoS One e0204417 (2018); Deborah Goodman et al., *De-identified genomic data sharing: the research participant perspective*, 8 J. Community Genet. 173 (2017).

7 Regulation for the Protection of Human Research Participants, General Requirements for Informed Consent, 45 CFR § 46.116 (2020).

8 Zhijun Yin et al., *A scalable framework to detect personal health mentions on Twitter*, 17 J. Med. Internet Res. e138 (2015).

9 Nikos Tsirakis et al., *Large scale opinion mining for social, news and blog data*, 127 J. Syst. Softw. 237 (2017).

10 Zhijun Yin et al., *#PrayForDad: learning the semantics behind why social media users disclose health information*, 2016 Proc. Int. AAAI Conf. Weblogs Soc. Media 456 (2016).

11 Yongtai Liu et al., *Biomedical research cohort membership disclosure on social media*, 2019 AMIA Annu. Symp. Proc. 607 (2019).

12 Alessandro Acquisti, Laura Brandimarte & George Loewenstein, *Privacy and human behavior in the age of information,* 347 Science 509 (2015). https://www.ftc.gov/system/files/documents/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014/140527databrokerreport.pdf

habits, and home address, which are increasingly collated by data aggregators and made available to varying degrees.[13]

I.B. Research data

Biomedical researchers collect growing amounts of information about individuals, potentially including genetic test results, clinical information, and an array of socio-cultural and behavioral variables, some of which are derived from public records. At times, data about the people to whom information pertains (whom we call *data sources*) are collected in the research setting with their knowledge and some consent, which increasingly includes permission for broad data sharing required by federal funders such as the National Institutes of Health and National Science Foundation.[14] Once gathered, these data are held, frequently with identifiers placed in escrow or removed entirely, by various entities, (whom we call *data holders*) ranging from universities and healthcare organizations, to free-standing collections, such as the Coriell Institute for Medical Research,[15] and private companies, such as 23andMe[16] or Human Longevity Inc.[17] These data collections are used by a variety of people (*data users* or *recipients*), including academic personnel, investigators affiliated with for-profit companies, and citizen scientists. Different data holders and users/recipients are subject to their own distinctive ethical norms and legal requirements. In addition, data do not flow in one direction from participants to researchers as investigators increasingly return individual research results.[18] Thus, research data, which can contain a very diverse collection of information unparalleled in other environments, do not stay in one place, but rather may flow quite broadly (see Figure 1). And importantly, data used in research do exist not in isolation but rather in the sea of publicly available data.[19]

This complex ecology increases risks for both research participants and those responsible for protecting their data because pinpointing an individual in a research dataset can provide ready access to all the other information about that person in the collection. Even when some, if not all, of the data is available at other data sites, finding a person's record obviates the need to search elsewhere. And even if directly identifying information has been removed from the data (e.g., personal names or Social Security numbers), the existence of identified data in the public domain opens up the possibility of re-identifying the person to whom it pertains through residual unique combinations

13 Federal Trade Commission, Data Brokers: A Call for Transparency and Accountability 88 (2014); Matthew Crain, *The limits of transparency: Data brokers and commodification*, 20 New Media Soc. (2018).

14 National Institutes of Health, *NIH Data Sharing Policy* (2020), https://grants.nih.gov/grants/policy/data_sharing/; National Science Foundation, *Dissemination and Sharing of Research Results* (2020), https://nsf.gov/bfa/dias/policy/dmp.jsp.

15 *Coriell Institute for Medical Research*, https://coriell.org.

16 *23andMe*, www.23andme.com.

17 *Human Longevity, Inc.*, https://humanlongevity.com/.

18 National Academies of Sciences, Engineering, and Medicine; Health and Medicine Division; Board on Health Sciences Policy; Committee on the Return of Individual-Specific Research Results Generated in Research Laboratories; Jeffrey R. Botkin, Michelle Mancher, Emily R. Busta & Autumn S. Downey, Editors, Returning Individual Research Results to Participants: Guidance for a New Research Paradigm (National Academies Press, 2018).

19 Latanya Sweeney, *The Data Map*, https://thedatamap.org/.

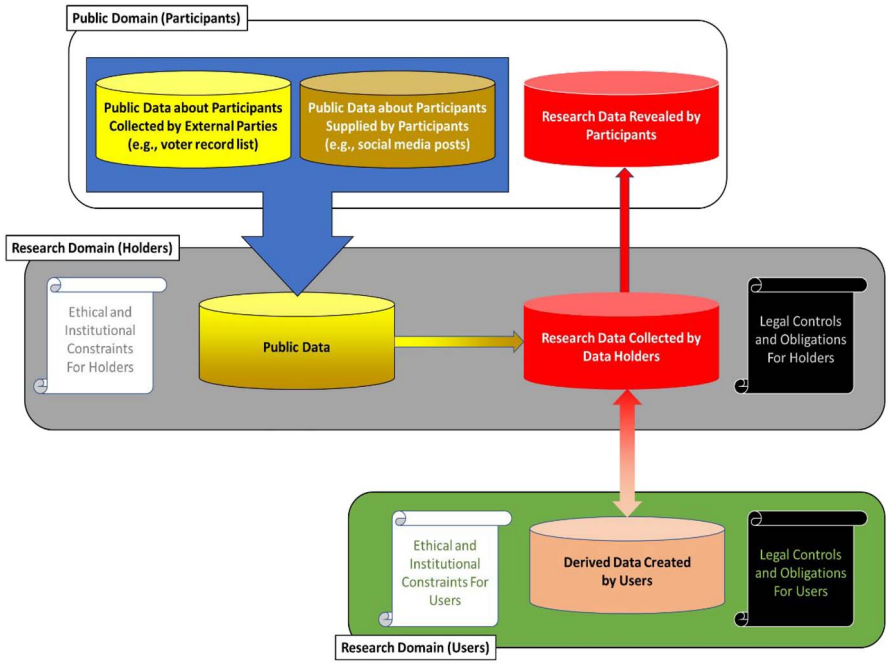

Figure 1. This diagram represents potential data flow among data sources, holders, and users/recipients as well as ethical and legal controls on these actors.

of demographics (e.g., date of birth, gender, and 5-digit ZIP code of residence).[20] Such revelations potentially pose risk to the identified individual, particularly if the data are stigmatizing. Identifying specific research participants is also likely to harm the researchers and institutions involved, raising questions about their ability to protect participants, thereby undermining trust in investigators, institutions, and the research enterprise generally.

Most of the literature to date has focused on the case in which the data recipient seeks to identify someone who has not disclosed participation.[21] Several investigations have demonstrated that re-identification of various types of data (e.g., demographics, clinical phenomena, and genomic records) is possible using publicly accessible records.[22] These examinations typically assume that the data recipient would use data from public records (e.g., voter registration databases),[23] or third party information aggregators

20 Kathleen Benitez & Bradley Malin, *Evaluating re-identification risks with respect to the HIPAA privacy rule*, 17 J. Am. Med. Inform. Assn 169 (2010).

21 Luc Rocher et al., *Estimating the success of re-identifications in incomplete datasets using generative models*, 10 Nat. Commun. 3069 (2019).

22 Khaled El Emam et al., *A systematic review of re-identification attacks on health data*, 6 PLoS One e28071 (2011); Khaled El Emam et al., *Correction: a systematic review of re-identification attacks on health data*, 10 PLoS One e0126772 (2015); Yaniv Erlich & Arvind Narayanan, *Routes for breaching and protecting genetic privacy*, 15 Nat. Rev. Genet. 409 (2014).

23 Benitez & Malin, *supra* note 20.

(e.g., Intellius.com), or social media (e.g., Twitter).[24] The likelihood, however, that the recipient would succeed in re-identification depends on numerous factors, including whether the targeted person is actually in any of the collected resources and, if so, whether the person is readily distinguishable from other people in the set.[25] A number of studies suggest, further, that even while possible, re-identification is unlikely to occur in most settings, due in part to the costs of the effort and the limited gain to the person who is seeking to make the identification.[26]

If participants publicly disclose their role, however, the possibility of their being identified in a research data set can increase significantly because the recipient knows that they are present therein, making other possible matches less likely. Put another way, imagine that there is one record in a research database with a specific combination of demographics. Further imagine that there are 100 identified people in the general population who have the demographics in question. Without additional knowledge, this implies that there is a 1/100 (or 1%) chance that the data recipient could guess the identity of the record in the research dataset. However, if the person whose record is in the research dataset publicly reveals that they are a member of the research study, then it is evident that the other 99 people could not be in the study. As a result, the risk of re-identification shifts from 1 to 100 per cent. Since the residual risk is higher if individuals reveal their role, the efficacy of ethical norms and legal rules in deterring efforts to identify the record about that person become more important.

II. PROMOTION OF TRUST IN RESEARCH REQUIRES PROTECTING PARTICIPANTS WHO SELF-DISCLOSE

While some people reveal their research participation on their own, others may be encouraged to do so by the sponsors or managers of the research project,[27] often to inspire others to enroll.[28] Certainly, all participants should be informed that publicly disclosing their role could expose them to a greater risk that research data about them could be identified and revealed. This possibility can be even greater if the person already is widely known or has a large media presence. Informing participants about this possibility gives them the opportunity to make decisions that conform more fully to their values and expectations.

Some might contend that those who self-disclose after being informed have assumed the risk of having the data in the research dataset about them revealed, arguing that researchers can never fully eliminate privacy risks and that as long as a self-disclosing participant is fully informed, the increased risk is acceptable to the individual. This argument, however, fails for at least three reasons. First, at a minimum, the research enterprise has a special responsibility to protect those whom it encourages to go public.

---

24 Liu et al., *supra* note 11.

25 Weiyi Xia et al., *Process-Driven Data Privacy*, Proceedings of the 24th ACM International on Conference on Information and Knowledge Management 1021 (2015).

26 Zhiyu Wan et al., *Expanding Access to Large-Scale Genomic Data While Promoting Privacy: A Game Theoretic Approach*, 100 Am. J. Hum. Genet. 316 (2017); Janice Branson et al., *Evaluating the re-identification risk of a clinical study report anonymized under EMA Policy 0070 and Health Canada Regulations*, 21 Trials 200 (2020).

27 *Supra* note 4

28 *Supra* note 11.

A second and broader argument is that data apply to people, and those who use these data are ethically required to consider the impact of their actions on the individuals to whom the data pertain. Third, failing to protect participants from harm resulting from their involvement also undermines trust in research. Thus, it is crucial to examine what data holders and users do with data and their ethical and legal obligations to protect participants.

III. WHAT SHOULD DATA HOLDERS DO TO PROTECT PARTICIPANTS?

Researchers and staff who collect data contribute them to specific repositories, which are responsible for their orderly and secure maintenance and protection. While simply storing data would incur little risk (except in the event of a security breach), a primary function of data holders is to distribute data to downstream users, for without use, data have no societal value. As steward, data holders may assess the merit of a proposed use and how much data to release and in what form, taking into account the location and capacity of the proposed user as well as the risk posed by external data that could be used for re-identification of participants. They may also consider whether the project is designed and sufficiently powered to be successful, as well as whether it may cause harm to others, for example, by stigmatization. This weighing of risks and benefits is required for projects subject to the Common Rule for the Protection Human Research Participants (Common Rule)[29] and the Food and Drug Administration (FDA) Regulations[30] as well as the Belmont Report on which these regulations rely.[31] Whether the use is consistent with the data source's consent is rarely at issue since participants frequently provide broad consent for use and data sharing, an option now permitted under the Common Rule and increasingly required as a condition of receiving funding from federal agencies.

The heightened risks experienced by self-disclosing participants should play a prominent role in a data holder's decisions regarding the collection and handling of data, particularly when participants are encouraged to make their role public. Numerous entities take part in overseeing the actions of the data holder, including Institutional Review Boards, privacy boards, data access committees, and more recently, community advisors. These entities, working with data holders, can and should carefully weigh proposals to seek self-disclosure for reasons such as encouraging wider research participation, particularly from individuals who may have particular privacy concerns. And while full disclosure of the particular risks faced by these participants is essential, their consent cannot bear all the weight, as noted above.

The data holder may also adopt certain technical approaches to defend data about participants. Removing identifiers from data is a primary strategy to balance the competing goals of promoting research use and minimizing privacy risks to individuals. Indeed, both HIPAA and the Common Rule, which apply to some but not all data holders, provide strategies, albeit through different mechanisms, which allow investigators to forgo obtaining consent and review. HIPAA, for example, defines the conditions

29 Common Rule for the Protection of Human Research Participants, 45 CFR part 46 et seq. (2021).

30 Food and Drug Administration, Regulations for the Protection of Human Subjects and Institution Review Boards, 21 CFR Parts 50 and 56 (2020).

31 Belmont Report, Ethical Principles and Guidelines for the Protection of Human Subjects of Research, Federal Register 1979 Apr. 18; 44(76) Part IV: 23192-23197, https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/read-the-belmont-report/index.html.

under which data can be deemed de-identified and hence be available for use without authorization.[32] In contrast, the Common Rule achieves a similar end by stating that it does not apply to data from which the 'identity of the subject is [not] or may [not] readily be ascertained by the investigator'.[33] Protection strategies based on removing identifiers, however, may be less effective when participants publicly disclose their role in the project. Safeguarding these individuals may require additional steps. One possible strategy—reducing the fidelity in data—can make re-identification more difficult but also has the unwanted byproduct of reducing the data's utility for research.[34]

Having received broad consent for data sharing and, in some cases, having encouraged people to 'go public', data holders and oversight bodies should 'step into the shoes' of data sources when considering protection strategies, particularly considering the additional risks potentially faced by self-disclosing participants. This is particularly important for protocols addressing stigmatizing conditions or involving vulnerable or distinctive groups, who may be more readily identifiable because of unique characteristics or who may face greater privacy harms than the average research participant. Addressing these concerns may be challenging since data users can vary in their motives to re-identify participants in research datasets. Some assert that they are simply demonstrating security gaps,[35] while others might wish to inflict specific harm on individuals—although instances of the latter are hard to come by in practice. The consequences for the person who pursues re-identification could vary as well. Investigators in academic institutions who use data inappropriately may damage not only their personal reputation and ability to obtain grant funding but also those of their institutional home, particularly if, as is usually the case, the latter is required to vouch for the researcher. These potential consequences lead many universities and medical centers to develop robust internal oversight mechanisms, enforced by non-trivial penalties for breach. Such institutional controls, however, may be less available for independent investigators.

Data holders can also require investigators to execute data use agreements (DUAs). These typically require that the user promise to undertake only the analyses proposed, not to re-identify the individuals to whom the data apply, and not to share the data with anyone else. However, careful drafting of DUAs is essential. For example, if there is a desire to maximize potential claims by participants, an agreement should explicitly state an intent to benefit them.[36] Even so, the damages from a violation of DUAs are likely to be difficult to prove for research participants seeking to recover or for data holders seeking to deter unwanted behavior. One solution may be to include liquidated

32 45 CFR § 164.514.

33 45 CFR §§ 46.102(e)(ii)&(5-6), 46.104(d)(4).

34 Weiyi Xia et al., *R-U policy frontiers for health data de-identification*, 22 J. Am. Med. Inform. Assn 1029 (2015) https://www.forbes.com/sites/adamtanner/2013/04/25/harvard-professor-re-identifies-anonymous-volunteers-in-dna-study/?sh=9ca931b92c9b; Fabian Prasser et al., *The Importance of Context: Risk-based De-identification of Biomedical Data*, 55 Methods Inf. Med. 347 (2016).

35 Adam Tanner, *Harvard Professor Re-Identifies Anonymous Volunteers in DNA Study*, Forbes (2013) https://www.forbes.com/sites/adamtanner/2013/04/25/harvard-professor-re-identifies-anonymous-volunteers-in-dna-study/?sh=9ca931b92c9b; Heather Murphy, *Most White Americans' DNA Can Be Identified Through Genealogy Databases*, New York Times, (2018) https://www.nytimes.com/2018/10/11/science/science-genetic-genealogy-study.html.

36 *Intent of Contracting Parties to Benefit Third Person*, 16 Am. Jur. Proof of Facts 2d 55 (2007).

damages in the contract, but if set too high these damages can run the risk both of being characterized by courts as an unenforceable penalty[37] as well as deterring users with legitimate goals. Setting damages based on an independent assessment of costs of breach response or reputational damage is possible, but even this may fail to deter some users if the perceived benefits they achieve through re-identification outweigh such costs.[38]

Careful oversight is necessary to protect the interests not only of the individuals to whom the data pertain but also of the data holders themselves, as inappropriate release or use can expose the holder to reputational damage as well as legal liability. Data holders are subject to numerous laws governing data protection, whose applicability varies depending on their physical and institutional location. Federally funded repositories containing patient information located in medical centers and their business associates are typically subject to some combination of the Common Rule, FDA regulations, and the Health Insurance Portability and Accountability Act of 1996 (HIPAA). Although these laws have significant gaps in their coverage,[39] failure to comply with their requirements can lead to a loss of federal funding (including sponsored research) or large civil and criminal penalties. However, fewer restrictions apply to data holders who are not subject to these laws.

Many states have enacted laws that offer a more comprehensive web of protection for participants.[40] California and Maryland, for example, provide criminal and civil penalties for failure to comply with Common Rule. Likewise, at least 16 states provide a private cause of action for a breach of state medical privacy laws.[41] And while HIPAA applies only to covered entities and their business associates, a handful of states impose confidentiality requirements on any person who receives protected health information, a few containing provisions directed specifically at researchers. Other laws dealing with research, however, are inconsistent and only apply in specific circumstances. Genetic information often receives special attention at the state level, but while these laws are relevant to many biomedical research projects, they typically provide little additional coverage.[42] Thus, neither federal nor state laws offer comprehensive coverage for all types of personal data collected, and few statutory or regulatory restrictions exist regarding use by third parties once consent is obtained.[43]

37 *CVS Pharmacy, Inc. v Press America, Inc.* 377 F.Supp.3d 359 (S.D. N.Y. 2019).

38 Yongtai Wan, *supra* note 24; Weiyi Xia et al., *It's all in the timing: calibrating temporal penalties for biomedical data sharing*, 25 J. Am. Med. Inform. Assoc. 25 (2018).

39 Ellen Wright Clayton et al., *The law of genetic privacy: applications, implications, and limitations*, 6 J. Law Biosci. 1 (2019).

40 Leslie E. Wolf et al., *Protecting Participants in Genomic Research: Understanding the 'Web of Protections' Afforded by Federal and State Law*, 48 J. Law Med. Ethics 126 (2020).

41 Leslie E. Wolf et al., *The web of legal protections for participants in genomic research*, 29 Health matrix (Cleveland, Ohio: 1991) 3 (2019).

42 *Supra* note 39; Heather L. Harrell & Mark A. Rothstein, *Biobanking research and privacy laws in the United States*, 44 J. Law Med. Ethics 106 (2016).

43 Congressional Research Service, *Data Protection Law: An Overview* (2019), https://fas.org/sgp/crs/misc/R45631.pdf.

IV. WHAT SHOULD DATA USERS DO TO PROTECT PARTICIPANTS?

Data users should also consider the interests of the data sources as well as the need to support the integrity of the research enterprise. Although data users are not directly involved in research participants' decision to self-disclose, they must still grapple with the accompanying risks if they or a third party seek to re-identify individuals in the data they received. Data holders may impose certain requirements on data users through a DUA, but data users should apply best practices in information security as well. For instance, researchers affiliated with federally funded academic research centers, which likely have well-developed compliance procedures in place, could face reputational damage to their home institutions and professional censure in the event of a privacy violation. Data users from non-traditional research backgrounds, however, such as citizen scientists[44] or private sector institutions that do not receive federal funding, may not have internal privacy review processes in place but still should be mindful of these concerns. If a privacy violation occurs, these users may risk losing access to research data and affect broader public support for their own research endeavors as well as those of other private researchers.

Beyond these ethical concerns, breaching the scope of a DUA, in particular, could subject data users to penalties under certain federal laws, such as the federal Computer Fraud and Abuse Act (CFAA),[45] which provides that a person who accesses a website or restricted database without authorization can face criminal and civil penalties,[46] provisions intended to prevent hacking by outsiders. The Supreme Court, however, recently made clear in *Van Buren v U.S.*,[47] a case in which a police officer used police files for personal gain, that the law does not apply to misuse of data to which the user had otherwise been granted access.

State law also provides penalties to users who re-identify or disclose research data. Texas, for example, prohibits any individual from attempting to re-identify health data.[48] Several states have significantly expanded privacy protections in recent years, and many more have pending legislation.[49] Three notable examples are the California Consumer Privacy Act (CCPA), originally enacted in 2018, the Virginia Consumer Data Protection Act (VCDPA) [50] and the Colorado Privacy Act (CPA) [51], which were passed earlier this year. The CCPA, for example, provides a private cause of action for consumers whose 'nonencrypted and nonredacted personal information' is subject to an unauthorized disclosure as a result of a business's failure to 'implement reasonable security procedures and practices'. Under the broad terms of the Act, a business could include an individual accessing personal information on behalf of a for-profit California

44 Andrea Wiggins & John Wilbanks, *The rise of citizen science in health and biomedical research*, 19 Am. J. Bioethics 3 (2019).

45 Computer Fraud and Abuse Act, 18 U.S.C. §§1030(a)-(c).

46 Congressional Research Service, *Cybercrime: An Overview of the Federal Computer Fraud and Abuse Statute and Related Federal Criminal Laws* (2014), https://fas.org/sgp/crs/misc/97-1025.pdf.

47 *Van Buren v U.S.*, No. 19-783 (U.S., June 3, 2021).

48 Medical Records Privacy, Tex. Health & Safety Code Ann. § 181.151 (2020).

49 National Conference of State Legislatures, *2020 Consumer Data Privacy Legislation* (2020), https://www.ncsl.org/research/telecommunications-and-information-technology/2020-consumer-data-privacy-legislation637290470.aspx.

50 Virginia Act 35, *to be codified at* Va. Code Ann. ch. 52 §§59.1-571-59.1-581 (effective Jan. 1, 2023).

51 Colorado Senate Bill 21-190, to be codified at Colo. Rev. Stat. Tit. 6, Art. 1, part 13.

entity and performing even manual operations on personal data.[52] While the CCPA was recently amended to exclude coverage of data de-identified in accordance with federal policy,[53] it provides that any attempt to re-identify the data would make users once again subject to the Act.[54] The private cause of action under the CCPA, however, is limited to personal information as defined under the California Customer Records Act, which includes individually identifiable medical information in combination with an individual's encrypted or redacted name.[55] Thus, the court's holding in *In re Yahoo! Inc. Customer Data Security Breach Litigation*[56] suggests that a data holder could release demographic data that could increase the likelihood of attribute disclosure without incurring liability so long as identifiable medical information was not primarily disclosed. Ultimately, it is difficult to predict how courts will interpret the broad and somewhat vague terms of the CCPA. Moreover, CCPA applies only to companies that: (i) have gross revenue of at least $25,000,000, (ii) collect personal data from more than 50,000 California residents, households, or device each year, or (iii) earn at least 50% of their income from selling personal information about California residents.[57] Regardless, self-disclosing participants are at increased risk of being re-identified, and users should keep this in mind when designing experiments if they wish to avoid personal liability or liability for their home institution as California and other states take a more active role in protecting personal data.

At present, however, research participants are likely to have few legal remedies themselves if they are harmed by data users. Remedies under the tort law of privacy are notoriously difficult to obtain, and tort law has generally failed to keep the pace with changing data practices.[58] Contract law can play a role in shaping behavior, but contract damages for breach of a DUA, no matter how well constructed, would likely be minimal at best. Neither HIPAA nor the Common Rule provides a private right of action, and potential claims under the CFAA have been severely cut back against research data users. Thus, the injured party would at most be able to pursue claims under a few state statutes. These minimal protections heighten the need to put in place procedures to protect self-disclosing participants from additional risks.

V. CONCLUSION

People who talk publicly about their participation in research make it easier for others to try to find their research records, enabled by the large amount of information available outside the research dataset. While significant harm to an individual resulting from re-identification appears to be rare—even for those who disclose their participation—this problem may get worse as the depth and breadth of data grow. Moreover, the research enterprise is ethically required and has strong incentives to protect the data of participants, including those who self-disclose, whether they do so on their own or at the invitation of researchers. Our analysis shows that the law's defenses are incomplete.

52 California Consumer Privacy Act, Cal. Civil Code § 1798.140.
53 California Consumer Privacy Act of 2018, Assembly Bill 713, Cal. Civil Code §1798.146 (2020).
54 California Consumer Privacy Act of 2018, Assembly Bill 713, Cal. Civil Code § 1798.148 (2020).
55 Cal. Civil Code §§ 1798.150, 1798.81.5(d)(1)(2020).
56 In re Yahoo! Inc. Customer Data Security Breach Litigation 313 F.Supp.3d 1113, 1444-45 (N.D. Cal. 2018).
57 *Supra* note 39.
58 Daniel J. Solove & Neil M. Richards, *Prosser's Privacy Law: A Mixed Legacy*, 98 Cal. L. Rev. 1887 (2010).

Yet, the need to address the interests of participants, as well as the need to protect the integrity of the research enterprise, suggests some steps that should be taken. First, potential participants should be informed of the possibility of revelation so they can make informed choices about taking part and talking about their involvement in research. Second, holders must put systems of oversight and public accountability in place and devise mechanisms to evaluate users and inform them of their obligations to respect the interests of participants. Third, users should adhere to these requirements at minimum to ensure continued access to research data, but with the understanding that, although uncertain now, they may well face liability as jurisdictions increase efforts to regulate data use and protect privacy.

ACKNOWLEDGMENTS

This work was supported in part by the National Institutes of Health, National Human Genome Research Institute, 5RM1HG009034-04.

CONFLICT OF INTEREST

The authors have no conflicts of interest to disclose.