February, 2022

Not So Private

Stacey A. Tovino, *University of Oklahoma College of Law*

Duke Law Journal

VOLUME 71 FEBRUARY 2022 NUMBER 5

NOT SO PRIVATE

STACEY A. TOVINO†

ABSTRACT

Federal and state laws have long attempted to strike a balance between protecting patient privacy and health information confidentiality on the one hand and supporting important uses and disclosures of health information on the other. To this end, many health laws restrict the use and disclosure of identifiable health data but support the use and disclosure of de-identified data. The goal of health data de-identification is to prevent or minimize informational injuries to identifiable data subjects while allowing the production of aggregate statistics that can be used for biomedical and behavioral research, public health initiatives, informed health care decision making, and other important activities. Many federal and state laws assume that data are de-identified when direct and indirect demographic identifiers such as names, user names, email addresses, street addresses, and telephone numbers have been removed. An emerging reidentification literature shows, however, that purportedly de-identified data can—and increasingly will—be reidentified. This Article responds to this concern by presenting an original synthesis of illustrative federal and state

[†] Professor of Law and Faculty Lead, Healthcare Law Graduate Programs, University of Oklahoma College of Law, Norman, Oklahoma. The Author thanks Dean Katheleen Guzman for her financial support of this research project and Mr. Christopher Punto for his outstanding research assistance. The Author also thanks the organizers and participants of the following symposia and colloquia for comments on the ideas presented in this Article: The Privacy and Intellectual Property in the Age of Coronavirus Symposium, sponsored by the Institute for Privacy Protection and the Gibbons Institute of Law, Science & Technology at Seton Hall University School of Law (Sept. 17, 2020); The Childress Lecture and Symposium, sponsored by Saint Louis University School of Law (Oct. 2, 2020); The Center for Law and Biomedical Sciences Colloquium, sponsored by the University of Utah S.J. Quinney College of Law (Jan. 19, 2021); and Faculty Speaker Series, Syracuse University School of Law (Feb. 11, 2021).

identification and de-identification laws that expressly or potentially apply to health data; identifying significant weaknesses in these laws in light of the developing reidentification literature; proposing theoretical alternatives to outdated identification and de-identification standards, including alternatives based on the theories of evolving law, non-reidentification, non-collection, non-use, non-disclosure, and non-discrimination; and offering specific, textual amendments to federal and state data protection laws that incorporate these theoretical alternatives.

TABLE OF CONTENTS

INTRODUCTION

Consider a hypothetical involving a woman with a serious mental health condition. The woman, who wishes to obtain inpatient services at a nearby psychiatric hospital, provides sensitive health information to the hospital's admissions coordinator as part of the intake and assessment process.[1] Assume the woman is subsequently admitted to the hospital, where she is thoroughly evaluated[2] and treated pursuant to a comprehensive plan of care.[3] As required by federal and state law, substantial information regarding the woman's history, reasons for admission, social services, provisional and substantiated diagnoses, treatments, and progress is documented in her electronic medical record.[4]

After ten days of inpatient services,[5] the woman has made sufficient progress such that she is discharged from the hospital and referred to outpatient care. In accordance with state law, the hospital electronically discloses significant (although purportedly de-identified) data regarding the woman to a state health care database that was created to support research on the cost, utilization, and efficacy of

[1] *See, e.g.*, TEX. HEALTH & SAFETY CODE ANN. § 572.0025(h)(3) (West 2021) (defining "intake" as "the administrative process for gathering information about a prospective [psychiatric hospital] patient and giving a prospective patient information about the facility and the facility's treatment and services"); *id.* § 572.0025(h)(2) (defining "assessment" as "the administrative process a facility uses to gather information from a prospective [psychiatric hospital] patient, including a medical history and the problem for which the patient is seeking treatment, to determine whether a prospective patient should be examined by a physician to determine if admission is clinically justified").

[2] *See, e.g.*, 42 C.F.R. § 482.61(b)(1) (2020) (requiring each patient who is admitted to a Medicare-participating psychiatric hospital to receive a psychiatric evaluation within sixty hours of admission).

[3] *See, e.g.*, *id.* § 482.61(c)(1)(iii) (requiring each patient who is admitted to a Medicare-participating psychiatric hospital to have a "comprehensive treatment plan . . . based on . . . the patient's strengths and disabilities" and requiring such plan to include the patient's treatment modalities).

[4] *See, e.g.*, *id.* § 482.61(a)–(d), (f) (requiring medical records maintained by Medicare-participating psychiatric hospitals to include substantial data); 26 TEX. ADMIN. CODE § 568.101(a) (2021) (requiring medical records maintained by Texas psychiatric hospitals to include substantial data).

[5] *See, e.g.*, Sungkyu Lee, Aileen B. Rothbard & Elizabeth L. Noll, *Length of Inpatient Stay of Persons with Serious Mental Illness: Effects of Hospital and Regional Characteristics*, 63 PSYCH. SERVS. 889, 891–92, 892 fig.1 (2012) (reporting the mean length of psychiatric hospitalization as ten days).

health care provided in the state.[6] Assume a researcher accesses the database and reidentifies the woman and other patients with mental health conditions by matching their data to publicly available newspaper reports.[7] Further assume that a hacker subsequently gains unauthorized access to the researcher's reidentified records and makes them publicly available.[8]

Independent of its reporting obligations to the state health care database, assume the hospital also has an arrangement with a technology company pursuant to which the hospital discloses supposedly anonymous medical data to the technology company.[9] After receiving the data, the technology company uses the data to create machine learning tools capable of predicting adverse health outcomes,[10] such as heart attacks and strokes. Assume the technology

[6] *See, e.g.*, VT. STAT. ANN. tit. 18, § 9410 (2021) (creating a centralized health-care database and requiring hospitals and other health-care entities in Vermont to report certain health-care data to the database).

[7] *See, e.g.*, Latanya Sweeney, *Only You, Your Doctor, and Many Others May Know*, TECH. SCI. (Sept. 28, 2015), https://techscience.org/a/2015092903 [https://perma.cc/2KSS-MAH4] (reporting that researchers affiliated with Harvard's Data Privacy Lab reidentified the subjects of allegedly de-identified hospital discharge data purchased from the State of Washington for fifty dollars by matching the data to publicly available newspaper reports).

[8] *See, e.g.*, Adil Hussain Seh, Mohammad Zarour, Mamdouh Alenezi, Amal Krishna Sarkar, Alka Agrawal, Rajeev Kumar & Raees Ahmad Khan, *Healthcare Data Breaches: Insights and Implications*, 8 HEALTHCARE 133, 134, 136, 148 (2020) (reporting "the total number of individuals affected by healthcare data breaches" between 2005 and 2019 (249.09 million), implying that the healthcare industry has faced the highest number of data breaches among all industries, finding that hacking is the most prevalent type of healthcare data breach, and defining hacking to include "malware attack[s], ransomware attack[s], phishing, spyware," and stolen data).

[9] *See, e.g.*, Class Action Complaint and Demand for a Jury Trial at 2–3, 19–29, Dinerstein v. Google, LLC, 484 F. Supp. 3d 561 (N.D. Ill. 2020) (No. 1:19-cv-04311) [hereinafter Dinerstein Lawsuit] (alleging that The University of Chicago Medical Center impermissibly disclosed hospital data to Google to support Google's creation of machine learning tools capable of predicting patients' future health conditions, and arguing that Google could reidentify the data subjects by matching their allegedly de-identified data to Google's proprietary consumer data obtained through mobile applications such as Google Maps and Waze as well as through internet protocol addresses tied to individuals' Google searches), *appeal docketed*, No. 20-3134 (7th Cir. Nov. 2, 2020). For an additional example of a data-sharing arrangement between Google and the health-care industry, see generally Rob Copeland, *Google's 'Project Nightingale' Gathers Personal Health Data on Millions of Americans*, WALL ST. J., https://www.wsj.com/articles/google-s-secret-project-nightingale-gathers-personal-health-data-on-millions-of-americans-11573496790 [https://perma.cc/4WDW-4UVR] (last updated Nov. 11, 2019, 4:27 PM), which reports on a data-sharing relationship between Google and "Ascension, a Catholic chain of 2,600 hospitals, doctors' offices and other [health-care] facilities."

[10] *See, e.g.*, Dinerstein Lawsuit, *supra* note 9, at at 2 (explaining Google's intent to "partner[] with the University [of Chicago]" to research ways "to use 'machine learning' to. . . predict future medical events").

company is able to reidentify the subjects of the medical data by matching their data to proprietary consumer data in the company's possession.[11] Further assume that, shortly after the data subjects are reidentified, a disgruntled company employee makes an unauthorized disclosure of the reidentified health information.[12]

This fact pattern is based on recent lawsuits, research studies, and media reports showing that sensitive, but allegedly de-identified, health data can be reidentified by large technology companies,[13] researchers,[14] journalists,[15] and interested community members[16] through linkage with consumer data, obituary records, and other public or private data. This Article contends that current data protection laws, which allow the use and disclosure of de-identified health data without the prior written authorization of the data subjects,[17] insufficiently protect valuable health data[18] and that amending data protection laws and health status nondiscrimination laws is necessary.

Federal and state laws have long attempted to strike a balance between protecting patient privacy and confidentiality and supporting important uses and disclosures of health information.[19] To this end,

[11] *See, e.g.*, *id.* at 3 (explaining that Google has in its possession "'detailed geolocation information that it can use to pinpoint and match exactly when certain people entered and visited the University's hospital'").

[12] *See, e.g.*, Seh et al., *supra* note 8, at 141 fig.5, 148 (finding that unauthorized internal disclosures of data are the second most prevalent type of healthcare data breach).

[13] *See supra* note 9.

[14] *See supra* note 7.

[15] *See, e.g.*, Khaled El Emam, Fida K. Dankar, Angelica Neisa & Elizabeth Jonker, *Evaluating the Risk of Patient Re-Identification from Adverse Drug Event Reports*, BMC MED. INFORMATICS & DECISION MAKING, Oct. 5, 2013, at 1, 1 [hereinafter El Emam et al., *Evaluating the Risk of Patient Re-Identification*] (noting that a national reporter reidentified a deceased twenty-six-year-old woman from a Canadian adverse drug event database by matching her de-identified database data with publicly available obituary records).

[16] *See, e.g.*, Sangchul Park, Gina Jeehyun Choi & Haksoo Ko, *Information Technology-Based Tracing Strategy in Response to COVID-19 in South Korea—Privacy Controversies*, 323 JAMA 2129, 2129 (2020) (reporting that members of the South Korean public were able to reidentify purportedly anonymous COVID-19 data).

[17] *See, e.g.*, 45 C.F.R. § 164.500(a) (2020) (applying the HIPAA Privacy Rule to "protected health information"); *id.* § 160.103 (defining "protected health information" as "individually identifiable health information"); *id.* § 164.514(a) (stating that de-identified information is not "individually identifiable health information" and therefore is not protected by the HIPAA Privacy Rule).

[18] *See, e.g.*, Seh et al., *supra* note 8, at 133 ("Data from the healthcare industry is regarded as being highly valuable. This has become a major lure for the misappropriation and pilferage of healthcare data.").

[19] *See* U.S. DEP'T OF HEALTH & HUM. SERVS., OCR PRIVACY BRIEF: SUMMARY OF THE HIPAA PRIVACY RULE 1 (2003) ("A major goal of the Privacy Rule is to assure that individuals'

many health laws restrict the use and disclosure of identifiable health data but support the use and disclosure of de-identified data.[20] The goal of health data de-identification is to prevent or minimize informational injuries to identifiable data subjects[21] while allowing the production of aggregate statistics that can be used for biomedical and behavioral research, public health initiatives, and other important activities.[22]

Many federal and state laws assume that data are de-identified when direct and indirect demographic identifiers such as names, usernames, email addresses, street addresses, and telephone numbers have been removed.[23] The reidentification literature shows, however,

health information is properly protected while allowing the flow of health information needed to provide and promote high quality health care and to protect the public's health and well being."); Letter from William W. Stead, Chair, Nat'l Comm. on Vital & Health Stat., to Thomas E. Price, Sec'y, Dep't of Health & Hum. Servs. 1 (Feb. 23, 2017) [hereinafter NCVHS Letter], https://www.ncvhs.hhs.gov/wp-content/uploads/2013/12/2017-Ltr-Privacy-DeIdentification-Feb-23-Final-w-sig.pdf [https://perma.cc/8EDD-FW69] (explaining the "many important uses for de-identified" health data).

[20] *Compare* 45 C.F.R. § 164.508(a)(1), (b)(1)(i) (2020) (requiring prior written authorization from an individual before the individual's protected health information may be used or disclosed by certain covered entities), *with id.* § 164.514(a) ("Health information that does not identify an individual and with respect to which there is no reasonable basis to believe that the information can be used to identify an individual is not individually identifiable health information," allowing the free use and disclosure of such information), *and id.* § 164.514(b)(1)–(2) (setting forth two methods for de-identifying protected health information). *See generally* Gregory E. Simon, Susan M. Shortreed, R. Yates Coley, Robert B. Penfold, Rebecca C. Rossom, Beth E. Waitzfelder, Katherine Sanchez & Frances L. Lynch, *Assessing and Minimizing Re-identification Risk in Research Data Derived from Health Care Records*, 7 EGEMS, Mar. 29, 2019, at 1, 2 (explaining that "re-identification of any individual would be a serious breach of trust").

[21] Mehmet Kayaalp, *Modes of De-identification*, AMIA ANN. SYMP. PROC. 1044, 1044 (2017); *see* FED. TRADE COMM'N, FTC INFORMATIONAL INJURY WORKSHOP 1–3 (2018) (defining informational injuries as market-based and non-market-based injuries "that consumers . . . suffer [following] privacy and security incidents, [including] data breaches [and] unauthorized disclosure[s] of data"); Mark A. Rothstein, *Ethical Issues in Big Data Health Research: Currents in Contemporary Bioethics*, 43 J.L. MED. & ETHICS 425, 426–27 (2015) (discussing physical and dignitary harms associated with "[t]he loss of privacy" in the context of big data health research when patients are identified or reidentified); Seh et al., *supra* note 8, at 134 (noting that unauthorized access to healthcare data can lead to "data tampering[, which] can lead to faulty treatment, [resulting in] fatal and irreversible losses to patients"); NCVHS Letter, *supra* note 19, at 10, 12 (explaining that data subject reidentification can "produce[] harm [and] diminish[] trust" and that "[d]one poorly, de-identification can expose individuals, protected groups, and establishments to risk of harm to physical well-being, personal dignity, reputation, or financial position").

[22] NCVHS Letter, *supra* note 19, at 1; Joshua Rolnick, *Aggregate Health Data in the United States: Steps Toward a Public Good*, 19 HEALTH INFORMATICS J. 137, 137–38 (2013).

[23] *See, e.g.*, Kayaalp, *supra* note 21, at 1044 (defining de-identification as "a process of detecting identifiers (e.g., personal names and social security numbers) that directly or indirectly point to a person (or entity) and deleting those identifiers from the data"); *see infra* Part II

that purportedly de-identified data can—and increasingly will—be reidentified.[24] This Article responds to these concerns and proceeds as follows. Part I carefully reviews health data reidentification claims and concerns, providing specific examples of de-identified health data that were reidentified following matching with other public, semipublic, or private data.[25] Part I also reviews the scientific literature assessing the risk of reidentification and the efficacy of particular de-identification techniques to show that health data have a high risk of reidentification.[26] Part II, which is supported by an Appendix,[27] presents an original synthesis of illustrative federal and state data protection laws that expressly or potentially apply to health data, identifying trends and limitations therein.[28] Part II aims to assess whether federal and state standards for health data identification and de-identification reflect the reidentification risks and disproportionate reidentification burdens described in Part I.[29] To respond to calls for data de-identification reform,[30] Part III of this Article proposes theoretical alternatives to current identification and de-identification standards and implements them in specific textual amendments to federal and state data protection laws.[31] These alternatives are based on principles of evolving law as well as the concepts of non-reidentification, noncollection, nonuse, nondisclosure, and

(assessing federal and state standards governing data identification and de-identification and providing examples of common identifiers).

[24] *See infra* Part I.

[25] *See infra* Part I.A.

[26] *See infra* Part I.B.

[27] *See* Stacey A. Tovino, *Not So Private*, 71 DUKE L.J. 985, app. (2021), https://dlj.law.duke.edu/2022/01/appendix-not-so-private/ [https://perma.cc/JSK7-MP72] [hereinafter Appendix].

[28] *See infra* Part II.

[29] *See infra* Part II.

[30] For example, Matthew Fisher wrote,

> The one area for attention though is the ability to de-identify protected health information. . . . [W]hat may no longer be an academic argument is whether data can truly be de-identified given the plethora that exists and the variety of sets that can be combined to re-identify information that was previously believed to be completed de-identified. . . . [T]hat line of thinking hints at a gap in HIPAA . . . .

Matthew Fisher, *Google-Ascension: Why Is HIPAA Probably Not Being Violated?*, HIT CONSULTANT (Nov. 13, 2019), https://hitconsultant.net/2019/11/13/google-ascension-why-is-hipaa-probably-not-being-violated [https://perma.cc/7469-V7TZ]; *see also* INST. OF MED., BEYOND THE HIPAA PRIVACY RULE: ENHANCING PRIVACY, IMPROVING HEALTH THROUGH RESEARCH 190, 265, 281 (Sharyl J. Nass et al. eds., 2009) (calling for legal prohibitions against the unauthorized reidentification of health data subjects and the imposition of sanctions on such reidentification).

[31] *See infra* Part III.

nondiscrimination.[32] If adopted by federal and state lawmakers, this Article's proposals will protect health data subjects from discrimination and other injuries associated with the use, disclosure, and redisclosure of their identifiable—and potentially reidentifiable—health data.

I. NOT SO PRIVATE

This Part carefully reviews health data reidentification claims and concerns, providing specific examples of de-identified health data, including hospital medical record data, hospital discharge data, adverse drug event data, physical activity data, and infectious disease data, that were reidentified following matching with other public, semipublic, or private data.[33] This Part also reviews the scientific literature assessing the risk of reidentification and the efficacy of particular de-identification techniques showing that health data have a high reidentification risk.[34] As discussed below, health data from which as many as eighteen different identifiers have been removed still carry some reidentification risk.[35] In addition, this reidentification risk is not shared equally among data subjects.[36] Vulnerable individuals, including those with rare health conditions and members of minority racial and ethnic groups, bear disproportionate risk.[37] Data reidentification is concerning because of its significant human impact, including psychological distress, financial injury, loss of trust in the medical system, stereotyping, stigma, and discrimination.[38]

A. Reidentification Claims and Concerns

A number of lawsuits, research studies, and media reports show how the reidentification of health data occurs, including through the linkage (or matching) of supposedly de-identified health data with

[32] *See infra* Part III.

[33] *See infra* Part I.A.

[34] *See infra* Part I.B.

[35] *See infra* Part I.A.

[36] *See infra* Part I.B.

[37] *See infra* Part I.B.

[38] *See* sources cited *supra* note 21 (providing examples of injuries suffered by individuals whose data are reidentified or otherwise breached).

public, purchasable, or proprietary data.[39] *Dinerstein v. Google, LLC*[40] is illustrative.[41] On June 26, 2019, plaintiff Matt Dinerstein sued Google and the University of Chicago Medical Center on behalf of himself and other patients whose electronic health records ("EHRs") between 2010 and 2016 were disclosed to Google for predictive health research.[42] Dinerstein alleged several causes of action, including breach of contract, tortious interference with contract, intrusion upon seclusion, unjust enrichment, and violations of consumer law.[43]

These claims stemmed from the Medical Center's failure to de-identify the EHRs before disclosing them to Google without prior patient notification or authorization.[44] Dinerstein was particularly worried because he had not one, but two inpatient stays at the Medical Center.[45] During these stays, the hospital collected and recorded significant data about Dinerstein, including his "vital signs, diagnoses, procedures, and prescriptions."[46] The hospital also maintained timestamps indicating the exact date and time Dinerstein visited parts of the hospital and received certain health-care services.[47]

In his lawsuit, Dinerstein focuses heavily on the fact that the EHRs contained timestamps.[48] Because Google also obtains precise

[39] *See, e.g.*, Simon et al., *supra* note 20, at 3 (discussing re-identification using public data); Sweeney, *supra* note 7, at 2 (discussing re-identification using purchased data); Dinerstein Lawsuit, *supra* note 9, at 25 (discussing re-identification using proprietary location data).

[40] Dinerstein v. Google, LLC, 484 F. Supp. 3d 561 (N.D. Ill. 2020), *appeal docketed*, No. 20-3134 (7th Cir. Nov. 2, 2020).

[41] *See* Dinerstein Lawsuit, *supra* note 9, at 1–4.

[42] *Id.* at 1, 42; *Dinerstein*, 484 F. Supp. 3d. at 566, 568. Predictive health, sometimes called predictive modeling or predictive health analytics, is a form of modeling that uses statistical methods, data mining, and/or game theory to analyze current and historical health data collected from healthcare providers and/or health insurers. *Predictive Analytics in Health Care*, DELOITTE (July 19, 2019), https://www2.deloitte.com/us/en/insights/topics/analytics/predictive-analytics-health-care-value-risks.html [https://perma.cc/8B26-7VFH]. The goal of predictive health is to identify the type and timing of future health-care events and prevent such events from occurring. *See* Kenneth L. Brigham, *Predictive Health: The Imminent Revolution in Health Care*, 58 J. AM. GERIATRICS SOC'Y S298, S298 (2010).

[43] Dinerstein Lawsuit, *supra* note 9, at 41–42.

[44] *Id.* at 19–21, ¶¶ 67, 69–74.

[45] *Dinerstein*, 484 F. Supp. 3d at 566, 570.

[46] *Id.* at 566.

[47] *Id.* at 570, 586; Dinerstein Lawsuit, *supra* note 9, at 26–27.

[48] Dinerstein Lawsuit, *supra* note 9, at 20–21, ¶¶ 71, 75.

geolocation data through applications like Google Maps[49] and Waze,[50] and through IP addresses tied to individuals' Google searches,[51] Dinerstein argued that Google could reidentify him and the other EHR subjects in violation of their right to privacy.[52] Specifically, Dinerstein argued that Google could match the patients' timestamp information to Google's geolocation data.[53]

Although Google is a multinational company with a large number of internet-related services and products that collect considerable data,[54] Google is not uniquely capable of reidentifying health data. A

[49] *See id.* at 24. Google Maps is a mobile application that provides "real-time GPS navigation, traffic, and transit in[formation]" to individuals located in more than "220 countries and territories." *See Google Maps*, GOOGLE PLAY, https://play.google.com/store/apps/details?id=com.google.android.apps.maps [https://perma.cc/MRT6-B36G] (last updated Sept. 30, 2021).

[50] *See* Dinerstein Lawsuit, *supra* note 9, at 24. Waze is a mobile application that analyzes community-driven data for the purpose of helping drivers and riders arrive at their destinations "faster, smoother, safer, and happier." *Waze Carpool*, N. ORANGE CNTY. CHAMBER, https://business.nocchamber.com/list/member/waze-carpool-24774 [https://perma.cc/S2S3-NQBP]; *About Us*, WAZE, https://www.waze.com/about [https://perma.cc/4JCP-U8SD] (noting that Waze is a mobile navigation application that analyzes community-driven data).

[51] *See* Dinerstein Lawsuit, *supra* note 9, at 24. *See generally* John Herrman, *Google Knows Where You've Been, but Does It Know Who You Are?*, N.Y. TIMES MAG. (Sept. 12, 2018), https://nyti.ms/2NbjeV2 [https://perma.cc/QPG7-W7QW] (citing a report by the Associated Press, which noted that "[s]ome Google app[lications] automatically store time-stamped location data without asking").

[52] *See* Dinerstein Lawsuit, *supra* note 9, at 23–24, 36 ¶¶ 84–85, 134. The U.S. District Court for the Northern District of Illinois noted,

> [F]or a user of Google applications like Mr. Dinerstein, Google can track the specific University hospital buildings or departments he visited and the time of his visits. Plaintiff alleges that the combination of such geolocation information and the EHRs, which include the date and time of hospital services, "creates a perfect formulation of data points for Google to identify who the patients in those records really are."

Dinerstein v. Google, LLC, 484 F. Supp. 3d 561, 570 (N.D. Ill. 2020) (citations omitted).

[53] Dinerstein Lawsuit, *supra* note 9, at 28, ¶ 93. *See generally* Anya E.R. Prince, *Location as Health*, HOUS. J. HEALTH L. & POL'Y (forthcoming), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3767122 [https://perma.cc/PZ8D-YPAY] (examining how private companies can use location data "to infer health information" and proposing five data protections that will help safeguard individuals' health privacy as identified through location data). On September 4, 2020, the U.S. District Court for the Northern District of Illinois granted the motion to dismiss Google, The University of Chicago, and The University of Chicago Medical Center on grounds unrelated to the validity of the plaintiff's factual reidentification allegations. *See Dinerstein*, 484 F. Supp. 3d at 561, 580, 588, 593–95. On November 2, 2020, Dinerstein appealed this dismissal. *Id.*, *appeal docketed*, No. 20-3134 (7th Cir. Nov. 2, 2020). As of this writing, the U.S. Court of Appeals for the Seventh Circuit has yet to rule on Dinerstein's appeal.

[54] *See* Brian X. Chen, *It's Google's World. We Just Live in It*, N.Y. TIMES, https://nyti.ms/2IPqrIX [https://perma.cc/7DVA-YEPV] (last updated May 3, 2021) (reporting Google's dominance as a technology company); Dale Smith, *Google Collects a Frightening Amount of Information About You*, CNET.COM (June 28, 2020, 10:18 AM), https://www.cnet.com/how-to/google-collects-a-frightening-amount-of-data-about-you-you-can-

small research team affiliated with Harvard University was able to reidentify the subjects of allegedly de-identified hospital discharge data purchased from the State of Washington for fifty dollars.[55] Although the data lacked direct patient identifiers, it did have the diagnoses, medical and surgical procedures, physician names, hospital charges, payment methods, and ages of patients who received care in Washington hospitals.[56] After linking the hospital data with newspaper reports of hospitalizations following motor vehicle accidents, assaults, and events like house fires, thirty-five of eighty-one patients, or 43 percent, were reidentified.[57]

Lawmakers in other states attributed this study's success to the specifics of the Washington database, which contains the discharged patients' ages in months, rather than years.[58] Accordingly, the researchers repeated this study using Maine and Vermont hospital discharge data without patient ages in months.[59] Then they ran the Maine hospital discharge data against 177 local news stories and the Vermont data against thirty-eight local news stories.[60] Of the names from the 244 names in the 177 local news stories, 28.3 percent "uniquely matched" a single Maine hospitalization.[61] Maine had not completely de-identified its hospital discharge data in accordance with the Health Insurance Portability and Accountability Act ("HIPAA") Privacy Rule's de-identification Safe Harbor,[62] an industry standard for de-identification that requires the removal of eighteen different

find-and-delete-it-now [https://perma.cc/T9F5-6M5X] (describing the vast amounts of data Google has collected).

[55] Sweeney, *supra* note 7, at 2, 22.

[56] *Id.* at 2, 9–10.

[57] *Id.* at 7, 14, 16. When the Harvard researchers reported their reidentification results to the State of Washington, the State began restricting public access to its hospital discharge data, *id.* at 19, improving patient privacy going forward.

[58] Ji Su Yoo, Alexandra Thaler, Latanya Sweeney & Jinyan Zang, *Risks to Patient Privacy: A Re-identification of Patients in Maine and Vermont Statewide Hospital Data*, TECH. SCI. (Oct. 9, 2018), https://techscience.org/a/2018100901 [https://perma.cc/75TL-WQ55] ("[M]any states were not convinced that the same re-identification strategy would be successful on their datasets. One reason was a belief that Washington State was more vulnerable because it shared patient age in months, a practice not followed by many other states.").

[59] *Id.* at 2, 43–44.

[60] *Id.* at 3, 23.

[61] *Id.* at 3.

[62] *Id.* at 21. The Maine discharge data contained the quarter of the year of the patient's hospital admission as well as the dates of particular health-care procedures performed. *Id.* at 4, 7.

identifiers before information is considered properly de-identified.[63] However, when the researchers completely de-identified the Maine data in accordance with the Safe Harbor, eight matches (a reidentification rate of 3.2 percent) still resulted.[64] As discussed in more detail in Part II, HIPAA Privacy Rule's de-identification standard does not protect against patient reidentification.

With respect to the Vermont hospital discharge data, the researchers found that of the names from the forty-seven local news stories, 34 percent uniquely matched to one hospitalization.[65] When the Vermont hospital discharge data were completely de-identified in accordance "with the HIPAA Safe Harbor . . . , five matches" (a 10.6 percent reidentification rate) still resulted.[66] The researchers concluded: "[P]atients' personal information is vulnerable to re-identification even when hospital data is de-identified according to HIPAA Safe Harbor guidelines."[67] They also "call[ed] for more rigorous inquiry on the vulnerabilities that exist even when following HIPAA Safe Harbor as a standard for de-identification."[68]

Researchers at the Massachusetts Institute of Technology, the University of California, Berkeley, Tsinghua University, and the University of California, San Francisco, used machine learning[69] to reidentify 95 percent and 94 percent of the adult subjects and 87 percent and 86 percent of the minor subjects whose allegedly de-identified physical activity data were collected from wearable devices as part of National Health and Nutrition Examination Surveys in 2003

[63] *Id.* at 4 (stating, "[T]he Safe Harbor de-identification guidelines are still widely recognized as a standard for de-identifying all kinds of health data"); *see infra* Part II.B.2 (assessing the HIPAA Safe Harbor). The HIPAA Safe Harbor requires the removal of all elements of dates other than year. The regulations require removing

> [a]ll 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.

45 C.F.R. § 164.514(b)(2)(i)(C) (2020).

[64] Ji Su Yoo et al., *supra* note 58, at 3.

[65] *Id.*

[66] *Id.*

[67] *Id.*

[68] *Id.*

[69] Machine learning, a subset of artificial intelligence, is the scientific study of algorithms and statistical models that computer systems use to perform specific tasks without using explicit instructions and by relying on patterns and inference instead. *See generally* Rahul C. Deo, *Machine Learning in Medicine*, 132 CIRCULATION 1920 (2015) (summarizing applications of machine learning in medicine).

through 2004 and 2005 through 2006, respectively.[70] Although wearable device manufacturers such as Fitbit[71] and Strava[72] have long maintained that collecting and disclosing de-identified physical activity data do not raise privacy concerns, this study suggests otherwise.[73]

Journalists and news reporters also have succeeded in reidentifying health data. In Canada, a national broadcaster reidentified a deceased twenty-six-year-old woman whose de-identified data had been disclosed to a Canadian adverse drug event ("ADE") database after her drug-associated death.[74] The broadcaster matched her de-identified ADE data with publicly available obituary data.[75] The broadcaster then obtained additional information about the woman, contacted the woman's family, and released a story that used the woman to illustrate the adverse effects of the drug at issue.[76] The woman's parents thus lost the ability to keep the nature and cause of their daughter's death private.

Even lay community members have reidentified the subjects of supposedly anonymous health data. In the Republic of Korea, citizens with COVID-19 were quickly reidentified[77] after the South Korean

[70] Liangyuan Na, Cong Yang, Chi-Cheng Lo, Fangyuan Zhao, Yoshimi Fukuoka & Anil Aswani, *Feasibility of Reidentifying Individuals in Large National Physical Activity Data Sets from Which Protected Health Information Has Been Removed with Use of Machine Learning*, 1 JAMA NETWORK OPEN, Dec. 21, 2018, at 1, 2–3, 9 (using machine learning to demonstrate the "[f]easibility of [r]eidentifying [i]ndividuals in . . . [p]hysical [a]ctivity [d]ata . . . [f]rom [w]hich [p]rotected [h]ealth [i]nformation [h]as [b]een [r]emoved.").

[71] *See Fitbit Privacy Policy*, FITBIT, https://www.fitbit.com/global/us/legal/privacy-policy [https://perma.cc/ZG7U-49NJ] (last updated Aug. 16, 2021) ("We may share non-personal information that is aggregated or de-identified so that it cannot reasonably be used to identify an individual.").

[72] *See Strava Privacy Policy*, STRAVA, https://www.strava.com/legal/privacy [https://perma.cc/EC2G-S7TV] ("Information connected to you that is no longer necessary and relevant to provide our Services may be de-identified or aggregated with other non-personal data to provide insights which are commercially valuable to Strava, such as statistics of the use of the Services.").

[73] *See* Na et al., *supra* note 70, at 1 ("The findings of this study suggest that current practices for deidentifying physical activity data are insufficient for privacy and that deidentification should aggregate the physical activity data of many people to ensure individuals' privacy.").

[74] KHALED EL EMAM, GUIDE TO THE DE-IDENTIFICATION OF PERSONAL HEALTH INFORMATION 6 (2013) (reporting the story of the young woman's reidentification).

[75] El Emam et al., *Evaluating the Risk of Patient Re-Identification*, *supra* note 15, at 1 (reporting the same story).

[76] *Id.*

[77] *See, e.g.*, Park et al., *supra* note 16, at 2129 ("Some of these individuals were affected by unwanted privacy invasion and even became subject to public disdain.").

government publicized anonymous[78] data that included the citizens' "sex, nationality, and age."[79] Community members then linked the anonymous data with other data provided by local governments, including "detailed routes [traveled by infected persons] as well as the names of restaurants, shops, and other business premises that infected persons visited."[80]

In the United States, COVID-19 contact tracing and exposure notification have raised similar concerns.[81] Consider Google and Apple's jointly developed Exposure Notifications System ("ENS"), which uses mobile devices' Bluetooth technology to track users' proximity to other users who test positive for COVID-19.[82] Although Google and Apple claim that ENS user privacy is protected,[83] critics worry that users' COVID-19 data could be disclosed to government

[78] Anonymous data do not include the data subjects' names. *See, e.g.*, *Anonymous*, OXFORD LEARNER'S DICTIONARIES, https://www.oxfordlearnersdictionaries.com/us/definition/english/anonymous. [https://perma.cc/5L46-K85E] (defining anonymous as "(of a person) with a name that is not known or that is not made public").

[79] In their article published in JAMA, Park, Choi, and Ko state that

> [t]he current [public] disclosures on the [South Korean Ministry of Health and Welfare] home page include, in addition to [the path and means of transportation of infected persons, the medical institutions that treated infected persons, and the health status of those in contact with an infected person], the sex, nationality, and age of infected persons, although their names are not revealed. Certain municipal and local governments, however, went further and provided highly detailed routes as well as the names of restaurants, shops, and other business premises that infected persons visited.

Park et al., *supra* note 16, at 2129.

[80] *Id.*

[81] *See generally* I. Glenn Cohen, Lawrence O. Gostin & Daniel J. Weitzner, *Digital Smartphone Tracking for COVID-19: Public Health and Civil Liberties in Tension*, 323 JAMA 2371 (2020) (asking whether COVID-19 digital tracking "confer[s] sufficient public health benefit to justify adoption given privacy concerns" and arguing that "[w]idespread deployment would only be warranted if" public health efficacy balanced favorably against "privacy and other costs").

[82] *See* Apple & Google, *Exposure Notifications: Help Slow the Spread of COVID-19, with One Step on Your Phone*, GOOGLE: COVID-19 INFO. & RES., https://www.google.com/covid19/exposurenotifications [https://perma.cc/UY5A-FL3U] ("We understand how important your privacy is. Here's how we've built this system to respect your privacy and keep you in control.").

[83] *See* Apple & Google, *Privacy-Preserving Contact Tracing*, APPLE, https://covid19.apple.com/contacttracing [https://perma.cc/WM3R-JTB7] (referring to their technology solution as "privacy-preserving contact tracing"); APPLE & GOOGLE, EXPOSURE NOTIFICATION: BLUETOOTH SPECIFICATION 8 (2020) ("Maintaining user privacy is an essential requirement in the design of this specification."); Jack Nicas & Daisuke Wakabayashi, *Apple and Google Team Up To 'Contact Trace' the Coronavirus*, N.Y. TIMES, https://www.nytimes.com/2020/04/10/technology/apple-google-coronavirus-contact-tracing.html [https://perma.cc/R94L-S2YD] (last updated June 3, 2020) ("Apple and Google said they were discussing how much information to include in those [COVID-19 contact tracing] alerts with health officials, aiming to strike a balance between being helpful while also protecting the privacy of those who have the coronavirus.").

agencies, insurers, marketing companies, data brokers, and cybercriminals for purposes unrelated to public health.[84]

B. Research Assessing Reidentification Risk

A growing scientific literature aims to calculate the risk of health data reidentification as well as the efficacy of particular data de-identification and data perturbation methods.[85] In 2011, researchers affiliated with the CHEO Research Institute, the University of Ottawa, and Vanderbilt University systematically reviewed the "statistics, computer science, and health informatics" literature to locate data reidentification attacks and to determine the portion of data correctly

[84] *See, e.g.*, Editorial, *Privacy Cannot Be a Casualty of the Coronavirus*, N.Y. TIMES (Apr. 7, 2020), https://nyti.ms/2XiIwnG [https://perma.cc/PC2T-DNRB] ("Americans have lost control over a lot as a result of the coronavirus. At least they should be able to control what happens to their personal data."); Daniel R. Stoller & Lydia Wheeler, *Apple-Google Virus Tracking Plan Carries Data Security Risks*, BLOOMBERG L. NEWS (Apr. 16, 2020, 5:46 AM), https://news.bloomberglaw.com/privacy-and-data-security/apple-google-virus-tracking-plan-carries-data-security-risks [https://perma.cc/J86Y-BRLG] (reporting, in the context of COVID-19 cell-phone tracking, that "[c]ybercriminals might be able to tie the anonymized private data back to specific individuals, including corporate and government officials," "[s]ometimes anonymized data isn't a 'silver bullet' to guard information, because it can be combined with data on social media and other online records," and "Apple and Google could also face federal and state regulatory scrutiny if data is breached and is not properly de-identified"); Ben Brody & Naomi Nix, *Pandemic Data-Sharing Puts New Pressure on Privacy Protections*, BLOOMBERG L. NEWS (Apr. 5, 2020, 6:00 AM), https://www.bloomberglaw.com/bloomberglawnews/health-law-and-business/XDR76Q5G000000 [perma.cc/YFU7-UNNB] ("The sources of anonymous data can sometimes be exposed by combining datasets. Even when made anonymous, location points that come from phone apps, for instance, can be linked to a person by checking who lives at the address where the phone rests at night."); Shira Stein & Daniel R. Stoller, *NIH in Pursuit of Privacy-Protected Contact Tracing Tool (1)*, BLOOMBERG L. NEWS, https://www.bloomberglaw.com/product/blaw/bloomberglawnews/health-law-and-business/ [https://perma.cc/35SF-TCCS] (last updated May 27, 2020, 1:09 PM) ("An NIH-backed smartphone app could still raise privacy concerns even if it de-identifies or anonymizes the data it collects. Advocates and researchers have shown that such data can be traced back to individuals if not done properly, undercutting any privacy protections that may be built into the application."). *See generally* Clarisa Long, *Privacy and Pandemics*, *in* LAW IN THE TIME OF COVID-19, at 89 (Katharina Pistor ed., 2020) (discussing a range of privacy issues raised by the COVID-19 pandemic).

[85] *See generally* Nikunjkumar Patel, Data Mining: Privacy Preservation Using Perturbation Technique 12 (May 6, 2015) (M.S. Thesis, SUNY Polytechnic Institute), https://dspace.sunyconnect.suny.edu/bitstream/handle/1951/67640/NPatel-Final.pdf [https://perma.cc/ZM5S-CUTQ] (identifying and discussing data perturbation techniques, including "random noise addition methods, rotation perturbation, projection perturbation and k-anonymization model").

reidentified in those attacks.[86] The researchers identified fourteen distinct reidentification attacks and found that 26 percent of the attacked records (and 34 percent of the attacked health records) were successfully reidentified.[87] This evidence revealed a "high" rate of data reidentification.[88]

Two years later, a subset of the same University of Ottawa researchers published a second study assessing "the risk of patient re[]identification" in Canada's ADE database.[89] In particular, the Ottawa researchers developed and tested an updated reidentification risk model that was designed to mimic the behaviors of realistic reidentifiers, called adversaries, who were mildly or highly motivated to verify potentially successful data matches.[90] The researchers then applied their model to the risk of reidentifying ADE data.[91] Due to the public availability of mortality data (which could serve as a link to allegedly de-identified ADE data containing the date of the death report, which might be close to the date of death), the researchers focused specifically on the risk of reidentifying ADE data of drug-induced deaths.[92]

For mildly motivated adversaries, the risk of reidentification following the ADE database's disclosure of a patient's province, age at death, gender, and exact date of ADE report was high, at 18.44 to 30.78 percent, although removing the patient's province reduced the risk significantly to 1.95 to 5.05 percent.[93] The risk of reidentification was lower still at 0.21 to 0.63 percent if the ADE database only disclosed the month and year (versus the precise date) of the ADE report, even

[86] Khaled El Emam, Elizabeth Jonker, Luk Arbuckle & Bradley Malin, *A Systematic Review of Re-Identification Attacks on Health Data*, PLOS ONE, Dec. 2, 2011, at 1, 2 [hereinafter El Emam et al., *A Systematic Review*].

[87] *Id.* at 1, 5, 7.

[88] *Id.* at 1, 7.

[89] *See* El Emam et al., *Evaluating the Risk of Patient Re-Identification*, *supra* note 15, at 1. *See generally Canada Vigilance Adverse Reaction Online Database*, GOV'T OF CAN., https://www.canada.ca/en/health-canada/services/drugs-health-products/medeffect-canada/adverse-reaction-database.html [https://perma.cc/Y86V-2V2F] (last updated Sept. 1, 2021) (providing information about Canada Vigilance, formerly named CADRIS).

[90] El Emam et al., *Evaluating the Risk of Patient Re-Identification*, *supra* note 15, at 1, 4–8 (describing the difference in the identification verification efforts of mildly versus highly motivated adversaries, explaining how adversaries can verify potential matches, and explaining the necessity of the researchers' model).

[91] *Id.* at 8–9.

[92] *Id.* at 8–10.

[93] *Id.* at 9–10.

if all other data, including province, were disclosed.[94] For highly motivated adversaries, all ADE records had a high risk of reidentification, "but the plausibility of that scenario is limited because of the financial and practical deterrent even for highly motivated adversaries."[95]

Based on this, the researchers did not object to disclosing the ADE's province field, but did object to disclosing the date of the death report of the individuals reported to the database.[96] The researchers also recommended that: (1) "the public . . . be discouraged from [sharing] personal information"; (2) "commercial and government[al] organizations . . . be discouraged from collecting non-required personal information"; and (3) "data custodians . . . generalize their data to increase the costs of verification for an adversary."[97]

Although the Ottawa research shows that data containing indirect patient demographic identifiers (such as province, age at death, and gender) risk reidentification, few studies have assessed reidentification risk for data containing no direct or indirect demographic identifiers. Aware of this gap, a group of Vanderbilt Department of Biomedical Informatics researchers investigated reidentification risk for data in which only clinical features, such as standardized diagnostic codes, remained.[98] In particular, the researchers investigated the feasibility of linking otherwise de-identified electronic medical records from a population of 1,174,793 Vanderbilt University Medical Center patients that contained standardized diagnostic codes with external data deposited into the DataBase of Genotypes and Phenotypes containing the same diagnostic codes.[99] An example of a standardized diagnosis code is "ICD 493.00," which is the former International Classification of Diseases ("ICD") diagnostic code for a type of asthma.[100]

The Vanderbilt researchers then measured the likelihood of linking the Vanderbilt medical records with the External Data based

[94] *Id.* at 10.

[95] *Id.* at 1, 10.

[96] *Id.* at 10–11.

[97] *Id.* at 11.

[98] *See* Grigorios Loukides, Joshua C. Denny & Bradley Malin, *The Disclosure of Diagnosis Codes Can Breach Research Participants' Privacy*, 17 J. AM. MED. INFORMATICS ASS'N 322, 322 (2010) ("To the best of our knowledge, this work is the first to illustrate that clinical features released in the form of standardized codes can also facilitate privacy violations.").

[99] *See id.* at 322–24.

[100] *See id.* at 323 (using this diagnostic code example in Figure 1); *2012 ICD-9-CM Diagnosis Code 493.0: Extrinsic Asthma*, ICD9DATA.COM, http://www.icd9data.com/2012/Volume1/460-519/490-496/493/493.0.htm [https://perma.cc/XC8W-VJ8U].

on standardized diagnostic codes in two scenarios: (1) when "no [reidentification] protecti[ve] measures [were] applied and (2)" when the data "perturbation[101] techniques of suppression" (i.e., "the removal of patient records or particular quasi-identifier values") and generalization (i.e., "the replacement of quasi-identifier values with more general, but semantically consistent values") were applied.[102]

The researchers found that when five-digit ICD codes were used without any perturbation techniques, the reidentification risk was very high—a full 96 percent.[103] After suppressing all rare ICD codes[104] (i.e., "ICD-9 codes that appeared in at most 5%, 15%[,] and 25% of transactions"),[105] the researchers found that the reidentification risk was still 75 percent and 26 percent (applying "the 5% and 15% suppression threshold[s]," respectively).[106] The threshold at which no Vanderbilt Medical Record subject could be reidentified was 25 percent.[107] When the researchers tried generalizing the ICD codes "to their three-digit representation," there was a privacy gain of less than 2 percent.[108]

"[T]he re[]identification of patient-specific data through standardized [diagnostic codes] is a practical privacy threat"[109] because "the majority of patients' diagnosis codes are unique not only within the sample, but also with respect to the larger population of the 1.2 million patients from which they were derived."[110] The researchers concluded that "neither suppression of rare codes (a requirement of the HIPAA [Safe Harbor] privacy rule), nor generalization, sufficiently protects records while retaining clinically meaningful

[101] In this context, data perturbation is "the changing of patient-specific quasi-identifier values." Loukides et al., *supra* note 98, at 323.

[102] *Id.* at 323–24.

[103] *See id.* at 325 ("Notice that, over 96% of patients are vulnerable to re-identification when no perturbation is invoked. The attack may thus be successful in practice.").

[104] An example of a rare ICD code would be the ICD code (G31.01) for Pick's disease, *2022 ICD-10-CM Diagnosis Code G31.01: Pick's Disease*, ICD10DATA.COM, https://www.icd10data.com/ICD10CM/Codes/G00-G99/G30-G32/G31-/G31.01 [https://perma.cc/8FDN-HS8E], which is a rare form of dementia. *Pick's Disease*, PENN MED., https://www.pennmedicine.org/for-patients-and-visitors/patient-information/conditions-treated-a-to-z/picks-disease [https://perma.cc/H7CD-DSPS].

[105] Loukides et al., *supra* note 98, at 325.

[106] *Id.*

[107] *Id.*

[108] *Id.*

[109] *Id.* at 327.

[110] *Id.* at 322–23.

information."[111] Parts II and III will discuss the implications of these conclusions for the effectiveness of federal and state data identification and de-identification law.[112]

Finally, health researchers affiliated with Kaiser Permanente,[113] HealthPartners Institute, and Baylor Scott and White Research Institute describe a "framework for assessing [data reidentification] risk."[114] The risk of reidentifying the subject of a particular health record or other sensitive record depends on three factors: (1) the availability of public or other data[115] containing one or more elements that overlap with the sensitive record; (2) "the size of the class defined by those overlapping elements in which the [particular sensitive] record . . . falls;" and (3) "the overlap in the population covered by the [sensitive records] and the population covered by the external data source."[116]

For example, the reidentification risk is higher if "the population covered by [the] external data source" (e.g., a single state's motor vehicle crash database) is "congruent or completely overlapping" with the population covered by the health or other sensitive record set (e.g., the same state's "all-payer insurance claims database").[117] The reidentification risk is lower if a single state motor vehicle crash database represents a subset consisting of one particular insurer's claims in a state that licenses a number of other insurers.[118] The reidentification risk is lower still if the two data sources cover partially overlapping populations.[119] For example, consider a single state's motor vehicle crash database that is linked to the insurance claims

[111] *Id.* at 327; *see also id.* at 323 ("[P]opular techniques, such as suppression and generalization, fail to prevent re-identifying patients, or excessively distort the released information to the point that it loses its clinical utility for GWAS [(genome-wide association studies)].").

[112] *See infra* Parts II, III.

[113] "Kaiser Permanente is one of the nation's largest not-for-profit health plans, serving 12.5 million members." *Fast Facts: Our Company*, KAISER PERMANENTE, https://about.kaiserpermanente.org/who-we-are/fast-facts [https://perma.cc/4NB8-47DW].

[114] *See* Simon et al., *supra* note 20, at 1.

[115] Examples of external data include consumer location data (such as that referenced in *Dinerstein v. Google, LLC*), hospital discharge data (such as that referenced in the Harvard University studies), and public obituary data (such as that referenced in the Canadian national broadcaster reidentification example). *See supra* Part I.A (discussing these cases, studies, and anecdotes).

[116] Simon et al., *supra* note 20, at 3–4.

[117] *Id.*

[118] *Id.* at 4.

[119] *Id.*

records of a single insurer that insures only "[one] in [five state] residents in a [three]-state area."[120]

Given this, the researchers offered three steps for data stewards to follow to assess the risk of reidentification of a sensitive but purportedly de-identified record set, such as a health record set.[121] "First, a data steward should consider the range of external data sources containing overlapping data elements as well as explicit identifiers."[122] "Second, a data steward should examine the prevalence of unique or nearly unique records defined by those overlapping data elements."[123] The researchers offered the example of a sensitive, but purportedly de-identified, suicide risk prediction ("SRP") research dataset that had a number of data elements that overlapped with publicly available mortality records, including sex, age group range, race, "Hispanic ethnicity, calendar year of death, and" category of death (e.g., overdose).[124] "[U]nique or nearly unique records"[125] could then be suppressed.[126] "Third, a data steward should consider the likely pattern of population overlap with identified external data sources."[127] "If[, for example,] the population covered by the [health data] overlaps completely with that covered by an external data source, then each cell . . . in the [health] research database [could] be completely linked to the corresponding cell . . . in the [identifiable,] external data[source]."[128] The researchers then illustrated the application of their three steps to an allegedly de-identified SRP research dataset "to estimate re[]identification risk" following linkage with external data in a publicly available state mortality database.[129] In particular, common data elements, like "Hispanic females aged 13–17 dying in 2012 in Washington state by overdose judged to have undetermined intent," "identified only [three] individuals in the" allegedly de-identified SRP dataset.[130] Assuming that Kaiser Permanente provides health insurance to one-fifth of the State of Washington, "then those three records in

[120] *Id.*

[121] *Id.* at 5.

[122] *Id.*

[123] *Id.*

[124] *Id.* at 2, 5.

[125] Nearly unique records have some overlapping data elements.

[126] *See id.* at 5.

[127] *Id.*

[128] *Id.*

[129] *Id.* at 6.

[130] *Id.*

the research dataset would match . . . an estimated [fifteen] recordsin the Washington state mortality data[set]."[131] Thus, someone with both the sensitive SRP data and the public mortality data could find fifteen records in the SRP data and know that a specific person is among those fifteen.[132]

The researchers then offered measures designed to prevent reidentification, including whole record deletion and alteration of certain values within a record,[133] such as the patient's health system or state. However, such measures could reduce the dataset's value for suicide prevention research or another publicly beneficial activity.[134] This is because the less that is known about a patient, the less is known about why the patient committed suicide and the less valuable the data are for predicting or preventing future suicides.

The researchers emphasized that the "risk of re[]identification does not fall equally among all people included in a research dataset."[135] "Instead[,] it falls largely or exclusively on those in small cells or classes, usually defined by demographic or health-related characteristics," such as "vulnerable or traditionally disadvantaged groups, including people with rarer health conditions and members of minority racial or ethnic groups."[136] Therefore, they recommended that disproportionate reidentification burdens be considered alongside the benefits traditionally associated with including members of vulnerable populations in research.[137]

Part I has carefully reviewed health data reidentification claims and concerns, providing specific examples of de-identified health data that were reidentified following matching with public, semipublic, or private data. This Part has also reviewed illustrative studies investigating the risk of reidentification and the efficacy of particular de-identification or perturbation techniques. Four conclusions may be drawn from this Part. First, health data have a high risk of

[131] *Id.*

[132] *See id.*

[133] *Id.*

[134] *Id.*

[135] *Id.* at 8.

[136] *Id.*

[137] *Id.*; *see also* Heng Xu & Nan Zhang, *Implications of Data Anonymization on the Statistical Evidence of Disparity*, INFORMS PUBSONLINE, June 4, 2021, at 1, 2 (asking "whether data anonymization could mask gross statistical disparities between [vulnerable] subpopulations in the data," masking health disparities associated with "gender, race, ethnicity, income, [and] sexual orientation" (emphasis omitted)).

reidentification.[138] Second, health data from which as many as eighteen different identifiers have been removed still carry some risk of reidentification.[139] Third, the suppression of rare data elements and the generalization of other data elements—two approaches followed by some data protection laws—may be insufficient to prevent reidentification.[140] Finally, the risk of reidentification is not shared equally by data subjects.[141] Vulnerable individuals, including individuals with rare health conditions and individuals who are "members of minority racial [and] ethnic groups," may bear a disproportionate burden of reidentification.[142] Part II, below, assesses whether current federal and state data protection laws reflect—and respond to—these findings.

II. THE LAW OF (DE)IDENTIFICATION

A number of current and pending federal and state data protection laws expressly or potentially protect patient privacy, health information confidentiality, or health data security, and/or require notification of health data subjects in the event of a privacy or security breach.[143] As discussed below, these laws contain a wide variety of

[138] *See, e.g.*, *supra* notes 57, 61, 65, 70, 87, 93, 103 and accompanying text (reporting rates of health data reidentification from particular studies).

[139] As discussed in more detail at *infra* Part II.B.2, the HIPAA Safe Harbor requires a HIPAA-covered data steward to remove eighteen different direct and indirect identifiers relating to a patient (or the patient's relatives, employers, or household members) from the information in order for the information to be considered de-identified. *See* 45 C.F.R. § 164.514(b)(2)(i)(A)–(R) (2020) (listing the eighteen identifiers). As discussed at *supra* notes 64 and 66 and accompanying text, however, reidentification rates as high as 10.6 percent have been found with respect to data that has been de-identified in accordance with the HIPAA Safe Harbor.

[140] *See supra* note 111 and accompanying text.

[141] *See supra* note 135 and accompanying text.

[142] *See supra* notes 135–137 and accompanying text.

[143] The Appendix collects, and this Article originally synthesizes, a wide range of illustrative, current and pending, federal and state authorities that expressly or potentially protect the confidentiality and security of health data as well as the privacy of individuals who are the subjects of such data. Illustrative, not exhaustive, examples of health data protected by these authorities include general medical record data, *see, e.g.*, Appendix at Montana-2, Texas-2, HIPAA, infectious disease data, *see, e.g.*, Appendix at ENPA, PHEPA, health, fitness, or kinesthetic data, *see, e.g.*, Appendix at SMARTWATCH Data Act, genetic data, *see, e.g.*, Appendix at District of Columbia, Maryland, CCDPA, biometric data, *see, e.g.*, Appendix at DCA, HIPAA, PHEPA, and, importantly, general consumer data from which an individual's health data can be inferred, *see, e.g.*, Appendix at California-2, CalOPPA, MYOBA. This Article: (1) pinpoints the identification and de-identification standards within these illustrative authorities; (2) assesses the standards' strengths and weaknesses in light of the data reidentification literature; (3) shows how current and even pending standards for data identification and de-identification are insufficient to protect against reidentification; and (4)

approaches to health data identification—through legal protection—and health data de-identification—through deregulation. Most, if not all, of these standards insufficiently protect against reidentification.

A. Identification Standards

Most data protection laws begin by identifying the individuals and institutions that must comply with the law (sometimes called "covered entities"[144]) and the class or classes of data that are protected by the law (sometimes called "personal information," "sensitive personally identifying information," or "protected health information").[145] With respect to the protected classes of data, most data protection laws require information to relate to a particular individual—to be individually identifiable before the data will receive legal protection.[146] Stated another way, how a data protection law clarifies which data relate to a particular individual will affect whether the data receive

proposes illustrative changes to these standards. Beyond the focus of this Article are: (1) federal and state authorities that focus solely on non-health-related data where health data cannot be inferred, such as financial data, bank data, education data, tax or payroll data, criminal data, property data, telephone record data, or utility services data; and (2) federal and state professional and institutional certification and licensing authorities applicable to a wide range of health industry participants that may be preempted by the HIPAA Privacy Rule because they are not as stringent as the HIPAA Privacy Rule and/or that may not contain particular identification or de-identification standards.

[144] *See, e.g.*, 45 C.F.R. §§ 164.102(a), 160.103 (2020) (applying the HIPAA Privacy Rule to certain "covered entities," defined to include health plans, health-care clearinghouses, and those "health care provider[s] who transmit[] . . . health information in electronic form in connection with [standard] transaction[s]"); ALA. CODE § 8-38-2(2) (2021) (applying the Alabama Data Breach Notification Act to certain "covered entit[ies]," defined to include "[a] person, sole proprietorship, partnership, government entity, corporation, nonprofit, trust, estate, cooperative association, or other business entity that acquires or uses sensitive personally identifying information"). The question of who must comply with federal and state data protection laws was the focus of the Author's prior work. *See* Stacey A. Tovino, *Going Rogue: Mobile Research Applications and the Right to Privacy*, 95 NOTRE DAME L. REV. 155, 174–76 nn.108–23 (2019) (showing that many technology companies, citizen scientists, independent scientists, mobile health applications, and mobile research applications that collect, use, and/or disclose health data are not regulated by the HIPAA Privacy and Security Rules, although other data protection laws may apply).

[145] *See, e.g.*, FLA. STAT. § 501.171(1)(g) (2021) (applying protections to "personal information"); ALA. CODE § 8-38-2(6) (2021) (applying protections to "sensitive personally identifying information"); 45 C.F.R. § 164.500(a) (2020) (applying the protections set forth in the HIPAA Privacy Rule to "protected health information").

[146] *See, e.g.*, ARK. CODE ANN. § 4-110-103(5) (2021) (defining "[m]edical information" as "*individually* identifiable information" (emphasis added)); 45 C.F.R. § 160.103 (defining "[p]rotected health information," in relevant part, as "*individually* identifiable health information" (emphasis added)).

legal protection.[147] This Article refers to this clarification as a data identification standard.

Data protection laws incorporate a wide variety of data identification standards, including "reasonable basis to believe" standards, "capable (or reasonably capable) of being associated" standards, "linked (or reasonably linkable)" standards, "sufficient to perform identity theft" standards, "first name or first initial and last name" standards, "name, username, or email address plus password" standards, and "multiple data element" standards. Because a lack of identifiability, or initial legal protection, has the same end result as de-identification, or the loss of legal protection, identification standards must be assessed together with de-identification standards.

1. *"Reasonable Basis to Believe" Standards*. A number of data protection laws protect health data if there is a "reasonable basis to believe" that the data subject can be identified from the data. For example, the federal HIPAA Privacy Rule,[148] which regulates certain covered entities[149] and business associates[150] when using or disclosing protected health information ("PHI"),[151] contains a "reasonable basis

[147] These legal protections can take the form of a requirement for prior written authorization from the health data subject before the subject's information can be used or disclosed. *See, e.g.*, 45 C.F.R. § 164.508(a)(1), (b)(1)(i), (c)(1)(vi) (2020) (requiring a covered entity to obtain the prior written authorization of the individual who is the subject of the information (or the individual's legally authorized representative) before using the individual's information for certain activities). These legal protections can also take the form of administrative, physical, and technical safeguards (i.e., security standards) that must be implemented to protect the confidentiality, integrity, and availability of the health data. *See, e.g.*, *id*. §§ 164.308, 164.310, 164.312 (requiring covered entities to implement administrative, physical, and technical data safeguards). These legal protections also include breach notification standards; that is, patient notification and protection procedures that must be followed in the event of a privacy or security breach. *See, e.g.*, D.C. CODE § 28-3852 (2021) (requiring notification of Washington, D.C., residents whose personal information is part of a security system breach).

[148] The HIPAA Privacy Rule is codified at 45 C.F.R. Part 164, Subpart E. *See* 45 C.F.R. §§ 164.500–164.534 (2020).

[149] *See* 45 C.F.R. § 160.103 (defining "[c]overed entity" to include health plans, health care clearinghouses, and "health care provider[s] who transmit[] health information in electronic form in connection with a [standard] transaction"); *id.* § 160.102(a) (applying the HIPAA Rules to covered entities).

[150] *See id*. § 160.103 (defining business associate); *id.* § 160.102(b) (applying the HIPAA Rules to business associates).

[151] *Id.* § 160.103 (defining protected health information ("PHI")); *id.* (defining "[i]ndividually identifiable health information" as "a subset of health information" that "[i]s created or received by a health care provider, health plan, employer, or health care clearinghouse" and that "[r]elates to the past, present, or future physical or mental health or condition of an individual; the provision of health care to an individual; or the past, present, or

to believe" identification standard. Among other avenues, information is PHI under the HIPAA Privacy Rule if there is a "reasonable basis to believe" the information can be used to identify the individual.[152] The federal Protecting Personal Health Data Act ("PPHDA") similarly protects information "with respect to which there is a reasonable basis to believe that the information can be used to identify the individual."[153] Some states that extended the HIPAA Privacy Rule's protections to non-HIPAA covered entities[154] also follow the "reasonable basis to believe" standard. For example, the Texas Medical Records Privacy Act ("TMRPA") offers legal protections to "protected health information," internally referencing the HIPAA Privacy Rule.[155]

Of note, neither the HIPAA Privacy Rule, the PPHDA, nor the TMRPA specifies the identifiers that must be present before a

future payment for the provision of health care to an individual" and that either (i) "identifies the individual" or (ii) "[w]ith respect to which there is a reasonable basis to believe the information can be used to identify the individual"); *id.* (listing the four exclusions from the definition of PHI).

[152] *See id.* (defining individually identifiable health information); *id.* § 164.514(a) ("Health information that does not identify an individual and with respect to which there is no reasonable basis to believe that the information can be used to identify an individual is not individually identifiable health information."); Appendix at HIPAA.

[153] Protecting Personal Health Data Act, S. 24, 117th Cong. § 3(5) (2021); Appendix at PPHDA. Senator Amy Klobuchar introduced the PPHDA on June 13, 2019. *S.24 - Protecting Personal Health Data Act*, CONGRESS.GOV, https://www.congress.gov/bill/117th-congress/senate-bill/24 [https://perma.cc/UGM2-Q3VH]. The PPHDA would direct the Secretary of the U.S. Department of Health and Human Services to promulgate regulations that would "strengthen privacy and security protections for . . . personal health data that [are] collected, processed, analyzed, or used by consumer devices, services, applications, and software." S. 24, § 4; Appendix at PPHDA. The PPHDA did not become law.

[154] Although the HIPAA Administrative Simplification provisions and the privacy and security regulations adopted thereunder regulate covered entities and business associates thereof, these provisions do not constrain non-covered entities and non-business associates, including many technology companies, online service providers, mobile health applications, mobile research applications, and other individuals and institutions that collect, create, use, disclose, and redisclose health data. *See* Tovino, *supra* note 144, at 174–76 nn.108–23 (showing that many technology companies, citizen scientists, independent scientists, mobile health applications, and mobile research applications that collect, use, and/or disclose health data are not regulated by the HIPAA Privacy and Security Rules). As a result, some states have adopted HIPAA-like provisions, but have applied these provisions to a broader range of covered entities, including any person who "comes into possession of protected health information" as well as any person who "obtains or stores protected health information." *See* TEX. HEALTH & SAFETY CODE ANN. § 181.001(b)(2)(B)–(C) (West 2021). The Texas Medical Records Privacy Act is one example of such a state law. *Id*.

[155] HEALTH & SAFETY § 181.001(a) ("Unless otherwise defined in this chapter, each term that is used in this chapter has the meaning assigned by the [HIPAA Privacy Rule]"); Appendix at Texas-2.

"reasonable basis to believe" exists.[156] Consequently, a psychiatric hospital administrator, urgent care receptionist, or other data custodian not formally trained in statistics or biomedical informatics might release data like a patient's standardized diagnostic code without HIPAA- or state law-compliant[157] patient authorization—thinking there is no reasonable basis to believe the data are identifiable. As shown by the Vanderbilt researchers, however, a diagnostic code can be paired with external information to identify the subject of the diagnostic code.[158]

Lawmakers need to rethink the use of "reasonable basis to believe" identification standards, including those set forth in the HIPAA Privacy Rule, the PPHDA, and the TMRPA. Why? Data custodians might think that no reasonable basis exists when the opposite is true. Not all data custodians are familiar with the reidentification literature. Some sophisticated data custodians, including some university researchers, may be familiar with this literature but other data custodians, such as young clerks in small doctors' offices, may lack familiarity. The clerk might wrongly assume that health data that does not include a name could not be reidentified. Not all data custodians' beliefs regarding health data identification or reidentification are reasonable or scientifically supportable.[159]

2. *"Capable (or Reasonably Capable) of Being Associated" and "Linked (or Reasonably Linkable)" Standards.* Other data protection laws protect data elements that are "capable of being associated with a particular individual." For example, a Wisconsin law provides security protections to "[p]ersonally identifiable data about an individual's

[156] 45 C.F.R. § 160.103 (2021) (defining "individually identifiable health information" and PHI without reference to particular identifiers); *see infra* Part II.B.2 (discussing the identifiers that must be removed in order for protected health information to be considered de-identified under the HIPAA Safe Harbor).

[157] The HIPAA Privacy Rule generally preempts contrary state laws. 45 C.F.R. § 160.203 (2021). However, state laws that are more stringent than the HIPAA Privacy Rule survive preemption. *Id.* § 160.203(b). Depending on the results of a HIPAA-preemption analysis (i.e., a comparison of the HIPAA Privacy Rule and the contrary state law), either the HIPAA Privacy Rule or the state law must be followed. *See generally id.* §§ 160.201–160.205 (2021) (setting forth the HIPAA Privacy preemption provisions).

[158] *See supra* notes 98–111 and accompanying text.

[159] *See generally* NCVHS Letter, *supra* note 19, at 9 ("The lessons from de-identification research are not informing day-to-day practice. Practitioners responsible for de-identification and assessing risk of re-identification in non-research settings are often not adequately trained to apply critically the latest methods and research findings.").

medical condition."[160] The Wisconsin law defines "personally identifiable" as "capable of being associated with a particular individual through one or more identifiers or other information or circumstances."[161] The California Consumer Privacy Act of 2018 ("CCPA") adds an element of reasonableness to the standard set forth in the Second Wisconsin Act. That is, the CCPA protects information that "is reasonably capable of being associated with, or could reasonably be linked, directly or indirectly, with a particular consumer or household."[162] Virginia and Colorado have followed the CCPA's lead with new legislation. The Virginia Consumer Data Protection Act ("VCDPA") protects "information that is linked or reasonably linkable to an identified or identifiable natural person."[163] The Colorado Privacy Act ("CPA") also protects information that is "linked or reasonably linkable to an identified or identifiable individual."[164]

Recent federal bills contain similar standards. For example, one portion of the federal Data Care Act ("DCA") of 2021[165] would establish duties of care, loyalty, and confidentiality for online service providers that handle "individual identifying data."[166] The DCA defines "individual identifying data" with reference to data that are

[160] WIS. STAT. § 134.97(1)(e)(1) (2021); Appendix at Wisconsin-2.

[161] WIS. STAT. § 134.97(1)(f) (2021).

[162] CAL. CIV. CODE § 1798.140(v)(1) (West 2023) (effective Jan. 1, 2023); Appendix at California-2.

[163] VA. CODE ANN. § 59.1-571 (2021) (effective Jan. 1, 2023) (defining "personal data"); Appendix at Virginia-3. The VCDPA was signed into law on March 2, 2021, and takes effect on January 1, 2023. 2021 VA. ACTS 2307, https://lis.virginia.gov/cgi-bin/legp604.exe?211+ful+SB1392ES1+pdf [https://perma.cc/36QP-VZKL] (providing an effective date of January 1, 2023, in Section 3 of the bill); Cat Zakrzewski, *Virginia Governor Signs Nation's Second State Consumer Privacy Bill*, WASH. POST (Mar. 2, 2021, 8:17 PM), https://www.washingtonpost.com/technology/2021/03/02/privacy-tech-data-virgina/ [https://perma.cc/T6LA-SLRW] (stating that then Governor Ralph Northam signed the VCDPA into law on Tuesday, March 2, 2021).

[164] 2021 Colo. Sess. Laws 3445, 3448 (to be codified at COLO. REV. STAT. § 6-1-1303(17)(a)) (effective July 1, 2023); Appendix at Colorado-3. The CPA was signed into law on July 7, 2021, and takes effect on July 1, 2023. . *See* S.B. 21-190, 73d Gen. Assemb., 1st Sess. § 7 (Colo. 2021), https://leg.colorado.gov/sites/default/files/2021a_190_signed.pdf [https://perma.cc/F5EK-PBD2] (listing the general effective date in Section 7 of the bill and showing the date of the Governor's signature on the last page of the bill).

[165] *See generally* Data Care Act of 2021, S. 919, 117th Cong. (2021); Appendix at DCA. The DCA was introduced by Senator Brian Schatz in 2021. *S.919 - Data Care Act of 2021*, CONGRESS.GOV, https://www.congress.gov/bill/117th-congress/senate-bill/919 [https://perma.cc/Y95Y-MNSX].

[166] S. 919 § 3(b)(1)–(3) (establishing the duties of care, loyalty, and confidentiality); Appendix at DCA.

"linked, or reasonably linkable, to" certain persons and devices.[167] Similarly, the Mind Your Own Business Act ("MYOBA"), introduced by Senator Ron Wyden (D-OR) in 2019,[168] would require the Federal Trade Commission to promulgate regulations obligating certain entities to "implement reasonable cyber security and privacy policies, practices, and procedures to protect personal information."[169] MYOBA defines "personal information" as information "that is reasonably linkable to a specific consumer or consumer device."[170]

Legislation specific to the COVID-19 pandemic also includes the "linked or reasonably linkable" identification standard. For example, the COVID-19 Consumer Data Protection Act of 2020 ("CCDPA") protects, among other types of data, "persistent identifier[s]" and "personal health information."[171] The CCDPA defines these terms to include identifiers and information that are "linked or reasonably linkable to an individual."[172] Another example, the Public Health Emergency Privacy Act ("PHEPA") establishes certain privacy and security protections for "emergency health data," defined in relevant part as "data linked or reasonably linkable to an individual or device."[173] A final example is the Exposure Notification Privacy Act ("ENPA").[174] The ENPA imposes certain data privacy and security

[167] S. 919 § 2(3)(B); Appendix at DCA.

[168] Mind Your Own Business Act of 2019, S. 2637, 116th Cong. § 7 (2019)(stating "Mr. Wyden introduced the following bill; which was read twice and referred to the Committee on Finance"); *Wyden Introduces Mind Your Own Business Act of 2019*, COVINGTON: INSIDE PRIVACY (Oct. 21, 2019), https://www.insideprivacy.com/data-privacy/wyden-introduces-mind-your-own-business-act-of-2019/ [https://perma.cc/4365-8HLP] ("On October 17, Senator Ron Wyden introduced in the Senate a privacy bill that would expand the FTC's authority to regulate data collection and use, allow consumers to opt out of data sharing, and create civil and criminal penalties for certain violations of the Act.").

[169] S. 2637 § 7; Appendix at MYOBA.

[170] S. 2637 § 2(12); Appendix at MYOBA.

[171] COVID-19 Consumer Data Protection Act of 2020, S. 3663, 116th Cong. § 2(6)(A), (13)–(14) (2020); Appendix at CCDPA. The CCDPA was introduced by Senator Roger Wicker on May 7, 2020. *S.3663 - COVID–19 Consumer Data Protection Act of 2020*, CONGRESS.GOV, https://www.congress.gov/bill/116th-congress/senate-bill/3663 [https://perma.cc/K686-9W44].

[172] S. 3663 § 2(13)–(14); Appendix at CCDPA.

[173] Public Health Emergency Privacy Act, S. 81, 117th Cong. § 2(8) (2021); Appendix at PHEPA. The PHEPA was re-introduced by Senator Richard Blumenthal on January 28, 2021. *See generally* S. 81 (stating on the first page, "IN THE SENATE OF THE UNITED STATES; January 28, 2021; Mr. Blumenthal . . . introduced the following bill; which was read twice and referred to the Committee on Health, Education, Labor, and Pensions").

[174] ENPA was introduced by Senator Maria Cantwell on June 1, 2020. Exposure Notification Privacy Act, S. 3861, 116th Cong. (2020), (stating on the first page, "IN THE SENATE OF THE UNITED STATES; June 1, 2020; Ms. Cantwell . . . introduced the following

standards on operators of automated infectious disease exposure notification services with respect to "covered data."[175] "Covered data" is defined as information that is "linked or reasonably linkable" to individuals and certain devices.[176]

"Capable (or reasonably capable) of being associated" and "linked (or reasonably linkable)" standards such as those used in the above statutes are insufficient to protect patient privacy and health information confidentiality. Why? Because health data custodians who are not familiar with the reidentification literature presented in Part I might think that a data element is not capable of being associated with a particular individual (or is not linkable to a particular individual) when quite the opposite is true. Stated another way, some data custodians' beliefs regarding data association and data linkage may be unfounded and/or unsupported.

3. *"Sufficient to Perform or Attempt to Perform Identity Theft" Standards.* Other data protection laws apply to information that is "sufficient to perform or attempt to perform identity theft."[177] For example, the Georgia Personal Identity Protection Act ("Georgia Act") protects information that "would be sufficient to perform or attempt to perform identity theft."[178] Similarly, the Maine Notice of Risk to Personal Data Act ("Maine Act") protects information that "would be sufficient to permit a person to fraudulently assume or attempt to assume the identity of the person."[179] The Oregon Identity Theft Protection Act ("Oregon Act") applies when "[t]he data element or combination of data elements would enable a person to commit identity theft."[180] The Washington Notice of Security Breaches Act ("Washington Act") also applies when "[t]he data element or combination of data elements would enable a person to commit identity theft."[181]

bill; which was read twice and referred to the Committee on Commerce, Science, and Transportation").

[175] *Id.* §§ 3–7 (imposing standards on operators of automated exposure notification services).

[176] Exposure Notification Privacy Act, S. 3861, 116th Cong. § 2(6) (2020); Appendix at ENPA.

[177] GA. CODE ANN. § 10-1-911(6)(E) (2021); Appendix at Georgia-1.

[178] GA. CODE ANN. § 10-1-911(6)(E).

[179] ME. REV. STAT. ANN. tit. 10, § 1347(6)(E) (2021); Appendix at Maine.

[180] OR. REV. STAT. ANN. § 646A.602(12)(a)(C)(ii) (West 2021); Appendix at Oregon.

[181] WASH. REV. CODE ANN. § 19.255.005(2)(a)(iii)(B) (West 2021); Appendix at Washington.

"Sufficient to perform or attempt to perform identity theft" identification standards suffer from the same weakness as "reasonable basis to believe" and "reasonably capable of being associated with." Unsophisticated data custodians might think that data are insufficient to perform or attempt to perform identity theft when the data may easily be reidentified, thus leading to identity theft. These data custodians may not know or may underestimate the ability of a data element to be paired with external data and reidentified.

4. *"First Name or First Initial and Last Name" Standards.* A number of data protection laws protect health data when the data contain the first name or first initial and last name of the data subject. These laws may be referred to as "first name or first initial and last name" laws. The Alabama Data Breach Notification Act ("Alabama Act"), for example, only protects "sensitive personally identifying information," defined as "an Alabama resident's first name or first initial and last name in combination with" certain other classes of information, including "[a]ny information regarding an individual's medical history, mental or physical condition, or medical treatment or diagnosis by a health care professional."[182] The Alabama Act requires breach notification—a form of legal protection for a data subject[183]—when a breach occurs and the breached data include the subject's "first name or first initial and last name" in combination with other information.[184] Breach notification is unavailable, however, when a breach occurs and the breached data do not contain the subject's first name or first initial and last name.[185]

Similar "first name or first initial and last name" standards are set forth in the laws of Alaska,[186] Arkansas,[187] Connecticut,[188] Delaware,[189]

[182] ALA. CODE § 8-38-2(6)(a)(4) (2021); Appendix at Alabama.

[183] *Security Breach Notification Laws*, NAT'L CONF. OF STATE LEGIS. (Apr. 15, 2021), https://www.ncsl.org/research/telecommunications-and-information-technology/security-breach-notification-laws.aspx [https://perma.cc/S2R8-QFSK] (providing citations to state laws that offer legal protections afforded to individuals whose data is breached and briefly summarizing these laws).

[184] *See* ALA. CODE §§ 8-38-2(6)(a)(4), 8-38-5(a) (2021).

[185] *Id.*

[186] ALASKA STAT. § 45.48.090(7)(A)(i)–(ii) (2021); Appendix at Alaska.

[187] ARK. CODE ANN. § 4-110-103(7) (2021); Appendix at Arkansas.

[188] CONN. GEN. STAT. § 36a-701b(a) (2021); Appendix at Connecticut-1.

[189] DEL. CODE ANN. tit. 6, §§ 12B-100, 12B-101(7)(a) (2021); Appendix at Delaware-1.

Hawaii,[190]Idaho,[191] Kansas,[192] Kentucky,[193] Louisiana,[194] Massachusetts[195] Michigan,[196] Minnesota,[197] Mississippi,[198] Missouri,[199] Montana,[200] Nevada,[201] New Hampshire,[202] New Jersey,[203] New Mexico,[204] North Carolina,[205] North Dakota,[206] Ohio,[207] Oklahoma,[208] Pennsylvania,[209] Rhode Island,[210] South Carolina,[211] Tennessee,[212] Utah,[213] Vermont,[214] Virginia,[215] West Virginia,[216] Wisconsin,[217] and Wyoming.[218] These data protection laws all require (and sometimes only require) the presence of the data subject's first name or first initial and last name before legal protections attach.

"First name or first initial and last name" identification standards present a higher than necessary risk of data reidentification and should not be used in data protection laws. One reason is because not one of the datasets described in Part I of this Article contained the patients' first names or first initials and last names. That is, not one of those

[190] HAW. REV. STAT. § 487N-1 (2021); Appendix at Hawaii-1; HAW. REV. STAT. § 487R-1 (2020); Appendix at Hawaii-2.

[191] IDAHO CODE § 28-51-104(5) (2021); Appendix at Idaho.

[192] KAN. STAT. ANN. § 50-7a01(g) (West 2021); Appendix at Kansas.

[193] KY. REV. STAT. ANN. § 365.732(1)(c) (West 2021); Appendix at Kentucky-2.

[194] LA. STAT. ANN. § 51:3073(4)(a) (2020); Appendix at Louisiana.

[195] MASS. GEN. LAWS ANN. ch. 93H, § 1(a) (West 2021); Appendix at Massachusetts.

[196] MICH. COMP. LAWS § 445.63(r) (2021); Appendix at Michigan.

[197] MINN. STAT. § 325E.61(1)(e) (West 2021); Appendix at Minnesota.

[198] MISS. CODE ANN. § 75-24-29(2)(b) (2021); Appendix at Mississippi.

[199] MO. ANN. STAT. § 407.1500(1)(9) (West 2021); Appendix at Missouri.

[200] MONT. CODE ANN. § 30-14-1704(4)(b)(i) (West 2021); Appendix at Montana-2.

[201] NEV. REV. STAT. ANN. § 603A.040(1) (LexisNexis 2021); Appendix at Nevada-1.

[202] N.H. REV. STAT. ANN. § 359-C:19(IV)(a) (2020); Appendix at New Hampshire.

[203] N.J. STAT. ANN. § 56:8-161 (West 2021); Appendix at New Jersey.

[204] N.M. STAT. ANN. § 57-12C-2(C) (2021); Appendix at New Mexico.

[205] N.C. GEN. STAT. § 75-61(10) (2021); Appendix at North Carolina.

[206] N.D. CENT. CODE § 51-30-01(4)(a) (2021); Appendix at North Dakota.

[207] OHIO REV. CODE ANN. § 1349.19(A)(7)(a) (LexisNexis 2021); Appendix at Ohio.

[208] OKLA. STAT. tit. 24, § 162(6) (2021); Appendix at Oklahoma.

[209] 2005 Pa. Laws 474; Appendix at Pennsylvania.

[210] 11 R.I. GEN. LAWS § 11-49.3-3(a)(8) (2021); Appendix at Rhode Island.

[211] S.C. CODE ANN. § 39-1-90(D)(3) (2020); Appendix at South Carolina.

[212] TENN. CODE ANN. § 47-18-2107(a)(4)(A) (2021); Appendix at Tennessee-1.

[213] UTAH CODE ANN. § 13-44-102(4)(a) (West 2021); Appendix at Utah-1.

[214] VT. STAT. ANN. tit. 9, § 2430(10)(A) (2020); Appendix at Vermont-2.

[215] VA. CODE ANN. § 18.2-186.6(A) (2021); Appendix at Virginia-1.

[216] W. VA. CODE § 46A-2A-101(6) (2021); Appendix at West Virginia.

[217] WIS. STAT. § 134.98(1)(b) (2021); Appendix at Wisconsin-1.

[218] WYO. STAT. ANN. § 40-12-501(a)(vii) (2021); Appendix at Wyoming.

datasets would be protected by the data protection laws listed in the preceding paragraph. Yet technology companies, research teams, news reporters, and lay community members successfully reidentified the nameless data based on the presence of less direct identifiers, such as a patient's geolocation,[219] the date of a patient's adverse drug event,[220] and a patient's diagnostic code.[221] "First name or first initial and last name" identification standards are vulnerable to reidentification when certain other non-name data elements are present.

5. *"Name, Username, or Email Address Plus Password" Standards.* Other data protection laws require the presence of either the first name or first initial and last name of the data subject *or* the subject's username or email address in combination with a password or security question and answer. These laws may be referred to as "name, username, or email address plus password" laws. These laws recognize that a person's identity (e.g., Stacey Tovino) may be determined from the person's username (e.g., STovino) or email address (e.g., Stacey.Tovino@ou.edu).

To illustrate, an Arizona law defines "personal information" as either "[a]n individual's first name or first initial and last name in combination with one or more specified data elements [or an] individual's user name or email address, in combination with a password or security question and answer, that allows access to an online account."[222] Other state laws containing the same "name, username, or email address plus password" standard include those in Florida,[223] Illinois,[224] Maryland,[225] Nebraska,[226] and California.[227]

[219] *See supra* notes 41–53 and accompanying text.

[220] *See supra* notes 89–95 and accompanying text.

[221] *See supra* notes 98–111 and accompanying text.

[222] ARIZ. REV. STAT. ANN. § 18-551(7)(a) (2021); Appendix at Arizona. The Arizona Act expressly applies to health information because the Act defines "[s]pecified data element" to include, among other items, "[i]nformation about an individual's medical or mental health treatment or diagnosis by a health care professional," and "[u]nique biometric data generated from a measurement or analysis of human body characteristics to authenticate an individual when the individual accesses an online account." *See* ARIZ. REV. STAT. ANN. § 18-551(11)(f), (i) (2021); Appendix at Arizona.

[223] FLA. STAT. § 501.171(1)(g)(1) (2021); Appendix at Florida.

[224] 815 ILL. COMP. STAT. 530/5 (2021); Appendix at Illinois.

[225] MD. CODE ANN., COM. LAW § 14-3501(e)(1) (West 2021); Appendix at Maryland.

[226] NEB. REV. STAT. § 87-802(5) (2021); Appendix at Nebraska.

[227] CAL. CIV. CODE § 1798.81.5(d)(1)(A)–(B) (West 2021); Appendix at California-1.

This Article argues that "name, user name, or email address plus password" standards are unlikely to minimize the risk of reidentification. Not one of the datasets from Part I contained the patients' names, usernames, or email addresses plus passwords. Yet, technology companies, research teams, news reporters, and lay community members successfully reidentified the (user)name-less data using other, less direct, identifiers, such as apatient's geolocation,[228] the date of a patient's adverse drug event,[229] and a patient's diagnostic code.[230]

6. *"Multiple Data Element" Standards.* Perhaps recognizing the weaknesses of each of the identification standards discussed above, some federal and state authorities allow a broader range of non-name data elements to be considered for legal protection. The CCPA assists data custodians by identifying twelve different paragraphs' worth of identifiers that qualify for legal protection.[231] Illustrative identifiers include, but certainly are not limited to, medical information,[232] biometric information,[233] and geolocation data.[234] The Colorado Consumer Protection Act also protects multiple, specific data elements such as "a social security number; a personal identification number; a password; a pass code; an official state or government-issued driver's license or identification card number; a government passport number; biometric data . . . ; an employer, student, or military identification number; or a financial transaction device."[235] The California Shine the Light Act protects a very long list of data elements including, but not limited to, "[a]n individual's name and address," email address, "age or date of birth," names of children, email "or other addresses of children," number of children, "age or gender of children," height, weight, race, religion, occupation, telephone number, education, political party affiliation, medical condition, drugs, and "[i]nformation

[228] *See supra* notes 41–53 and accompanying text.

[229] *See supra* notes 89–95 and accompanying text.

[230] *See supra* notes 98–111 and accompanying text.

[231] CAL. CIV. CODE § 1798.140(v) (West 2021) (effective Jan. 1, 2023).

[232] *Id.* § 1798.140(v)(1)(B) (referencing a definition in CAL. CIV. CODE § 1798.80(e) (West 2021) that includes "medical information"); Appendix at California-2.

[233] CAL. CIV. CODE § 1798.140(v)(1)(E) (West 2021) (effective Jan. 1, 2023) (specifically listing "[b]iometric information"); Appendix at California-2.

[234] CAL. CIV. CODE § 1798.140(v)(1)(G) (West 2021) (effective Jan. 1, 2023) (specifically listing "[g]eolocation data"); Appendix at California-2.

[235] COLO. REV. STAT. § 6-1-713(2)(b) (2021); Appendix at Colorado-1.

pertaining to creditworthiness, assets, income, or liabilities."[236] State laws in Connecticut,[237] the District of Columbia,[238] Kentucky,[239] Nevada,[240] and Vermont[241] also contain "multiple data element" standards. The federal PHEPA also contains a "multiple data element" identification standard.[242] That is, the PHEPA protects geolocation data, proximity data, "genetic data, biological samples, . . . biometric[]" data, demographic data, contact data, and "any other data collected from a personal device."[243]

One benefit of multiple data element identification standards is that an unsophisticated data custodian can review the data for one or more enumerated elements and know the data are legally protected when one element is discovered. A second benefit—at least from the perspective of those who value the individual right to privacy over research and activities that benefit the wider population—is that identification standards containing high numbers of data elements will increase the amount of data receiving legal protection.

However, multiple data element identification standards that include high numbers of protected identifiers do run the risk of reducing or eliminating the usefulness of unregulated data for health-related records research, public health initiatives, informed health-care decision-making, and other public benefit activities.[244] In addition, "multiple data element" identification standards require lawmakers to make difficult decisions regarding which data elements to include in the standard, with consequences that they may not understand.

As background for this last point, recall the Kaiser Permanente researchers whose work was highlighted in Part I.B of this Article. These researchers found that the risk of health data reidentification depends on three factors: (1) the availability of public, semi-public, or private external data containing one or more elements that overlap with the health data; (2) "the size of the class defined by those

[236] CAL. CIV. CODE § 1798.83(e)(7) (West 2021); Appendix at California-3.

[237] *See* CONN. GEN. STAT. § 42-471(c)(1) (2021); Appendix at Connecticut-2.

[238] *See* D.C. CODE § 28-3851(3)(A) (2021); Appendix at District of Columbia.

[239] *See* KY. REV. STAT. ANN. § 365.720(4) (West 2021); Appendix at Kentucky-1.

[240] *See* NEV. REV. STAt. ANN. § 603A.320 (LexisNexis 2021); Appendix at Nevada-2.

[241] *See* VT. STAT. ANN. tit. 9, § 2445(a)(3) (2021); Appendix at Vermont-1.

[242] *See* Public Health Emergency Privacy Act, S. 81, 117th Cong. § 2(8) (2021); Appendix at PHEPA.

[243] S. 81; Appendix at PHEPA.

[244] *See, e.g.*, *supra* notes 111, 137 and accompanying text (explaining that too-strong data perturbation techniques can minimize the clinical and other usefulness of health data).

overlapping elements in which the [health data] falls;" and (3) "the overlap in the population covered by the [health data] and the population covered by the external data."[245] The difficulty with applying this three-factor framework in the context of health data is that there is not just one type of health data.

Health data comes in many forms, including general medical record data, hospital discharge data, health and fitness data, infectious disease data, suicide risk data, genetic data, biometric data, prescription data, and even general consumer data. Each type's reidentification risk varies depending on the availability of external data containing elements common to that type. A multiple data element standard that includes one data element, such as a patient's diagnostic code, a patient's age in months, or a patient's geolocation, may reduce the risk of reidentifying some forms of health data but not others. Thus, a lawmaker who drafts a data protection law containing a multiple data element standard must (1) consider all types of health data that are created, collected, used, and/or disclosed in the jurisdiction; (2) decide which classes of health data are worthy of or are in need of legal protection in that jurisdiction; and (3) consider which data elements will reduce the risk of reidentification of each such class of data. This is not an easy task.

In summary, multiple data element identification standards offer the benefit of clarity and ease of application for unsophisticated data custodians. Multiple data element identification standards, to the extent they include high numbers and types of data elements, also increase the amount of data receiving legal protection. However, long multiple data element standards will reduce or eliminate the clinical, research, and other public utility of unprotected health data. Moreover, lawmakers will struggle to draft perfect multiple data element standards because including particular data elements will affect the reidentification risk of some forms of health data but not others.

B. De-Identification Standards

Because de-identification, or the loss of legal protection, has the same end result as a lack of identification, or a lack of initial legal protection, de-identification standards also must be assessed to understand the scope of data protection laws. Data protection laws

[245] *See supra* notes 114–116 and accompanying text.

incorporate a wide range of de-identification standards, including "truncated, modified, or redacted" ("TMR") standards, "laundry list" safe harbors, "expert determination" standards, "public commitment plus contractual obligation" standards, and "delegation to administrative agency" standards.

Understanding the application of these de-identification standards is important. In the context of a patient privacy or health information confidentiality law, losing legal protection may mean that a data custodian is permitted to use or disclose a patient's data without the patient's prior written authorization.[246] The patient will lose the opportunity to make an autonomous decision about the use and disclosure of sensitive, and perhaps stigmatizing, health information. In the context of a breach notification law, losing legal protection may mean that a patient will not be notified when a privacy or security breach involves their data.[247] The patient will lose the ability to implement a credit freeze or otherwise respond proactively to the privacy or security breach.[248]

1. *"Truncated, Modified, or Redacted" Standards.* Many data protection laws contain very general "truncated" or "modified" de-identification standards, where data lose legal protection when truncated or modified. For example, an Alabama law does not apply if the data are "truncated . . . or modified by any other method or

[246] *See, e.g.*, Press Release, U.S. Department of Health and Human Services, Unauthorized Filming for "NY Med" Results in $2.2 Million Settlement with New York Presbyterian Hospital (Apr. 21, 2016), https://wayback.archive-it.org/3926/20170128230744/https://www.hhs.gov/about/news/2016/04/21/unauthorized-filming-ny-med-results-22-million-settlement-new-york-presbyterian-hospital.html [https://perma.cc/F9AD-Z3HD] (imposing a $2.2 million resolution agreement amount on New York Presbyterian Hospital, which had allowed ABC film crews to enter its hospital and film patients receiving care as part of a medical documentary series without obtaining the prior written authorization of the patients filmed).

[247] *See, e.g.*, Press Release, U.S. Department of Health and Human Services, First HIPAA Enforcement Action for Lack of Timely Breach Notification Settles for $475,000 (Jan. 9, 2017), http://wayback.archive-it.org/3926/20170127111957/https://www.hhs.gov/about/news/2017/01/09/first-hipaa-enforcement-action-lack-timely-breach-notification-settles-475000.html [https://perma.cc/NE66-SRVF] (imposing a $475,000 resolution agreement amount on Presence Health, which failed to notify 836 individuals that their information had been breached, and explaining that "[i]ndividuals need prompt notice of a breach of their unsecured PHI so they can take action that could help mitigate any potential harm caused by the breach").

[248] *See* Stacey A. Tovino, *Health Privacy, Security, and Information Management*, in LAWS OF MEDICINE: CORE LEGAL ASPECTS FOR THE HEALTH CARE PROFESSIONAL (Springer Nature Switzerland forthcoming 2022) (manuscript at 9–10) (on file with author) (discussing cases in which patients have suffered privacy and security breaches as well as the legal consequences of those breaches).

technology that removes elements that personally identify an individual."[249] By further example, a Florida law does not apply to "information that is . . . modified by any other method or technology that removes elements that personally identify an individual or that otherwise renders the information unusable."[250]

Other states use "redacted." Redacted information is not protected under laws in Alaska,[251] Arkansas,[252] California,[253] Colorado,[254] Georgia,[255] Illinois,[256] Kentucky,[257] Louisiana,[258] Maryland,[259] New Mexico,[260] South Carolina,[261] or Wisconsin.[262] Not one of these laws defines "redacted." Thus, unsophisticated data custodians might think that they have redacted sufficient data when the remaining data are reidentifiable.

[249] ALA. CODE § 8-38-2(6)(b)(2) (2021); Appendix at Alabama.

[250] FLA. STAT. § 501.171(1)(g)(2) (2021); Appendix at Florida.

[251] *See* ALASKA STAT. §§ 45.48.010, 45.48.090(7) (2021) (defining protected "personal information" to exclude "encrypted or redacted" information); Appendix at Alaska.

[252] *See* ARK. CODE ANN. §§ 4-110-103(7), -104 to -105 (2021) (defining protected "personal information" to exclude "encrypted or redacted" data elements); Appendix at Arkansas.

[253] *See* CAL. CIV. CODE § 1798.81.5(a)–(d)(1)(A) (West 2021) (defining protected "personal information" to exclude "encrypted or redacted" information); Appendix at California-1.

[254] *See* COLO. REV. STAT. § 6-1-716(1)(g), (2)(a) (2021) (defining protected "personal information" to exclude "encrypted, redacted, or secured" data elements); Appendix at Colorado-2.

[255] *See* GA. CODE ANN. §§ 10-1-911(6), -912(a)–(b) (2021) (defining protected "personal information" to exclude "encrypted or redacted" data elements); Appendix at Georgia-1.

[256] *See* 815 ILL. COMP. STAT. 530/5–530/10 (2020) (defining protected "personal information" to exclude "encrypted or redacted" data elements); Appendix at Illinois.

[257] *See* KY. REV. STAT. ANN. § 365.732(1)(c)–(3) (West 2021) (defining protected "personally identifiable information" to exclude redacted data elements); Appendix at Kentucky-2.

[258] *See* LA. STAT. ANN. §§ 51:3073(4)(a), :3074 (2021) (defining protected "personal information" to exclude "encrypted or redacted" data elements); Appendix at Louisiana.

[259] *See* MD. CODE. ANN., COM. LAW §§ 14-3501(e)(1)(i), -3502 (West 2021) (defining protected "personal information" to exclude "encrypted, redacted, or otherwise protected" data elements); Appendix at Maryland.

[260] *See* N.M. STAT. ANN. §§ 57-12C-2(C) to -6 (2021) (defining protected "personal identifying information" to exclude data elements "protected through encryption or redaction"); Appendix at New Mexico.

[261] *See* S.C. CODE ANN. § 39-1-90(A)–(D)(3) (2021) (defining protected "[p]ersonal identifying information" to exclude "encrypted [or] redacted" data elements); Appendix at South Carolina.

[262] *See* WIS. STAT. § 134.98(1)(b), (2) (2021) (defining protected "[p]ersonal information" to exclude "encrypted, redacted, or altered" data elements); Appendix at Wisconsin-1.

"Redacted" information also is not protected under laws in Arizona,[263] Hawaii,[264] Indiana,[265] Iowa,[266] and Virginia.[267] But these laws do define "redacted," "redact," or "redaction." Under the Virginia law, for example, "[r]edact" means the "alteration or truncation of data such that no information regarding an individual's medical history, mental or physical condition, or medical treatment or diagnosis, or no more than four digits of a health insurance policy number, subscriber number, or other unique identifier are accessible as part of the medical information."[268] Under the Arizona law, "'[r]edact' means to alter or truncate a number so that no more than the last four digits are accessible and at least two digits have been removed."[269] Under the Hawaii law, "'[r]edacted' means the rendering of data so that it is unreadable or is truncated so that no more than the last four digits of the identification number are accessible as part of the data."[270] "Redacted," "redact," or "redaction" are similarly (although not identically) defined in laws in Indiana,[271] Iowa,[272] Kansas,[273] Michigan,[274] Missouri,[275] Nebraska,[276] North Carolina,[277] Ohio,[278]

[263] *See* ARIZ. REV. STAT. ANN. §§ 18-551(1), -552 (2021) (protecting against "[b]reach[es]" or "security system breach[es]," defined to include only the compromising of "unencrypted and unredacted . . . personal information"); Appendix at Arizona.

[264] *See* HAW. REV. STAT. §§ 487N-1 to -2 (2021) (protecting against "[s]ecurity breach[es]," defined to include the unauthorized access of only "unencrypted or unredacted records or data containing personal information"); Appendix at Hawaii-1.

[265] *See* IND. CODE §§ 24-4.9-2-10 to -3-3.5 (2021) (defining protected "[p]ersonal information" to exclude "encrypted or redacted" data elements); Appendix at Indiana.

[266] *See* IOWA CODE §§ 715C.1(11), 715C.2 (2021) (defining protected "[p]ersonal information" to exclude "encrypted, redacted, or otherwise altered" data elements); Appendix at Iowa.

[267] *See* VA. CODE ANN. § 32.1-127.1:05(B) (2021) (protecting against breaches of "unencrypted or unredacted medical information"); Appendix at Virginia-2.

[268] VA. CODE ANN. § 32.1-127.1:05(A).

[269] ARIZ. REV. STAT. ANN. § 18-551(9) (2021); Appendix at Arizona.

[270] HAW. REV. STAT. § 487N-1 (2021); Appendix at Hawaii-1.

[271] IND. CODE § 24-4.9-2-11 (2021); Appendix at Indiana.

[272] IOWA CODE § 715C.1(12) (2021); Appendix at Iowa.

[273] KAN. STAT. ANN. § 50-7a01(d) (West 2021); Appendix at Kansas.

[274] MICH. COMP. LAWS § 445.63(t) (2020); Appendix at Michigan.

[275] MO. ANN. STAT. § 407.1500(1)(10) (West 2020); Appendix at Missouri.

[276] NEB. REV. STAT. ANN. § 87-802(6) (West 2021); Appendix at Nebraska.

[277] N.C. GEN. STAT. § 75-61(13) (2021); Appendix at North Carolina.

[278] OHIO REV. CODE ANN. § 1349.19(A)(9) (LexisNexis 2021); Appendix at Ohio.

Oklahoma,[279] Oregon,[280] Pennsylvania,[281] Vermont,[282] Virginia,[283] and West Virginia.[284]

This Article contends that lawmakers need to avoid the use of TMR de-identification standards in which the words truncated, modified, and redacted are undefined. Unsophisticated data custodians attempting to implement these standards might think that redacting a patient's first name, last name, and email address is sufficient even though the remaining information could easily be matched with available external data. TMR de-identification standards that include definitions would be more helpful to unsophisticated data custodians. Consider a medical records custodian who is instructed to remove date stamps, patients' ages in months, and the dates of patients' adverse health events. Complying with this instruction would seem easier than interpreting a vague instruction to "truncate all identifiable data elements."

However—and as with the multiple data element identification standard—a defined TMR standard that includes a high number of data elements will reduce or eliminate the clinical, research, and other utility of the health data. Lawmakers will struggle to draft perfectly defined TMR standards. Recall that removing particular data elements will affect the risk of reidentification of some forms of health data but not others.[285]

2. *"Laundry List" Safe Harbors.* Some data protection laws contain what may be referred to as "laundry list" de-identification safe harbors. For example, the HIPAA Privacy Rule identifies eighteen particular data elements that must be removed from information.[286] If these data elements are removed and the HIPAA-covered data custodian "does not have actual knowledge that the information could be used alone or in combination with other information to identify an individual who is a subject of the information," the information is considered de-identified under the HIPAA Safe Harbor and is no

[279] OKLA. STAT. tit. 24, § 162(8) (2021); Appendix at Oklahoma.

[280] OR. REV. STAT. ANN. § 646A.602(16) (West 2021); Appendix at Oregon.

[281] 2005 Pa. Laws 474; Appendix at Pennsylvania.

[282] VT. STAT. ANN. tit. 9, § 2430(12) (2021); Appendix at Vermont-2.

[283] VA. CODE ANN. § 18.2-186.6(A) (2021); Appendix at Virginia-1.

[284] W. VA. CODE § 46A-2A-101(8) (2021); Appendix at West Virginia.

[285] *See supra* Part I.B.

[286] 45 C.F.R. § 164.514(b)(2)(i) (2020); Appendix at HIPAA.

longer protected by the HIPAA Privacy Rule.[287] Some states incorporate by reference the HIPAA Safe Harbor. For example, the CPA provides that personal data is not regulated by the CPA if it is "[d]e-identified in accordance with the [Safe Harbor]."[288]

The eighteen identifiers that must be removed under the Safe Harbor include (1) names, (2) "[a]ll geographic subdivisions smaller than a [s]tate," (3) "[a]ll elements of dates," including years or dates of birth, (4) telephone numbers, (5) fax numbers, (6) email addresses, (7) social security numbers, (8) medical record numbers, (9) "[h]ealth plan beneficiary numbers," (10) account numbers, (11) certificate and license numbers, (12) vehicle identifiers and license plate numbers, (13) device identifiers, (14) "[u]niversal [r]esource [l]ocators URLs," (15) "[i]nternet [p]rotocol (IP) address numbers," (16) biometric identifiers, (17) full facial photographs and comparable images, and (18) "[a]ny other unique identifying number, characteristic, or code."[289]

One problem with the "actual knowledge" portion of the Safe Harbor is that it can insulate unsophisticated HIPAA covered entities. Consider an isolated doctor, dentist, psychologist, or other covered health-care professional who is not up-to-date on the reidentification literature and lacks actual knowledge that facially de-identified data can still be combined with external data to reidentify the data subject. Requiring actual knowledge may also shield sophisticated covered entities who may have, but do not wish to make public, their actual knowledge.

One problem with the "laundry list" portion of the HIPAA Safe Harbor is the eighteenth provision. It requires the removal of "[a]ny other unique identifying number, characteristic, or code."[290] Again, unsophisticated data custodians who are not familiar with the reidentification literature may think that a particular number, characteristic, or code (such as a standardized diagnostic code) is not uniquely identifying, when the code could easily be matched with external data for reidentification.

Moreover, the laundry list shares the strengths and limitations of the multiple data element identification standard. That is, the "laundry list" offers the benefit of clarity for unsophisticated data custodians and

[287] 45 C.F.R. § 164.514(a)–(b); Appendix at HIPAA.

[288] 2021 Colo. Sess. Laws 3445, 3451 (to be codified at COLO. REV. STAT. § 6-1-1304(1)(g)(I)) (effective July 1, 2023); Appendix at Colorado-3.

[289] 45 C.F.R. § 164.514(b)(2)(i); Appendix at HIPAA.

[290] 45 C.F.R. § 164.514(b)(2)(i)(R); Appendix at HIPAA.

will increase the amount of data receiving legal protection due to the long list of identifiers that must be removed. However, the laundry list that must be removed may also reduce or eliminate the data's clinical, research, and other public utility.[291] In addition, lawmakers who wish to adopt a similar list must carefully attend to the selection of the particular identifiers requiring removal and be aware of how particular elements will impact the risk of reidentification.[292]

Recall that the HIPAA Safe Harbor has been empirically tested. Harvard University researchers de-identified Maine hospital discharge data in accordance with the Safe Harbor.[293] Yet reidentification was still possible in 3.2 percent of cases.[294] De-identified Vermont hospital discharge data in accordance with the Safe Harbor yielded a reidentification rate of 10.6 percent.[295] The researchers' conclusions were startling but consistent with this Article's analyses: (1) "patients' personal information is vulnerable to re-identification even when hospital data is de-identified according to HIPAA Safe Harbor guidelines[;]"[296] (2) "[the] HIPAA Safe Harbor's framework is often considered the de facto standard for protecting patient privacy even though HIPAA has not been rigorously confirmed to guarantee privacy[;]"[297] (3) "[t]he de-identification checklist that HIPAA Safe Harbor promotes is the bare minimum protection against re-identification[;]"[298] (4) "[a] more rigorous inquiry on the vulnerabilities that exist even when following HIPAA Safe Harbor as a standard for de-identification" is needed;[299] (5) "[p]olicy-makers and data-sharing centers should consider scientifically tested protocols that guarantee

[291] *See generally* NCVHS Letter, *supra* note 19, at 8 ("[D]e-identification reduces the quality and utility of data, the consequence of which must be judged against the characteristics of the dataset and the intended uses."). The HIPAA Privacy Rule recognizes this point through its creation of an easier-to-satisfy "limited data set" ("LDS") provision for uses and disclosures of protected health information for "research, public health, [and] health care operations." *See* 45 C.F.R. § 164.514(e). Compared to information de-identified in accordance with the HIPAA Safe Harbor, an LDS contains a few additional identifiers. *See id.* § 164.514(e)(2) (listing only sixteen identifiers that must be excluded).

[292] *See generally* NCVHS Letter, *supra* note 19, at 12 (noting that the HIPAA "de-identification standard is too often executed with inadequate attention to the unique characteristics of the dataset to which it is applied and its intended uses").

[293] *See supra* notes 59–64 and accompanying text.

[294] Ji Su Yoo et al., *supra* note 58, at 3.

[295] *Id.*

[296] *Id.*

[297] *Id.* at 44.

[298] *Id.*

[299] *Id.* at 3.

privacy protections to patients, especially since they cannot opt out of inclusion in hospital records[;]"[300] and (6) "states should revisit de-identification practices and reassess risks to patient privacy when determining data sharing protocol."[301]

3. *"Expert Determination" Standard.* The HIPAA Privacy Rule also permits health information to be freely used and disclosed if an expert has determined that the information is de-identified ("Expert Determination").[302] In an Expert Determination, "[a] person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable" must "apply[] such principles and methods" and "determine[] that the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information."[303] In theory, an expert could determine that data elements required to be removed by the Safe Harbor could remain or that additional data elements not listed in the Safe Harbor must be removed.[304]

Expert Determinations have the benefit of overcoming the challenges posed by unsophisticated data custodians. An average data custodian may struggle to interpret or apply vague identification and de-identification standards, such as the "reasonable basis to believe" standard, the "reasonably linkable" standard, or an undefined TMR standard. However, the Expert Determination removes this responsibility from the unsophisticated data custodian and places it on the shoulders of an individual who knows and has experience with relevant statistical and scientific principles for rendering information not individually identifiable.

One limitation of the Expert Determination standard is that, as currently written, it does not specify the minimum education, training, experience, skills, or competencies necessary for an individual to

[300] *Id.* at 44.

[301] *Id.* at 2.

[302] *See* 45 C.F.R. § 164.514(b)(1) (2020) (setting forth requirements for de-identification of protected health information by an expert); NCVHS Letter, *supra* note 19, at 2 (summarizing the Expert Determination standard).

[303] 45 C.F.R. § 164.514(b)(1).

[304] Simon, *supra* note 20, at 3.

qualify as an expert.[305] Some covered entities may think a particular employee, contractor, or consultant qualifies when the individual is less knowledgeable than what the U.S. Department of Health and Human Services ("HHS") intended. Thus, HHS should amend the Expert Determination standard in the HIPAA Privacy Rule to specify minimum expert skills and competencies.

A second limitation is that regulated entities must have access to qualifying experts or the resources necessary to acquire such access. Academic medical centers likely have in-house statisticians or scientists who are or could become sufficiently knowledgeable in rendering information not individually identifiable. If not, these centers likely can afford to contract with qualifying experts. It is less likely that the hundreds of thousands of small physician offices and clinics across the United States have the time and resources to obtain a proper Expert Determination before using or disclosing health data.[306]

4. *"Public Commitment Plus Contractual Obligation" Standards.* Other de-identification standards require regulated entities, among other measures, to (1) publicly commit to maintaining the information in de-identified form and not reidentify the information and (2) impose downstream contractual obligations on information recipients, pursuant to which the recipients agree to adhere to the same data protection requirements that apply to the regulated entity. The CCPA,[307] CCDPA,[308] ENPA,[309] VCDPA,[310] CPA,[311] and the Information Transparency and Personal Data Control Act[312] follow this "public commitment plus contractual obligation" de-identification

[305] *See* 45 C.F.R. § 164.514(b)(1) (not specifying any particular education, training, experience, skills, or competencies); Appendix at HIPAA.

[306] *See generally* NCVHS Letter, *supra* note 19, at 8 (noting that the Expert Determination is "more expensive, and there are too few experts available for hire").

[307] CAL. CIV. CODE § 1798.140(m)(2)–(3) (West 2021) (effective Jan. 1, 2023); Appendix at California-2.

[308] COVID-19 Consumer Data Protection Act of 2020, S. 3663, 116th Cong. § 2(9)(C)–(D) (2020); Appendix at CCDPA.

[309] Exposure Notification Privacy Act, S. 3861, 116th Cong. § 2(2)(B), (D) (2020); Appendix at ENPA.

[310] VA. CODE ANN. § 59.1-575 (2021) (effective Jan. 1, 2023); Appendix at Virginia-3.

[311] 2021 Colo. Sess. Laws 3445, 3448 (to be codified at COLO. REV. STAT. § 6-1-1303(11)(b)–(c)) (effective July 1, 2023); Appendix at Colorado-3.

[312] Information Transparency & Personal Data Control Act, H.R. 1816, 117th Cong. § 7(6)(C)–(F) (2021); Appendix at ITPDCA.

standard. From a privacy standpoint, the public commitment portion of this standard would help companies like Google determine which patients have reidentification worries.[313] However, it is unclear how the downstream contractual obligation portion would be applied when the information recipient is a state agency that operates a database designed to foster health-care cost, utilization, and efficacy research. Would the state agency be responsible for ensuring that every researcher who accesses the database also agrees to the data protection requirements? If so, would there be an exception for reidentification research, like that done by the Harvard, Berkeley, and Vanderbilt researchers in Part I?[314] Moreover, privacy regimes that initially relied on downstream privacy contracting, such as the HIPAA Privacy Rule's business associate agreement provisions, ultimately moved to direct regulation when downstream contracting proved insufficient.[315] It is a real fear that downstream contractual obligations are an insufficient first step when direct regulation would be simpler and more effective. Finally, it is not clear the downstream contractual obligations will impact or otherwise deter non-researcher adversaries who illicitly obtain health data or try to reidentify health data for personal, professional, or financial gain.

5. *"Delegation to Administrative Agency" Standard.* Perhaps given the limitations of the de-identification standards discussed above, one

[313] *See, e.g.*, Dinerstein Lawsuit, *supra* note 9, at 24–30 (explaining how Google could reidentify the plaintiff's University of Chicago Medical Center visit data because he used Google products and services that tracked his geolocation).

[314] *See supra* notes 55–73, 98–111 and accompanying text. The CCPA has considered this point, permitting businesses to "attempt to reidentify the information solely for the purpose of determining whether its deidentification processes satisfy the requirements of [the CCPA]."]." CAL. CIV. CODE § 1798.140(m)(2) (West 2021) (effective Jan. 1, 2023); Appendix at California-2.

[315] Before President Obama signed the American Recovery and Reinvestment Act ("ARRA") into law in 2009, the HIPAA Privacy Rule directly (and only) regulated covered entities that included health plans, health care clearinghouses, and certain health care providers. Standards for Privacy of Individually Identifiable Health Information, 65 Fed. Reg. 82462, 82798 (Dec. 28, 2000) (to be codified at 45 C.F.R. § 160.102) (making the HIPAA Privacy Rule applicable only to health plans, health care clearinghouses, and certain health care providers). These covered entities were required to enter into downstream business associate agreements with their business associates ("BAs"), thus contractually obligating the BAs to maintain the confidentiality of the covered entities' protected health information. American Recovery and Reinvestment Act of 2009, Pub. L. No. 111-5, 123 Stat. 115 (2009) (codified as amended in scattered sections of the U.S. Code). Recognizing the insufficiency of downstream privacy contracting, the Health Information Technology for Economic and Clinical Health Act within the ARRA changed this scheme such that BAs became directly regulated by the HIPAA Privacy Rule. *See id.* § 13404(a), (c), 123 Stat. at 264 (applying the HIPAA Privacy Rule provisions and the violation penalties directly to BAs).

recent federal bill simply directs an administrative agency to promulgate de-identification regulations. The PPHDA would direct HHS to "consider appropriate standards for the de-identification of personal health data."[316] The PPHDA does not, however, contain further direction regarding permissible regulatory methods of de-identification. In addition, the PPHDA does not recognize the growing risk of reidentification other than one provision directing a task force to "study the long-term effectiveness of de-identification methodologies" in two limited contexts: "genetic data and biometric data."[317] Although genetic and biometric data are frequently noted for their privacy concerns,[318] this Article has intentionally highlighted reidentification success stories that do not involve genetic or biometric data to illustrate the many contexts where reidentification can occur. A final concern with PPHDA's approach is that federal de-identification regulations could take years for HHS to promulgate (if HHS even promulgates them),[319] during which the science of reidentification marches on.

III. PROPOSALS

Part II identified the weaknesses of current and pending standards for health data identification and de-identification. Is there a way that these weaknesses can be minimized? Is there a way to strengthen legal protections for identifiable as well as potentially reidentifiable health data? Six theoretical alternatives are offered below.

[316] Protecting Personal Health Data Act, S. 24, 117th Cong. § 4(b)(2)(F) (2021); Appendix at PPHDA.

[317] S. 24 § 5(b)(1); Appendix at PPHDA.

[318] *See, e.g.*, Nora von Thenen, Erman Ayday & A. Ercument Cicek, *Re-Identification of Individuals in Genomic Data-Sharing Beacons Via Allele Inference*, 35 BIOINFORMATICS 365, 365 (2019) ("We show that countermeasures such as hiding certain parts of the genome or setting a query budget for the user would fail to protect the privacy of the participants."); Zachary Shapiro, *Big Data, Genetics, and Re-Identification*, BILL OF HEALTH (Sept. 14, 2015) https://blog.petrieflom.law.harvard.edu/2015/09/24/big-data-genetics-and-re-identification [https://perma.cc/P2HD-7K76] (focusing on the re-identification of genetic data); Md Shopon, Sanjida Nasreen Tumpa, Yajurv Bhatia, K. N. Pavan Kumar & Marina L. Gavrilova, *Biometric Systems De-Identification: Current Advancements and Future Direction*s, J. CYBERSECURITY & PRIVACY 470, 470–71 (2021) (focusing on the privacy issues raised by biometric data).

[319] *See generally* Stacey A. Tovino, *A Timely Right to Privacy*, 104 IOWA L. REV. 1361, 1394 (2019) [hereinafter Tovino, *A Timely Right to Privacy*] (explaining that certain privacy penalty sharing regulations, delegated by Congress to HHS with a promulgation deadline of February 17, 2012, have yet to be promulgated by HHS). As of the writing of this Article, these regulations remain overdue, as they have been for almost ten years.

A. *Evolving Law*

The first theoretical alternative to fixed or outdated identification and de-identification standards is based on the concept of evolving law. As background, it is important to note that most of the identification and de-identification standards assessed in Part II are static. At the time they were written, they specified particular tests or particular data elements that implicated legal protection or resulted in deregulation, and these chosen tests or elements have not changed over time. It is also important to note that de-identification in accordance with the de-identification standards referenced in Part II produces, if anything, temporary de-identification. After data are de-identified, new external data may be created. This new data may be linked with the temporarily de-identified health data to reidentify the data subjects.

The problem with these identification and de-identification standards is that they neither evolve with the science of data reidentification nor recognize the temporariness of data de-identification. Although most of the standards could illustrate this point, consider the eighteen-element HIPAA Safe Harbor, promulgated through a final rule published on December 28, 2000.[320] Although research has shown that data de-identified in accordance with the Safe Harbor are vulnerable to reidentification,[321] HHS has not updated the Safe Harbor in over two decades.[322]

Further consider the California Shine the Light Act, signed into law on September 23, 2003, which protects twenty-seven categories of personal information.[323] The California State Legislature is to be

[320] Standards for Privacy of Individually Identifiable Health Information, 65 Fed. Reg. 82,462, 82,818 (Dec. 28, 2000) (codified at 45 C.F.R. § 164.514(b)(2)(i)(A)–(R) (2020)) (promulgating the eighteen-element Safe Harbor that remains in effect as of this writing).

[321] *See supra* notes 64, 66 and accompanying text (reporting 3.2 percent and 10.6 percent reidentification rates, respectively, with respect to Maine and Vermont discharge data de-identified in accordance with the HIPAA Safe Harbor).

[322] This does not mean that HHS necessarily should repeal its Safe Harbor. Research does show that application of the Safe Harbor can reduce the risk of reidentification. *See supra* notes 61–66 and accompanying text (reporting that the risk of re-identification of Maine and Vermont hospital discharge data fell from 28.3 percent to 3.2 percent and 34 percent to 10.6 percent, respectively, after applying the HIPAA Safe Harbor).

[323] California Shine the Light Act, 2003 Cal. Stat. 3891. The categories of personal information protected under the law are:

> (A) An individual's name and address. (B) Electronic mail address. (C) Age or date of birth. (D) Names of children. (E) Electronic mail or other addresses of children. (F) Number of children. (G) The age or gender of children. (H) Height. (I) Weight. (J) Race. (K) Religion. (L) Occupation. (M) Telephone number. (N) Education. (O) Political party affiliation. (P) Medical condition. (Q) Drugs, therapies, or medical products or equipment used. (R) The kind of product the customer purchased, leased,

commended for including several data elements not commonly included in other data protection laws, such as race, religion, occupation, education, political party, and creditworthiness.[324] As noted by the Kaiser Permanente researchers, individuals who are members of vulnerable populations (such as racial minority groups, religious minority groups, those in uncommon occupations, small or independent political parties or groups, and socioeconomically disadvantaged groups) may bear a disproportionate burden of health data reidentification.[325] That said, the Shine the Light Act's definition of personal information could be improved by adding a provision requiring lawmakers or regulators to periodically review and update the data elements listed therein.

There is precedent for applying a theory of evolving law to data protection. For example, the Massachusetts Security Breaches Act includes the following sentence: "The [Massachusetts Office] of [C]onsumer [A]ffairs and [B]usiness [R]egulation may adopt regulations, from time to time, to revise the definition of 'encrypted', as used in this chapter, to reflect applicable technological advancements."[326] This Article proposes that data protection law makers consider including similar evolving law language at the conclusion of their identification and/or de-identification tests or standards, as appropriate. Options for implementing evolving law include statutory sunset provisions, which provide for the expiration (and, therefore, perhaps renewal with amendments reflecting advances in science) of the law after a certain period of time, as well as provisions delegating the promulgation of regulations that would update the law from time to time.

> or rented. (S) Real property purchased, leased, or rented. (T) The kind of service provided. (U) Social security number. (V) Bank account number. (W) Credit card number. (X) Debit card number. (Y) Bank or investment account, debit card, or credit card balance. (Z) Payment history. (AA) Information pertaining to creditworthiness, assets, income, or liabilities.

CAL. CIV. CODE § 1798.83(e)(7) (2020); Appendix at California-3.

[324] CAL. CIV. CODE § 1798.83(e)(7)(J)–(L), (N)–(O), (AA); Appendix at California-3. *But see* Xu & Zhang, *supra* note 137, at 2–3 (asking "whether data anonymization could mask gross statistical disparities between [vulnerable] sub-populations in the data" and proposing that certain mechanisms of data anonymization may do so, which could preclude identification of health disparities (emphasis omitted)).

[325] *See supra* notes 135–136 and accompanying text (noting that the risk of reidentification falls "largely or exclusively on those in small cells or classes" such as "vulnerable or . . . disadvantaged groups" (quoting Simon et al., *supra* note 20, at 8)).

[326] MASS. GEN. LAWS ANN. ch. 93H, § 1(b) (West 2021); Appendix at Massachusetts.

Using the California Shine the Light Act as an example, the following new statutory provision (in italics) could be added to the end of current Cal. Civ. Code section 1798.83(e)(7) so that the data elements selected in 2003 for the definition of "personal information" keep pace with advances in data reidentification:

> (e) For purposes of this section, the following terms have the following meanings: . . .
>
> > (7) "Personal Information" . . . *The California Department of Consumer Affairs shall adopt regulations from time to time revising the definition of "personal information" to reflect advances in data identification and/or data reidentification. In promulgating such regulations, the Department shall consider: (1) the availability of public, semi-public, or private external data containing one or more elements that overlap with one or more of the data elements listed in the definition of "personal information"; (2) the size of the class defined by those overlapping elements in which the personal data element(s) fall(s); and (3) the overlap in the population covered by the personal data element(s) and the population covered by the external data source(s). See, e.g., Gregory E. Simon et al., Assessing and Minimizing Re-identification Risk in Research Data Derived from Health Care Records, 7(1) EGEMS 1 (2019).*

As applied to this California law, an evolving law theory promotes privacy by instructing lawmakers to become familiar with the reidentification literature and to consider how such literature impacts the selection of (or the failure to select) particular data elements for inclusion in the definition of "personal information." An evolving law theory also maintains and supports the concrete prescription of a multiple data element identification approach, which may help unsophisticated data custodians' compliance. One limitation of the evolving law theory as applied to the California Shine the Light Act is that it requires an administrative agency to act, which usually, but does not always, occur.[327]

B. *Reidentification Prohibition*

A second theoretical alternative is based on reidentification prohibition. For background, most of the data protection laws collected

[327] *Cf.* Tovino, *A Timely Right to Privacy*, *supra* note 319, at 1394 (explaining that certain federal privacy penalty sharing regulations remain overdue years after Congress delegated them to HHS).

and reviewed by this Article include health data identification standards (which are standards that specify which data will receive legal protection) and de-identification standards (which are standards that specify when data loses legal protection). Few laws expressly prohibit data reidentification. As Professor Daniel Solove notes in a clever cartoon, the science of health data reidentification has advanced to the point where it is difficult to eliminate the risk of reidentification through removing additional data elements.[328] Although the focus on identification and de-identification may have made sense when health industry privacy schemes, chiefly the HIPAA Privacy Rule, were being developed in the late 1990s and early 2000s,[329] different or additional theoretical approaches are currently necessary.[330]

Consider an alternative to data protection law's current focus on identification and de-identification. That is, what about a prohibition against reidentification? There is precedent for the use of a reidentification prohibition in data protection law in Texas law,[331]

[328] *See* Daniel Solove, *Cartoon: De-Identifying PHI Under HIPAA*, PRIV. + SEC. BLOG (May 18, 2020), https://teachprivacy.com/cartoon-de-identifying-phi-under-hipaa [https://perma.cc/D4KQ-JUJP].

[329] President Clinton signed HIPAA into law on August 21, 1996. *See generally* Health Insurance Portability and Accountability Act of 1996 (HIPAA), Pub. L. No. 104-191, 110 Stat. 1936 (1996) (codified as amended in scattered sections of 42 U.S.C.). As directed by HIPAA, HHS published its proposed HIPAA Privacy Rule on November 3, 1999, and its final HIPAA Privacy Rule on December 28, 2000. *See generally* Standards for Privacy of Individually Identifiable Health Information, 64 Fed. Reg. 59,918, 59,919 (Nov. 3, 1999) (codified at 45 C.F.R. §§ 160–64); Standards for Privacy of Individually Identifiable Health Information, 65 Fed. Reg. 82,462 (Dec. 28, 2000). HHS published proposed and final modifications to the HIPAA Privacy Rule on March 27, 2002, and August 14, 2002, respectively. *See generally* Standards for Privacy of Individually Identifiable Health Information, 67 Fed. Reg. 14,776 (Mar. 27, 2002) (codified at 45 C.F.R. §§ 160–64); Standards for Privacy of Individually Identifiable Health Information, 67 Fed. Reg. 53,182 (Aug. 14, 2002). The HIPAA Privacy Rule—which, at least before President Obama signed the ARRA into law, directed additional changes to be made—originated in the time frame of 1996 to 2002. *See supra* note 315 (explaining the changes made by ARRA to the HIPAA Privacy Rule).

[330] Although the HIPAA Privacy Rule has its origins in the time frame of 1996 to 2002, *see supra* note 329, the health data reidentification studies cited in Part I of this Article were published in 2010, Loukides et al., *supra* note 98, 2011, El Emam et al., *A Systematic Review*, *supra* note 86, 2013, El Emam et al., *Evaluating the Risk of Patient Re-Identification*, *supra* note 15, 2015, Sweeney, *supra* note 7, 2018, Ji Su Yoo et al., *supra* note 58, 2018, Na et al., *supra* note 70, and 2019, Simon et al., *supra* note 20.

[331] "A person may not reidentify or attempt to reidentify an individual who is the subject of any protected health information without obtaining the individual's consent or authorization if required under this chapter or other state or federal law." TEX. HEALTH & SAFETY CODE § 181.151 (2020); Appendix at Texas-2.

Arkansas law,[332] Hawaii law,[333] California law and the CCPA, and a federal research regulation. In the context of health-care procedure data submitted by California hospitals to the State of California that then may be re-shared with submitting hospitals in the form of risk-adjusted outcome rates,[334] a California law provides: "In no case shall a hospital, contractor, or subcontractor reidentify or attempt to reidentify any information received pursuant to this section."[335] The CCPA as well as recently introduced federal data protection bills that incorporate a "public commitment plus contractual obligation" de-identification standard also include prohibitions against reidentification.[336] Federal human subjects research regulation also offers precedent for a prohibition against reidentification. A regulation permits "[s]econdary research uses of identifiable private information or identifiable biospecimens" when, among other requirements, the researcher "will not re-identify" the subjects.[337]

This Article proposes a similar reidentification prohibition that would apply to all health data. In terms of placing such a prohibition, note that many current data protection laws have limited applicability. For example, the HIPAA Privacy Rule only applies to covered entities and business associates.[338] Even if Congress directed HHS to amend the HIPAA Privacy Rule to include a general prohibition against data reidentification, the end result would still be that only covered entities (health plans, health care clearinghouses, and certain health care providers) and their business associates would be subject to the

[332] This Arkansas law establishing an all-payer health insurance claims database prohibits data in the database from being used to "[r]eidentify or attempt to reidentify an individual who is the subject of any submitted data without obtaining the individual's consent." ARK. CODE ANN. § 21-61-907(a)(2)(B) (2020).

[333] "Under no circumstances shall a person attempt to re-identify [the] subjects of [health insurance data submitted to the state health insurance claims database.]" HAW. REV. STAT. § 323D-18.5(i) (2020).

[334] CAL. HEALTH & SAFETY CODE §§ 128735-37, 128745, 128748 (West 2021) (requiring hospitals to disclose a variety of health data to the State of California and requiring the state to use the data to publish risk-adjusted outcome rates).

[335] *Id.* § 128766(b) (West 2021).

[336] CAL. CIV. CODE § 1798.140(m)(2) (West 2021) (effective Jan. 1, 2023); Appendix at California-2; COVID-19 Consumer Data Protection Act of 2020, S. 3663, 116th Cong. § 2(9)(B), (C), (D) (2020); Appendix at CCDPA; Exposure Notification Privacy Act, S. 3861, 116th Cong. § 2(2)(B) (2020); Appendix at ENPA.

[337] 45 C.F.R. § 46.104(d)(4)(ii) (2020).

[338] *See supra* notes 148–152 and accompanying text (explaining the limited application of the "Reasonable Basis to Believe" standard in the HIPAA Privacy Rule).

prohibition.[339] Part I.A shows, however, that patients are concerned that their data will be reidentified by technology companies, nonclinical research teams, news reporters, lay community members, and other non-health industry participants that may use their information for financial or other personal gain.[340]

For this reason, this Article prefers (with some modifications) the approach of the TMRPA. The TMRPA broadly applies to any person who "comes into possession of" or "obtains or stores" certain information.[341] The TMRPA thus regulates nonhealth industry participants, including the technology companies and other individuals and institutions beyond the reach of the HIPAA Privacy Rule.

One limitation of the TMRPA is that the Texas Attorney General only enforces the law on behalf of Texas residents.[342] In response, this Article proposes a federal statutory prohibition against health data reidentification that would apply to any individual or any institution that collects, purchases, obtains, maintains, stores, or otherwise comes into possession of health data, regardless of the subject's residency.[343] A uniform law adopted by all states, or as many states as possible, would also serve this purpose.

[339] *See supra* notes 148–152 and accompanying text ; *infra* Part III.C (showing workaround involving Congress directing HHS to expand the application of the HIPAA Privacy Rule to non-health-industry participants).

[340] *See supra* Part I.A.

[341] TEX. HEALTH & SAFETY CODE ANN. § 181.001(b)(2)(B)–(C) (West 2021).

[342] *About the Attorney General*, KEN PAXTON: ATT'Y GEN. OF TEX., https://www.texasattorneygeneral.gov/about-office [https://perma.cc/SHV6-PBV2] (noting that the Attorney General "is focused on protecting Texans and upholding Texas laws").

[343] This proposal is supported (by analogy to the research context) by Nass, Levit & Gostin. *See* BEYOND THE HIPAA PRIVACY RULE: ENHANCING PRIVACY, IMPROVING HEALTH THROUGH RESEARCH, *supra* note 30, at 34, 190 (arguing that "unauthorized reidentification of individuals from DNA sequences, by anyone, should be strictly prohibited" and that "[t]o further protect privacy, unauthorized reidentification of information that has had direct identifiers removed should be prohibited by law, and violators should face legal sanctions"). The European Union ("EU") has a nonsectoral General Data Protection Regulation that may serve as a model for this federal (versus state) recommendation. *See* Regulation 2016/679, of the European Parliament and of the Council of 27 Apr. 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation), 2016 O.J. (L 119) 22 (establishing a general regulation to protect individuals from health data reidentification that applies to any "controller or . . . processor" of personal data). *See generally* Stacey A. Tovino, *The HIPAA Privacy Rule and the EU GDPR: Illustrative Comparisons*, 47 SETON HALL. L. REV. 973 (2017) (reviewing the non-sectoral GDPR and explaining how it compares to the health industry-specific HIPAA Privacy Rule); Tovino, *supra* note 144, at Parts II.A–D (discussing provisions within the GDPR as applied to mobile health research applications).

A second limitation of the TMRPA is that the prohibition against reidentification technically only applies to PHI, which is based on the definition of individually identifiable health information ("IIHI").[344] IIHI, in turn, is based on whether "there is [a] reasonable basis to believe" the information can be used to identify an individual.[345] To be effective, a reidentification prohibition must also apply to health information thought by some not to be identifiable. Otherwise, adversaries will defend themselves by saying that the prohibition did not apply to their extraordinary work because there was no ordinary belief that the information could be used to identify an individual.

The draft text that may be used for a general health data reidentification prohibition is: "In no case shall any individual or institution reidentify or attempt to reidentify the subject of any health data collected, received, purchased, or otherwise obtained by the individual or institution." Note that it uses the catchall phrase "health data," not "protected health information" or "individually identifiable health information." This is intentional. The law must protect information that some might think is not individually identifiable from being reidentified. During the legislative committee process, particular attention should be paid to defining "health data."[346]

C. Noncollection

A third theoretical alternative is based on noncollection. Noncollection refers to a prohibition against gathering, or collecting, health data by any person other than the data subject without the prior written notification and/or authorization of the data subject. Some federal bills and some state laws already incorporate health data

[344] TEX. HEALTH & SAFETY CODE ANN. § 181.151 (West 2021).

[345] *See* 45 C.F.R. § 164.514(a) (2020) (defining non-individually identifying health information as health information for "which there is no reasonable basis to believe that the information can be used to identify an individual").

[346] The HIPAA Privacy Rule defines "health information" as

> any information, including genetic information, whether oral or recorded in any form or medium, that: (1) Is created or received by a health care provider, health plan, public health authority, employer, life insurer, school or university, or health care clearinghouse; and (2) Relates to the past, present, or future physical or mental health or condition of an individual; the provision of health care to an individual; or the past, present, or future payment for the provision of health care to an individual.

Id. § 160.103. This Article dislikes the first clause, which is too limiting in terms of who can create or receive the information for the information to qualify as "health information." For example, the first clause does not include technology companies, mobile health applications, nonclinical research teams, non-university-based research teams, journalists, news reporters, and lay community members. Instead, this Article recommends the broadly written second clause.

noncollection principles in some contexts. For example, the newly introduced federal ITPDCA would require a data controller to notify an individual "through a privacy and data use policy of a specific [intent] to collect . . . information" and to give the individual the right and ability to opt-out of such collection.[347] The privacy and data use policy would be required to include, among other notifications, (1) the "[i]dentity and contact information of the entity collecting" the individual's personal information; (2) the purpose, or reason why, the entity wants to collect the individual's personal information; and (3) "[t]he process by which individuals may withdraw [their] consent to the collect[ion of their] personal information."[348]

As of this writing, the ITPDCA has not been signed into law.[349] In addition, federal regulations like the HIPAA Privacy Rule and some state statutes like the TMRPA only regulate the use, disclosure, and sale of health data, not the collection of health data.[350] Because individuals voluntarily give significant health data to health industry participants, especially health care providers, while obtaining health care, prohibiting providers from generally collecting health data would not make sense unless the collection prohibition exempted diagnostic and treatment purposes. That said, prohibiting non-health-care providers, including mobile health applications, wearable devices, and infectious disease ENS, from collecting an individual's health data (as well as general consumer data from which health may be inferred) without the individual's prior notification and authorization does make sense.

D. Nonuse and Nondisclosure

This Article recommends evolving law, non-reidentification, and noncollection to reduce the vulnerability of health data to reidentification, to prohibit reidentification, and to reduce the amount

[347] Information Transparency & Personal Data Control Act, H.R. 1816, 117th Cong. § 3(a)(1)(A), (4)(A) (2021); Appendix at ITPDCA.

[348] H.R. 1816 § 3(a)(3)(A)–(D); Appendix at ITPDCA.

[349] *H.R.1816 - Information Transparency & Personal Data Control Act*, CONGRESS.GOV, https://www.congress.gov/bill/117th-congress/house-bill/1816/actions [https://perma.cc/2Z2V-BHDZ] (showing that the ITPDCA was introduced on March 11, 2021, but has not been enacted).

[350] *See* 45 C.F.R. § 164.502–164.514 (2020) (codifying the HIPAA Privacy Rule's "uses and disclosures" requirements and not setting forth any "collection" prohibitions); TEX. HEALTH & SAFETY CODE ANN. § 181.152 (West 2021) (regulating marketing uses and disclosures of protected health information); *id.* § 181.153 (regulating "[s]ale[s] of [p]rotected [h]ealth [i]nformation"); Appendix at Texas-2.

of data that may be reidentified, respectively. Nonuse and nondisclosure would provide checks and balances should incorporating evolving law or non-reidentification provisions be limited or fail. Nonuse is the prohibition of the use[351] of health data, including reidentified health data, by any person other than the data subject without the subject's prior written authorization. Nondisclosure is the prohibition of the disclosure[352] of health data, including reidentified health data, by any person other than the data subject without the subject's prior written authorization.

Federal and state health laws already incorporate nonuse and nondisclosure principles in some contexts. For example, the federal HIPAA Privacy Rule prohibits covered entities and business associates from using or disclosing a patient's PHI unless the patient has given prior written authorization or the use or disclosure falls into an exception.[353] Similarly, the TMRPA prohibits any person who "comes into possession of" or "obtains or stores" PHI[354] from (1) "disclos[ing] an individual's [PHI] . . . in exchange for direct or indirect remuneration" (i.e., selling an individual's PHI)[355] or (2) "electronically disclos[ing] an individual's [PHI] to any person without a separate[, prior] authorization from the individual . . . for each disclosure."[356] Nonuse and nondisclosure laws like the HIPAA Privacy Rule are premised on findings that health data subjects are concerned about how their information is used and how their information may be later transmitted or disclosed.[357] At the time that HHS was drafting the 1999 proposed HIPAA Privacy Rule, a contemporaneous *Wall Street*

[351] *See, e.g.*, 45 C.F.R. § 160.103 (defining "use" as "the sharing, employment, application, utilization, examination, or analysis of such information within an entity that maintains such information").

[352] *See, e.g.*, *id.* (defining "disclosure" as "the release, transfer, provision of access to, or divulging in any manner of information outside the entity holding the information").

[353] *Id.* § 164.508(a)(1), (b)(1)(i).

[354] TEX. HEALTH & SAFETY CODE ANN. § 181.001(b)(2)(B)–(C) (West 2021) (defining a "[c]overed entity").

[355] *Id.* § 181.153(a).

[356] *Id.* § 181.154(b).

[357] In the Department of Health and Human Services' Federal Register notice proposing the HIPAA Privacy Rule, the agency noted,

> Individuals who provide information to health care providers and health plans increasingly are concerned about how their information is used within the health care system. Patients want to know that their sensitive information will be protected not only during the course of their treatment but also in the future as that information is maintained and/or transmitted within and outside of the health care system.

Standards for Privacy of Individually Identifiable Health Information, 64 Fed. Reg. 59,918, 59,919 (Nov. 3, 1999) (codified at 45 C.F.R. §§ 160–64).

Journal poll showed that 29 percent of respondents cited "[l]oss of personal privacy" as a first or second response to the question of "what concerned them most in the coming century."[358] Other concerns, including "terrorism, . . . war, and global warming," received first or second responses at 23 percent or less.[359]

Nonuse and nondisclosure laws are also premised on findings showing that patients preemptively respond to concerns about loss of privacy by (1) lying about (or minimizing) their symptoms, health-related behaviors, and/or health conditions when seeking care; (2) visiting multiple health care providers to avoid shame associated with repeated unhealthy behaviors and the documentation thereof; and (3) refusing to seek health care altogether.[360] HHS explained that it designed the 1999 proposed rule to respond to these privacy concerns: "The use of these standards will help to restore patient confidence in the health care system, providing benefits to both patients and those who serve them."[361]

Nonuse and nondisclosure laws are further premised on findings showing that an increasing number of individuals and institutions were becoming involved in the collection, use, disclosure, and redisclosure of health data. The 1999 proposed rule stated: "The number of entities who are maintaining and transmitting individually identifiable health information has increased significantly over the last 10 years."[362] The same statement could be made today, more than two decades later. Technology companies, nonclinical researchers, journalists, and lay community members are increasingly acquiring, using, and sharing health information—sometimes to the ridicule, shame, detriment, embarrassment, or unwanted publicity of the data subjects and their families.[363] Whereas health information used to be created, used, and

[358] *Id.* (referencing *Wall Street Journal* poll).

[359] *Id.*

[360] *Id.* at 59,920 (referencing a California HealthCare Foundation survey reporting that "one-sixth of respondents indicated that they had taken some form of action to avoid the misuse of their information, including providing inaccurate information, frequently changing physicians, or avoiding care").

[361] *Id.*; *see also* BEYOND THE HIPAA PRIVACY RULE: ENHANCING PRIVACY, IMPROVING HEALTH THROUGH RESEARCH, *supra* note 30, at 57, 59, 60, 65, 83–84 (explaining that some individuals do not participate in research because of privacy and confidentiality concerns as well as a lack of trust in researchers).

[362] Standards for Privacy of Individually Identifiable Health Information, 64 Fed. Reg. at 59,920.

[363] *See, e.g.*, Park et al., *supra* note 16, at 2129 (describing members of the South Korean public reidentifying COVID-19 data, resulting in some reidentified people being "affected by

disclosed within the somewhat closed loop of the health-care delivery and health insurance industries, the story is quite different today.[364]

Unfortunately, both the HIPAA Privacy Rule and the TMRPA are limited in application. As mentioned earlier, the HIPAA Privacy Rule only applies to certain health industry participants and their business associates.[365] Although the TMRPA technically applies to anyone who comes into possession of, or collects or stores, protected health information,[366] the law is enforced only on behalf of Texas residents.[367] This Article therefore proposes that Congress direct HHS to (1) expand the application of the HIPAA Administrative Simplification rules (of which the HIPAA Privacy Rule is one part)[368] to anyone who comes into possession of, collects, or stores PHI or reidentified health data and (2) apply the nonuse and nondisclosure provisions within the HIPAA Privacy Rule to such newly regulated persons and to such data. The result would be that anyone who comes into possession of, collects, or stores PHI or reidentified data would be prohibited from using or disclosing that data without the data subject's prior written authorization. To achieve this result, the following new statutory language (in italics) should be added to the end of § 1320d-1(a) of Title 42 of the U.S. Code as follows:

> (a) Applicability. Any standard adopted under this part shall apply, in whole or in part, to the following persons:

unwanted privacy invasion" and facing "public disdain"); El Emam et al., *Evaluating the Risk of Patient Re-Identification*, *supra* note 15, at 1 (reporting a journalist's reidentification of a twenty-six-year-old woman from a de-identified adverse event database, and noting that the woman died while taking a particular medication and the journalist subsequently contacted her family and broadcasted the unfortunate story).

[364] *See, e.g.*, Tovino, *supra* note 144, at 208 (explaining that, "a quarter of a century ago, . . . most data originated in the industry to which it pertained"; that is, health data originated and stayed within the offices of health care providers, not technology companies like Google; arguing for more generally applicable forms of data protection rather than industry-specific laws to respond to the increasing number of non-health-industry participants who collect, use, and disclose health data).

[365] *See supra* notes 149–150 and accompanying text (explaining the application of the HIPAA Privacy Rule).

[366] TEX. HEALTH & SAFETY CODE ANN. § 181.001(b)(2)(B)–(C) (West 2021).

[367] *See* KEN PAXTON: ATT'Y GEN. OF TEX., *supra* note 342 ("Attorney General Paxton is focused on protecting Texans and upholding Texas laws and the Constitution.").

[368] *See* 45 C.F.R. §§ 164.500–164.534 (2020) (codifying the HIPAA Privacy Rule, which is one part of the Administrative Simplification Rules codified at 45 C.F.R. §§ 160.101–164.534). The HIPAA Administrative Simplification Rules also include a security rule and a breach notification rule. *See generally supra* note 143 (citing to the HIPAA Security Rule); 45 C.F.R. § 164.410 (2020) (codifying the HIPAA Breach Notification Rule).

> (1) A health plan.
>
> (2) A health care clearinghouse.
>
> (3) A health care provider who transmits any health information in electronic form in connection with a transaction referred to in section 1320d-2(a)(1) of this title.
>
> *(4) Any other natural or legal person who comes into possession of, or collects or stores, health data.*

The bill should also direct the Secretary of HHS to promulgate regulations within six months of the date of enactment of the bill making conforming changes to the HIPAA Administrative Simplification Rules. Illustrative, not exhaustive, examples of conforming regulatory changes would expand the definition of "[c]overed entity"[369] as well as expand the application of the Administrative Simplification Rules.[370] To the extent Congress does not enact such a bill, this Article recommends a uniform state law based on the TMRPA as next best.

E. Nondiscrimination

A sixth and final approach is necessary to minimize informational injuries to health data subjects: nondiscrimination. In this context, nondiscrimination refers to prohibiting undesirable discrimination against an individual based on the individual's health data by any person or institution.

A range of health laws already incorporate principles of nondiscrimination. For example, one portion of the Genetic Information Nondiscrimination Act of 2008 ("GINA")prohibits health insurers from using genetic information about an individual to adjust a group plan's premiums, or, for individual plans, to deny coverage, adjust premiums, or impose a preexisting condition exclusion.[371] One Affordable Care Act ("ACA") provision prohibits "group health plan[s] and . . . health insurance issuer[s] offering group or individual health insurance coverage [from] establish[ing] rules for [enrollment]

[369] 45 C.F.R. § 160.103.

[370] *Id.* §§ 160.102(a), 164.104(a). Additional changes to the HIPAA Statute and Administrative Simplification Rules recommended by the Author in prior works, including those relating to a qui tam process and a private right of action, would further support the arguments made in this Article. *See, e.g.*, Tovino, *A Timely Right to Privacy*, *supra* note 319, at Parts IV–V.

[371] Genetic Information Nondiscrimination Act of 2008, Pub. L. No. 110-233, §§ 101(a), 102(b), 122 Stat. 881, 883, 893.

eligibility" based on a long list of overlapping items, including health status, medical condition, claims experience, "[r]eceipt of health care," medical history, genetic information, evidence of insurability, disability, and "[a]ny other health status-related factor determined appropriate by the Secretary" of HHS.[372]

These illustrative nondiscrimination provisions are limited in application, however. Neither the GINA provision nor the ACA provision applies to disability insurers, life insurers, or long-term care insurers.[373] To understand why this regulatory gap is problematic, consider the fact that racial and ethnic minorities are disproportionately represented among confirmed COVID-19 cases,

[372] 42 U.S.C. § 300gg-4(a). Although some state laws do apply to other forms of insurance, these state laws are in the minority. *See, e.g.*, COLO. REV. STAT. § 10-3-1104.7(2)–(3) (2021) (prohibiting "entit[ies] that provide[] group disability insurance" from using information from genetic tests for nontherapeutic or underwriting purposes); *see also* Yann Joly, Charles Dupras, Miriam Pinkesz, Stacey A. Tovino & Mark A. Rothstein, *Looking Beyond GINA: Policy Approaches To Address Genetic Discrimination*, 21 ANN. REV. GENOMICS & HUM. GENETICS 491, 495–96 (2020) (stating that a minority of state laws (approximately 25 percent, 20 percent, and 10 percent, respectively) prohibit genetic discrimination "in the context of disability insurance," life insurance, and long-term care insurance).

[373] *See, e.g.*, *Genetic Discrimination*, NAT'L HUM. GENOME RSCH. INST., https://www.genome.gov/about-genomics/policy-issues/Genetic-Discrimination [https://perma.cc/X6UG-ZKRJ] (explaining that because GINA does not apply to disability insurers, life insurers, or long-term care insurers, some states have passed laws that do cover the insurers); 42 U.S.C. § 300gg-4(a) (only regulating health plans and health insurance issuers, not disability insurers, life insurers, or long-term care insurers). Disability nondiscrimination law, such as the Americans with Disabilities Act and section 1557 of the Affordable Care Act ("ACA"), also must be analyzed to determine the extent to which individuals may be discriminated against in the context of insurance on the basis of data recording a disability. For example, as the Ninth Circuit stated in *Schmitt v. Kaiser Foundation Health Plan of Washington*,

> Section 1557 of the [ACA] . . . prohibits covered health insurers from discriminating based on various grounds, including disability. Prior to the ACA's enactment, an insurer could generally design plans to offer or exclude benefits as it saw fit without violating federal antidiscrimination law—in particular, the Rehabilitation Act—so long as the insurer did not discriminate against disabled people in providing treatment for whatever conditions it chose to cover. The primary issue before us is whether the ACA's nondiscrimination mandate imposes any constraints on a health insurer's selection of plan benefits. We hold that it does.

965 F.3d 945, 948 (9th Cir. 2020); *see, e.g.*, Valarie Blake, *Rethinking the Americans with Disabilities Act's Insurance Safe Harbor*, 6 LAWS 1, 1 (2017) (explaining that an ADA safe harbor "explicitly permits insurers to discriminate on the basis of disability in health insurance so long as the differential treatment is supported by actuarial data and is not just intended to disadvantage the disabled" and arguing for the removal of the safe harbor). More recent nondiscrimination laws also must be considered. For example, one provision within the Coronavirus Aid, Relief, and Economic Security Act prohibits entities from "discriminat[ing] against . . . individual[s] on the basis of information received . . . pursuant to an inadvertent or intentional disclosure of [substance use disorder] records," including in the health care and employment contexts. Coronavirus Aid, Relief, and Economic Security Act, Pub. L. No. 116-136, § 3221(g), 134 Stat. 281, 377–78 (2020) (codified at 42 USC § 290dd-2(i)(1)).

hospitalizations, long-haul symptoms, and deaths.[374] As a result, racial and ethnic minorities are disproportionately represented among recorded COVID-19 data that are undesirable from the perspective of disability insurers, life insurers, and long-term care insurers.[375] Thus, these vulnerable populations may have greater difficulty obtaining the financial security associated with these forms of insurance, further reinforcing the socioeconomic determinants of health that contributed to higher rates of COVID-19 to begin with.[376] Disadvantage is then heaped upon disadvantage because individuals who are members of racial and ethnic minority groups also bear a disproportionate burden of data reidentification.[377]

For these reasons, this Article joins the chorus of health law scholars who have written extensively about the concerns associated

[374] *See, e.g.*, U.S. DEP'T OF HEALTH & HUM. SERVS., HHS INITIATIVES TO ADDRESS THE DISPARATE IMPACT OF COVID-19 ON AFRICAN AMERICANS AND OTHER RACIAL AND ETHNIC MINORITIES 2 (2020) (noting that the CDC data show higher hospitalizations of African Americans for COVID-19); *COVID-19 in Racial and Ethnic Minority Groups*, U.S. CTRS. FOR DISEASE CONTROL & PREVENTION, https://stacks.cdc.gov/view/cdc/89820/cdc_89820_DS1.pdf? [https://perma.cc/6ECD-6RM2] ("Long-standing systemic health and social inequities have put some members of racial and ethnic minority groups at increased risk of getting COVID-19 or experiencing severe illness, regardless of age."); Rachel R. Hardeman, Eduardo M. Medina & Rhea W. Boyd, *Stolen Breaths*, 383 NEW ENG. J. MED. 197, 197 (2020) ("In Minnesota, . . . [B]lack Americans account for 6% of the population but 14% of Covid-19 cases and 33% of Covid-19 deaths . . . ."); Khiara M. Bridges, *The Many Ways Institutional Racism Kills Black People*, TIME (June 11, 2020, 6:31 AM), https://time.com/5851864/institutional-racism-america [https://perma.cc/F28H-KZBL] ("COVID-19 has disproportionately killed [B]lack people in the U.S."); Kyle Yomogida, Sophie Zhu, Francesca Rubino, Wilma Figueroa, Nora Balanji & Emily Holman, *Post-Acute Sequelae of SARS-CoV-2 Infection Among Adults Aged ≥18 Years — Long Beach, California, April 1–December 10, 2020*, 70 MORBIDITY & MORTALITY WKLY. REP. 1274, 1277 (2021) (reporting higher rates of post-acute sequelae among Black persons). *See generally* Ruqaiijah Yearby & Seema Mohapatra, *Law, Structural Racism, and the COVID-19 Pandemic*, 7 J.L. & BIOSCIENCES, Jan.–June 2020, at 1 (explaining how employment, housing, health care, and COVID-19 relief laws have been manipulated to disadvantage racial and ethnic minorities, making them more susceptible to COVID-19 infection and death).

[375] *See, e.g.*, AM. ACAD. OF ACTUARIES, ISSUE BRIEF: THE USE OF GENETIC INFORMATION IN DISABILITY INCOME AND LONG-TERM CARE INSURANCE 2–3 (2002) (discussing medical underwriting in the contexts of disability insurance and long-term care insurance and noting that both physical and mental health conditions are major causes of insurance claims).

[376] *See, e.g.*, Emily A. Benfer & Lindsay F. Wiley, *Health Justice Strategies To Combat COVID-19: Protecting Vulnerable Communities During a Pandemic*, HEALTH AFFS. BLOG (Mar. 19, 2020), https://www.healthaffairs.org/do/10.1377/hblog20200319.757883/full [https://perma.cc/3MW7-5PQP] (citing recent census report showing that "communities of color and low-income neighborhoods" have more concentrated poverty risk and that "[m]any low-income individuals and families face significant challenges that prevent them from protecting themselves and others from COVID-19").

[377] *See supra* notes 135–137 and accompanying text.

with health status (and underlying health data) discrimination and argues strongly against undesirable forms of discrimination against health data subjects.[378] This Article also supports the scholars who promote universal health insurance.[379] Universal access regimes prohibit or limit discrimination based on health status and therefore health data, making health data less consequential.[380] Decreased consequence may, in turn, decrease efforts to collect, reidentify, use, or disclose health data.

CONCLUSION

This Article carefully reviews health data reidentification claims and concerns, providing specific examples of de-identified health data that were reidentified following matching with other public, semipublic, or private data. This Article also reviews the scientific literature assessing the risk of reidentification and the efficacy of particular de-identification techniques. Four conclusions result: First, health data have a high reidentification risk. Second, health data from which as many as eighteen different identifiers have been removed still carry some reidentification risk. Third, two popular data perturbation techniques—suppressing rare data elements and generalizing other data elements—do not always prevent reidentification. Finally, data subjects do not equally share reidentification risk. Vulnerable individuals, including individuals with rare health conditions and individuals in racial and ethnic minority groups, bear a disproportionate burden of reidentification.

To determine whether the law recognizes these reidentification risks and burdens and appropriately responds, this Article collects and originally synthesizes a wide range of illustrative current and pending federal and state laws that expressly or potentially protect health data's confidentiality and security as well as data subjects' privacy. Inexhaustive examples of the types of health data protected by the

[378] *See, e.g.*, JESSICA L. ROBERTS & ELIZABETH WEEKS, HEALTHISM: HEALTH-STATUS DISCRIMINATION AND THE LAW 180–97 (2018) (explaining the difference between desirable health status differentiation and undesirable health status discrimination).

[379] *See, e.g.*, Nicolas P. Terry & Christine Coughlin, *A Virtuous Circle: How Health Solidarity Could Prompt Recalibration of Privacy and Improve Data and Research*, 74 OKLA. L. REV. 51, 52 (2021) ("[I]f health care continues to robustly prohibit health discrimination and continues to grow closer to universal access, the need for health data protection should decrease. This decrease would result not because privacy declines as a value but because exposures of health information would be less consequential.").

[380] *See id.* at 77.

collected legal authorities include general medical record data, physical health condition data, mental health condition data, health insurance data, infectious disease data, suicide risk data, health and fitness data, genetic data, biometric data, biological samples, prescription data, geolocation data, and proximity data, as well as general consumer data from which health can be inferred. This Article identifies a number of significant weaknesses associated with the identification and de-identification standards reviewed that render both unprotected data and formally de-identified data vulnerable to reidentification.

To protect against the reidentification risks and disproportionate reidentification burdens described in this Article and to respond to calls for data de-identification reform, this Article proposes six theoretical alternatives to the current focus on identification and de-identification. These alternatives are based on the concepts of evolving law, non-reidentification, noncollection, nonuse, nondisclosure, and nondiscrimination. This Article also offers specific textual amendments to federal and state data protection laws that would incorporate these theoretical alternatives. If adopted by federal and state lawmakers, this Article's proposals will help protect health data subjects from discrimination and other informational injuries associated with the use, disclosure, and redisclosure of their identifiable—and potentially reidentifiable—health data.