Cleveland-Marshall
College of Law Library

Journal of Law and Health

Volume 34 | Issue 1 Article 6

11-30-2020

Hacking HIPAA: "Best Practices" for Avoiding Oversight in the Sale of Your Identifiable Medical Information

Riyad A. Omar

Recommended Citation

Riyad A. Omar, *Hacking HIPAA: "Best Practices" for Avoiding Oversight in the Sale of Your Identifiable Medical Information*, 34 J.L. & Health 30 (2020)
available at https://engagedscholarship.csuohio.edu/jlh/vol34/iss1/6

HACKING HIPAA: "BEST PRACTICES" FOR AVOIDING OVERSIGHT IN THE SALE OF YOUR IDENTIFIABLE MEDICAL INFORMATION

RIYAD A. OMAR, J.D, CISP/US, CISP/E

I. INTRODUCTION

"Your medical data is for sale – all of it."[1] This warning comes from Adam Tanner, of Harvard's Institute for Quantitative Social Science, who has published extensively on the topic of the business of selling medical records.[2] When you visit your doctor, you may think "I'm telling my doctor my most intimate medical secrets, and only my doctor knows about it."[3] Frequently, however, your medical records are being sold,[4] including your "[p]rescription records, blood tests, doctor notes, hospital visits and insurance records."[5]

This is a big business. Three quarters of all retail pharmacies in the U.S.[6] sell their patients' prescription records and healthcare information, as do major health insurers, such as UnitedHealth, Anthem and Blue Cross Blue Shield.[7] Your medical records are often sold to data brokers who consolidate them into a comprehensive profile about you. One data broker, for example, boasts of having "500 million comprehensive, longitudinal anonymous patient records" sourced from "over 100,000 data suppliers."[8] Another advertises the ability to create "healthcare journeys"[9] about patients created from a "collection of claims data for

[1] Sam Thielman, *Your private medical data is for sale – and it's driving a business worth billions*, THE GUARDIAN (Jan. 10, 2017), https://www.theguardian.com/technology/2017/jan/10/medical-data-multibilliondollar-business-report-warns.

[2] *See*, *e.g.*, Adam Tanner, *Our Bodies, Our Data: How Companies Make Billions Selling Our Medical Records* (2017); Adam Tanner, *Strengthening Protection of Patient Medical Data* (Jan. 10, 2017) [hereinafter *Strengthening Protection of Patient Medical Data*], https://production-tcf.imgix.net/app/uploads/2017/01/11165252/strengthening-protection-of-patient-medical-data-1.pdf; Adam Tanner *, The Hidden Global Trade in Patient Medical Data*, YALEGLOBAL ONLINE (Jan. 24, 2017) [hereinafter *The Hidden Trade in Medical Data*], https://yaleglobal.yale.edu/content/hidden-global-trade-patient-medical-data; Adam Tanner*, How Your Medical Data Fuels a Hidden Multi-Billion Dollar Industry*, TIME (Jan. 9, 2017) [hereinafter *Your Medical Data Fuels Hidden Industry*], http://time.com/4588104/medical-data-industry; Adam Tanner , *How Data Brokers Make Money Off Your Medical Records*, SCI. AM. (Feb. 1, 2016) [hereinafter *How Data Brokers Make Money Off Your Medical Records*], https://www.scientificamerican.com/article/how-data-brokers-make-money-off-your-medicalrecords.

[3] Thielman, *supra* note 1.

[4] *Id*.

[5] Tanner, *Your Medical Data Fuels Hidden Industry*, *supra* note 2.

[6] Tanner, *How Data Brokers Make Money Off Your Medical Records*, *supra* note 2.

[7] Tanner, *Strengthening Protection of Patient Medical Data*, *supra* note 2, at 7.

[8] IMS HEALTH HOLDINGS, INC. ANN. REP. (2016), at 8.

[9] Tanner, *Strengthening Protection of Patient Medical Data* , *supra* note 2, at 8.

280 million patients."[10] A New York City-based start-up claims to possess, as of the date of this writing, 29 billion lab records of 250 million patients sourced from leading national clinical labs, such as LabCorp and Quest Diagnostics, as well as oncology and genetic testing labs.[11] A San Francisco-based start-up claims to have a "health map" that "links 150 complete real-time datasets for more than 320 million patients."[12]

These profiles allow data brokers to track your diagnoses, prescriptions, lab tests and more, as you interact with the healthcare system. Brokers advertise their ability to create "patient journeys,"[13] fine-tuned enough that if you visit a CVS in Cleveland one day, and a Walgreens in Miami the next, or visit different doctors in those cities, the broker will know.[14] Data brokers seeking to downplay the risks to patient privacy may refer to a patient's medical records as a "byproduct," "exhaust," or an "asset" of the healthcare organization to be sold.[15] But this "exhaust" is the medical records of millions of patients containing the categories of sensitive information one reasonably expects in medical records. Brokers can not only track a patient's use of prescription drugs, they can also glean "insights" about that patient based on sensitive portions of his medical history, such as his psychiatric history, substance dependency, STDs or history of physical or sexual abuse.

Under the Health Insurance Portability and Accountability Act of 1996 and its data protection regulations (collectively referred to hereinafter as

[10] *PRA Health Sciences acquires Symphony Health, scrip and prescriber data provider*, PHARM. COMMERCE (Aug. 17, 2017), https://pharmaceuticalcommerce.com/business-and-finance/pra-health-sciences-acquires-symphony-health-solutions-scrip-prescriber-data-provider.

[11] PROGNOS HEALTH INC., https://prognoshealth.com/ (last visited Jan. 27, 2020).

[12] Press Release, Komodo Health, Veradigm and Komodo Health Partner to Create the Largest Linked HER and Claims Dataset for Life Sciene Research (Aug. 21, 2019), https://www.komodohealth.com/insights/2019/08/veradigm-and-komodo-health-partner-to-create-the-largest-linked-ehr-and-claims-dataset-for-life-science-research.

[13] *See*, *e.g.*, KOMODO HEALTH INC., https://www.komodohealth.com (last visited Nov. 8, 2020) (advertising the ability to "unlock the truth about the patient journey through the U.S. healthcare system.").

[14] Tanner, *Strengthening Protection of Patient Medical Data* , *supra* note 2, at 11.

[15] *Id*. at 7 (quoting former IMS executive: "We used to say, 'Look, you are creating data as a byproduct. It's an exhaust from your system. Why don't you take that thing and turn it into an asset and sell it?' That is the way we would get people to think about data as an asset …").

HIPAA),[16] it is illegal for a healthcare organization[17] to sell a patient's[18] medical information without first obtaining the patient's written authorization.[19] Healthcare organizations and their data brokers may be seeking to bypass HIPAA's prohibition by describing the patient medical information they transact in as "anonymized" or "de-identified." Such assertions, however, are rarely – if ever – verified by regulators or independent standards-setting bodies.

This lack of oversight may be coming at a price being paid by patients who lose their privacy in the process. "Data scientists," Tanner notes, can link these patient profiles with consumer profiles "with a surprising degree of accuracy."[20] This is a natural consequence of the fact that when enough information is added to *any* patient's profile, a broker will eventually obtain the ability to identify that patient. Prominent security researcher, Ross Anderson, noted this phenomenon when evaluating a proposal for creating a database of

[16] Health Insurance Portability and Accountability Act of 1996, Pub. L. No. 104–191, adding a new Part C to Title XI of the Social Security Act, comprising sections 1171–1179 of the Social Security Act, codified at 42 U.S.C. 1320d–8 (Aug. 21, 1996); as amended by the Health Information Technology for Economic and Clinical Health Act or HITECH Act, Pub. L. No. 111-5 (Feb. 17, 2009), and the regulations promulgated thereunder, including Standards for Privacy of Individually Identifiable Health Information, codified at 45 C.F.R. Parts 160 and 164 (Dec. 28, 2000; last amended, Jan. 25, 2013) [hereinafter *Privacy Rule*]; Health Insurance Reform: Security Standards, codified at 45 C.F.R. Part 160 and Subparts A (General Provisions) and C (Security Standards for the Protection of Electronic Protected Health Information) of Part 164 (Security and Privacy) (Feb. 20, 2003; last amended, Jan. 25, 2013) [hereinafter *Security Rule*]; Notification in the Case of Breach of Unsecured Protected Health Information, codified at 45 C.F.R. Subpart D (Aug. 24, 2009; last amended Jan. 25, 2013) [hereinafter *Breach Notification Rule*]; and Civil Money Penalties; Procedures for Investigations, Imposition of Penalties, and Hearings, Department of Health and Human Services, codified 45 C.F.R. Parts 160 and 164 (Apr. 17, 2003, last amended, Jan. 25, 2013) [hereinafter *Enforcement Rule*].

[17] HIPAA's provisions apply to organizations that it calls "covered entities" and "business associates." Covered entities includes a wide range of "health care providers," including hospitals, medical practices, pharmacies, clinical labs and many others; to health insurers, which HIPAA calls "health plans;" and to entities that process medical claims, known as "health care clearinghouses." The term "business associate" applies to organizations that provide services to covered entities that require the organization to receive access to identifiable information about the covered entities' patients or beneficiaries. See definitions for "covered entity," "health care provider," "health plan," "health care clearinghouse," and "business associate" in 45 C.F.R. § 160.103. For brevity, this Article refers to covered entities and business associates collectively as "healthcare organizations."

[18] HIPAA uses the term "individual" to refer to a "person who is subject of protected health information," which is, subject to limited exceptions, individually identifiable health information transmitted or maintained by healthcare organizations. *See* 45 C.F.R. § 160.103. Most frequently "individuals" are patients of a healthcare provider and/or beneficiaries or members of a health plan. Outside of HIPAA, the term "individual" has many other meanings. Because the term "individual" has ambiguous meanings outside of HIPAA, for clarity this Article refers to "individuals" as "patients."

[19] 45 C.F.R. §§ 164.502(a)(5)(ii), 164.508(a)(4).

[20] Thielman, *supra* note 1.

Iceland's medical, genealogy and genetic data.[21] In examining the proposal, Anderson noted that "it is effectively impossible to de-identify … records … which link together all (or even many) of the health care encounters in a patient's life."[22] "For this reason," Anderson concluded, "a database of [such] medical records *must be considered to be personal health information*."[23]

Despite the red flags, observers often acquiesce to the notion that these large volumes of sensitive medical information are "de-identified" in accordance with HIPAA's requirements. Tanner, for example, summarizes his belief that "IMS and other data brokers are not restricted by medical privacy rules in the U.S., because their records are designed to be anonymous:"[24]

> "On the surface, it might seem impossible for a data miner to link anonymized information about a patient from separate sources—CVS at home in Cleveland today, but at Walgreens while on vacation in Miami Beach next month—or from different doctors in these cities. Yet data miners are able to match these files by getting pharmacies, insurers, testing labs, electronic health record systems, and other suppliers to all install the same de-identification software (for which they compensate the data suppliers).
>
> This software removes the personal details for each individual—such as name, address, telephone number, and Social Security number—but assigns that person the same anonymous patient identification key across all locations using that de-identification system. 'If they install that de-ID engine at every source and it has the same algorithm, that means everyone with the same PHI (personal health information) will get the same IMS patient key,' says Mark Degatano, who has advised IMS Health and worked at rival data miner Symphony Health.
>
> The 'De-ID engine' allows data miners to assemble a patient dossier with thousands of data points spanning back years. The file does not include a name, but lists age and gender, as well as what section of Cleveland she lives in." [25]

[21] Ross Anderson, *The DeCODE Proposal for an Icelandic Health Database* (Oct. 20, 1998), https://www.cl.cam.ac.uk/~rja14/Papers/iceland.pdf.

[22] *Id.* at 3.

[23] *Id.* at 4 (emphasis added).

[24] Tanner, *How Data Brokers Make Money Off Your Medical Records*, *supra* note 2.

[25] Tanner, *Strengthening Protection of Patient Medical Data* , *supra* note 2, at 11.

This process, Tanner concludes, complies with HIPAA because “[HIPAA] governs only the transfer of medical information that is tied *directly* to an individual's identity.”[26]

Tanner synopsis of HIPAA, however, is incorrect. HIPAA’s protections have *never* been limited to information that is “tied *directly* to an individual’s identity.” On the contrary, HIPAA’s protections have *always* applied to “*any* information … [with respect to which there is a reasonable basis to believe] … *can be used* to identify a patient.[27] Simply removing your *direct identifiers* (such as your *name*, *telephone number*, *address* or *social security number*) from your medical records has *never* been viewed of as sufficient to allow your doctor to sell your medical records. As noted by Judge Posner when considering whether the medical records of forty-five women who had received abortions where adequately protected because their direct identifiers had been “redacted:”

> “Some of these women will be afraid that … persons of their acquaintance, or skillful ‘Googlers,’ sifting the information contained in the medical records concerning each patient's medical and sex history, will put two and two together, ‘out’ the 45 women, and thereby expose them to threats, humiliation, and obloquy. As the court pointed out in *Parkson v. Central DuPage Hospital* … ‘whether the patients’ identities would *remain confidential by the exclusion of their names and identifying numbers is questionable at best*. The patients’ admit and discharge summaries arguably *contain histories of the patients' prior and present medical conditions, information that in the cumulative can make the possibility of recognition very high*.”[28]

In addition to including much of a patient’s medical history, such as a patient’s medical appointments, care plans, medical claims, medications, lab and radiology tests and results, history of psychiatric care, pregnancy care and dietary services,[29] medical records also often include a lot of demographic information that “in the cumulative can make the possibility of recognition very high,” such as the patient’s date of birth, gender, geography of residence, languages spoken and marital status,[30] as well as the patient’s birth place, adoption information, citizenship, nationality, disabilities, religion and places of religious

[26] Tanner, *How Data Brokers Make Money Off Your Medical Records*, *supra* note 2 (emphasis added).

[27] 42 U.S.C § 1320d–6 (emphasis added).

[28] Nw. Mem’l Hosp. v. Ashcroft, 362 F.3d 923, 929 (7th Cir. 2004) (emphasis added).

[29] *Specification 8.1, Resource Patient – Content*, HL7 FHIR RELEASE 4 (Nov. 1, 2019), https://www.hl7.org/fhir/patient.html.

[30] *Id*.

congregation.[31] They may also include demographic information and medical histories of the patient's family members, including family members' ages, locations and medical conditions.[32]

HIPAA recognizes this reality by protecting medical records that describe aspects of your life that *can be used* to identify you – such as what town you were born in, where you grew up, where you work, or when you were married or got divorced. To *de-identify* your medical records, therefore, your doctor must remove *all* the information that *can be used* to identify you, not just directly but also *indirectly*.

Tanner's confusion about HIPAA's requirements highlights a rarely discussed ambiguity in how the label "de-identified" is currently applied to medical information in the United States. On the one hand, there is HIPAA's definition of *de-identified*, which applies to health information that is devoid of identifiable information and, therefore, can be disseminated free of HIPAA's comprehensive data protection safeguards. This is the definition used by Federal agencies when they *de-identify* their own medical records for use by researchers and the public. It is also the definition discussed by the US Department of Health and Human Services (hereinafter referred to as *HHS*) and HHS's Office of Civil Rights (*OCR*) in their commentary and guidance regarding *de-identification*. Because HIPAA's form of *de-identified* information is devoid of identifying information that *can be used* to harm patients, HIPAA contemplates that it can be used and disseminated free of the restrictions HIPAA otherwise applies to *individually identifiable heath information*.[33]

In the private sector, on the other hand, certain permissive "de-identification guidelines" give parties significant flexibility in how they apply the label "de-identified." This flexibility can be stretched so far as to label information "de-identified" even in circumstances where there is a substantial risk that it *can be used* to identify *many*, *most* or *all* of the patients involved. Because this nominally "de-identified" information often *can be used* to identify patients, it presents the *same* risks to patients as *individually identifiable health information*. If the information is hacked, for example, it *could be used* to discriminate against the patients or blackmail them.[34] So it is not a coincidence that these permissive "de-identification guidelines" anticipate that their form of

[31] *Id.*; *Specification 8.1.16, Resource Patient – Extensions & Profiles*, HL7 FHIR RELEASE 4 (Nov. 1, 2019), https://www.hl7.org/fhir/patient-profiles.html.

[32] *Id.*; *Specification 94.1 Resource Family Member History – Content*, HL7 FHIR RELEASE 4 (Nov. 1, 2019), https://www.hl7.org/fhir/familymemberhistory.html#FamilyMemberHistory.

[33] *See* 45 C.F.R. § 164.502(d)(2) ("The requirements of [the Privacy Rule] do not apply to information that has been de-identified in accordance with the applicable requirements of § 164.514 …").

[34] Tanner, *The Hidden Global Trade in Medical Data*, *supra* note 2.

"de-identified" information will be secured in a manner that, in certain respects, echoes what HIPAA requires for *individually identifiable health information*.

Healthcare organizations and their data broker customers are notoriously secretive about their medical records transactions. "Pharmacies prefer not to announce that they sell their prescription information," Tanner notes, and "even [] employees are often unaware of the trade."[35] An industry insider notes that, "[i]t was forbidden to ever mention that topic … It was the big secret."[36] "The trade in patient data is so opaque," Tanner points out, "that many even in health care and government do not know about it."[37]

Given this conspicuous secrecy, it is impossible to say which healthcare organizations follow HIPAA's express requirements for *de-identification* and which utilize permissive "de-identification guidelines." Tanner's exposé raises a number of red flags suggesting HIPAA's requirements are not always followed.

In light of the confusion invited by applying the label "de-identified" to information that *can be used* to identify patients, it is paramount that regulators, compliance professionals, patient advocates and the general public understand the significant differences between the standards applied by HIPAA and those applied by permissive "de-identification guidelines." This Article discusses those differences in detail.

The discussion proceeds in four Parts. Part II (HIPAA's Heartbeat: Why HIPAA Protects Identifiable Patient Information) examines Congress's motivations for defining *individually identifiable health information* broadly, which included to stop the harms patients endured prior to 1996 arising from the commercial sale of their medical records. Part III (Taking the "I" Out of Identifiable Information: HIPAA's Requirements for De-Identified Health Information) discusses HIPAA's requirements for *de-identification* that were never intended to create a loophole for identifiable patient information to escape HIPAA's protections. Part IV (Anatomy of a Hack: Methods for Labeling Identifiable information "De-Identified") examines the goals, methods and results of permissive "de-identification guidelines" and compares them to HIPAA's requirements. Part V (Protecting Un-Protected Health Information) evaluates the suitability of permissive "de-identification guidelines," concluding that the vulnerabilities inherent in their current articulation render them ineffective as a data protection standard. It also discusses ways in which compliance professionals, regulators and advocates can foster accountability and transparency in the utilization of health information that *can be used* to identify patients.

[35] Tanner, *Strengthening Protection of Patient Medical Data*, *supra* note 2, at 6.

[36] Tanner, *How Data Brokers Make Money Off Your Medical Records*, *supra* note 2.

[37] Tanner, *The Hidden Trade in Medical Data*, *supra* note 2.

II. HIPAA'S HEARTBEAT: WHY HIPAA PROTECTS IDENTIFIABLE PATIENT INFORMATION

When Congress passed the Health Insurance Portability and Accountability Act of 1996,[38] it defined "individually identifiable health information" to include "*any* information … that … [either]:

> (i) *identifies* the individual; or
>
> (ii) with respect to which there is a reasonable basis to believe … *can be used* to identify the individual."[39]

Clause (i) covers *directly*-identifying information in your medical records – often called "direct identifiers" or "obvious identifiers" – your *name*, *addresses*, *social security number*, your *telephone number*, and the like.

Your medical history also includes other information that *could be used* to identify you. Even if your *direct identifiers* are removed, you might be identifiable from other information, such as where you were born, where you grew up, your education history, where you work, when or if you were married or divorced. For that reason, Congress included clause (ii) that applies to *any* other information that reasonably *can be used* to identify you.

The ability of this "other information" to identify people is widely recognized by researchers. Two decades ago, for example, Latanya Sweeney[40] authored a highly-cited study where researchers reported that 87% of the U.S. population were likely identifiable if you had access to their *gender*, *date of birth* and *zip code*.[41] These are sometimes called "indirect identifiers" because they only *indirectly* identify patients. On their own, *indirect* identifiers are benign. But when they are *combined* within one or more documents (such as a medical record or your consolidated medical history), they often *can be used* to identify the patient as easily as if you had her direct identifiers.

Gender, *date of birth* and *zip code* are included in *most* medical records.[42] But they are not the only categories of *indirect identifiers*. HHS recognized that a wide range of information *could be combined* to identify patients, as is the case

[38] Health Insurance Portability and Accountability Act of 1996, Pub. L. No. 104–191 (adding a new Part C to Title XI of the Social Security Act, comprising sections 1171–1179 of the Social Security Act, codified at 42 U.S.C. 1320d–8 (Aug. 21, 1996)).

[39] 42 U.S.C §1320d–6 (emphasis added).

[40] *See* Latanya Sweeney, *Simple Demographics Often Identify People Uniquely* (Carnegie Mellon Univ., Data Privacy, Working Paper 3, Pittsburgh 2000).

[41] *Id.* at 2.

[42] *See, e.g.*, *Specification 8.1.14, Resource Patient – Detailed Descriptions*, HL7 FHIR RELEASE 4 (Nov. 9, 2019), www.hl7.org/fhir/patient-definitions.html (demographics data includes, among other categories, gender, date of birth, home address, marital status).

when combining a patient's *gender*, *date of birth*, *zip code*, *languages spoken*, *race*, *diagnoses*, *dates of service*, among other variables:

> "It is not always obvious when information identifies the subject. If the name and identifying numbers (e.g., SSN, insurance number, etc.) are removed, a person could still be identified by the address. With the address removed, *the subject of a medical record could be identified based on health and demographic characteristics (e.g., age, race, diagnosis)*.[43]

Accordingly, "removing only the direct identifiers" has never been recognized as an effective means of *de-identifying* medical records because "the resulting information would often remain identifiable, and its dissemination could result in significant violations of privacy."[44] As HHS noted:

> "Congress [] intended to go beyond 'direct' identification and to encompass circumstances in which a reasonable likelihood of identification exists. Even after removing 'direct' or 'obvious' identifiers of information, a risk or probability of identification of the subject of the information may remain; in some instances, the risk will not be inconsequential."[45]

The government had compelling reasons to protect all forms of identifiable patient information. Before Congress enacted HIPAA in 1996, medical information was treated like any asset to be sold on the open market. A 1994 report, for example, found that *40%* of health insurers sold their patients' medical information to lenders, employers and marketers without the patients' permission, or even knowledge.[46]

Purchasers of patients' medical information often used that information to discriminate against those patients or otherwise harm them. A 30-year FBI veteran, for example, was placed on administrative leave because his pharmacy informed his employer that he was being treated for depression.[47] In a separate case, a banker used patient records to cancel mortgages of recently-diagnosed cancer patients.[48] In another case, a businessman purchased a medical practice and held its patient medical records ransom until the patients paid a bounty to

[43] Standards for Privacy of Individually Identifiable Health Information, 64 Fed. Reg. 59,918, 59,935 (Nov. 3, 1999) (to be codified at 45 C.F.R. pt. 160-164) [hereinafter the *1999 Proposed Rule*].

[44] Standards for Privacy of Individually Identifiable Health Information, 65 Fed. Reg. 82,462, 82,708 (Dec. 28, 2000) (to be codified at 45 C.F.R. pt. 160-164) [hereinafter *2000 Final Rule*].

[45] *Id.* at 82,611.

[46] *Id.* at 82,468.

[47] *Id.*

[48] *Id.*

purchase their medical records.[49] These were not isolated incidents. A 1990 survey, for example, found that *35%* of Fortune 500 companies reviewed "... people's medical records before making hiring and promotion decisions."[50]

The commodification of medical information put patients in a "catch-22." On one hand, a patient cannot obtain effective medical care without being candid with their doctors, hospitals and pharmacists. On the other hand, the price of that candor could result in the patient losing her job, a mortgage, or suffering other forms of discrimination or personal embarrassment when her medical records were sold to current or future employers, banks and whoever else was interested in her medical history.

The harms caused by the free flow of patient health information were so widespread, they became a recognized public health concern:

> A 1993 Lou Harris poll found that 75 percent of those surveyed worry that medical information from a computerized national health information system will be used for many non-health reasons, and 38 percent are very concerned … An ACLU Poll in 1994 also found that 75 percent of those surveyed are concerned a 'great deal' or a 'fair amount' about insurance companies putting medical information about them into a computer information bank to which others have access.[51]

The lack of privacy protection reduced patients' "trust in the health care system and institutions that serve them."[52] As HHS noted, "[i]ndividuals cannot be expected to share the most intimate details of their lives unless they have confidence that such information will not be used or shared inappropriately."[53] In the environment where patients knew their health information was for sale to the highest bidder, many felt compelled to withhold critical information from the healthcare system:

> "[O]ne in six Americans reported that they have taken some sort of evasive action to avoid the inappropriate use of their information by providing inaccurate information to a health care provider, changing physicians, or avoiding care altogether."[54]

A 1999 study found that "[t]o protect their privacy and avoid embarrassment, stigma, and discrimination, some people withhold information

[49] *Id.* at 82,467.

[50] *Id.*

[51] *Id.* at 82,467 (*citing* Harris Equifax, HEALTH INFO. PRIVACY STUDY, at 2, 33 (1993)).

[52] *Id.*

[53] *Id.* at 82,467-68.

[54] *Id.* at 82,468.

from their health care providers, provide inaccurate information, doctor-hop to avoid a consolidated medical record, pay out-of-pocket for care that is covered by insurance, and—in some cases—avoid care altogether."[55]

The abuse was so widespread that clinicians often felt compelled to protect their patients by censoring what they recorded in their patients' medical records:

> "[T]he Association of American Physicians and Surgeons reported 78 percent of its members reported withholding information from a patient's record due to privacy concerns and another 87 percent reported having had a patient request to withhold information from their records."[56]

The lack of privacy protection had become a danger to public health, to which Congress responded to by passing HIPAA. The new law not only prohibited the commercial exploitation of medical information, but also included significant penalties for its violations. If, for example, a violation is committed with the intent to "sell, transfer or use" medical information "for commercial advantage, personal gain, or malicious harm," a violator can be imprisoned for up to 10 years and fined up to $250,000.[57] Congress tasked HHS with implementing and monitoring specific rules on how medical information can be appropriately used, maintained and disclosed that are today memorialized in four significant and holistic data protection regulations known as HIPAA's Privacy Rule,[58] Security Rule,[59] Breach Notification Rule, [60] and Enforcement Rule.[61] All four of these regulations protect health information that can be used to identify a patient – whether directly or *indirectly*.

III. Taking the "I" Out of Identifiable Information: HIPAA's Requirements for De-Identified Health Information

A. Defining Information that Does Not Need to be Protected by HIPAA

Although HIPAA sought to curb the sale of patient health information to data brokers, it did not intend to bar legitimate uses of health information that present no risk to patients. When a hospital, for example, notifies the public of the

[55] *Id*. (*citing Best Principles for Health Privacy*, Health Privacy Working Group (Jul. 1999)).

[56] *Id*. at 82,468.

[57] *See* 42 U.S.C. § 1320d–6(b)(3).

[58] *See Privacy Rule*, *supra* note 16.

[59] *See Security Rule*, *supra* note 16.

[60] *See Breach Notification Rule*, *supra* note 16.

[61] *See Enforcement Rule*, *supra* note 16.

number of COVID-19 patients it has treated in the previous three months, that number is incapable of identifying any of the patients. The Privacy Rule sought to authorize such benign uses of medical information by introducing the concept of *de-identified health information*.

The purpose of HIPAA's definition of *de-identified health information* is to describe a category of health information that can be freely disseminated by healthcare organizations without restrictions, wholly unprotected by HIPAA's comprehensive data protection requirements.[62] Because this category of information is intended to be unprotected by HIPAA's Privacy Rule, Security Rule or Breach Notification Rule, the *de-identification process* itself is responsible for safeguarding patient privacy. This can only be accomplished by removing all information that *can be used* to identify those patients. So long as the information *cannot be used* to identify any of the patients involved, it also *cannot be used* to *harm* those patients.

Section 514(a) of the Privacy Rule defines "de-identified health information" as "[h]ealth information that does not identify an individual *and* with respect to which there is no *reasonable basis to believe … can be used* to identify an individual."[63] This definition is in harmony with HIPAA's definition of *individually identifiable information*, which is "*any* information … that … [either]: (i) *identifies* the individual; or (ii) with respect to which there is *a reasonable basis to believe … can be used* to identify the individual." [64]

The phrase "reasonable basis to believe" appears in both definitions. HIPAA's implementation specifications for *de-identification* – stated in Sections 514(b) and (c) of the Privacy Rule – give healthcare organizations procedures for adhering to Section 514(a)'s standard. Those procedures start with the removal of *all direct identifiers*. If a healthcare organization wants to replace those direct identifiers with *identification codes* – what Tanner called a "anonymous patient identification key" – then it must comply with the requirements of Section 514(c). After the *direct identifiers* have been removed or replaced with *identification codes*, the healthcare organization must comply with Section 514(b) to confirm that it has removed all other information that *could be used* to identify any of the patients.

[62] *See* 45 C.F.R. § 164.502(d)(2) ("The requirements of [HIPAA's Privacy Rule] do not apply to information that has been de-identified in accordance with the applicable requirements of § 164.514 …").

[63] *Standard: de-identification of protected health information*, 45 C.F.R. § 164.514(a).

[64] 42 U.S.C §1320d–6 (emphasis added).

B. The Inherent Dangers of Identification Codes and HIPAA's Safeguards Against Them

Under HIPAA, *de-identification* starts with the removal of *direct identifiers*. In many instances, there is no need to use *identification codes*. If a hospital, for example, wants to release statistics about the number of COVID-19 patients it has treated in the preceding three months, because this information is simple statistics, there is no need to replace *patient identifiers* with *identification codes*.

There are circumstances, however, where using *identification codes* is called for. A researcher studying the comparative effectiveness of a particular drug may need to look at relevant information regarding each patient in the study. This patient-level medical data is also known as "microdata." The healthcare organization could provide a *de-identified* version of this microdata in order to allow the researcher to conduct her study. If the study uncovers a surprising side effect of the drug, the researcher could alert the healthcare organization about that danger. The healthcare organization could then use the *identification code* to re-identify the patient's records and take appropriate steps to address the health risk to the patient.

Identification codes are very susceptible to compromise.[65] The entity that assigns the *identification codes*, for example, typically also possesses the *ability* to identify the patients represented by those codes. This could occur *directly*, by using the coding technology to *reverse* the *code* back into the patient's identity. But even if there are safeguards against this direct misuse, a party with the ability to utilize the coding technology could use it to identify patients *indirectly*. The party could accomplish this, for example, by using the technology to create *identification codes* for all people within a certain location, or even for the entire US. It could then compare those codes to the *identification codes* provided by the healthcare organization to unlock the identities of *all* of the healthcare organization's patients.

Any compromise of the *identification codes* results in a compromise of the patients' identities and the medical records associated with those codes.[66] In light of the inherent risks of *identification codes*, HIPAA is very cautious about their

[65] *See*, *e.g.*, Tanner, *The Hidden Trade in Medical Data*, *supra* note 2 ("… it turned out that a simple code could unlock the patients' national ID numbers.").

[66] 45 C.F.R. § 164.502(d) ("(i) Disclosure of a code or other means of record identification designed to enable coded or otherwise de-identified information to be re-identified constitutes disclosure of protected health information; and (ii) [i]f de-identified information is re-identified, a covered entity may use or disclose such re-identified information only as permitted or required by this subpart.").

use with *de-identified information*. Section 514(c)[67] places strict limitations on their utilization:

> A covered entity *may* assign *a code or other means of record identification to allow information* de-identified under this section *to be re-identified by the covered entity*, provided that:
>
> > (1) Derivation. The *code* or *other means of record identification* is *not derived from* or related to information about the individual *and is not otherwise capable of being translated so as to identify the individual*; and
> > (2) Security. The covered entity does *not use* or *disclose* the code or other means of record identification *for any other purpose*, and *does not disclose the mechanism for re-identification*.

Section 514(c) places four requirements on the implementation of *identification codes*. Each requirement corresponds to a way in which *identification codes* could result in compromising patients' identities:

1. The *identification code* must be *assigned by the healthcare organization.*[68] This is because the ability to assign the code leads to the ability to directly or indirectly identify the patients represented by the code.

2. The *identification code* must <u>not</u> be "derived from or related to information about the [patient], or "capable of being translated so as to identify the patient." [69] If the code were *capable* of being translated to identify a patient, this could easily result in the identification of the patient.

3. The healthcare organization cannot *disclose the mechanism for re-identification* to any third party [70] because disclosing that mechanism enables third parties to identify the patients. This echoes Section 502(d)(2)(i) of the Privacy Rule, which states that "[d]isclosure of a code or other means of record identification designed to enable coded or otherwise de-identified information to be re-identified constitutes disclosure of protected health information."[71]

[67] *Id.* at § 164.514(c).

[68] *Id.* ("A covered entity may assign a code or other means of record identification …").

[69] *See id.* § 164.514(c)(1) ("The code or other means of record identification is not *derived from* or *related to information* about the [patient] and is not otherwise capable of being translated so as to identify the [patient]; …") (emphasis added).

[70] *See id.* at § 164.514(c)(2) ("The covered entity does not … disclose the mechanism for re-identification.")

[71] *Id.* at § 164.502(d).

4. The healthcare organization can only use the identification code to *re-identify its own de-identified health information*.[72] This is because, if a healthcare organization gave a third party the ability to re-identify its patients, it would also be giving that party the ability to re-identify the patients' medical records.

In light of the notorious secrecy healthcare organizations and their data broker customers maintain regarding their medical records transactions, it is impossible to assess whether they are complying with Section 514(c). Tanner's exposé, however, raises a number of red flags. Recalling Tanner's discussion of the data brokers' software that replaces direct identifiers with identification codes, he says:

> [D]ata miners are able to match these files by getting pharmacies, insurers, testing labs, electronic health record systems, and other suppliers to all install the same de-identification software (for which they compensate the data suppliers).
>
> This software removes the personal details for each individual—such as name, address, telephone number, and Social Security number—but assigns that person the *same anonymous patient identification key* across all locations using that de-identification system. 'If they install that de-ID engine at every source and it has the same algorithm, *that means everyone with the same PHI (personal health information) will get the same IMS patient key*,' says Mark Degatano, who has advised IMS Health and worked at rival data miner Symphony Health.[73]

Based on this passage, it appears that healthcare organizations, rather than assigning the *identification codes* to their own patients' records, are allowing the software of the data brokers or the brokers' vendors to assign the *identifications codes* – that may be violating Section 514(c)'s first control.

It also appears that the healthcare organizations may be violating Section 514(c)'s fourth control. Rather than using their patients' *identification codes* solely for the purpose of re-identifying their *own* de-identified information, Tanner's passage suggests that healthcare organizations are using them to "sweeten the deal" for their data broker customers. The data broker customers want the *identification codes* so that *they* can build medical histories about each of the healthcare organizations' patients and track those patients throughout their lives. The healthcare organizations sell that ability by allowing the broker (or its vendor) to assign *identification codes* to the healthcare organization's patients identifiers.

[72] *See id.* at § 164.514(c)(2) ("The covered entity does not use or disclose the code or other means of record identification *for any other purpose* …") (emphasis added).

[73] Tanner, *Strengthening Protection of Patient Medical Data*, *supra* note 2, at 11.

The passage gives fewer indications regarding Section 514(c)'s second and third controls. Use of the term "patient identification key" suggests that the broker, or its agent who operates the "de-ID engine," may be using cryptographic algorithms to create the *identification codes*. As HHS has made clear, however, the use of cryptographic algorithms to create *identification codes* does not comply with HIPAA if the implementation otherwise violates Section 514(c).

In discussing a proposed use of a cryptographic hashing technique known as "Keyed-Hash Message Authentication Code" (or *HMAC*), for example, HHS rejected implementations that allowed a data broker to track patients over time:

> "… it appears *the key is shared with or provided by the recipient of the data in order for that recipient to be able to link information about the [patient] from multiple entities or over time*. Since the HMAC allows identification of individuals by the recipient, disclosure of the HMAC violates the [Privacy] Rule." [74]

The use of the cryptographic technology *per se* wasn't the problem. If the implementation of the HMAC *identification codes* complied with Section 514(c)'s requirements, it would be permissible. The proposed implementation, however, was configured to allow the "recipient to be able to link information about [patients] from multiple [sources] over time."[75] This can only be accomplished either if the healthcare organization shares the cryptographic key with the recipient, or if the recipient (or its agent) provides the key for creating the HMAC *identification codes*. Either case violates Section 514(c) and results in giving the recipient the ability to identify the patients represented by those *identification codes*.

Tanner's work hints at another way *identification codes* can be compromised. The reason data brokers want the patients' *identification codes* is to create comprehensive profiles or "dossiers" *about* those patients. "Data scientists," however, are capable of "… marrying anonymized patient dossiers with named consumer profiles available elsewhere – with a surprising degree of accuracy."[76] This capability, in turn, lets the broker compromise the *identification codes* without even needing to compromise the cryptographic algorithm used to create them.

In the Sweeny study, for example, researchers indicated that 87% of the U.S. population could be identified using only their *gender*, *date of birth* and *zip*

[74] Standards for Privacy of Individually Identifiable Health Information, 67 Fed. Reg. 53,233 (Aug. 14, 2002) [hereinafter the *2002 Final Rule*] (emphasis added).

[75] *Id.*

[76] Thielman, *supra* note 1.

code.[77] If a data broker possessed *identification codes* for 200 million patients that are linked to their *gender*, *date of birth* and *zip code*, the broker would possess the *ability* to compromise 174 million of those *identification codes*. For those 174 million *identification codes*, the broker's ability to link those codes to the correct patient is *100%*. Furthermore, the number of *identification codes* that join this "100% club" will inevitably increase with time as more and more medical information is added to the other 36 million codes.

This issue exists even if no single data source contains all of the necessary indirect identifiers, for example, where a broker has one dataset that only includes the patients' *gender* and *age* and then purchases a second dataset that has their *zip code*s. When a broker has the ability to use *identification codes* to link those two data sources, it also possesses the ability to use those codes to identify a vast majority of the patients represented by them by merging the patients' indirect identifiers.

Healthcare organizations that give data brokers enough *indirectly*-identifying information to identify a patient – whether that's in a single data file, or as a piece of the puzzle – appear to violate the spirit of Section 514(c)'s third control, if not the letter. By disclosing those indirect identifiers together with an *identification code* that the broker *can use* to link that patient's other indirect identifiers, the healthcare organization is disclosing a "mechanism for re-identification" that *can be used* to identify a substantial majority of its patients.[78]

C. Removing Indirect Identifiers

Removing direct identifiers is a first step in the process of *de-identification*, but it is far from sufficient to create *de-identified health information*. As previously discussed, "removing only the direct identifiers" is inadequate because "the resulting information would often remain identifiable, and its dissemination could result in significant violations of privacy."[79] Thus, if a healthcare organization removes "only the direct identifiers," the resulting information continues to be *individually identifiable health information* that must be protected in accordance with HIPAA's Privacy, Security and Breach Notification Rules.[80]

[77] *See* Sweeney, *Simple Demographics Often Identify*, *supra* note 35, at 2.

[78] *See* 45 C.F.R. § 164.514(c)(2) ("The covered entity does not … disclose the mechanism for re-identification.")

[79] *2000 Final Rule*, *supra* note 44, at 82,708.

[80] One example of this kind of health information is a *limited data set*. *Limited data sets* are created by removing 16 of the patients. *direct identifiers* from their medical records. *See* 45 C.F.R. § 164.514(e)(2) ("A limited data set *is* protected health information …"). Even though limited data sets contain no direct identifiers, they continue to be *individually identifiable health information*

Removing all of the potential *indirect identifiers* can be a daunting task. First, there are many pieces of information that could potentially be used to identify a patient. Sweeney and HHS discussed *gender*, *date of birth*, *zip code*, *race* and *diagnoses*, but medical records can include many other potential indirect identifiers, such as *place of birth*, *ethnic origin*, *religion*, *languages spoken*, *profession*, *event dates* (such as admission to a hospital, discharge, length of stay), *number of children*, *living parents*, *education*, and so on. Second, different combinations of indirect identifiers can be used to identify different types of people. Some individuals, for example, may be identifiable by combing their *dates of birth*, *religion* and *education*, while another group may be identifiable by *criminal history*, *languages spoken*, *location of birth*. This, in turn, leads to a third challenge – because individuals can be identified by combinations of indirect identifiers, it takes significantly more effort to address *all* of the potential combinations. Any of those combinations can be linked to or combined with *external* sources of information that can be used to identify the patients. As HHS noted:

> … the existence of *external sources of records* with matching data elements which *can be used to link with the de-identified information* and identify individuals (e.g., voter registration records or driver's license records). The *risk of disclosure increases as the number of variables common to both types of records increases, as the accuracy or resolution of the data increases, and as the number of external sources increases*."[81]

Notwithstanding this analytical complexity, there are *many* situations where a disclosure is *very unlikely* to identify a patient. When, for example, a hospital wants to inform the public that "last month we treated fifteen COVID-19 patients," it is very unlikely that statement *can be used* to identify any of the hospital's patients.

In light of these two aspects, Section 514(b) of HIPAA's Privacy Rule[82] gives healthcare organizations *two* options for confirming that a potential release of medical information is sufficiently devoid of information so that "there is *no reasonable basis to believe* [that it] *can* be used to identify [a patient]."[83]

The first, stated in Section 514(b)(1),[84] embraces the analytical approach, by allowing healthcare organizations to confirm that the information *cannot be*

that must be protected in accordance with HIPAA; *see id.*, because they contain *indirect identifiers* that can be used to identify patients.

[81] *2000 Final Rule*, *supra* note 44, at 82,709 (emphasis added).

[82] *See* 45 C.F.R. § 164.514(b).

[83] *Id.* at § 164.514(a).

[84] *Id.* at § 164.514(b)(1).

used to identify their patients using recognized statistical and scientific methods. HHS's model for this process is what Federal agencies use when they *de-identify* their records before distributing them to the public.

The second, specified in Section 514(b)(2),[85] helps healthcare organizations in more straightforward situations, such as if a hospital wants to make a public statement such as "Last month we treated 25 COVID-19 patients." It does this by giving healthcare organizations an easy-to-follow checklist.

1. Statistical Confirmation Method

a. The Language of Section 514(b)(1)

For health information to be "de-identified" in accordance with Section 514(b)(1), a healthcare organization must remove all patient-identifying information until:

> A person with *appropriate knowledge of and experience* with *generally accepted statistical and scientific principles and methods for rendering information not individually identifiable*:
>
> > (i) *Applying such principles and methods*, determines 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; *and*
> >
> > (ii) *Documents* the *methods and results* of the analysis that justify such determination …[86]

Section 514(b)(1), therefore, has *substantive* and *procedural* requirements.

Substantively, the healthcare organization must remove enough information from a patient's health record as to enable a qualified statistician to confirm, using "generally accepted statistical and scientific principles and methods," that the resulting information *cannot be used* to identify the patient. The statistician's assessment is an objective analysis based on "generally accepted statistical and scientific principles and methods." This analysis must take into account the ways the output information can be *combined* with *other available information* to identify patients. If, for example, a patient record includes her *gender*, *date of birth* and *zip code* that could be correlated with publicly available voting records, the statistician must assess the likelihood that the patient *could be identified* if her medical records and voting records were combined.

[85] *Id*. at § 164.514(b)(2).

[86] *Id*. at § 164.514(b)(1).

Procedurally, the qualified expert must document the statistical and scientific methods she used, her results, and the analysis she used to justify her results.

This two-step process mirrors how federal agencies confirm they have removed all identifying information before they disseminate reports or data to the public. To aid healthcare organizations in following the government's approach, HHS identified two key documents used by federal agencies.[87] The first is the STATISTICAL POLICY WORKING PAPER 22[88] (hereinafter referred to as *WORKING PAPER 22*) that describes the "generally accepted statistical and scientific principles and methods for rendering information not individually identifiable" used by federal agencies and described in clause (i) of Section 514(b)(1). With respect to the requirements of clause (ii), HHS identified the CHECKLIST ON DISCLOSURE POTENTIAL OF PROPOSED DATA RELEASES[89] (hereinafter referred to as the *CHECKLIST*), which describes the documentation federal agencies use to memorialize that they utilized the appropriate statistical methods before releasing it.

b. Very Small Risk of Identification

Under Section 514(b)(1), health information can be labeled "de-identified" *only if* there is a "very small" risk that it *could be used* to identify any of the patients represented by such information.[90] "Very small" is synonymous with "very low probability"[91] that is defined and measured by "generally accepted statistical and scientific principles and methods" described in WORKING PAPER 22.[92] It is *not* something that can be *selected*; nor is it a "judgement call" incapable of *bona fide* measurement in controlled settings. Rather, it is a threshold based on objectively established statistical or scientific principles for protecting

[87] *See 2000 Final Rule*, *supra* note 44, at 82,708 (discussing consultation with the Confidentiality and Data Access Committee, Federal Committee on Statistical Methodology, Office of Management and Budget and objectives for selection of guidance documentation for confirming health information is de-identified).

[88] SUBCOMMITTEE ON DISCLOSURE LIMITATION METHODOLOGY, FEDERAL COMMITTEE ON STATISTICAL METHODOLOGY, *Report on Statistical Disclosure Limitation Methodology*, (Office of Info. & Regulatory Affairs, Office of Mgmt. & Budget, Working Paper No. 22, 1994) [hereinafter Working Paper 22].

[89] CHECKLIST ON DISCLOSURE POTENTIAL OF PROPOSED DATA RELEASES, Interagency Confidentiality and Data Access Committee, Federal Committee on Statistical Methodology, Statistical Policy Office, Office of Info. & Regulatory Affairs, Office of Mgmt. & Budget (July 1999) [hereinafter CHECKLIST], https://nces.ed.gov/FCSM/doc/checklist_799.doc.

[90] 45 C.F.R. § 164.514(b)(1)(i)–(ii) (emphasis added).

[91] *2000 Final Rule*, *supra* note 44, at 82,709 ("In this context, we do not view the difference between a very low probability and a very small risk to be substantive.").

[92] *Id*.

the confidentiality of patients' identities against known risks. Because those risks will "change over time to keep up with technology and the current availability of public information from other sources,"[93] what counts as "very small" in one decade will be different from what constitutes "very small" in the next.

HHS's definition of *de-identification* is premised on the idea that information qualifying as "de-identified information" is no longer covered by HIPAA and can be disclosed to the public without any restrictions.[94] HHS's selection of WORKING PAPER 22 reflects this view because its methods are used by "federal agencies that routinely de-identify and anonymize information for public release."[95] As noted by WORKING PAPER 22, its techniques protect privacy solely by removing identifiable information until the outputs can safely be released to the public "without restrictions on use or other conditions."[96]

The level of certainty, therefore, is measured by the understanding that being "wrong" will violate the law and result in a data breach. For example, under 42 U.S.C. §1306(a), it is illegal under most circumstances to release any portion of a tax return filed with the Internal Revenue Service (*IRS*).[97] Thus, when the IRS releases statistical information based on tax returns it has collected, it must remove all information that would violate Section 1306(a). If the IRS's process for "de-identifying" a tax return resulted in releasing identifiable information to the public, that would be a violation of the law. Before releasing "de-identified" tax return information, therefore, the IRS must have a very high level of confidence that none of the information disclosed will result in a violation of Section 1306(a). Similarly, to cryptography's concept of "infeasibility," the risk does not need to be *zero*. But it must be small enough, based on known vulnerabilities, to provide a very high level of confidence based on generally accepted statistical principles and methods. [98]

[93] *Id.*

[94] *See* 45 C.F.R. § 164.502(d)(2) ("The requirements of [HIPAA's Privacy Rule] do not apply to information that has been de-identified in accordance with the applicable requirements of § 164.514 …").

[95] *2000 Final Rule*, *supra* note 44, at 82,709.

[96] Working Paper 22, *supra* note 88, at 3 ("The statistical disclosure limitation techniques described in this paper are applied whenever confidentiality is required and data or estimates are to be publicly available.").

[97] 42 U.S.C. § 1306(a)(1) ("No disclosure of any return or portion of a return (including information returns and other written statements) filed with the Commissioner of Internal Revenue … shall be made except as the head of the applicable agency may by regulations prescribe and except as otherwise provided by Federal law. Any person who shall violate any provision of this section shall be deemed guilty of a felony …").

[98] For a discussion of the measure of "infeasibility" for approved cryptographic hash algorithms, *see* discussion of "Hash Function Properties" in NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY, SPECIAL PUBLICATION 800-107 REVISION 1 RECOMMENDATION FOR APPLICATIONS

This overlaps with HIPAA's obligations under its Breach Notification Rule. Under the Rule, a "breach" is defined as "the acquisition, access, use, or disclosure of [identifiable] health information in a manner not permitted under [the Privacy Rule] which compromises the security or privacy of [such identifiable] health information."[99] Under the Breach Notification Rule, this is "*presumed to be a breach unless* the [healthcare organization] *demonstrates* that there is a *low probability* that the [identifiable] health information has been compromised *based on a risk assessment*" of four factors.[100] The *first* of those factors is the "nature and extent" of the information involved. [101] If the healthcare organization has removed patient identifiers, it must consider the remaining identifiers and "the likelihood of re-identification."[102] If, for example, those medical records contain *indirect identifiers* that could not be safely released to the public, this is *individually identifiable patient information* and the healthcare organization must perform the required risk assessment and/or warn patients that their information has been released to an unauthorized party.

HHS's own *de-identification* practices offer a useful illustration. In the late 2000's, HHS's Office of the National Coordinator for Health Information Technology (hereinafter referred to as *ONC*) commissioned a research team to attempt to re-identify the health records of approximately 15,000 individuals that had the 18 identifiers removed in accordance with Section 514(b)(2), the so-called "safe harbor" method of de-identification. The research team compared those records with consumer data provided by a national data broker and was able to re-

USING APPROVED HASH ALGORITHMS, Computer Security Division, Information Technology Laboratory, National Institute of Standards and Technology (Aug. 2012) [hereinafter *NIST SP 800-107*], 6-9. Cryptographic algorithms protect the confidentiality of personal information by rendering it indecipherable unless you have the "key" to unlock it. All deterministic algorithms, however, can be compromised. If an attacker has an infinite amount of time and computing power, she could break the algorithm simply by guessing new passwords until she unlocks the personal information. This inherent vulnerability is known as a "brute force attack" or "exhaustive key search," and it is a fundamental risk that all cryptographic algorithms are measured against. This risk is objectively measured as a function of (i) the number of mathematical operations it takes to make each guess, multiplied by (ii) the number of mathematical operations that current computers are capable of processing in a given time period. Based on these metrics, cryptographers have defined "infeasibility" – their version of "very small" or "very low probability" in terms of the physical limits on what an attacker could utilize based on current technologies. To be able to compromise an algorithm based on a brute force attack, it should take many years of guesses before the attack would be successful. Converted into probabilities, the chance that an attacker succeeds on her first "lucky guess" is measured in the range of less than *one-in-a-trillion trillion trillion*. This probability is *not* zero. Rather, it is a *very low probability* that is sufficiently robustly that it accomplishes the intended goal against known vulnerabilities.

[99] *See* 45 C.F.R. § 164.402.

[100] *See* 45 C.F.R. § 164.402 (emphasis added).

[101] *Id.* (emphasis added).

[102] *Id.* (emphasis added).

identify two of approximately 15,000 individuals, or *0.013%* of the population.[103] As low as this percentage is, it has not been recognized as an "acceptable error rate" when it comes to HHS's own *de-identified health information*. A release of the records of the two patients' medical records would constitute a data breach because their entities have been compromised.

Accordingly, when HHS's Centers for Medicare and Medicaid Services (hereinafter referred to as *CMS*) releases *de-identified* medical claims data for its beneficiaries in its "public use files," CMS utilizes the full gamut of disclosure limitation techniques described in WORKING PAPER 22,[104] including:

- Drastically reducing the number of variables in a claim record;
- Suppressing rare diagnosis and procedure codes;
- Substituting claims from donor beneficiaries using an actual beneficiary as the seed, or pattern, for the synthetic beneficiaries; donor claims were found using a key variable from the seed and donor claims;
- Restricting the amount of information coming from any one donor and always using multiple donors; a minimum of three donors contributing to each single synthetic beneficiary claim set;
- Synthesizing secondary variable sets within the donated claims conditioned on key variables, for added disclosure protection;
- Perturbing various claim dates by altering the start date of the claim set used as the seed and proportionally altering the number of days between claims;
- Coarsening expenditure variables, so that larger values were coarsened into larger bins, and truncating both tails of the distribution (top and bottom);
- Synthesizing provider information (institution and physician) by drawing from empirical distribution conditioned on the synthesized geography of the beneficiary;
- Suppressing rare combinations of institution and physician codes from the data used to create synthetic claims. [105]

In order for information to be considered *de-identified* under Section 514(b)(1), therefore, the application of *bona fide* statistical and scientific methods must be able to objectively demonstrate that there is a *very low probability* that the resulting information *can be used* to identify patients by any

[103] *See* Deborah Lafky, *The Safe Harbor Method of De-Identification: An Empirical Test*, ONC PRESENTATION (Oct. 9, 2009) [hereinafter *The Safe Harbor Method: An Empirical Test*], http://www.ehcca.com/presentations.

[104] *See* USER MANUAL CENTERS FOR MEDICARE AND MEDICAID SERVICES (CMS) LINKABLE 2008-2010 MEDICARE DATA ENTREPRENEURS' SYNTHETIC PUBLIC USE FILES (DE-SYNPUF) (2013), https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/SynPUFs/Downloads/SynPUF_DUG.pdf.

[105] *Id*. at 14-15.

party who is not authorized to obtain identifiable patient information under HIPAA's Privacy Rule. Unless that can be demonstrated, the information in question does not satisfy Section 514(b)(1)'s standard for *de-identification*.

c. Anticipated Recipients

Although HHS defines *de-identified information* as information that can be safely disclosed to the public free of all restrictions in a manner contemplated by Section 502(d)(2) of HIPAA's Privacy Rule,[106] HIPAA does not *require* every instance of *de-identified information* to be released to the public. Healthcare organizations can release *de-identified information* to specific entities rather than the public. Section 514(b)(1) calls any entity anticipated to receive that medical information an "anticipated recipient."

The reference to "anticipated recipient" reflects important practical differences between the ways in which *de-identified medical information* is released to the public versus private parties. Healthcare organizations often release information to the public on their own initiative. A hospital's report regarding its COVID-19 procedures is an example of such a release, and its motivation is to provide the public with information about an important public health issue.

In contrast to public reports, *private* data releases are often initiated by the *recipient* and are intended to benefit the recipient, not the public, the healthcare organization, or any of its patients. A data broker or researcher, for example, may approach a healthcare organization and offer to pay substantial premiums for granular patient medical records that go far beyond what the healthcare organization would release on its own initiative. A data broker, for example, may be looking for the *full medical records* – including specific diagnoses, medications and even the doctors' notes – about tens of thousands of patients with an eye to analyzing that data or *combining it* with other data about those patients.

Because these requests are often unique, Section 514(b)(1) requires the statistician's analysis to address the unique aspects of information the data broker – the *anticipated recipient* – is requesting. The analysis must also take into consideration what additional information the data broker may have at its disposal that *could be used* to identify any of the patients. If, for example, the broker is a consumer reporting agency with access to significant volumes of detailed personal information about the patients whose medical records are being provided, the analysis must take that into account when applying the appropriate techniques from WORKING PAPER 22. The statistician, for example, may conclude that the data needs to be fictionalized to a greater extent to ensure that the medical records

[106] *See* 45 C.F.R. § 164.502(d)(2) ("The requirements of [HIPAA's Privacy Rule] do not apply to information that has been de-identified in accordance with the applicable requirements of § 164.514 …").

being provided *cannot be linked or correlated* with any personal information the broker can access about those same patients.

2. The (*Semi*-) Safe Harbor Method

a. Redacting Identifiers in Accordance with Section 514(b)(2)(i)

The requirements of Section 514(b)(1) allow healthcare organizations to use accepted statistical methods to ensure that the health information they are releasing *cannot be used* to identify their patients; or, more precisely, to ensure that "there is no *reasonable basis to believe* [that the information] *can* be used to identify [their patients]." Those requirements, however, are extraordinarily burdensome if a hospital simply wants to confirm the number of COVID-19 patients it treated the previous month. For those cases, Section 514(b)(2) of the Privacy Rule[107] gives healthcare organizations a simpler alternative.

If a hospital wants to confirm whether a simple statement is *de-identified* – such as "Last month we treated 15 COVID-19 patients" – Section 514(b)(2) allows the hospital to presumptively confirm that this statement is *de-identified* by confirming that the following eighteen identifiers have been removed for the patient, as well as her relatives, employers and household members:[108]

(A) names;

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

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

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

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

(D) telephone numbers;

(E) fax numbers;

(F) electronic mail addresses;

(G) social security numbers;

(H) medical record numbers;

(I) health plan beneficiary numbers;

(J) account numbers;

[107] *Id*. at § 164.514(b)(2).

[108] *Id*. at §164.514(b)(2)(i).

(K) certificate/license numbers;
(L) vehicle identifiers and serial numbers, including license plate numbers;
(M) device identifiers and serial numbers;
(N) web Universal Resource Locators (URLs);
(O) Internet Protocol (IP) address numbers;
(P) biometric identifiers, including finger and voice prints;
(Q) full face photographic images and any comparable images; and
(R) any other unique identifying number, characteristic, or code, except as permitted by [Section 514(c)] […].

Given Section 514(b)(2)'s "paint-by-numbers" approach, it is frequently referred to as the "safe harbor" method of *de-identification*.

The use of the term "safe harbor" is misleading. Although the 18 identifiers offer a good rule-of-thumb for a wide range of routine disclosures, it is still possible for identifying information to slip through the cracks. If the hospital described above, for example, wanted to raise awareness about the infectiousness of COVID-19 by informing the public that even young, healthy athletes can contract it, it may be tempted to issue the following announcement:

> "Even young, healthy athletes can get COVID-19! Just last week we treated the most famous seven-foot baller who has three championship rings to his name! If he can get it, so can *you*!"

This statement does not include any identifiers listed in Section 514(b)(2)(i). Nevertheless, the statement likely *could be used* to identify the patient either on its own, or in combination with other information.[109] Members of the public, for example, could look up which athletes are seven feet tall, and infer that the hospital is referring to a basketball player. They could then review sports websites to find out which seven-footers have won three championships and see what teams play in the vicinity of the hospital.

This situation is easy to detect in isolated statements. But it is often more difficult to detect if the healthcare organization intends to release individual medical records for thousands of individuals. Some of the more famous breaches of identifiable patient information have occurred from the release of medical records that complied with the redaction requirements of Section 514(b)(2)(i).

A famous example occurred when Professor Sweeney identified Massachusetts Governor, William Weld, using hospital records that had been redacted in accordance with Section 514(b)(2)(i).[110] This was followed by a wave

[109] *Id*. at §164.514(b)(2)(ii) ("The [healthcare organization] 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.").

[110] Latanya Sweeney, *Weaving Technology and Policy Together to Maintain Confidentiality*, 25 J.L. MED. & ETHICS 99 (1997).

of studies that produced similar results. In a 2015 study, for example, researchers identified patients in purportedly de-identified hospital records released by the State of Washington[111] that described incidents of venereal diseases, drug dependency, alcohol use, tobacco use.[112] Similar studies successfully identified patients from redacted hospital data from California, Maine and Vermont.[113]

Section 514(b)(2)(i)-redaction is particularly vulnerable where a patient's longitudinal records are being assembled. In the ONC study described above, the research team was able to identify approximately *0.013%* of the population.[114] When researchers have a chance to compare a patient's medical records over a period of time, the percentages can rise dramatically. A 2013 study examined this phenomenon through evaluating the identification risks associated with a database of biometric information that had been redacted in accordance with Section 514(b)(2)(i).[115] The study found that the ability to identify patients represented in the database increased dramatically when researchers could compare it to the patient's longitudinal health information.[116] When researchers could utilize the known results for *four* consecutive PCV panels, for example, they had *19.5%* chance of uniquely identifying a patient in the redacted biomedical database.[117] When researchers had access to *six* consecutive panels, the rate jumped to *89%*.[118]

For these reasons, redacting medical records in accordance with Section 514(b)(2)(i) is *not* an effective means for confirming that "there is no *reasonable basis to believe* [that it] *can* be used to identify [a patient]" when the healthcare organization has "actual knowledge that the information could be used alone or in combination with other information to identify an individual who is a

[111] Latanya Sweeney, *Only You, Your Doctor, and Many Others May Know*, TECH. SCI. (Sept. 29, 2015), https://techscience.org/a/2015092903/.

[112] *Id.*

[113] *See*, *e.g*., Latanya Sweeney et al., *Re-identification Risks in HIPAA Safe Harbor Data: A study of data from one environmental health study*, TECH. SCI. (Aug. 28, 2017), https://techscience.org/a/2017082801/ (noting data from California); Ji Su Yoo et al., *Risks to Patient Privacy: A Re-identification of Patients in Maine and Vermont Statewide Hospital Data*, TECH. SCI. (Oct. 8, 2018), https://techscience.org/a/2018100901/ (noting data from Maine and Vermont).

[114] *See* Lafky, *supra* note 103, at 19.

[115] *See* Ravi V. Atreya et al., *Reducing patient re-identification risk for laboratory results within research datasets*, 20 J. AM. MED. INFORMATICS ASS'N 95 (2013), https://pubmed.ncbi.nlm.nih.gov/22822040/.

[116] *Id.* at 98.

[117] *Id.*

[118] *Id.*

subject of the information."[119] Thus, even if all 18 of the required identifiers have been removed, the data is not *de-identified* if the healthcare organization knows that the information identifies one or more individuals.

b. Healthcare Organizations' Duty to Be Informed About Known Risks to their Patients' Confidentiality under HIPAA's Security Rule

Section 514(b)(2)(ii)'s reference to the phrase "actual knowledge" can be confusing. "Actual knowledge" is a legal term that describes the mental state of an individual or an organization. For an individual, the phrase refers to the contents of her mind – what specific facts or information was she consciously aware of? Given the diversity of human experience, those contents can vary from person to person. Accordingly, someone's "actual knowledge" cannot be presumed. There are many situations where a person may be ignorant of facts that are well-known to others. The legal term "actual knowledge" is focused on a person's specific state of mind, *not* what that person *should* have known, regardless of how commonly known or easy to learn the facts in question are.[120]

Assessing an *organization's* actual knowledge is trickier still. Organizations are made up of many individuals, often numbering in the hundreds, thousands or tens of thousands. How does one assess what a company of 10,000 employees "actually knows?" Does the company "know" a fact if a single employee is aware of that fact? Or if a majority of employees do? Or if certain categories of employees know, such as corporate officers or employees with specific job responsibilities?

HIPAA's Security Rule and Breach Notification Rules answer some of these questions by *requiring* healthcare organizations to *know* about potential risks and harms to patient confidentiality. Each is particularly focused on uses or disclosures of patient information that are not permitted under HIPAA's Privacy Rule. Healthcare organizations cannot choose to be ignorant about those risks and harms.

HIPAA's Security Rule, for example, requires healthcare organizations to protect against "any reasonably anticipated uses or disclosures" of their patient information that would not be permitted by the Privacy Rule.[121] This includes protecting against any "reasonably anticipated threats or hazards to the security

[119] 45 C.F.R. § 164.514(b)(2)(ii).

[120] *See*, *e.g.*, Intel Corp. Inv. Policy Comm. v. Sulyma, 140 S. Ct. 768, 776 (2020) ("Legal dictionaries give 'actual knowledge' the same meaning: '[r]eal knowledge as distinguished from presumed knowledge or knowledge imputed to one.'; BALLENTINE'S LAW DICTIONARY 24 (3d ed. 1969); accord, BLACK'S LAW DICTIONARY 1043 (11th ed. 2019) (defining 'actual knowledge' as '[d]irect and clear knowledge, as distinguished from constructive knowledge'). The qualifier 'actual' creates that distinction.").

[121] 45 C.F.R. § 164.306(a)(3).

[...] of such information."[122] In furtherance of this obligation, healthcare organizations must conduct "an *accurate* and *thorough* assessment of the *potential risks* and *vulnerabilities* to the confidentiality" of the health information they hold.[123] Healthcare organizations, therefore, are required to *inform* themselves *thoroughly* about potential risks and vulnerabilities to the confidentiality of their patients' information.

To aid healthcare organizations in conducting such risk assessments, HHS commends special publications from the National Institute of Standards and Technology (hereinafter referred to as *NIST)*, [124] such as NIST Special Publication 800-30 (2002).[125] Documenting all potential human, environmental and natural threats is a significant undertaking. NIST SP 800-30 (2002) aids organizations in creating an inventory by listing examples of such threats ranging from corporate espionage and criminal hacking, on one hand, to poorly trained employees or agents, on the other.[126] NIST's guidance regarding compromises of *de-identified information* is located in NIST Interagency Report 8053 (hereinafter referred to as *NIST IR 8053)*.[127] NIST IR 8053 identifies a number of such risks that could result in the identification of individuals, such as patients, which it calls "re-identification attacks."[128] As with any other human threat, the motivations for re-

[122] *Id.* at § 164.306(a)(2).

[123] *Id.* at § 164.308(a)(1) (emphasis added).

[124] *See*, *e.g.*, HIPAA Security Rule: Health Insurance Reform: Security Standard, 68 Fed. Reg. 8,334, 8,346 (Feb. 20, 2003) (to be codified at 45 C.F.R. pt. 160, 162, 164), https://www.cms.gov/Regulations-and-Guidance/Regulations-and-Policies/QuarterlyProviderUpdates/downloads/cms0049f.pdf; OFFICE OF CIVIL RIGHTS, GUIDANCE ON RISK ANALYSIS REQUIREMENTS UNDER THE HIPAA SECURITY RULE 1, 3–4 (2010), https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/administrative/securityrule/rafinalguidancepdf.pdf; DEP'T OF HEALTH AND HUMAN SERVS., HIPAA SECURITY SERIES: BASICS OF RISK ANALYSIS AND RISK MANAGEMENT 3–5, (2005, rev. 2007), https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/administrative/securityrule/riskassessment.pdf.

[125] NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY, RISK MANAGEMENT GUIDE FOR INFORMATION TECHNOLOGY SYSTEMS, NIST SPECIAL PUBLICATION 800-30 (2002) (superseded by NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY, GUIDE FOR CONDUCTING RISK ASSESSMENTS, NIST SPECIAL PUBLICATION 800-30 REVISION 1 (2012)), https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-30.pdf.

[126] *Id.* at 14 (listing categories of threats, motivations and threat actions).

[127] *See*, *e.g.*, Simson L. Garfinkel, NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY, DE-IDENTIFICATION OF PERSONAL INFORMATION, NIST INTERNAL REPORT 8053 (2015), https://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf.

[128] *Id.* at 9.

identification attacks vary. NIST IR 8053[129] lists six threat motivations that overlap with those described in NIST SP 800-30.[130]

The ongoing risk assessments required by the Security Rule have important implications for the level of "actual knowledge" that healthcare organizations are required to possess. Their "actual knowledge" must be informed by the "*accurate* and *thorough* assessment[s] of the potential risks and vulnerabilities" they must conduct[131] in order to protect the health information in their possession from "any reasonably anticipated threats or hazards[…]."[132] These assessments should document the types of risks, threats and hazards described by NIST in its various publications, including NIST SP 800-30 and NIST IR 8053. The Security Rule does not permit healthcare organizations to simply "plead ignorance" or be "willfully blind" about well understood risks to the confidentiality of their patients' information. They cannot simply choose to be ignorant of the many potential vulnerabilities of *identification codes* or of the ways longitudinal health records can be compromised even if they are redacted in accordance with Section 514(b)(2)(i).

c. Healthcare Organizations' Duty to Be Informed under the Breach Notification Rule

HIPAA's Breach Notification Rule also requires healthcare organizations be informed about potentially harmful disclosures of their patients' health information. Under the Breach Notification Rule, a "breach" is defined as "the acquisition, access, use, or disclosure of [identifiable] health information in a manner not permitted under [the Privacy Rule] which compromises the security or privacy of [such identifiable] health information."[133] Section 402 of the Breach Notification Rule goes on to say that the "acquisition, access, use, or disclosure of [identifiable] health information in a manner not permitted under [the Privacy Rule] is *presumed to be a breach unless* the [healthcare organization] *demonstrates* that there is a low probability that the [identifiable] health information has been compromised based on a risk assessment of four factors." [134] The first of those factors is assessing the "nature and extent of the [identifiable]

[129] *Id.*

[130] *Id.* at 10 (discussing a range of threats, including entities attempting to harm the individuals re-identified, financial gain).

[131] *Id.* at § 164.308(a)(1).

[132] *Id.* at § 164.306(a)(2).

[133] *Id.* at § 164.402(1).

[134] *Id.* at § 164.402(2).

health information involved, including the types of identifiers and *the likelihood of reidentification*."[135]

Even if the health information has had *identifiers* removed, any disclosure of that information *is a breach unless* the covered entity can demonstrate that either (i) there is a low probability that such health information is re-identifiable; or (ii) if it is re-identifiable, that such health information has only been provided to entities otherwise entitled to receive it under the Privacy Rule. Unless the healthcare organization can demonstrate that that release presents a low probability of re-identification by the recipient, patients must be warned that their information has been released to an unauthorized party so that they can protect themselves from the potentially harmful impact.[136]

As with the Security Rule, the Breach Notification Rule requires healthcare organizations to be *knowledgeable* about relevant risks when health information is released or used in ways that would violate the Privacy Rule. Under Section 402, it is *presumed* that any disclosure of health information that would not be permitted by the Privacy Rule is a "breach" – even for health information that has had identifiers removed – *unless* the covered entity (or business associate) can demonstrate that there is a low probability of compromise.[137] That documented demonstration must include reviewing the identifiers that have been removed from the health information and assessing the likelihood that such information – even with the removed identifiers – can nevertheless be identified.[138] The Breach Notification Rule, therefore, requires that healthcare organizations possess this level of "actual knowledge" when they release health information for purposes that would otherwise violate HIPAA's Privacy Rule.

D. When De-Identified Information is Unsuitable for a Research Study

Not every research study can be conducted using *de-identified* patient information. If a study protocol requires analyzing patients' *gender*, *dates of birth* and *zip codes*, for example, Sweeney's study indicates that approximately 87% of those patients will be identifiable. If the study requires analyzing 10,000 patients, the researcher would obtain identifiable patient information for at least 8,700 of those patients.

[135] *Id.* at § 164.402(2)(i) (emphasis added).

[136] *See*, *e.g.*, *id.* § 164.404(a) ("A covered entity shall, following the discovery of a breach of unsecured protected health information, notify each individual whose unsecured protected health information has been, or is reasonably believed by the covered entity to have been, accessed, acquired, used, or disclosed as a result of such breach.")

[137] *Id.*

[138] *Id.*

This problem is well-understood by federal agencies. As noted in WORKING PAPER 22, "[a]ll disclosure limitation methods result in some loss of information, and sometimes the publicly available data may not be adequate for certain statistical studies."[139] As further noted by the Confidentiality and Data Access Committee (hereinafter referred to as *CDAC*), "[f]requently, the results of the Checklist or other considerations, will mean that a file cannot be released for public use, yet there is an acute need in the research community for detailed data."[140]

In these situations, the solution is *not* for the researcher or agency to "stretch" the definition of *de-identification*. If the information *can be used* to identify individuals, it *is* identifiable information regardless of the interests or motivations of the researchers or the agency's desires to release that identifiable information.

Instead, the answer is to acknowledge the identifiability of the information in question and to impose restrictions on how the researcher can access that information in accordance with applicable law. CDAC refers to these restrictions as "Restricted Access" arrangements.[141]

Broadly speaking there are two forms of "Restricted Access." CDAC calls the first "Licensing," in which a "researcher must sign an agreement with the agency which permits the installation of the restricted data on their computer in return for meeting the agency's conditions relating to maintaining confidentiality of the data."[142] The contract a researcher must sign includes the following types of provisions:

- Demonstration of a need for detailed data;
- Designation of those who have access;
- Statement of legal provision;
- Data security and enforcement/provision for inspection;
- Restrictions on use (prohibition against linking with other files);
- Restrictions on release of research results/adherence to agency policies; and
- Return/Destruction of data provided.[143]

[139] Working Paper 22, *supra* note 88, at 6.

[140] Alvan Zarate et al., *Paper presented at the FCSM Statistical Methodology Seminar: Integrating Federal Statistical Information and Processes*, at 5 (Nov. 8-9, 2000), https://nces.ed.gov/FCSM/pdf/CDAC_paper2000.pdf.

[141] *Id.* at 5.

[142] *Id.* at 6.

[143] *Id.*

The second form of "Restricted Access" supplements "Licensing" by limiting the physical access to identifiable information to a "Research Data Center," where the data is housed in a monitored environment that restricts the researcher's access in various ways, such as:

- Review of research protocol;
- Formal agreement covering work to be done, data used, and types of output;
- In-house files without identifiers;
- Limitations on types of analysis;
- No outside (linkable) data brought in by researcher;
- Dedicated computers;
- Review of data outputs;
- Inspection of material removed from site; and
- Physical presence of agency staff.[144]

HIPAA follows this general framework. When a research study cannot be effectively conducted with *de-identified health information*, the answer is *not* to stretch the definition of "de-identified." Instead, the answer is to acknowledge the *identifiability* of the information in question and to impose restrictions on how the researcher can access that information in light of the risks to patient privacy.

In the hypothetical study described above, for example, the research protocol required access to patients' *gender*, *dates of birth* and *zip codes* for approximately 10,000 patients. As per Sweeney's study, this would result in giving researchers access to identifiable information that *can be used* to identify approximately 8,700 of those patients. Consequently, the researchers need access to *identifiable* patient information that must be protected as such.

In these cases, it may be possible for researchers to use what HIPAA calls a "limited dataset." Under HIPAA a *limited dataset* is created by removing all of the following sixteen *identifiers* about the patients, their relatives, household members and employers:

(i) Names;
(ii) Postal address information, other than town or city, state, and zip code;
(iii) Telephone numbers;
(iv) Fax numbers;
(v) Electronic mail addresses;
(vi) Social security numbers;
(vii) Medical record numbers;
(viii) Health plan beneficiary numbers;
(ix) Account numbers;
(x) Certificate/license numbers;

[144] *Id.*

(xi) Vehicle identifiers and serial numbers, including license plate numbers;
(xii) Device identifiers and serial numbers;
(xiii) Web Universal Resource Locators (URLs);
(xiv) Internet Protocol (IP) address numbers;
(xv) Biometric identifiers, including finger and voice prints; and
(xvi) Full face photographic images and any comparable images.[145]

There is a significant overlap between the identifiers removed for *limited datasets* and those removed under Section 514(b)(2)(i) under the "safe harbor" method of *de-identification*. However, limited datasets can include a patient's *date of birth*, whereas Section 514(b)(2)(i) requires that the healthcare organization only list the patient's *year* of birth.[146] Furthermore, Section 514(b)(2)(i) requires that *all* specific dates be removed from *de-identified health information*, which often makes it impossible to conduct studies looking at impacts over shorter time periods. *Limited datasets*, on the other hand, can include the full dates.

Because *limited datasets can be used* to identify patients, they are *individually identifiable health information* protected by HIPAA.[147] But HIPAA allows *limited datasets* to be used for research, public health and health care operations purposes.[148] They can also be disclosed to a researcher provided that the researcher enters into a "data use agreement" that conforms to the requirements of Section 514(e)(4)(ii).[149] This resembles CDAC's "Licensing" approach to "Restricted Access" by requiring researchers to enter into a "data use agreement" with a number of patient privacy protections. Researchers, for example, are required to report any unauthorized use or disclosure of the *limited dataset*.[150] Researchers also are not permitted to attempt to identify or contact any of the patients.[151] If the healthcare organization becomes aware of violations by the researcher, the healthcare organization *must* take steps to cure the violation. If

[145] *See* 45 C.F.R. § 164.514(e)(2)(i)-(xvi).

[146] *See id.* at § 164.514(b)(2)(i) ("The following identifiers of the individual … are removed: … All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death …").

[147] *See id.* at § 164.514(e)(2) ("A limited data set *is* protected health information …") (emphasis added).

[148] *See id. at* § 164.514(e)(3)(i).

[149] *See id*. at § 164.514(e)(4)(ii).

[150] *Id.* at § 164.514(e)(4)(ii)(C)(3).

[151] *Id*. at § 164.514(e)(4)(ii)(C)(5).

those steps are unsuccessful, the healthcare organization *must* terminate the relationship and *report the matter* to HHS.[152]

In situations where *limited datasets* are insufficient for the research in question, HIPAA provides additional pathways for healthcare organizations to make their *identifiable* patient information available. One is under Section 508 of the Privacy Rule that authorizes healthcare organizations and researchers to use health information for research purposes pursuant to a patient authorization.[153] This is a common pathway in clinical research involving human subjects that would be used to test a drug or medical device. Such research is often conducted in accordance with the FDA's "Common Rule," which governs how federally funded clinical research is conducted, and is the *de facto* standard for clinical research in the United States that involves human subjects.[154] In such research, "valid authorizations" are often obtained at the same time that a patient executes the "informed consent" required under the Common Rule.

When the research requires the health information of a very large number of patients and obtaining valid authorizations is not feasible, Section 512(i) of the Privacy Rule authorizes healthcare organizations to conduct research under the supervision of an Institutional Review Board (also called an *IRB*) or a Privacy Board. These IRBs or Privacy Boards provide independent oversight over the research activity in question, including the researcher's data management practices.[155]

HHS's CMS utilizes three of these approaches for releases of its Medicare claims information. CMS releases properly de-identified health information in its "Public Use Files." If a researcher has a *bona fide* need for identifiable patient information for her research question, CMS gives researchers two options. One is to obtain access to CMS's *limited dataset* files that it has prepared in accordance with Section 514(e)(2).[156] In situations where the research requires the fully-identifiable patient information, the researcher can request CMS's "Research Identifiable Files" (or *RIFs*). In conformance with Section 512(i) of the Privacy Rule, access to RIFs is "reviewed by CMS's Privacy Board to ensure that the

[152] *Id.* at § 164.514(e)(4)(iii).

[153] *Id.* at § 164.508(a)(1).

[154] The "Common Rule" is a familiar name for subpart A of Federal Policy for the Protection of Human Subjects, codified at 45 C.F.R. part 46. The purpose of the Common Rule is to protect human research subject who have agreed to participate in government funded research. The main elements of the Common Rule are requirements for researchers obtaining and documenting informed consent, and compliance requirements for research stakeholders, including research institutions and Institutional Review Boards.

[155] *See* 45 C.F.R. § 164.512(i).

[156] *See*, *e.g.*, *id.*

beneficiaries' privacy is protected and only the minimum data necessary is requested and justified."[157]

IV. ANATOMY OF A HACK: METHODS FOR LABELING IDENTIFIABLE INFORMATION "DE-IDENTIFIED"

A. Overview of Permissive "De-Identification Guidelines"

HIPAA's requirements for *de-identification* are strict. Information that *can be used* to identify patients is entitled to HIPAA's safeguards, and the label "de-identified" is not meant to be a "work-around" to use or disclose patient information in ways otherwise prohibited by the Privacy Rule.

Removing all of the information that *can be used* to identify patients protects the patients from the misuse of that information by entities not entitled to possess it, including by entities who believe they are not subject to HIPAA's jurisdiction. As discussed in Section D of Part IV, however, removing identifiable information will often impair the use of that information for certain purposes. If a research protocol studies the relationship between specific health conditions and patients' *gender*, *dates of birth* and *zip codes*, it may be impossible for that research to be done with information that *cannot be used* to identify those patients. As also discussed, HIPAA recognizes this reality by allowing multiple pathways for researchers to access this *identifiable* patient information subject to HIPAA's protections.

In the private sphere, however, some organizations have taken a different approach. When a purchaser is willing to pay a premium for medical records that potentially *could be used* to identify patients, some organizations appear to have chosen to downplay the *identifiability* of those records by labeling them "de-identified." To justify labeling *identifiable* information "de-identified," parties have looked to certain permissive "de-identification guidelines." This Part IV examines in detail one of those guidelines, *Concepts and Methods for De-*

[157] *Id*. Note also that of the most important and widespread uses of patient information authorized under HIPAA is for "quality improvement activities" pursuant to paragraph (1) of HIPAA's definition of "health care operations." *See*, paragraph (1) of definition of "health care operations" at 45 C.F.R. § 164.501. Under paragraph (1), healthcare organizations can utilize identifiable patient information for a variety of healthcare quality improvement and assessment activities, including evaluating clinical outcomes, developing clinical guidelines, developing protocols and patient safety activities.

Identifying Clinical Trial Data[158] (hereinafter *Concepts*). A number of similar permissive guidelines exist that share *Concepts'* objective, methods and results.[159]

As will be discussed in detail below, the goal of these permissive "de-identification guidelines" is *not* to ensure that information labeled "de-identified" is sufficiently devoid of identifying information so that it can be safely disseminated as contemplated by Section 502(d)(2).[160] Rather, it is to provide a rationale for applying the label "de-identified" to patient information that often *can be used* to identify many, most or even all of the patients involved. Because their objectives are fundamentally different than those of Section 514(a) of HIPAA's Privacy Rule, these permissive guidelines abandon HIPAA's implementation specifications in Sections 514(b) and (c).

If, for example, a data broker is willing to pay a healthcare organization to replace its patients' *direct identifiers* with *identification codes* controlled by the broker (or its agent), these permissive guidelines ignore most of Section 514(c)'s requirements. As a result, these guidelines allow the parties to label the healthcare organization's patients' *identification codes* as "de-identified" even in situations where the codes *can be used* by the broker (or its agent) to identify all of those patients.

These guidelines wholly disregard Section 514(b)'s requirements. Rather than requiring healthcare organizations to follow Section 514(b)(1)'s "statistical confirmation" method or Section 514(b)(2)'s "safe harbor" method, these permissive guidelines allow healthcare organizations to label *identifiable* information as "de-identified" if a number of ambiguously defined or administered conditions are satisfied.

The first requirement is that their approach applies only to "nonpublic data disclosures." These are disclosures a healthcare organization may make to any commercial partner, for example, when a pharmacy sells its patients' prescription records to a data broker.

[158] Khaled El Emam & Bradley Malin, *Concepts and Methods for De-identifying Clinical Trial Data*, in APPENDIX B of SHARING CLINICAL TRIAL DATA, MAXIMIZING BENEFITS, MINIMIZING RISK 203 (Institute of Medicine of the National Academies 2015) [hereinafter *Concepts*].

[159] *See, e.g.*, Khaled El Emam, GUIDE TO THE DE-IDENTIFICATION OF PERSONAL HEALTH INFORMATION, (CRC Press, 2013); *De-identification Guidelines*, INFO. & PRIVACY COMMISSIONER OF ONTARIO (June 2016) [hereinafter *De-identification Guidelines*], https://www.ipc.on.ca/wp-content/uploads/2016/08/Deidentification-Guidelines-for-Structured-Data.pdf.

[160] *See* 45 C.F.R. § 164.502(d)(2) ("The requirements of [HIPAA's Privacy Rule] do not apply to information that has been de-identified in accordance with the applicable requirements of § 164.514 …").

The second requirement is that the data broker must contractually promise not to abuse its ability to identify patients. The guidelines allow the parties to decide for themselves how strict or enforceable this promise needs to be.

The third requirement is that the healthcare organization ensures that all of its patients' *direct identifiers* are removed. As noted above, however, *Concepts* allows the data broker (or its agent) to replace the patients' identifiers with *identification codes* in contravention of Section 514(c). Also in contravention of Section 514(c), *Concepts* allows the parties to label those *identification codes* "de-identified" even when the broker (or its agent) have enough control over how those *identification code*s are deployed and utilized so that they *can be used* to identify the healthcare organization's patients.

As this stage, the permissive guidelines apply the label "de-identified" to describe the *identification codes* and *indirect identifiers* the broker is purchasing. This vocabulary is employed regardless of how easily the broker can use those *identification codes* or *indirect identifiers* to identify *some*, *many*, *most* or even *all* of the healthcare organization's patients.

Because this form of "de-identified" information *can be used* to identify patients, the permissive guidelines no longer require healthcare organizations to apply "generally acceptable statistical and scientific principles for rendering information not individually identifiable." Instead, a healthcare organization need only engage in informed speculation as to whether its patients' *identification codes* and *indirect identifiers* being supplied to the data broker *will be used* to "re-identify" the patients.

There are no "generally accepted statistical and scientific principles and methods" for predicting whether or not a data broker will abuse its ability to identify patients. For this reason, permissive guidelines do not require healthcare organizations use "generally recognized statistical or scientific methods" in their informed speculation. Instead, healthcare organizations are only required to perform (i) a set of loosely-described "examinations" of their data broker customers; and (ii) a calculation that purports to indicate the likelihood that this "de-identified" information *will be used* to "re-identify" the patients.

The "examinations" the healthcare organization is required to perform are of the data broker's security and privacy practices, its conflicts-of-interests and its commercial motivations. Although they purport to be probative of addressing the likelihood that the patient identifiers *will be used* to identify patients, the guidelines do not require healthcare organizations to use evidence-based or industry-recognized auditing standards or assessment criteria. As a result, the healthcare organization getting paid to sell its patients' medical records is free to decide for itself what it deems to be adequate. Moreover, even if the healthcare organization's examination identifies "defects" in the data broker's practices, conflicts-of-interests or motivations, such defects do not disqualify the data broker from obtaining patient data. Rather, the healthcare organization can simply treat those defects as remedied if the broker is willing to contractually agree to addressing them based on whatever criteria the healthcare organization deems adequate.

The calculation that healthcare organizations are required to perform is described by the guidelines as a calculation of "actual risk" or "actual re-identification risk." The inputs to this calculation, however, are *not* based on generally accepted statistically or scientifically validated principles or methods. Rather, the guidelines allow healthcare organizations to use *any* numbers that are "defensible" based on the ability to find those numbers in any publication. Those numbers do not need to be validated by generally accepted scientific or statistical methods. As will be discussed below, the numbers do not need to be plausible or even accurately reflect the published source. To be "defensible," it appears all that is required is that the numbers are published in any "body of work."

Regardless of whatever numbers the healthcare organization decides are "defensible," the final calculation of "actual re-identification risk" does not need to correspond to HIPAA's definition of "very low probability." The guidelines give healthcare organizations broad discretion to select its own "acceptable risk threshold" that can be *orders of magnitude higher* than HIPAA's definition of "very low probability." In contrast to the ONC example,[161] where *0.013%* corresponds to a roughly *1-in-7,500* chance of identification, these guidelines allow parties to deem values as high as *33%* or *1-in-3* as "acceptable," a figure that is over *2,550 times greater* than the probability described in the ONC example. To be clear, this risk is only deemed "acceptable" by the healthcare organization and the data broker. No patients, security researchers or regulators have deemed a *1-in-3* an "acceptable" threshold for *de-identification*.

As malleable as the definitions of "acceptable risk" and "actual re-identification risk" are, the permissive "de-identification" guidelines allow healthcare organizations to significantly understate the actual identifiability of the *identification codes* and *indirect identifiers*. This is because these permissive guidelines only require the parties to assess the "identifiability" of each individual release of health information regardless of what other "de-identified" information the broker may also have available. So long as the broker promises not to abuse its ability to identify patients, the healthcare organization can disregard any other "de-identified" information the broker may possess, even if that information *can be used* to identify all of the healthcare organization's patients. These guidelines allow parties to calculate the "actual risk" of identification to be *33%*, even in circumstances where the broker may possess the *100%* ability to identify *many*, *most* or *all* of the patients involved.

[161] *See* Lafky, *supra* note 103.

B. Concepts and Methods for De-Identifying Clinical Trial Data

Concepts and Methods for De-Identifying Clinical Trial Data,[162] is among a number of "de-identification guidelines" that purport to describe a method for "de-identifying" patient information.

Concepts recognize two categories of "de-identified" information. One category is *public* data releases. "For public data," *Concepts* notes that "the [healthcare organization] needs to make a worst-case assumption and protect against an adversary who is targeting the [patients] with the highest risk of re-identification."[163] Accordingly, "[f]or a public data release, we assume … it is necessary to manage maximum risk."[164] The way *Concepts* "manage[s] maximum risk" is by applying the disclosure limitation techniques described in WORKING PAPER 22, such as "generalization," "suppression," "randomization" and "subsampling,"[165] until the information *can no longer be used* to identify patients. Because "there are no other controls"[166] that protect patients' identities, the data itself must protect the patients. This appears to be similar to the objectives and methods described in Section 514(a) and (b)(1) of the Privacy Rule.

The second category of "de-identified" information *Concepts* recognizes is what it calls "nonpublic data disclosures." These include patient information that a healthcare organization supplies to a private party, for example, when a pharmacy sells its patients' prescription records to a data broker. *Concepts*' methods for applying the label "de-identified" to "nonpublic data disclosures" differ significantly from its requirements for *public* data disclosures.

Concepts' approach to the "de-identification" of *non-public patient information* is made up of the following eleven steps:

Step 1: Determine direct identifiers in the dataset.
Step 2: Mask (transform) direct identifiers.
Step 3: Perform threat modeling.
Step 4: Determine minimal acceptable data utility.
Step 5: Determine the re-identification risk threshold.

Step 6: Import (sample) data from the source database.
Step 7: Evaluate the actual re-identification risk.
Step 8: Compare the actual risk with the threshold.
Step 9: Set parameters and apply data transformations.

[162] El Emam & Malin, *Concepts*, *supra* note 158.

[163] *Id.* at 207.

[164] *Id.* at 228-29.

[165] *Id.* at 236-37.

[166] *Id.* at 229.

Step 10: Perform diagnostics on the solution.
Step 11: Export transformed data to external dataset.[167]
As discussed below, Steps 2, 3, 5 and 7 of *Concepts'* approach depart significantly from a number of Section 514(a)-(c)'s requirements.

C. Step 2: Mask (Transform) Direct Identifiers

The job of *Step 1* in *Concepts'* approach is to review the medical records in question to inventory all of the patients' *direct identifiers*, such as the patients' *names*, *residential addresses*, *government identification numbers*, and the like.

The job of *Step 2* is to "mask" or "transform" those *direct identifiers*. *Concepts* defines "masking" or "transforming" as the "replacement of direct identifiers with pseudonyms."[168] If, for example, a medical record originally included the patient's *name* and *social security number*, those *direct identifiers* would be replaced with a pseudonym or *identification code* – such as "7iZw4M2k1p." The *identification code* would then serve as a proxy for patients' *names* and *social security numbers*.

Under Section 514(b), as well as WORKING PAPER 22 and the CHECKLIST,[169] the first step of *de-identification* requires the removal of *direct identifiers*. If, however, a healthcare organization wants to replace its patients' *direct identifiers* with *identification codes* that represent those patients, Section 514(c) requires healthcare organizations to comply with the following four requirements:

1. The *identification code* must be *assigned by* the healthcare organization;[170]

2. The *identification code* must *not* be "derived from or related to information about the [patient], or "capable of being translated so as to identify the patient;" [171]

[167] *Id.* at 240-43.

[168] *Id.* at 241.

[169] Working Paper 22, *supra* note 88, at 78 ("The first step to protect the respondent's confidentiality is to remove from the microdata all directly identifying information such as name, social security number, exact address, or date of birth."); CHECKLIST, *supra* note, at 8 ("Names, addresses, and other unique numeric identifiers such as Social Security, Medicare, or Medicaid numbers ***must*** be removed from the file.") (emphasis in original).

[170] 45 C.F.R. § 164.514(c) ("A covered entity may assign a code or other means of record identification …").

[171] *See* 45 C.F.R. § 164.514(c)(1) ("The code or other means of record identification is not *derived from* or *related to information* about the [patient] and is not otherwise capable of being translated so as to identify the [patient]; …") (emphasis added).

3. The healthcare organization cannot *disclose the mechanism for re-identification* to any third party; [172] and

4. The healthcare organization can only use the *identification code* to *re-identify its own de-identified health information*.[173]

Concepts' interpretation of Section 514(c), on the other hand, disregards most of its requirements. Rather, it describes Section 514(c) as follows:

> … in the HIPAA Privacy Rule at §164.514(c), it is stated that any code that is derived from information about an individual is considered identifiable data. However, such pseudonyms are practically important for knowing which records belong to the same clinical trial participant and constructing the longitudinal record of a data subject. Not being able to create derived pseudonyms means that random pseudonyms must be created. To be able to use random pseudonyms, one must maintain a crosswalk between the individual identity and the random pseudonym. The crosswalk allows the sponsor to use the same pseudonym for each participant across datasets and to allow re-identification at a future date if the need arises. These crosswalks, which are effectively linking tables between the pseudonym and the information about the individual, arguably present an elevated privacy risk because clearly identifiable information must now be stored somehow. Furthermore, the original regulations did not impose any controls on this crosswalk table.[174]

In this passage *Concepts* describes a hypothetical crosswalk table that links each patient's name to her "pseudonym" (*i.e.*, *identification code*). Anyone possessing this crosswalk table, therefore, is able to identify *all* of the patients and their medical records. *Concepts* acknowledges that the crosswalk is "clearly identifiable information" that "arguably present[s] an elevated privacy risk."[175] But *Concepts* states that the "original regulations did not impose any controls on this crosswalk table."[176] As a result, *Concepts* is unable to articulate any specific restrictions that apply to it.

[172] *See id.* at § 164.514(c)(2) ("The covered entity does not … disclose the mechanism for re-identification.")

[173] *See id.* at § 164.514(c)(2) ("The covered entity does not use or disclose the code or other means of record identification *for any other purpose …*") (emphasis added).

[174] El Emam & Malin, *Concepts, supra* note 158, at 212.

[175] *Id.* ("These crosswalks, which are effectively linking tables between the pseudonym and the information about the individual, arguably present an elevated privacy risk because clearly identifiable information must now be stored somehow.").

[176] *Id.*

Concepts' conclusion that the "original regulations did not impose any controls" is manifestly incorrect. Section 514(c) imposes *multiple* controls on how that crosswalk table is maintained. If, for example, the healthcare organization assigned the pseudonyms, but provided the researcher with the crosswalk table, this would violate each of Section 514(c)'s *second*, *third* and *fourth* controls. As HHS noted in its commentary to the "original regulations," healthcare organizations are "prohibited from disclosing the mechanism for re-identification, *such as tables*, algorithms, *or other tools that could be used to link the code with the subject of the information*."[177] Crosswalk tables are not exempt from Section 514(c)'s requirements merely because they are named "crosswalk tables."

As a result of its erroneous interpretation, *Concepts* expresses concern that a crosswalk table is "clearly identifiable information." Because the "original regulations did not impose any controls," *Concepts* recommends that any means to "reverse th[e] pseudonym[s] [should be] tightly controlled."[178]

Concepts, however, does not articulate specific criteria for what it means by "tightly controlled." It appears sufficient that a healthcare organization and a data broker enter into any form of agreement or adopt policies to protect this "clearly identifiable information" in a manner similar to how researchers supervised by an Institutional Review Board (*IRB*) operate:[179]

> Under the Common Rule, which guides IRBs, if the data recipient has no means of getting the key, for example, through an agreement with the sponsor prohibiting the sharing of keys under any circumstances or through organizational policies prohibiting such an exchange, then creating such derived pseudonyms is an acceptable approach.[180]

Concepts' advice here, however, confuses HIPAA's regulations that apply to *de-identified information* with its regulations that apply to *identifiable information*. Those regulations are very different.

If *identification codes* are used with *individually identifiable health information*, such as with a *limited dataset* or research information under Section 512(i), then the parties are *not* governed by Section 514(c). Rather, they are governed by HIPAA's comprehensive safeguards for *individually identifiable health information*, which include the protections and oversight set forth in HIPAA's Privacy, Security, Breach Notification and Enforcement Rules, and could include policies and procedures along the lines referenced in *Concepts*' passage. If, on the other hand, *identification codes* are to be used with properly

[177] *2000 Final Rule*, *supra* note 44, at 82,537 (emphasis added).

[178] El Emam & Malin, *Concepts, supra* note 158, at 212-13.

[179] *Id*. at 212.

[180] *Id*.

de-identified information that is utilized in ways that fall *outside* of HIPAA's safeguards and limitations, then Section 514(c)'s strict controls apply.

HHS discussed this distinction when considering the proposed use of HMAC encryption technology to create patient *identification codes*. Although the proposed arrangement violated the controls set forth in Section 514(c) (as discussed in Section B of Part III above), those codes *could* be used with *limited datasets*, which are otherwise protected by HIPAA:

> "The HMAC methodology, however, may be used in the context of the limited dataset …. The limited dataset contains individually identifiable health information and is not a de-identified dataset. Creation of a limited dataset for research with a data use agreement, as specified in § 164.514(e), would not preclude inclusion of the keyed-hash message authentication code in the limited dataset." [181]

Under *Concepts'* approach, however, *identification codes* are not protected by *either* HIPAA or Section 514(c). Rather, all that is required is that the parties enter into an agreement or adopt a nebulous set of policies that facially appear to manifest "tight controls."

Concepts does recognize *one* of Section 514(c)'s *four* controls, namely that the *identification code* must *not* be "derived from or related to information about the [patient], or "capable of being translated so as to identify the patient;" [182] Once again, however, *Concepts* misstates what HIPAA requires.

Concepts' understanding is based on its interpretation of guidance from HHS's Office of Civil Rights (hereinafter referred to as *OCR*) regarding cryptographic algorithms:

> "… in the recent guidelines from OCR, this is clarified to state that 'a covered entity may disclose codes derived from PHI (protected health information) as part of a de-identified dataset if an expert determines that the data meets the de-identification requirements at §164.514(b)(1).' (HHS, 2012, p. 22). *This means that a derived code, such as an encryption or hash function, can be used as a pseudonym as long as there is assurance that the means to reverse that pseudonym are tightly controlled*."[183]

[181] *2002 Final Rule*, *supra* note 74, at 53233 (emphasis added).

[182] *See* 45 C.F.R. § 164.514(c)(1) ("The code or other means of record identification is not *derived from* or *related to information* about the [patient] and is not otherwise capable of being translated so as to identify the [patient]; …") (emphasis added).

[183] El Emam & Malin, *Concepts, supra* note 158, at 212-13.

The referenced guidance is taken from OCR's "*Guidance Regarding Methods for De-identification*,[184] which states in relevant part:

> "The re-identification provision in §164.514(c) does not preclude the transformation of PHI into values derived by *cryptographic hash functions* using the expert determination method, *provided the keys associated with such functions are not disclosed, including to the recipients of the deidentified information*."[185]

Notably, *Concepts* omits a critical caveat in the OCR guidance, namely, that cryptographic *identification codes* can only be used if "the keys associated with such functions *are not disclosed*, *including to the recipients of the de-identified information*."[186] This echoes requirements of Section 514(c) as well as HHS's guidance regarding the HMAC algorithm.

Concepts' omission of this caveat, along with the other controls in Section 514(c), allows the parties to choose for themselves the ways they want to "tightly control" the *identification codes* and the technology used to create them. *Concepts* does not prohibit, for example, healthcare organizations from allowing their paying customers to use their patients' *identification codes* to aggregate all of their medical information, even if that information *could be used* by the broker or its agent to identify all of the healthcare organization's patients. Nor does it place any restrictions on who can assign the *identification codes*, even if that assignment could result in the data broker or its agent having the ability to identify *100%* of all of the healthcare organization's patients. In either case, *Concepts* allows the parties to label the *identification codes* as "de-identified."

D. Step 3: Perform Threat-Modeling

Step 3 consists of two components:

(1) identification of the plausible adversaries and what information they may be able to access; and

(2) determination of the quasi-identifiers.[187]

The second component of this Step – the determination of *quasi-identifiers* – refers to creating an inventory of "quasi-identifiers," another name for *indirect*

[184] Office of Civil Rights, U.S. Dep't of Health and Human Services, Guidance Regarding Methods For De-Identification of Protected Health Information in Accordance with the Health Insurance Portability And Accountability Act (HIPPA) Privacy Rule 22 (2015) [hereinafter OCR Guidance], https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf.

[185] *Id.* (emphasis added).

[186] *Id.*

[187] El Emam & Malin, *Concepts, supra* note 158, at 242.

identifiers. This process is similar to what is required by WORKING PAPER 22 when preparing *de-identified health information* in accordance with Section 514(b)(1). The *first* element of *Concepts' Step 3*, however, diverges significantly from Section 514(b)(1)'s requirements.

This departure begins with the vocabulary *Step 3* uses: *Concepts* presupposes that the *indirect identifiers* and *identification codes* resulting from *Step 2* are already "de-identified." From HIPAA's perspective, these *indirect identifiers* and *identification codes* continue to be *individually identifiable health information* because:

1. The medical records continue to include all of the patients' *indirect identifiers*; and
2. The *identification codes* do not comply with the requirements of Section 514(c).

From *Step 3*'s perspective, on the other hand, these identifiable patient records can be labeled "de-identified" if the data purchaser promises not to abuse its ability to use the *indirect identifiers* and *identification codes* to "re-identify" the patients.

The second departure involves the assessment the healthcare organization is required to perform. Under Section 514(b)(1)'s express language, the healthcare organization must remove all potential indirect identifiers until it can be demonstrated – by applying "generally accepted statistical and scientific principles and methods for rendering information not individually identifiable" – that there is a *very low probability* that the remaining *indirect identifiers* (and *identification codes*) *could be used* by the data purchaser to identify any of the patients.

Step 3, on the other hand, is *not* focused on ensuring that this information *cannot be used* to identify patients. Instead, *Step 3* requires the healthcare organization only to document its informed speculation regarding whether or not *indirect identifiers* (and *identification codes*) *will be used* to "re-identify" healthcare organization's patients.

Attempting to predict what a data broker *would do* with identifiable patient information is very different than assessing whether information *could be used* to identify patients. There is a large volume of "generally accepted statistical and scientific principles and methods for rendering information not individually identifiable," many of which are discussed in WORKING PAPER 22. But there are no "generally accepted statistical and scientific principles and methods" for predicting whether or not a data broker will abuse its ability to identify patients.

For this reason, *Concepts* does not require that the healthcare organization utilize "generally recognized statistical or scientific methods," much less methods for "rendering information not individually identifiable." Rather, *Concepts* only requires that the healthcare organization use a method that is "generally known" and "justified" by a "body of work." The "body of work" does not need to be scientifically or statistically valid, and it does not need to be generally accepted by scientists or statisticians. It simply needs to be "generally known."

Concepts' rationale for this approach is based on a puzzling misreading of Section 514(b)(1)'s express language:

> "The de-identification must be based on generally accepted statistical and scientific principles and methods for rendering information not individually identifiable. *This means that the [healthcare organization] needs to ensure that there is a body of work that justifies and evaluates the methods that are used for the de-identification and that these methods must be generally known.*"[188]

Similar to how *Concepts*' misreading of Section 514(c) omits three-fourths of Section 514(c)'s express requirements,[189] *Concepts*' interpretation of Section 514(b)(1) omits three of Section 514(b)(1)'s express requirements.

First, *Concepts'* method for labeling information "de-identified" no longer requires that it must be a "method for rendering information not individually identifiable." Second, it no longer needs to be a *statistical* or *scientific* method. Although the method needs to be "justified" by a "body of work," that justification does not need to be statistically or scientifically valid, and the "body of work" does not need to be *bona fide* statistical or scientific methods or principles. Third, the method no longer needs to be "generally *accepted*." Instead, it is sufficient if the "body of work" is "generally *known*." *Concepts*' interpretation, for example, does not disqualify a method even if it is "generally known" to be ineffective. Being "generally known" is sufficient.

Thus, so long as a data purchaser promises not to use the *indirect identifiers* and *identification codes* it will be receiving to "re-identify" a healthcare organization's patients, *Concepts* allows the parties to label the information "de-identified" notwithstanding the fact that it *can be used* to identify those patients. And because *Concepts* permits the parties to apply the label "de-identified" to information that *can be used* to identify patients, it no longer requires the parties to apply "generally acceptable statistical and scientific principles for rendering information not individually identifiable." Instead, the healthcare organization needs only to engage in informed speculation as to whether the *indirect identifiers* and *identification codes* being supplied to the data purchaser *will be used* to "re-identify" healthcare organization's patients.

Step 3 contemplates three potential ways that the identifiable patient information could be "re-identified:"

[188] *Id.* at 211.

[189] *See* discussion *supra* Section IV.C.

- A *deliberate attack*, which happens when the data broker (or agent) deliberately attempts to "re-identify" patients in the dataset;
- An *inadvertent attack*, which transpires when a data analyst working with the data broker (or the data broker itself) inadvertently "re-identifies" someone in the dataset; and
- A *breach*, which occurs if there is a data breach at the broker's facility.[190]

Concepts characterizes these risks as "cover[ing] the universe of attacks."[191]

The inconsistency between *Concepts'* vocabulary and HIPAA's language makes it difficult to assess precisely what *Concepts* means by "re-identify" or "breach." For example, because the *indirect identifiers* and *identification codes* are not *de-identified* in accordance with HIPAA's requirements, there is no unambiguous interpretation of what *Concepts* means by "re-identifying" information that was never *de-identified* to begin with. Presumably re-inserting the patients' *direct identifiers* would qualify. But without a rigorous definition, parties looking to profit from using patients' medical records in ways that would otherwise violate HIPAA could define "re-identification" narrowly to suit those objectives. They could, for example, create "de-identified" profiles about patients to generate "risk scores" about them using the patients' *identification codes*. Because those "predictive models" were created with patient profiles purporting to be "de-identified," a broker could decide that they could sell those "risk scores" to the patients' employers or insurers who wouldn't be allowed to receive the same substantive information if it were acknowledged to be derived from the patients' medical records.

Concepts' definition of "a breach" is likewise ambiguous. It, therefore, fails to define with any specificity how its identifiable form of "de-identified" information should be assessed in determining whether a breach has occurred and the extent to which it needs to be reported.

1. What Kind of Contract Does a Data Broker Need to Sign to Allow Patient Identifiers to Be Considered "De-Identified?"

In order to label *indirect identifiers* and *identification codes* "de-identified," *Concepts* requires data purchaser "to sign a contract that contains the relevant prohibitions."[192] *Concepts*, however, describes only three of these "relevant prohibitions:"

- Prohibition on "re-identification;"
- Restrictions on linking the data with other datasets; and

[190] El Emam & Malin, *Concepts, supra* note 158, at 229 ("For a nonpublic data set, we consider three types of attacks that cover the universe of attacks …").

[191] *Id.*

[192] *Id.* at 230.

- Disallowing the sharing of the data with other third parties.[193]

As discussed in the previous section, *Concepts* does not offer an unambiguous definition of what it means by "re-identification." *Concepts* does not discuss, for example, whether the use of its identifiable form of "de-identified" information in ways that would violate HIPAA if it were acknowledged to be identifiable constitutes "re-identification." Nor does it provide clarity about what kinds of "restrictions on linking" and "sharing" are specifically required. Nor does *Concepts* discuss other provisions that often would be considered if the information in question were acknowledged to be *individually identifiable health information*, such as:

- What are the consequences if the purportedly "de-identified" information is deliberately or inadvertently "re-identified?" Should patients be notified of the breach? Should HHS and State Attorneys General be notified?
- Should healthcare organizations be required to note such disclosures in their accounting of disclosures to their patients?
- What kinds of disclaimers or limitations of liability are appropriate given the identifiability of the information?
- Should patients be included as third-party beneficiaries in the event the patient identifiers are "re-identified?"
- What security standards are required for safeguarding such identifiers?
- What data breach standards should be applied to assessing whether or not a data breach has occurred?
- Who is authorized to evaluate the substance, suitability and effectiveness of the data purchaser's agreement?

Concepts does not condition its application of the label "de-identified" on how the parties answer any of these questions.

Moreover, *Concepts* does not discuss the inherent weakness of contracts when they are the *only* tool that is protecting patient privacy. In practice, contracts often offer *weak incentives* to comply with their obligations. Contracts frequently significantly limit the parties' financial liability and disclaim liability to "third party beneficiaries," such as patients, who may be adversely impacted by the parties' behavior. Enforcing contracts in court is expensive and time-consuming. As a result, many companies do not enforce their contracts and other companies calibrate their compliance efforts against the "practical risk" hoping they can get away with contract violations indefinitely. The 1978 version of Working Paper, known as "WORKING PAPER 2,"[194] discussed the inherent weakness of contracts in describing the results of a compliance audit conducted by Bureau of Census. The Bureau noted that:

[193] *Id.* at 235.

[194] *Report on Statistical Disclosure and Disclosure-Avoidance Techniques* (Office of Federal Statistical Policy and Standards, U.S. Dep't of Commerce, Working Paper 2, 1978) [hereinafter Working Paper 2].

> *"…* it was apparent that the sample purchasers either did not take their signed agreement seriously, forgot it after a period of time, or were not able to control handling of the file at their institutions. In a few cases the agreement had been signed by a university purchasing agent and was unknown to the actual users.[195]

This notorious compliance risk is why HIPAA extends its protections *regardless of what parties agree to in their contracts*. Even if the parties attempt to shield themselves from liability, the protections of HIPAA's Privacy, Security, Breach Notification and Enforcement Rules supply protections the parties may omit.

Step 3's approach leaves it entirely to the discretion of the healthcare organization to decide how it wants to protect its patients' "de-identified" patient identifiers. The contract can have fulsome protective provisions accompanied by rigorous compliance monitoring. Or it could have ambiguous prohibitions that give the data purchaser significant flexibility to decide for itself how it will use the healthcare organization's patients' medical records.

E. Step 5: Determine the Re-Identification Risk Threshold

Under Section 514(b)(1), medical records qualify as "de-identified" only if there is a *very low probability*, measured using *bona fide* statistical and scientific methods, that a broker could use the medical records to identify patients. As previously discussed, for example, the HHS's ONC commissioned a research team to attempt to identify approximately 15,000 individuals whose medical records were redacted in accordance with Section 514(b)(2)(i). The research team compared those records with consumer data provided by a national data broker and was able to identify *two* of 15,000 individuals, or *0.013%* of the population.[196] This is equivalent to a *1-in-7,500* chance that the data could be used to identify a patient. Although this low percentage has not been recognized as an "acceptable error rate," it illustrates that patient identifiability is measured using *bona fide* "statistical and scientific principles and methods for rendering information not identifiable."

Concepts' Step 5, in contrast, allows healthcare organizations and their customers to *select* their own "acceptable risk threshold." Because this is a *selection* process, not a *statistical* or *scientific* process, the parties are free to select "acceptable risk thresholds" that are orders of magnitude higher than the ONC example. *Concepts* allows the parties to agree to "acceptable risk thresholds" that are as high as a *1-in-3* chance that a patient *could* be identified from the medical records.[197]

[195] *Id*. at 31.

[196] *See* Lafky, *supra* note 103.

[197] El Emam & Malin, *Concepts, supra* note 158, at 233 (depicting a chart of "commonly used risk thresholds" of *1-in-11*, *1-in-5* and *1-in-3* chance that the data could be used to identify patients).

In a chart depicting 27 sample patients, for example, *Concepts* shows "probability of re-identification" for each patient in the sample.[198] On one end of the spectrum, there are nine patients listed, each of whose medical records have a "probability of re-identification" of *33%*, or approximately a *one-in-three* chance.[199] On the other end are eight patients, each of whose records have a "probability of re-identification" of *12.5%*, or a *one-in-eight* chance.[200] The average "probability of re-identification" for each of the 27 patients is *22%*,[201] or approximately a *one-in-five* chance that each patient's medical records could be used to identify her. There is no assessment of the likelihood that "at least 1 patient" can be identified, or "at least n patients[202]" can be identified, numbers which naturally increase with every new patient added to the analysis.

Because *Step 5* is a *selection* process, there is no statistical or scientific justification for why *Concepts*' "acceptable risk threshold" is approximately *2,500 times* less secure than the ONC's re-identification demonstration. *Concepts*, however, offers non-scientific rationales based on its misinterpretation of Section 514(b)(1)'s express requirements.

As discussed in Section D above, *Concepts*' interpretation of Section 514(b)(1) omits three express requirements. First, *Concepts'* method for labeling information "de-identified" no longer requires that healthcare organizations use "methods for rendering information not individually identifiable." Second, it no longer requires healthcare organizations to use statistical or scientific methods. Third, it no longer requires healthcare organizations to use methods that are "generally *accepted*" rather than simply "known."

Consequently, the "acceptable risk threshold" no longer needs to be equivalent to a *very low probability* as defined by the application of "generally accepted statistical and scientific principles and methods for rendering information not individually identifiable." Instead, the parties are free to use any way of "measuring re-identification risk in a defensible way and have a repeatable process to follow that allows for the definition of very small risk."[203]

[198] *Id.* at 227 ("Table B-5: The Generalized Data Set with No Uniques or Doubles").

[199] *Id.*

[200] *Id.*

[201] *Id.* at 237 ("…. It was possible to further reduce the average risk to 0.22 in Table B-5.").

[202] The author uses "n" here to refer to an arbitrary sample size number.

[203] *Id.* at 211.

This "repeatable process" does not need to be validated by *bona fide* statistical or scientific methods or principles. It simply needs to be included in a "body of work" and "generally known," even if unscientific. *Concepts* definition of "generally known," itself, appears to be remarkably lenient. Although *Concepts* acknowledges that it would "difficult" to classify "undocumented methods or proprietary methods that have never been published" as "generally known," it does not categorically prohibit healthcare organizations from doing so."[204]

Concepts, therefore, deems a risk threshold of *33%* acceptable because other permissive "de-identification guidelines" also use the word "acceptable" to describe *33%*,[205] and is "commonly used … based on the review/references in the text."[206] *Concepts* does not assert that a *33%* risk is an "acceptable" threshold based on "generally accepted statistical and scientific principles and methods for rendering information not individually identifiable" or that there is any statistically or scientifically validated methods or principles that would assign a *one-in-three* risk as "acceptable." Nor does *Concepts* assert that a *one-in-t* risk is acceptable to regulators, data protection researchers or the patients whose medical records are being sold without their knowledge.

In order for the parties to select *Concepts*' notably high "acceptable risk thresholds," *Concepts* requires an examination of the "context of the data"[207] or the "factors characterizing the [data broker] and the data themselves."[208] What *Concepts* means by "context" and "factors" are the following:

- The data purchaser's privacy and security practices;
- The data purchaser's commercial motivations; and
- The data purchaser's conflicts of interest.[209]

[204] *Id.* ("… undocumented methods or proprietary methods that have never been published would be difficult to classify as 'generally accepted.'").

[205] *Id.,* at 230-31 ("There are quite a few precedents for what can be considered an acceptable amount of risk …") (*citing* El Emam, GUIDE TO THE DE-IDENTIFICATION OF PERS. HEALTH INFO., *supra* note 159).

[206] *Id.* at 237 (describing identification risk thresholds of .33 as "commonly used").

[207] *E.g.*, *id.* at 234 ("Selecting an acceptable threshold … requires an examination of the *context of the data* themselves."); at 236 ("… the amount of data transformation needed will be a function of these other contextual factors. For example, if the [data broker] has good security and privacy practices in place, the threshold chosen will be higher, which means that the data will be subjected to less de-identification.").

[208] *Id.* at 234 ("The re-identification risk threshold is determined based on factors characterizing the [data broker] and the data themselves.").

[209] *Id.* at 233-34.

If the healthcare organization has a favorable impression of the data purchaser's "factors," *Concepts* permits the healthcare organization to apply the notably high "acceptable risk thresholds" described above.

Like the "acceptable risk thresholds" themselves, however, the healthcare organization's "examination" of these "factors" has not been validated by statistical or scientific methods. Rather, these factors are highly subjective in nature, and are described as only being "in use informally."[210] Consequently, the healthcare organization getting paid to sell its patients' medical information and selecting its own "acceptable risk threshold" is also being tasked with selecting how stringently it wants to examine the "context of the data" justifying both.

1. Examining a Data Purchaser's Security and Privacy Practices

In order to "examine" a data purchaser's security and privacy practices, *Concepts* references "a collection of practices used by large data custodians"[211] that are listed in a separate article written by one of *Concepts*' authors.[212] The article identifies approximately 40 privacy and security controls, partially excerpted below:[213]

Checklist of Practices That Must Be in Place at a Higher Threshold for Re-identification Risk, as Detailed in Policies, Guidelines, and Application Forms of Various Bodies

Controlling access, disclosure, retention, and disposition of personal data

- Requestor allows only "authorized" staff to access and use data on a "need-to-know" basis (i.e., when required to perform their duties).
- Data-sharing agreement between collaborators and subcontractors has been or will be implemented.
- Nondisclosure or confidentiality agreement (pledge of confidentiality) is in place for all staff, including external collaborators and contractors.
- Requestor will only publish or disclose aggregated data that do not allow identification of individuals.
- Long-term retention of personal data will be subject to periodic audits and oversight by independent bodies.
- Data will be disposed of after a specified retention period.

[210] *Id.* at 233.

[211] *Id.* at 234.

[212] Khaled El Emam, *et. al, Evaluating the Risk of Re-identification of Patients from Hospital Prescription Records*, 62 CAN. J. OF HOSP. PHARMACY, 307 (July–Aug. 2009) [hereinafter *Evaluating the Risk of Re-identification*].

[213] *Id.* at 318.

- Information will not be processed, stored, or maintained outside of [the country], and parties outside of [the country] will not have access to the data.
- Data will not be disclosed or shared with third parties.

Neither *Concepts* nor the article it references describe *how* the checklist is to be used to "examine" a given data broker. Because the checklist corresponds to a list of data protection principles that can be found on the Internet, most of the "practices" *Concepts* expressly requires could be auto-generated using off-the-shelf software.[214] Without a validated auditing protocol, the list on its own is incapable of distinguishing between *bona fide* data protection controls from *pro forma* documents that have been auto-generated.

On one extreme, the parties could interpret *Step 5*'s examination to require a *thorough* review of the suitability and effectiveness of the broker's privacy and security controls. Such an audit, for example, could require a thorough review of every documented policy, security incident report and breach assessments, and complaints from personnel, contractors and customers. To ensure that the broker is complying with its controls, the audit could require a review of sales and supplier contracts, marketing materials, sales proposals and product requirements. To ensure that the broker has appointed qualified personnel, the credentials and qualifications of all security and privacy personnel could be reviewed, followed by interviews to ensure that the individuals possess the requisite knowledge of applicable regulations and industry standards, as well as the broker's own internal controls. To ensure that the broker's compliance function is not subject to undue influence, the broker's organizational chart could be reviewed, along with the employment agreements of compliance personnel to ensure that they do not have a financial incentive to overlook significant compliance violations. To ensure that these matters are reviewed by appropriately qualified professionals who do not have any conflicts-of-interest, the statistician could require that the foregoing matters are reviewed by independent security and law firms whose fees must be paid regardless of the outcome of their assessments.

On the other extreme, the healthcare organization is free to interpret *Step 5*'s "informal" "examination" to require only that the broker produce a list of auto-generated policies and procedures. Because *Concepts* does not require that these documents be read or understood by a qualified professional, the healthcare organization would never realize if the auto-generated policies are inconsistent with contemplated transaction. Further, because the healthcare organization has never compared the policies to the broker's actual practices, it would never realize if the broker has never complied with those auto-generated policies. Because the healthcare organization has never reviewed the broker's sales proposals or

[214] *See*, *e.g.*, APTIBLE, https://www.aptible.com/documentation/comply/reference/pdf-exports.html (last visited Nov. 8, 2020) (Aptible's documentation solution, that allows policies to be auto-generated based on user inputs.).

product requirements, the statistician would never learn if the broker intended to breach its data use agreement. Because the healthcare organization has never reviewed the qualifications of the broker's security and privacy professionals, the statistician is in no position to assess whether the broker has appointed qualified personnel supervising those functions. Despite the fact that such an "examination" only requires that the broker produce a set of unreviewed policies, *Concepts* allows the parties to deem such an examination sufficient to warrant an "acceptable risk threshold" of *33%.*

2. Evaluating Conflicts-of-Interest and Commercial Motivations

In addition to making sure that "security and privacy practices [have been] put in place," *Step 5* calls for an examination of "issues [such] as conflicts of interest, the potential for financial gain from re-identification.[215] *Concepts,* however, does not provide any framework for evaluating such conflicts or commercial motivations. This is notable given that *Concepts*' process itself gives rise to a number of inherent conflicts, including the following:

- The data broker has a financial interest in labeling as much detailed patient information as possible as "de-identified" because it enhances the broker's ability to commercialize it and reduce compliance costs;
- The healthcare organization supplying the patient information has a financial interest in labeling medical records "de-identified" in situations where the broker is willing to pay a premium for records that are classified as "de-identified;" and
- Any "de-identification expert" hired by the parties has a financial interest in obtaining the fees by assisting healthcare organizations and data brokers in classifying as much of the detailed patient information as possible as "de-identified." If the "expert" is deemed "uncooperative" in supporting the parties' commercial objectives, the statistician puts future engagements at risk.

Because *Concepts* does not identify its own conflicts-of-interest, it provides no guidance on any of the following questions: (i) what it specifically means by "conflicts-of-interest," (ii) who is authorized to examine these conflicts, or (iii) what criteria should be used to assess whether or not such conflicts are problematic. The healthcare organization is free to conduct these examinations based on any standard it deems suitable.

Examining the "potential for financial gain" is similarly problematic. A principal motivation for HIPAA was the inherent financial interest associated with selling patient information. The "potential for financial gain from re-identification" applies to *all* data brokers and healthcare organizations. Without any protocols or assessment criteria for defining what *Concepts* means by

[215] El Emam & Malin, *Concepts, supra* note 158, at 234-35.

"potential for financial gain," healthcare organizations are free to decide for themselves the scope of their examinations and the weight they give to them.

3. No Disqualifying Defects: Remedying Defects Found in Informal Examinations

Despite not providing any evidence-based auditing standards or assessment criteria for their informal "examinations," *Concepts* recognizes that it is possible for healthcare organizations to find one or more defects in a broker's mitigating controls or commercial motivations.

None of those defects, however, appear to ever disqualify data brokers from receiving patient identifiers. Rather, it appears that *any* defect in a data broker's controls or motivation can be deemed cured if the healthcare organization beefs up its contract with the broker:

> "The security and privacy practices of the [broker] can be manipulated through contracts. The contract signed by the [broker] can impose a certain list of practices that must be in place, which are the basis for determining the threshold. Therefore, they must be in place by the [broker] to justify the level of transformation performed on the data."[216]

A real-world example of this is seen in *Evaluating the Risk of Re-identification*, written by one of *Concepts*' authors, which involved the "de-identification" of patient prescription records that a hospital was providing to a company developing a "database of prescription records."[217]

The patient information in question was protected by Ontario's Personal Health Information Privacy Act, 2004 (hereinafter referred to as "PHIPA")[218] rather than HIPAA. Similar to HIPAA, however, PHIPA defines patient information as "information that identifies a [patient] or for which it is reasonably foreseeable in the circumstances that it *could* be utilized, either alone or with other information, to identify a [patient]."[219] Also like HIPAA, PHIPA mandates that a "health information custodian" cannot disclose "personal health information" about a patient without obtaining the patient's consent.[220]

[216] *Id.* at 236.

[217] *See* El Emam, *supra* note 211, at 308.

[218] Pers. Health Info. Prot. Act, 2004, S.O. 2004, c. 3, Sched. A.

[219] *See*, *id.*, definitions of "identifying information" at 2004, c. 3, Sched. A, s. 4 (1), and "personal health information" at 4 (2) (emphasis added).

[220] *Id.* at 2004, c. 3, Sched. A, s. 29 ("A health information custodian shall not collect, use or disclose personal health information about an individual unless … it has the individual's consent under this Act …").

The project ran into a snag when the application of *bona fide* statistical techniques was deemed “unacceptable” to the data broker.[221] Rather than acknowledging the possibility that the patient information being requested was *identifiable*, the parties’ strategy was to *increase* the “acceptable risk threshold” to ensure that the medical records would be labeled “de-identified.”[222] As noted in the article, “[i]f it is not possible to obtain a *good de-identification* … the threshold is *increased* […].”[223]

Increasing the threshold, however, came at a price: “[…] the higher probability threshold must be balanced with greater security and privacy practices by the data recipient.”[224] The article then lists the controls described above that “need to be in place” in order for the medical records to be labeled “de-identified.”[225]

A second snag arose when it was determined that the broker did not actually have to have the relevant security or privacy practices in place. This defect, however, did not disqualify the data broker from obtaining the “good de-identification” or the higher “acceptable risk threshold.” Instead of disqualifying the data broker, the broker’s practices could be deemed adequate so long as it agreed to implement additional controls in its data use agreement.[226] There is no requirement that the healthcare organization directly confirm that the practices have been put into place, or that the healthcare organization actively monitor the broker to ensure that none of its patients’ medical records are misused. As discussed above, in practice, contracts often offer weak incentives to comply with obligations viewed as cumbersome.

F. Step 7: Evaluate the Actual Re-Identification Risk

Concepts entitles *Step 7* “evaluate the actual re-identification risk.” *Use* of the word “actual” implies *Step 7* seeks to *objectively measure* the real-world risk that the nominally “de-identified” patient identifiers given to the data broker will be compromised. In reality, *Step 7* only requires the parties *perform a calculation* using numbers the parties are largely free to select. This calculation does not *measure* anything “actual” about the risk to the data. Instead, it is a method for generating numerical values that will be lower than the “acceptable risk

[221] El Emam, *supra* note 211, at 314.

[222] *Id.* at 314.

[223] *Id.* at 313.

[224] *Id.* at 313.

[225] *Id.* at 313 (“Appendix 1 lists the practices that need to be in place at the higher threshold.”).

[226] *Id.* at 307.

threshold" *selected* in *Step 5* without the need to consider the *actual* risk to the data itself.

Step 7 is yet another instance of where *Concepts* abandons Section 514(b)(1)'s requirement to apply "generally accepted statistical and scientific methods for rendering information not individually identifiable." As discussed in Sections D and E above, *Concepts* only requires that the parties use a method that is "generally known" in a "body of work" that gives "repeatable" results. The "body of work" does not need to be statistically or scientifically valid, nor does it need to be a *generally accepted* statistical or scientific method. The "repeatability" of the numbers used does not need to be the result of a statistically or scientifically validated measurement. Instead, it is sufficient if the "repeatability" is a byproduct of the fact that the parties copy numbers from other permissive "de-identification guidelines."

The calculation contemplated by *Step 7* echoes *Step 3*'s "threat modeling." As discussed in Section D above, *Step 3* permits the parties to label patients' *indirect identifiers* and *identification codes* as "de-identified" if the data broker promises not to use that identifiable information to "re-identify" the healthcare organization's patients. Because this form of "de-identified" information *can be used* to identify patients, *Concepts* no longer requires the parties to apply "generally acceptable statistical and scientific principles for rendering information not individually identifiable." Instead, the healthcare organization need only engage in informed speculation as to whether the *indirect identifiers* and *identification codes* being supplied to the data purchaser *will be used* to "re-identify" healthcare organization's patients. Under *Step 3*, this informed speculation contemplates three potential ways that the identifiable patient information could be "re-identified:"

- A deliberate attack, where the adversary deliberately attempts to "re-identify" individuals in the dataset;
- An inadvertent attack, where a data analyst working with the data broker (or the data broker itself) inadvertently "re-identifies" someone in the dataset; or
- A breach, where there is a data breach at the broker's facility.[227]

Step 7's calculation of "actual re-identification risk" builds on these assumptions by requiring the parties to assign numerical values to each of these possibilities by using the following formulas:

(1) Pr(re-id, attempt) = Pr(re-id | attempt) × Pr(attempt), where the term Pr(attempt) captures the probability that a deliberate attempt to re-identify the data will be made by the data recipient;

(2) Pr(re-id, attempt) = Pr(re-id | attempt) × Pr(attempt), which evaluates the probability of that data broker's personnel may inadvertently re-identify someone in the dataset.

[227] *Id.* at 229.

> (3) Pr(re-id, breach) = Pr(re-id | breach) × Pr(breach), where the term Pr(breach) captures the probability that a data breach occurs at the broker's facility.[228]

Concepts anticipates that the inputs to these formulas will be informed by the informal "examinations" described in *Step 5*. For formula (1), for example, *Concepts* states:

> "The actual value for Pr(attempt) will depend on the security and privacy controls that the data recipient has in place and the contractual controls that are being imposed as part of the data sharing agreement."[229]

However, as discussed in Section E above, the examination of a data purchaser's "security and privacy controls" is *not* statistically or scientifically validated process, and there are no objective assessment criteria or auditing standards for conducting those examinations. As a result, the substance, quality and results of this "examination" are wholly determined by the party financially benefiting from it.

Moreover, *Step 7* does not require the parties to use objective numerical measurements. Instead, the parties can use any numbers they can locate in a published article. For example, in discussing which numbers should be used to calculate the "actual risk" of a data breach, *Concepts* says:

> "Data for 2010 show that 19 percent of health care organizations suffered a data breach within the previous year (HIMSS Analytics, 2010); data for 2012 show that this number rose to 27 percent (HIMSS Analytics, 2012). These organizations were all following the HIPAA Security Rule. Note that these figures are averages and may be adjusted to account for variation."[230]

The passage allows the parties to use *any* of the numbers that can be found in a published article. If the parties want to use the *19%* figure, they can. They can also use the 27% figure. Or if they want to make an adjustment "to account for variation," they can do that as well.

Notably, *Concepts* does *not* require healthcare organizations to confirm that the cited numbers are accurate. The 2012 HIMSS ANALYTICS REPORT[231] cited in *Concepts*, for example, does *not* state that "Data for 2010 show that 19 percent of health care organizations suffered a data breach within the previous year

[228] *Id.* at 229.

[229] El Emam & Malin, *Concepts, supra* note 158, at 229.

[230] *Id.* at 230.

[231] Healthcare Information and Management Systems Society, 2012 HIMSS ANALYTICS REPORT: SECURITY OF PATIENT DATA (Apr. 2012) [hereinafter *2012 HIMMS ANALYTICS REPORT*].

(HIMSS Analytics, 2010); data for 2012 show that this number rose to 27 percent (HIMSS Analytics, 2012)."[232] Rather, the 2012 HIMSS ANALYTICS REPORT states:

> "In 2012, *27 percent of all respondents* to this survey indicated their organization has had a security breach in the past 12 months (up from 19 percent in 2010 and 13 percent in 2008); of those who reported a breach, 69 percent experienced more than one."[233]

The percentages reported in the 2012 HIMSS ANALYTICS REPORT, therefore, did *not* apply to "health care organizations." Rather, they applied to *250* individuals who responded to a survey.[234] There is an enormous difference between these two numbers. According to OCR's ANNUAL REPORT TO CONGRESS ON BREACHES OF UNSECURED PROTECTED HEALTH INFORMATION,[235] for example, OCR received 222 reports of data breaches involving 500 or more patients in 2012.[236] The CMS's MEDICARE PHYSICIAN AND OTHER SUPPLIER NPI REPORT FOR 2012[237] lists over 900,000 physicians or suppliers, which does not include hospitals and many other "health care organizations." Using OCR's reported numbers in the numerator and CMS's numbers in the denominator would give a percentage of approximately *0.025%*. *Concepts* does not define what it means by the terms "health care organization" or "data breach," but it is clear that *Concepts*' estimates of *19%* and *27%*, respectively, misrepresents what was reported in 2012 HIMSS ANALYTICS REPORT, and are orders of magnitude off from the *actual* percentages of healthcare organizations that experienced data breaches in those respective time periods.

Notably, notwithstanding the inaccuracy of the statistics, *Concepts* appears to sanction their use regardless of what a healthcare organization discovers during its informal "examination" of a data broker's security and privacy practices. So long as there is a "body of work" that includes the numbers, regardless of how inaccurate those numbers may be, *Concepts* allows the healthcare organization to

[232] El Emam & Malin, *Concepts, supra* note 158, at 230.

[233] *See* 2012 HIMSS ANALYTICS REPORT, *supra* note 230, at 5 (emphasis added).

[234] See 2012 HIMSS ANALYTICS REPORT, *supra* note 230, at 10 ("HIMSS Analytics invited a variety of individuals with experience in their healthcare organization's privacy and security environment to participate in this telephone-based survey. The 250 respondents included …").

[235] OFFICE FOR CIVIL RIGHTS, U.S. DEP'T OF HEALTH & HUMAN SERVICES, ANNUAL REPORT TO CONGRESS ON BREACHES OF UNSECURED PROTECTED HEALTH INFORMATION (2011-2012).

[236] *Id.* at 10 ("For breaches occurring in calendar year 2012, OCR received 222 reports of these larger breaches …").

[237] CENTERS FOR MEDICARE AND MEDICAID SERVICES, U.S. DEP'T OF HEALTH & HUMAN SERVICES, MEDICARE PHYSICIAN AND OTHER SUPPLIER NATIONAL PROVIDER IDENTIFIER (NPI) AGGREGATE REPORT, Calendar Year 2012 (Oct. 31, 2017), https://data.cms.gov/Medicare-Physician-Supplier/Medicare-Physician-and-Other-Supplier-National-Pro/i587-8mbi.

use those numbers without taking into account any specific risks applicable to the healthcare organization being "examined."

Concepts does not give an example of how it converts informal examinations of a data purchaser's security and privacy practices into numerical values, but other permissive "de-identification guidelines" do. *De-identification Guidelines,*[238] which closely follows *Concepts'* blueprint and was reviewed by one of *Concepts'* authors, does.[239]

Like *Concepts*, *De-identification Guidelines* starts with the removal of *direct* identifiers.[240] Also like *Concepts*, *De-identifications Guidelines* allows a healthcare organization to label *indirect* identifiers as "de-identified" if the broker agrees to restrictions in its "data sharing agreement."[241] Following *Concepts*' blueprint, *De-identification Guidelines* requires examinations of a data broker's "privacy and security controls"[242] and "motives."[243] These examinations are not based on statistically or scientifically validated criteria or auditing standards. Rather, they are merely "qualitative assessments, resulting in values typically in the range of 'low,' 'medium,' or 'high.'"[244]

There is no scientific or mathematical foundation for converting the informally determined grades into empirically validated predictions. In order to perform the calculation, the parties are free to use whatever numbers they want if they can be found in another permissive "de-identification guideline." *De-identification Guidelines* provides the following table that "may be used as a guideline in determining what may be considered an acceptable estimate for the probability:"[245]

[238] *De-identification Guidelines*, *supra* note 159.

[239] *Id.* at 2, n.3, (acknowledging that *De-identification Guidelines* "is based largely on the risk-based de-identification methodology developed by Dr. Khaled El Emam") (*citing*, among other texts, El Emam, GUIDE TO THE DE-IDENTIFICATION OF PERSONAL HEALTH INFORMATION, *supra* note 158).

[240] *Id.* at 8.

[241] *Id.* at 14.

[242] *Id.* at 14.

[243] *Id.* at 15.

[244] *See*, *e.g.*, *id.* at 14.

[245] *See id*. at 15, (*citing* El Emam, GUIDE TO THE DE-IDENTIFICATION OF PERSONAL HEALTH INFORMATION, *supra* note 158, at 208).

Chart of “Acceptable Estimates” Appearing in De-identification Guidelines

Privacy and Security Controls	**Motives and Capacity**	**Probability of Re-Identification Attack**
High	High	0.05
	Medium	0.1
	Low	0.2
Medium	High	0.2
	Medium	0.3
	Low	0.4
Low	High	0.4
	Medium	0.5
	Low	0.6

Under *Concepts,* using these numbers is “acceptable” because they have been previously published in a “de-identification guideline” written by one of *Concepts*’ authors.[246] And the process is “repeatable” because copying numerical values from another “de-identification guidelines” will always result in the same conclusion.

Similar to *Concepts*’ misapplication of the numerical values from 2012 HIMSS ANALYTICS REPORT discussed above,[247] *Concepts* does not require these numerical values to be accurate or reflect empirically validated facts or risks applicable to the data broker. It is sufficient that the numbers used to calculate “actual re-identification risk” are simply copied from another publication in order to ensure the method is “repeatable.”

V. PROTECTING UN-PROTECTED HEALTH INFORMATION

A. Do Permissive “De-Identification Guidelines” Adhere to HIPAA’s Requirements?

Concepts’ “de-identification” methods are substantially similar to those described in other permissive “de-identification guidelines.” These guidelines frequently cite one another as independent support and are often written or reviewed by the same authors. *De-identification Guidelines*,[248] discussed above, is

[246] *See supra* Section IV.F (regarding *Concepts’* use of numerical values being misquoted from the *2012 HIMMS* ANALYTICS REPORT, *supra* note 230, at 5, 10.

[247] *See* Part IV, Section F (regarding *Concepts*’ use of numerical values being misquoted from the *2012 HIMMS ANALYTICS REPORT*, *supra* note 230, at 5, 10).

[248] *See supra* Section IV.F (regarding the similarities between the methods described in *De-identification Guidelines* and *Concepts*).

one such example.[249] It acknowledges that its "approach to de-identification … is based largely on the risk-based de-identification methodology developed by [one of *Concepts*' authors]."[250] And *De-identification Guidelines* justifies using the unsubstantiated numerical values in its calculation of "actual re-identification risk" because those numbers were published in GUIDE TO THE DE-IDENTIFICATION OF PERSONAL HEALTH INFORMATION,[251] also written by the same author. *Concepts* and *De-identification Guidelines* both reference *Evaluating the Risk of Re-identification* as supplying the list of security and privacy practices that must be in place before a healthcare organization can apply the notably high "acceptable risk thresholds." [252] The eleven-step process described in Part IV above, is substantially similar to the method described in GUIDE TO THE DE-IDENTIFICATION OF PERSONAL HEALTH INFORMATION.[253]

The goals, methods and results of these permissive "de-identification guidelines" depart significantly from HIPAA's requirements for *de-identification*. For *nonpublic* data releases, their objective is *not* to ensure that information labeled "de-identified" is sufficiently devoid of identifying information that it can be safely disseminated as contemplated by Section 502(d)(2).[254] As a result, their form of "de-identified" information often *can be used* to identify *many*, *most* or even *all* of the patients involved, and can present the same category of risks to patients as *individually identifiable health information*. Accordingly, these guidelines assume that such "de-identified" information will be protected in a manner that, in certain respects (but not all), echoes what HIPAA requires for *individually identifiable health information*.

Because their goals are different than HIPAA's, permissive "de-identification guidelines" do not adhere to HIPAA's specifications for *de-identified health information*. *Concepts*, for example, disregards three-quarters of Section 514(c)'s controls on how *identification codes* can be deployed. Indeed, permissive "de-identification guidelines" often expressly authorize healthcare organizations to let data brokers apply patient *identification codes*:

[249] *See* Part IV, Section F (regarding the similarities between the methods described in *De-identification Guidelines* and *Concepts*).

[250] *De-identification Guidelines*, *supra* note 159, at 2, n.3.

[251] *See id.*, at 15 (*citing* El Emam, GUIDE TO THE DE-IDENTIFICATION OF PERSONAL HEALTH INFORMATION, *supra* note 159, at 208).

[252] *Cf.* El Emam & Malin, *Concepts*, *supra* note 158, at 234, *with De-Identification Guidelines*, *supra* note 159, at 14.

[253] *See e.g.*, El Emam, GUIDE TO THE DE-IDENTIFICATION OF PERSONAL HEALTH INFORMATION, *supra* note 159, at 155.

[254] *See* 45 C.F.R. § 164.502(d)(2) ("The requirements of [HIPAA's Privacy Rule] do not apply to information that has been de-identified in accordance with the applicable requirements of § 164.514 …").

> The [healthcare organizations] may use the same algorithm to generate the pseudonyms for the [patients] so that the same [patient] in multiple datasets will have the same pseudonym. This way the [data broker] can perform anonymous linking of the datasets. Another example is where [patients] need to be tracked longitudinally, and there will be multiple data disclosures over time. To facilitate the anonymous linking of the different datasets, it would be desirable to have the same pseudonym used for the same [patients] over time.[255]

The use of the term "anonymous" in this section does *not* require that the information is anonymous in the way it's commonly understood. As with "de-identified," it only means that the patients' medical records have had the *direct identifiers* removed and that the broker has agreed not to abuse its ability to use *identification codes* and *indirect identifiers* to identify patients.

Nor do these "de-identification guidelines" adhere to the requirements of Section 514(b)(1)'s statistical confirmation method. They do *not*, for example, require healthcare organizations to apply "generally accepted statistical and scientific principles and methods for rendering information not identifiable."[256] Nor are healthcare organizations required to consider all information reasonably available to a broker if the broker promises that it will not "re-identify" the information. Nor are healthcare organizations required to demonstrate that there is a *very low probability* that the information *could be used* to identify patients using generally accepted statistical and scientific principles and methods for rendering information not identifiable.

In place of Section 514(b)(1)'s requirements, these "de-identification guidelines" allow healthcare organizations to use *non*-statistical or *non*-scientific processes so long as they have been published in any "body of work" and give "repeatable" results. This is permitted even when those "repeatable" results are simply the result of copying unvalidated or misquoted numerical values from other publications.[257] Healthcare organizations and their data broker customers are also given wide latitude to select their own "acceptable risk thresholds," which can be orders of magnitude higher what generally accepted statistical and scientific methods and principles would consider a *very low probability*. Although these "de-identification guidelines" nominally require healthcare organizations to examine their customers' security and privacy practices, commercial motivations

[255] Khaled El Emam & Anita Fineberg, *An Overview of Techniques for De-identifying Personal Health information,* CHEO RESEARCH INSTITUTE 19 (Aug. 14, 2009).

[256] *See supra* Section IV.F (regarding *Concepts'* use of numerical values being misquoted from the *2012 HIMMS* Analytics Report, *supra* note 230, that are likely orders of magnitude off of an accurate estimate of the real-world phenomena measured using empirically supported methods).

[257] *See* Part IV, Section F (regarding *Concepts*' use of numerical values being misquoted from the *2012 HIMMS ANALYTICS REPORT*, *supra* note 230, that are likely orders of magnitude off of an accurate estimate of the real world phenomena measured using empirically supported methods).

and conflicts-of-interest, those examinations are not conducted in accordance with independently validated assessment criteria or auditing standards. As a result, the substance, quality and results of these "examinations" are left solely to the healthcare organization whose payment from the data broker is dependent on favorable conclusions. Furthermore, the calculation of "actual re-identification risk" utilizes inputs that are not the result of any statistically or scientifically validated measurement instrument.

These "de-identification guidelines" also depart from Section 514(b)(1)'s requirements regarding *who* participates in the process. Because Section 514(b)(1)'s express language protects *de-identified health information* solely through the application of *bona fide* statistical and scientific methods, Section 514(b)(1) only contemplates that a qualified de-identification statistician is involved in its process.[258] The scope of the "informal examinations" required by permissive "de-identification guidelines," on the other hand, covers a wide range of topics far beyond the professional competence of a statistician. In order for those "examinations" to be effective, the statistician would also need legal expertise to meaningfully evaluate the data broker's data use agreement, information security expertise to evaluate the broker's security practices, expertise in privacy laws to assess the broker's privacy programs, expertise in assessing corporate conflicts-of-interest and corporate and human motivations, and expertise in effectively conducting audits or internal investigations of each of these very different domains. The "de-identification guidelines" do not articulate if and which experts must be engaged to perform these wide-ranging examinations. Nor do they discuss the professional qualifications or independence of such professionals. Nor do the guidelines discuss that Section 514(b)(1) makes no reference to having lawyers, security engineers, corporate ethics professionals or auditors involved in determining whether patient information *could be used* to identify patients.

The "de- identification guidelines" depart from Section 514(b)(2)'s requirements as well. They do not require, for example, that healthcare organizations remove all of the identifiers listed in Section 514(b)(2)(i). Because they allow the parties to label information "de-identified" even in circumstances where the healthcare organization *knows* the information *can be used* by the broker to identify patients, the guidelines cannot be used in their current form without violating Section 514(b)(2)(ii).

[258] 45 C.F.R. § 164.514(b)(1) (describing the qualified statistician as a "person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable.").

B. Do Permissive "De-Identification Guidelines" Encourage Healthcare Organizations to Disregard HIPAA's Requirements?

The permissive "de-identification guidelines" at times suggest that their methods comply with HIPAA's requirements. As discussed in Section C of Part IV, for example, *Concepts* describes 514(c) of HIPAA's Privacy Rule as not imposing "any controls" on a spreadsheet that links patient identifiers with *identification codes*. And as discussed in Section D of Part IV, *Concepts* interprets Section 514(b)(1) as permitting healthcare organizations to use any method justified by a body of work that is generally known. Although both are incorrect interpretations of HIPAA's requirements, they evince an aspiration to comply with HIPAA.

There are other times, however, where permissive "de-identification guidelines" acknowledge that their use of the word "de-identified" is distinct from *any* legal definition of the term, including HIPAA's. For example, GUIDE TO THE DE-IDENTIFICATION OF PERSONAL HEALTH INFORMATION states:

> "It should also be noted that the amount of de-identification that is applied is influenced by external factors that must be taken into account: precedents, regulatory requirements and signals, and the public's expectations … Regulators may indicate preferences for certain amounts of de-identifications through regulations, orders, and guidance documents." [259]

In this passage "regulatory requirements" are characterized as "external factors" that exist independently of what it means for medical records to be "de-identified." Further, the passage suggests that "regulatory requirements" are simply "preferences" for "certain amounts of de-identification." Both suggestions indicate that these significantly permissive guidelines fail to understand HIPAA's framework.

Under HIPAA, *de-identification* is *defined by* its regulatory requirements. Sections 514(b) and (c) are not "external factors," and there is no such thing as "de-identified health information" that does not comply with *all* of HIPAA's requirements. Information that does not meet *all* of those requirements remains *individually identifiable health information*. Even when a large number of patient identifiers has been removed – such as with a *limited dataset* – that information *must* be safeguarded in accordance with HIPAA's requirements. "De-identification" is *not* something that can be "stretched" to meet the needs of a healthcare organization or its financial objectives. As discussed in Section D of Part III, HIPAA offers many options for using *individually identifiable health information* for *bona fide* research purposes that do not involve misclassifying patient information as "de-identified."

[259] El Emam, GUIDE TO THE DE-IDENTIFICATION OF PERSONAL HEALTH INFORMATION, *supra* note 159, at 153.

Despite the significant discrepancies between the guidelines' definition of "de-identified" and that of HIPAA, the guidelines do not explicitly acknowledge them. Thus, by using the homonym "de-identified" to describe the results of their process, these guidelines invite parties to disregard HIPAA's express requirements in favor of their own. This, in turn, invites healthcare organizations to make the following presumptions:

- Patient identifiers do not need to be protected in accordance with the security standards set forth in HIPAA's Security Rule;
- Patients do not need to be notified in accordance with HIPAA's Breach Notification Rule when the patient identifiers have been accessed in a hack or any other manner not authorized under the Privacy Rule;
- Neither the patient identifiers, nor the healthcare organization's or data broker's disclosure, receipt or maintenance of such patient identifiers, are subject to the jurisdiction of HIPAA's Enforcement Rule;
- The patient identifiers can be used, disclosed or sold to any third party for *any* purpose, including those that would violate HIPAA's Privacy Rule were the information acknowledged to be identifiable;
- The parties do not need to obtain patients' authorizations in accordance with Section 508[260] of the Privacy Rule, even when the patient identifiers are neither limited datasets[261] nor obtained pursuant to a privacy board's waiver of authorization in accordance with Section 512(i) of the Privacy Rule;[262] and
- Patients do not need to be notified in an accounting of disclosures under Section 528(a)(1) of the Privacy Rule,[263] even when the patient identifiers are neither limited datasets[264] nor received pursuant to a valid authorization.[265]

This confusion regarding the homonym "de-identified" can easily result in harms to patient privacy. Patients, for example, have no assurance that any of HIPAA's safeguards will be applied to their medical records once they are labeled "de-identified," regardless of how easily those "de-identified" records *can be used* to identify them. The permissive "de-identification guidelines" ostensibly require that identifiable forms of "de-identified" information be protected with "stringent

[260] 45 C.F.R. § 164.508(a).

[261] *Id.* at § 164.528(a)(1)(viii).

[262] *Id.* at § 164.512(i).

[263] *Id.* at § 164.528(a)(1).

[264] *Id.* at § 164.528(a)(1)(viii).

[265] Brief of Electronic Privacy Information Center (EPIC) *et al.* as Amici Curiae Supporting Petitioners, Sorrell v. IMS Health Inc., 564 U.S. 522 (2011) (No.10-799), http://epic.org/amicus/sorrell/EPIC_amicus_Sorrell_final.pdf [hereinafter *EPIC Brief*] at 20, 23-4.

controls," but they leave it to financially conflicted parties to determine how "stringent" those "controls" must be.

Pleadings in *Sorrell* offer a glimpse of the potential impact of misclassifying identifiable information as "de-identified." An amicus brief filed in the case alleged that data broker, IMS, used a cryptographic algorithm known as "MD5" to generate *identification codes* that IMS used to track healthcare organizations' patients.[266] It is also alleged that IMS continued to use the MD5 for many years after MD5 had been publicly compromised. After over a decade of warnings by security researchers, in the mid-2000s two teams of researchers published that they had "cracked" MD5 using ordinary desktop computers. This led to MD5 to be declared "cryptographically broken" in 2005[267] and "unsuitable for further use" by Department of Homeland Security's Computer Emergency Readiness Team in 2008.[268] It appears, however, that IMS may have been unaware that MD5 had been compromised. An IMS general manager, for example, testified that its *identification codes* were secure and "there is no way that you can actually reverse engineer the data back to a patient."[269]

If these allegations are correct and IMS was unaware of MD5's vulnerabilities, this could be because it viewed patients' *identification codes* as "de-identified." Although there are many ways that any *identification code* can be compromised in a way that results in patient identification, IMS may have concluded that it had no obligation under HIPAA to safeguard those identifiers. IMS also may not have implemented appropriate systems to detect when the patients' *identification codes* are used or disclosed in a way that compromises patient privacy. And it is unclear whether IMS would notify patients if such a compromise occurred.

This contrasts sharply with how HIPAA treats *identification codes*. If, for example, the *identification codes* were generated in accordance with Section 514(c), HIPAA's Security Rule would *require* healthcare organizations assigning those codes to conduct thorough and accurate risk assessments of the potential risks and vulnerabilities.[270] It would also require healthcare

[266] Brief of Amici Curiae Electronic Privacy Information Center (EPIC) *et al*. Supporting Petitioners, Sorrell v. IMS Health Inc., 564 U.S. 552 (2011) (No. 10-799), https://epic.org/amicus/sorrell/EPIC_amicus_Sorrell_final.pdf [hereinafter *EPIC Brief*], at 20, 23-24.

[267] *Id*. at 24.

[268] *Id*.

[269] Testimony of General Manager for IMS Health's Business Line Management, quoted in *id*., at 20.

[270] 45 C.F.R. § 164.308(a)(1)(ii).

organizations to protect against "any reasonably anticipated threats or hazards."[271] Healthcare organizations would be required to be informed about MD5's well-known defects. They also would be required to implement systems to detect any compromise of their patients' *identification codes* and to notify patients if a compromise results in the use or disclosure of their patients' medical information in unauthorized ways or to unauthorized parties.

C. Are Permissive "De-Identification Guidelines" Being Utilized by Healthcare Organizations in Lieu of HIPAA's Requirements?

Given the notorious secrecy surrounding the sale of patient medical information,[272] it is impossible to be certain whether or how many healthcare organizations utilize the permissive "de-identification guidelines" in lieu of HIPAA's express requirements. Tanner's exposé, however, indicates that at least some healthcare organizations may.

Recalling Tanner's discussion of data broker software that replaces direct identifiers with *identification codes*,[273] it appears that a number of healthcare organizations allow data brokers to assign their patients' *identification codes*. Tanner's account is bolstered by information in the *Sorrell* litigation, where testimony from an employee of one of IMS's agents, Verispan, revealed the way IMS' agent used "linking codes" to allow IMS to follow patients throughout their lives:

> "What we do is … strip out all of the identifiable information, and replace it with the serial *linking code* […] so that every time an entity comes into the database, it's replaced with the same code. *So you can follow an individual over time* […]"[274]

On its face, this practice appears to violate a number of Section 514(c)'s controls, as more fully discussed in Section B of Part III above. Not only do healthcare organizations allow data brokers (or their agents) to apply the patients' *identification codes*, they also allow the brokers (on their own or through agents) to *control* how those codes are used. They *can be used* by a broker and its agent, for example, to aggregate an unlimited amount of additional information about

[271] *See* discussion *supra* Part I (regarding the secrecy surrounding the sales of patient medical information).

[272] *See* discussion *supra* Section III.B.

[273] *See* discussion in Section B of Part III above.

[274] *See* EPIC Brief, *supra* note 265, at 20,26-27 (quoting the trial testimony of Jody Fisher, Vice President of Verispan's Product Management, C.A. App. A99 (emphasis added). Notably, the first clause of this quote is unlikely to be wholly accurate. As noted in *Concepts*, if you replaces "all of the identifying information" with a code, "the resulting data, in almost all cases, will not be useful for analytic purposes."[274] More likely, the only information that has likely been "strip[ped] out" by the code are the patients' direct identifiers and a limited number of indirect identifiers.).

patients. That ability, in turn, gives the broker (or its agent) the ability to use those *identification codes* – alone or in combination with the technology used to generate them or other patient information – to identify *many*, *most* or *all* of the healthcare organizations' patients. Regardless of whether or not the broker (or its agent) ever abuses this ability, the healthcare organization has disclosed patient information that *could be used* – alone or in combination with other information – to identify its patients.

The permissive "de-identification guidelines" place no specific restrictions on how *identification codes* are utilized. Consequently, they permit healthcare organizations to apply the label "de-identified" to their patients' *identification codes* even in circumstances where they *can be used* – directly or indirectly – to identify *100%* of the healthcare organizations' patients.

With respect to HIPAA's requirements in Section 514(b), Tanner notes that "[d]ata scientists can now circumvent HIPAA's privacy protections by … marrying [data brokers'] anonymized patient dossiers with named consumer profiles available elsewhere – with a surprising degree of accuracy."[275] On its face, this appears to contradict Section 514(b)(1)'s requirement that "the risk is very small that the information *could be used*, alone or *in combination with other reasonably available information*, by the [data broker] to identify [a patient]."[276] It also appears to contradict Section 514(b)(2)(ii) that does not allow information to be considered *de-identified* when the healthcare organization knows that "the information *could be used* alone or *in combination with other information* to identify [a patient]."[277]

Permissive "de-identification guidelines" take a markedly different approach. Once a data broker enters into a contract, the healthcare organization can disregard any "other information" available to the data broker that could be combined with patient medical information to identify patients. As a result, even when the *identification codes* and "other information" *can be used* by the data broker to identify *100%* of the healthcare organization's patients, these guidelines allow the label "de-identified" to be applied.

D. Do Permissive "De-Identification Guidelines" Provide an Effective Data Protection Alternative to HIPAA's Definition of De-Identified Information?

In their current form, permissive "de-identification guidelines" incorporate too many vulnerabilities to operate as an alternative to HIPAA's data protection framework. These guidelines, for example, provide no unambiguous or measurable requirements for what kinds of data use agreements warrant allowing

[275] Thielman, *supra* note 1.

[276] 45 C.F.R. § 164.514(b)(1)(i) (emphasis added).

[277] *Id.* at § 164.514(b)(1)(ii) (emphasis added).

identifiable information to be labeled "de-identified." They also provide no unambiguous or measurable requirements for assessing a data broker's security or privacy practices, conflicts-of-interests or commercial motivations. Nor do they provide unambiguous or measurable auditing criteria to ensure that the examinations of those topics are likely to uncover information relevant to the purported inquiries. Nor is there any method for converting the results of those examinations into numerical values that have been scientifically or statistically demonstrated to have any predictive accuracy. This precludes the calculation of "actual re-identification risk" from serving as an objective measure. Likewise, there is no objective criteria for defining the limits of what the parties can select for their "acceptable risk thresholds."

The vulnerabilities arising from the dearth of unambiguous or measurable requirements are compounded by manifest conflicts-of-interest created by these permissive guidelines. The guidelines, for example, allow healthcare organizations who are financially benefiting from labeling their identifiable patient information as "de-identified" to play a dispositive role in selecting its own "acceptable risk threshold" and in determining the substance and effectiveness of the broker's data use agreement, security and privacy practices, conflicts-of-interest and commercial motivations. This, in turn, gives the healthcare organizations dispositive influence over the purportedly objective calculation of "actual re-identification risk." The guidelines do not acknowledge the conflicts-of-interest they create, and thereby include no mechanisms for mitigating their influence. As such, they are ripe for abuse by parties who seek the imprimatur of a "process" for labeling *identifiable* medical records as "de-identified." These guidelines are incapable of distinguishing between healthcare organizations and brokers who have no desire to abuse patient privacy from those that do.

Analytically, the permissive "de-identification guidelines" will remain a viable differential privacy framework until they effectively address identification risks arising from combining multiple data sources. The types of data routinely available to data brokers about specific patients includes information from other care settings, longitudinal data, information about relatives and household members who may have overlapping healthcare histories. Any method that only measures the "identifiability" of each individual data source on its own – or solely in combination with public records – fails to consider the *many* situations where combinations of nominally "de-identified" information *can be used* to identify patients.

As previously discussed, for example, the Sweeny study indicated that 87% of the U.S. population could be identified using only their *gender*, *date of birth* and *zip code*.[278] If a data broker possessed *identification codes* for 200 million patients that are linked to their *gender*, *date of birth* and *zip code*, the

[278] *See* Sweeney, *supra* note 40, at 2.

broker would possess the *ability* to compromise 174 million of those *identification codes*. The broker's ability to link those 174 million *identification codes* to the correct patients is *100%*. This vulnerability exists even if no single data source contains all of the necessary indirect identifiers. If the broker has one dataset that only includes each of the patients' *gender* and *age*, and then purchases a second dataset that has their *zip code*s, the broker now possesses enough information that *could be used* to identify 174 million individuals.

This is a general phenomenon, and thus can arise from longitudinal records about the same patient that come from a single source, as was shown in the lab results study discussed earlier where researchers identified patients in a database of biometric information that had been redacted in accordance with Section 514(b)(2)(i). [279] The study found that the ability to identify patients represented in the database increased dramatically when researchers could compare it to the patient's longitudinal health information.[280] When researchers could utilize the known results for *four* consecutive PCV panels, for example, they had *19.5%* chance of uniquely identifying a patient in the redacted biomedical database.[281] And when researchers had access to *six* consecutive panels, the rate jumped to *89%*.[282]

These "de-identification guidelines" also lack a way to effectively model the identification risk of databases comprising multiple individuals. Averaging is incapable of describing the systemic risk arising from including *increasing* numbers of individuals in a single database. The likelihood of being able to identify "at least one" patient – or "at least n-number of patients" – from an ever-increasing database, for example, cannot be adequately modeled by simply averaging all of the patients' isolated risk scores. Nor can averaging effectively describe identification risk arising from patients who may be related to one another in a way that increases their respective risk of identification. Nor can averaging depict how the compromise of one patient's identity increases the identification risk to other patients in the same database, creating cascading identification scenarios.

The American Medical Association's recently issued AMA PRIVACY PRINCIPLES describes "appropriate de-identification" as "using techniques that are demonstrably robust, scalable, transparent, and provable."[283] The current

[279] *See* Atreya, *supra* note 115, at 95.

[280] *Id.* at 98.

[281] *Id.*

[282] *Id.*

[283] AM. MED. ASS'N, MEDICAL PRIVACY PRINCIPLES 1 (May 11, 2020), ("Privacy rights should be honored unless they are waived by an individual in a meaningful way, the information is

articulations of permissive "de-identification guidelines" fail *all four* of those criteria.

E. The Elephant in the Room

Independent standards bodies often partner with regulators and academic researchers to create thoroughly vetted industry standards for handling sensitive information. These standards address legal requirements and continuously evolving use-cases and security risks. They are also used by independent testing labs to audit companies to ensure that they are complying with those practices. The Payment Card Industry Security Standards Council is a standards body that maintains standards for technologies that protect credit card numbers and validates auditing procedures for ensuring compliance with those standards.[284] The National Institute of Standards and Technology plays a similar role for technologies protecting sensitive information, such as cryptographic algorithms.[285]

There is no equivalent oversight over the protection of medical records that have been labeled "de-identified." On the contrary, "[t]he trade in patient data is so opaque that many even in health care and government do not know about it."[286] The moment that healthcare organizations apply the label "de-identified" to their patients' medical records – regardless of how easily they *can be used* to identify the patients – all proactive oversight appears to vanish. This has led to the remarkable circumstance where a patient's credit card number is currently given *substantially greater protection* than her identifiable medical records that have been labeled "de-identified."

If your cardholder data is compromised, for example, you have reasonable assurance that this will be detected and that you will be notified about what data was compromised and what actions you should take to protect yourself. This contrasts sharply with how your medical information is protected the moment the label "de-identified" is slapped onto it.

It is unclear whether there is proactive oversight once your medical records are labeled "de-identified," regardless of how easily those records *can be*

appropriately de-identified (using techniques that are demonstrably robust, scalable, transparent, and provable …") [hereinafter *AMA PRIVACY PRINCIPLES*].

[284] *See*, *e.g.*, THE PAYMENT CARD INDUSTRY SECURITY STANDARDS COUNCIL, PCI DATA SECURITY STANDARD (PCI DSS), REQUIREMENTS AND SECURITY ASSESSMENT PROCEDURES, Version 3.2.1, (May 2018).

[285] *See*, *e.g.*, NIST SP 800-107, *supra* note 98; NIST, NIST SPECIAL PUBLICATION 800-57 PART 1 REVISION 4, RECOMMENDATION FOR KEY MANAGEMENT, PART 1: GENERAL (Jan. 2016).

[286] Tanner, *The Hidden Trade in Medical Data*, *supra* note 2.

used to identify you. Compromise of your identifiable "de-identified" medical records may or may not be detected. Even if it is detected, you may not be informed. Because your medical records have been labeled "de-identified," a broker may decide to use and disclose your information in ways that would be prohibited under the Privacy Rule if the *identifiability* of those records were acknowledged. Even if a broker does not directly sell your medical records to your health insurer or your employer, the broker may decide to "train" algorithms and create and then sell a "propensity model" or a "risk score" about you based on your medical records and *identification code*. They would be sold as "predictions" about you notwithstanding the fact that those "predictions" convey information derived from your *actual* medical records.

The current gaps in the proactive enforcement of HIPAA's requirements have prompted the AMA to call for greater effective oversight – in its recent AMA PRIVACY PRINCIPLES, it called for entities to "make their de-identification processes and techniques publicly available."[287]

Four brief provisions of HIPAA's Privacy Rule – Sections 514(a)-(c) and 502(d) – are all that safeguards the *petabytes* worth of patient information for *hundreds of millions* of Americans. It is critical, therefore, that the requirements of these provisions are understood, enforced and complied with.

For healthcare organizations, compliance starts with being informed of their obligations under the Privacy Rule. Unless *all* of Section 514(b)-(c)'s requirements are *fully* satisfied, the health information in question remains *individually identifiable health information* subject to HIPAA's comprehensive data protection requirements. Furthermore, even after information has been *de-identified*, healthcare organizations cannot forget that it is always possible it will revert back into *identifiable health information* and, thereby, once again be subject to HIPAA's protections.[288] Nor can healthcare organizations disregard widely known risks to patient privacy arising from implementations of *identification codes* or the aggregation of patient information that *can be used* to identify patients. HIPAA's Security and Breach Notification Rules do not permit healthcare organizations to remain "willfully ignorant" of "potential risks and vulnerabilities" to information that *can be used* to identify patients.

VI. CONCLUSION

Regulators have a central role to play in fostering transparency and accountability. There are many red flags indicating that a significant amount of health information labeled "de-identified" *can be used* to identify many, most or even all of the patients involved. The first step regulators should take, therefore, is

[287] AMA PRIVACY PRINCIPLES, *supra* note 22, at 3.

[288] 45 C.F.R. § 164.502(d)(2)(ii).

to stay informed about how "de-identification" is actually being conducted in commercial settings by exercising their compliance review authority.[289] Second, regulators should report their findings so that "de-identification" practices can be independently reviewed by data protection researchers to ensure that they align with HIPAA's express requirements and utilize up-to-date safeguards to address current uses and accompanying threats. Third, regulators should use those learnings to update existing guidance to ensure that *all* of requirements listed in Sections 514(b)-(c) and 502(b) are appropriately accounted for. Given the extraordinary sensitivity of health information, regulators should promote the adoption of documented industry standards for the use of *identification codes* and de-identification that are no less sophisticated than what exists in cryptography or other mature confidentiality domains.

Even if regulatory bodies are slow to act, advocates have been successful in bringing private claims against healthcare organizations that violate HIPAA's requirements. Although HIPAA does not give patients an express private right of action, courts have routinely found that HIPAA and its implementing regulations may be utilized to inform the standard of care applicable to claims arising from the inappropriate disclosure of patient information.[290] Healthcare organizations that disclose health information that *can be used* to identify their patients to entities that would not be otherwise entitled to receive it under HIPAA's Privacy Rule violate HIPAA's requirements. And healthcare organizations that receive a payment in exchange for disclosing such information may also be violating HIPAA's prohibition against selling it without the patient's written authorization.[291] In situations where laws remain unenforced, direct litigation may be the final avenue available to patients seeking to protect their privacy against the mislabeling of their clearly identifiable information as "de-identified."

[289] *Id.* at § 160.308(b).

[290] *See*, *e.g.*, Byrne v. Avery Center for Obstetrics & Gynecology, P.C., 314 Conn. 433, 458-59 (2014), *aff'd* Byrne v. Avery Center for Obstetrics and Gynecology, P.C., 327 Conn. 540 (2018) ("… HIPAA and its implementing regulations may be utilized to inform standard of care applicable to allegations of negligence in the disclosure of patients' medical records pursuant to a subpoena."); R.K. v. St. Mary's Medical Center, Inc., 229 W. Va. 712, 719-21, 735 S.E.2d 715 (2012), *cert. denied*, U.S., 133 S. Ct. 1738, 185 L. Ed. 2d 788 (2013) (concluding that state law claims arising from defendant hospital staff's unauthorized disclosure of plaintiff's psychiatric treatment records were not preempted by HIPAA and that goals of common-law remedies and HIPAA both protect the privacy of an individual's health care information); Acosta v. Byrum, 180 N.C. App. 562, 571-73, 638 S.E.2d 246 (2006) (concluding that HIPAA may serve as evidence of the appropriate standard of care); Sorensen v. Barbuto, 143 P.3d 295, 299 n.2 (Utah App. 2006) (concluding that the trial court improperly dismissed the plaintiff's claim for breach of professional duties that violations of HIPAA may inform the proper standard of care in a cause of action against a healthcare organization).

[291] 45 C.F.R. §§ 164.502(a)(5)(ii), 164.508(a)(4).