De-Identified and Unregulated: How Data Brokers Outpace State Privacy Laws

ABSTRACT

State consumer privacy laws, though increasingly important in the absence of a comprehensive federal privacy framework, fail to effectively regulate the practices of data brokers who exploit de-identified data. Laws like the Tennessee Information Protection Act (TIPA) exempt de-identified data from key protections, leaving significant gaps in oversight.

While the Health Insurance Portability and Accountability Act (HIPAA) establishes standards for de-identification, advanced analytics and linkage techniques employed by data brokers render this data increasingly susceptible to re-identification. The Federal Trade Commission (FTC) has taken steps to address these risks, but its limited authority highlights the need for comprehensive solutions.

This Note proposes two key approaches to addressing the privacy risks posed by data brokers and the re-identification of de-identified data: enacting federal privacy legislation and adopting synthetic data generation to mitigate re-identification risks to close regulatory loopholes. Together, these measures aim to address the shortcomings of state and federal privacy frameworks, ensuring stronger protections for de-identified data in an evolving data ecosystem.

TABLE OF CONTENTS

I. THE LEGAL LANDSCAPE OF DE-IDENTIFICATION AND BEYOND 867
A. *Data Brokers and their Ecosystems* 867
B. *HIPAA De-identification Standards* 870
C. *Differences Between De-identified, Pseudonymous and Anonymous Data* 873
D. *Section 5 of The Federal Trade Commission (FTC) Act* 877
E. *State Consumer Privacy Laws, Tennessee Definitions* 878
1. The General Data Protection Regulation (GDPR) 881

II. REGULATORY GAPS AND RE-IDENTIFICATION RISKS BY DATA BROKERS 883
A. *HIPAA's Treatment of De-identified Data and Its Consequences in TIPA's Regulatory Framework* 883
1. Covered Entities Sell De-identified Data, While States Focus on Definitions, Not Risks 885
B. *There Is No Incentive for Data Brokers to Utilize Privacy-Preserving Record Linkage* 886
C. *FTC's Role in Regulating Data Brokers* 888
III. TOWARD COMPREHENSIVE PRIVACY PROTECTIONS 890
A. *Regulate the Sale, Sharing, and Licensing of De-Identified Data* 891
B. *Synthetic Data Generation as a Broad-Reaching Solution* 893
IV. CONCLUSION 897

Data has become one of the most valuable assets in the global economy, rivaling traditional commodities like oil.[1] It fuels innovation, powers artificial intelligence (AI), and drives decision-making across industries like advertising and healthcare services.[2] However, as data grows in economic importance, so do the risks associated with its misuse.[3]

The emergence of data brokers as key players in the information economy has outpaced the development of effective legal frameworks to regulate their activities.[4] These entities aggregate and sell vast amounts of consumer data, often operating in the shadows away from public scrutiny.[5] Current data privacy laws, both federal and state, struggle to keep pace with the advanced analytics and linkage techniques employed by data brokers, which make it faster and easier to combine de-identified datasets with other sources of auxiliary data like phone records or voter registration to re-identify individuals.[6] In

[1] Lawrence Teixeira, *The New Black Gold: How Data Became the Most Valuable Asset in Tech*, MEDIUM (Feb. 12, 2024), https://medium.com/@lawrenceteixeira/the-new-black-gold-how-data-became-the-most-valuable-asset-in-tech-9e4541262ddf [https://perma.cc/27D8-YNMB].

[2] *Id.*

[3] *Id.*

[4] *See* Urbano Reviglio, *The Untamed and Discreet Role of Data Brokers in Surveillance Capitalism: A Transnational and Interdisciplinary Overview,* 11 INTERNET POL'Y REV. 8 (2022).

[5] Justin Sherman, *How Shady Companies Guess Your Religion, Sexual Orientation, and Mental Health*, SLATE (Apr. 26, 2023, 12:55 PM), https://slate.com/technology/2023/04/data-broker-inference-privacy-legislation.html [https://perma.cc/RP9D-2VPT].

[6] Reviglio, *supra* note 4, at 12 n.20 ("[D]ata broker Acxiom offers LiveRamp IdentityLink, an identity graph that matches directly identifiable data – like emails, postal

2023, the Pew Research Center found that most Americans are concerned with how companies use their information, reflecting the need for further legal developments in the United States.[7] The General Data Protection Regulation (GDPR), the EU's data privacy and security law, offers the strictest standard for protecting potentially re-identifiable consumer data,[8] in contrast to US consumer privacy laws, such as the Tennessee Information Protection Act (TIPA), the model law referenced throughout this Note, that do not impose comparable safeguards, particularly for de-identified health data.[9]

Recent findings reveal that data brokers have aggregated and sold information originally derived from mental health platforms—such as antidepressant usage and PTSD diagnoses—without the knowledge or consent of individuals who utilize such platforms.[10] Typically, mental health data, like speaking with a therapist, is protected by the Health Insurance Portability and Accountability Act (HIPAA) if the data originates from a covered entity or business associate.[11] However, most mental health platforms are not covered entities, thereby evading HIPAA's regulations, which then allows for the collection, sharing, and licensing of mental health data.[12] Problematically, data brokers—who

addresses, and phone numbers – with pseudonymous identifiers – like cookies and device IDs."); *see* Protecting Americans from Harmful Data Broker Practices, 89 Fed. Reg. 101402 (Dec. 3, 2024) (to be codified at 12 C.F.R. pt. 1022).

[7] Colleen McClain, Michelle Faverio, Monica Anderson & Eugenie Park, *How Americans View Data Privacy*, PEW RSCH. CTR. (Oct. 18, 2023), https://www.pewresearch.org/internet/2023/10/18/how-americans-view-data-privacy/ [https://perma.cc/KM46-GR2Q].

[8] STEPHEN P. MULLIGAN & CHRIS D. LINEBAUGH, CONG. RSCH. SERV., R45631, DATA PROTECTION LAW: AN OVERVIEW 50 (2019).

[9] *See* 2023 Tenn. Pub. Acts 408.

[10] Joanne Kim, *Data Brokers and the Sale of Americans' Mental Health Data*, DUKE CYBER POL'Y PROGRAM 1, 5 (Feb. 2023), https://techpolicy.sanford.duke.edu/wp-content/uploads/sites/4/2023/02/Kim-2023-Data-Brokers-and-the-Sale-of-Americans-Mental-Health-Data.pdf [https://perma.cc/7PAR-EXUQ].

[11] Under HIPAA, if a covered entity

> "engages with a business associate to help it carry out its health care activities and functions, the covered entity must have a written business associate contract or other arrangement with the business associate that establishes specifically what the business associate has been engaged to do and requires the business associate to comply with the Rules' requirements to protect the privacy and security of protected health information."

Covered Entities and Business Associates, U.S. DEP'T OF HEALTH & HUM. SERVS., https://www.hhs.gov/hipaa/for-professionals/covered-entities/index.html [https://perma.cc/F463-YS8S] (last visited Jan. 27, 2025).

[12] Kim, *supra* note 10, at 2; Nicole Martinez-Martin, Ishan Dasgupta, Adrian Carter, Jennifer A Chandler, Philipp Kellmeyer, Karola Kreitmair, Anthony Weiss & Laura Y Cabrera, *Ethics of Digital Mental Health During COVID-19: Crisis and Opportunities*, 7 J. MED. INTERNET RSCH. MENTAL HEALTH 2 (2020).

commonly are not covered entities under HIPAA—are often among the recipients of such information.[13] As such, de-identified mental health data from mental health applications remains susceptible to re-identification by data brokers who utilize advanced analytics, a process that combines de-identified and other datasets, such as machine learning, to reveal patterns and insights that provide highly targeted predictions and recommendations about individuals.[14] Mental health data represents only a small fraction of the vast amount of health data available for use by data brokers; this Note specifically focuses on de-identified data under HIPAA (i.e., data handled by covered entities and their business associates).[15] Outside of mental health data, the risk of data brokers exploiting data de-identified under HIPAA is largely unstudied by scholars, with current discussions and proposed regulations focusing on, among others, financial and sensitive personal data.[16]

This Note, therefore, attempts to examine the gaps in comprehensive consumer privacy laws, particularly TIPA, (including, briefly, data broker-specific laws), that enable data brokers to collect, share, and license de-identified data–gaps which are further exploited through exemptions for data de-identified under HIPAA. It then proposes two solutions to address this gap: first, a renewed call for comprehensive federal privacy legislation focused on preventing data brokers from receiving de-identified data as classified under HIPAA; and second, a shift toward synthetic data generation—artificially created datasets that mimic real-world data without being tied to actual

[13] Kim, *supra* note 10, at 3.

[14] Cole Stryker, *What is Advanced Analytics?*, IBM (July 10, 2024), https://www.ibm.com/think/topics/advanced-analytics [https://perma.cc/ER49-ZMUE]; *Data Brokers: Key Players in the Data Selling Ecosystem*, DATACY (Jan. 30, 2024), https://datacy.com/business/blog/data-brokers-key-players-in-the-data-selling-ecosystem [https://perma.cc/6YY5-BG5W].

[15] *See Guidance Regarding Methods for De-Identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule*, HHS (Nov. 26, 2012) [hereinafter *Guidance Regarding Methods for De-Identification*], https://www.hhs.gov/hipaa/for-professionals/special-topics/de-identification/index.html [https://perma.cc/SZB8-URVK]; Andrea Gadotti, Luc Rocher, Florimond Houssiau, Ana-Maria Crețu & Yves-Alexandre de Montjoye, *Anonymization: The Imperfect Science of Using Data While Preserving Privacy*, 10 SCI. ADVANCES 1, 5 (2024).

[16] "The proposed rule would limit the sale of personal identifiers like Social Security Numbers and phone numbers collected by certain companies and make sure that people's financial data such as income is only shared for legitimate purposes." *See CFPB Proposes Rule to Stop Data Brokers from Selling Sensitive Personal Data to Scammers, Stalkers, and Spies*, CONSUMER PROTECT. FIN. BUREAU (Dec. 3, 2024) [hereinafter *CFPB Proposes Rule*], https://www.consumerfinance.gov/about-us/newsroom/cfpb-proposes-rule-to-stop-data-brokers-from-selling-sensitive-personal-data-to-scammers-stalkers-and-spies/ [https://perma.cc/BS6X-EQYT].

individuals—to replace the dissemination of de-identified data.[17] By addressing the legislative and technological shortcomings of current consumer privacy laws (broad laws that regulate a wide range of personal data across industries), these solutions aim to balance data utility with privacy protections.[18]

I. THE LEGAL LANDSCAPE OF DE-IDENTIFICATION AND BEYOND

A. Data Brokers and their Ecosystems

The data broker industry operates largely unregulated, quietly driving the collection, analysis, and sale of vast amounts of consumer data.[19] The data broker market is estimated to be a multibillion dollar industry by 2034.[20] Data brokers can be classified into two categories: first-party data brokers, organizations that collect personal information directly from individuals through their own platforms or services (e.g., telehealth companies and mental health applications), and third-party data brokers, who still engage in information collection and sharing, but do not have a direct relationship with individuals.[21] For example, Axiom, one of the largest data brokers in the world, has thousands of data points on individuals without maintaining direct relationships with them.[22]

Although states like California and Texas have enacted legislation targeting the regulation of data brokers through security and disclosure mandates,[23] these laws exclude significant parts of the

[17] *See* Moez Ali, *Synthetic Data Is the Future of Artificial Intelligence*, MEDIUM (Jan. 18, 2023), https://moez-62905.medium.com/synthetic-data-is-the-future-of-artificial-intelligence-6fcfd2ce1a14 [https://perma.cc/F3B6-FN6L].

[18] *See* discussion *infra* Section III.A–B.

[19] *Data Brokers*, ELEC. PRIV. INFO. CTR., https://epic.org/issues/consumer-privacy/data-brokers/ [https://perma.cc/DLC2-G59G] (last visited Feb. 1, 2025).

[20] *Data Broker Market Overview*, MKT. RSCH. FUTURE, https://www.marketresearchfuture.com/reports/data-broker-market-11676#:~:text=Data%20Broker%20Market%20is%20projected,USD%20284.20%20billion%20in%202024 [https://perma.cc/A8N8-EX32] (last visited Jan. 31, 2025).

[21] Stephanie Pell, Justin Sherman & Jen Patja, *Lawfare Daily: Justin Sherman on the Benefits and Limits of a New Law Governing Data Brokers*, LAWFARE (Apr. 29, 2024, 8:00 AM), https://www.lawfaremedia.org/article/lawfare-daily-justin-sherman-on-the-benefits-and-limits-of-a-new-law-governing-data-brokers [https://perma.cc/Y4J4-HKUQ].

[22] *Id.*

[23] *See Texas Data Broker Act*, ATT'Y GEN. OF TEX., https://www.texasattorneygeneral.gov/consumer-protection/file-consumer-complaint/consumer-privacy-rights/texas-data-broker-act [https://perma.cc/N2FD-W5M9] ("The Act requires these entities to register as a data broker with the Texas Secretary of State; post a conspicuous notice on its website or app disclosing that it is a data broker; implement comprehensive information security safeguards to protect personal data.") (last visited May 5, 2025); Ca. Civ. Code § 1798.99.80(c).

data ecosystem.[24] These data broker-specific laws largely focus on three areas: requiring registration with the state, ensuring disclosure, and imposing security requirements.[25] The California Data Broker Act narrowly defines a data broker as a "business that knowingly collects and sells to third parties the personal information of a consumer with whom the business does not have a direct relationship."[26] Similarly, the Texas Data Broker Act defines a data broker as "a business entity whose principal source of revenue is derived from the collecting, processing, or transferring of personal data that the entity did not collect directly from the individual linked or linkable to the data."[27] While these laws purport to regulate personal data, both exempt de-identified data from their purview, though they do so in different ways.[28] Texas explicitly excludes "de-identified data" from its definition of personal data.[29] California, by contrast, does not use the term "de-identified" but exempts data processed or covered by entities or their business associates to the extent the data is already governed by existing federal health privacy regulations (e.g., HIPAA).[30] Rather than addressing the full breadth of data broker practices, these statutes only capture a small portion of data broker activities.[31] As a result, these laws fail to target data brokers' interactions with data that qualifies as de-identified under HIPAA.[32] States like Tennessee without data

[24] Justin Sherman, *Data Brokers and Sensitive Data on U.S. Individuals*, DUKE CYBER POL'Y PROGRAM 1, 1–2 (2021), https://techpolicy.sanford.duke.edu/wp-content/uploads/sites/4/2021/08/Data-Brokers-and-Sensitive-Data-on-US-Individuals-Sherman-2021.pdf [https://perma.cc/C3NE-JPG2].

[25] *See generally* TEX. BUS. & COM. CODE § 509.001(11) (2023) (The Act mandates that entities register as data brokers with the Texas Secretary of State, display a clear notice on their website or app, implement robust data security measures, train employees and contractors, and ensure third-party providers uphold security standards.)

[26] CAL. CIV. CODE § 1798.99.80(c) (2024).

[27] TEX. BUS. & COM. CODE § 509.001(11) (2023).

[28] *Id.* Vermont's data broker statute does not reference HIPAA or de-identified data directly, but defines "brokered personal information" to include any data that, "alone or in combination with other information sold or licensed, would allow a reasonable person to identify the consumer with reasonable certainty." VT. STAT. ANN. tit. 9, § 2430(3) (West 2025). Although this definition appears broader than those in California and Texas, it remains unclear whether data de-identified under HIPAA's standards—which allow for a limited risk of re-identification—would be considered identifiable under Vermont's "reasonable certainty" threshold. *See id.* Moreover, if the broker does not have a direct relationship with the consumer and the information is truly de-identified, it's uncertain whether such entities would even meet the definition of a data broker under the statute. *See id.*

[29] TEX. BUS. & COM. CODE § 509.001(11) (2023).

[30] *See* Ca. Civ. Code § 1798.146.

[31] *See* Tex. Bus. & Com. Code § 509.001(11) (2023); Ca. Civ. Code § 1798.146. Specifically (instead of as a result).

[32] *See id.*

broker-specific laws similarly exempt de-identified data from broader consumer privacy protections, such as the ability to opt out of data-sharing practices.[33]

The de-identified data at issue in this Note originates from various sources, including healthcare providers, insurance companies, and health information exchanges.[34] While this data is de-identified to comply with HIPAA standards, its handling by data brokers raises significant privacy concerns.[35] Consumers may reasonably expect that de-identified data will be used responsibly and remain protected, but this expectation often conflicts with how the data is ultimately handled.[36]

One technique used by entities required to comply with HIPAA data obligations, commonly referred to as "covered entities," and their business associates is privacy-preserving record linkage (PPRL), which allows multiple datasets to be linked to external datasets (e.g., linking vaccination records to external datasets to understand vaccination status across states) without exposing personal identifiers directly.[37] PPRL enables legitimate purposes, such as population health studies or research, while meeting HIPAA's de-identification standards. [38] However, it remains unclear what happens to de-identified data once it leaves these entities and enters the broader ecosystem.[39] Data brokers may acquire such data through licenses or sales, and because de-identified data is no longer subject to HIPAA or TIPA, there is no accountability for how it is subsequently used.[40]

While PPRL provides an example of linkage techniques that align with privacy regulations, data brokers likely have no incentive to use PPRL themselves. Instead, they rely on advanced analytics to combine de-identified datasets with other sources of data, such as public records, to reap the benefits associated with highly individualized data without the safeguards PPRL is designed to

[33] *E.g.*, 2023 Tenn. Pub. Acts 408.

[34] 45 C.F.R. §160.103; Office for Civil Rights (OCR), *Summary of the HIPAA Privacy Rule*, U.S. DEP'T OF HEALTH & HUM. SERVS. (Oct. 19, 2022), https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html [https://perma.cc/B89K-CYNE].

[35] *See id.*; *see also* Kim, *supra* note 10, at 10.

[36] *See* McClain et al., *supra* note 7.

[37] Emery Niemiec, *What Is Privacy Preserving Record Linkage (PPRL)?*, HEALTHVERITY (Sept. 1, 2022), https://blog.healthverity.com/what-is-privacy-preserving-record-linkage [https://perma.cc/YXD3-6NAS].

[38] *Id.*

[39] *See Data Brokers*, *supra* note 19.

[40] *See* Sherman, *supra* note 24; *Guidance Regarding Methods for De-Identification*, *supra* note 15.

ensure.[41] Thus, the lack of transparency in how de-identified data is shared or sold and the ability of brokers to link it with minimal oversight represent a critical gap in the regulatory framework and undermine consumer privacy protections.[42]

B. HIPAA De-identification Standards

HIPAA is a US federal law that mandates the protection and confidential handling of protected health information (PHI), like medical records.[43] The Privacy Rule, a key component of HIPAA, was established by the US Department of Health and Human Services (HHS) to enact national standards with safeguards and limitations on the uses and disclosures of PHI.[44] The Privacy Rule applies to health plans, healthcare clearinghouses, and healthcare providers that conduct certain healthcare transactions electronically, ensuring the consistent protection of individually identifiable health information.[45]

To uphold the Privacy Rule's protections, covered entities must implement appropriate safeguards to maintain compliance and protect the confidentiality of PHI.[46] "A covered entity cannot use or disclose PHI except either: (1) as the Privacy Rule permits; or (2) as the individual who is the subject of the information authorizes in writing."[47] Specifically, a covered entity may disclose PHI without a patient's authorization in the following circumstances: (1) disclosure to the individual or patient; (2) disclosure to healthcare providers, insurers, or other entities for treatment, payment, or healthcare operations; (3) disclosure to an individual or entity with an opportunity for the individual to agree or object; (4) disclosure to public authorities or entities for certain public interest and benefit activities; (5) disclosure incidental to an otherwise permitted use or disclosure; and (6) disclosure as part of a limited data set to researchers, public health authorities, or other authorized parties.[48] The process of

[41] DATACY, *supra* note 14.

[42] *See id.*

[43] Kim Theodos & Scott Sittig, *Health Information Privacy Laws in the Digital Age: HIPAA Doesn't Apply*, 18 PERSP. HEALTH INFO. MGMT. 1, 3 (2020).

[44] BEYOND THE HIPAA PRIVACY RULE: ENHANCING PRIVACY, IMPROVING HEALTH THROUGH RESEARCH 1–2 (Sharyl J. Nass, Laura A. Levit & Lawrence O. Gostin eds., 2009). The Privacy Rule is codified in Title 45 of the Code of Federal Regulations in section 160 and subparts A and E of section 164. 45 C.F.R. §§ 160.101, 164.104, 164.500 (2019).

[45] OCR, *supra* note 34.

[46] *Id.*

[47] *Id.*

[48] Steve Alder, *HIPAA Permitted Disclosures*, THE HIPAA J. (Dec. 6, 2023), https://www.hipaajournal.com/hipaa-permitted-disclosures/ [https://perma.cc/9LH8-WDV4].

de-identification removes identifiers, such as a person's social security number, from the data before sharing.[49] Specifically, direct identifiers are pieces of information that can reasonably identify an individual, such as name, birth date, or Social Security number,[50] whereas indirect identifiers, like an individual's city or state, are pieces of information that can, in combination with other information, potentially identify a specific individual.[51] It is important to note that de-identification under HIPAA removes the data from the scope of these restrictions and all other HIPAA requirements.[52] Once data is de-identified, it is no longer considered PHI and can be used or disclosed to recipients beyond the parameters described above. Thus, for example, a pharmaceutical company is enabled to use the data to analyze prescription trends to identify market opportunities[53] without limitation, including for purposes such as sales, research, and sharing.[54]

De-identification can be achieved through two methods.[55] First, the Expert Determination method is where an expert applies statistical or scientific principles (such as analyzing replicability, data source availability, and distinguishability) to reduce the risk of re-identification, which is "a process by which information is attributed to de-identified data in order to identify the individual to whom the de-identified data relate."[56] The method has three requirements: (1) the de-identification must be based on generally accepted statistical and scientific principles and methods for rendering information not individually identifiable, (2) the risk of re-identification needs to be *very small*, and (3) the methods and results of the analysis that justify such determination must be documented.[57]

Similarly, the second method of de-identification, the Safe Harbor method, is designed to both de-identify data and also provide a

[49] *See Guidance Regarding Methods for De-Identification*, *supra* note 15.

[50] Alder, *supra* note 48.

[51] *Quick Guide to HIPAA*, STAN. MED., https://med.stanford.edu/starr-tools/data-compliance/hipaa-primer.html#:~:text=HIPAA%20designates%20the%20following%20as%20indirect%20identifiers%3A%20city%2C%20state%2C,HIPAA%2Ddesignated%20as%20direct%20identifiers [https://perma.cc/ZRK6-TDET] (last visited Jan. 31, 2025).

[52] *See Guidance Regarding Methods for De-Identification*, *supra* note 15.

[53] Sophie Stalla-Bourdillon, *What Is Data De-Identification and Why Is It Important?*, IMMUTA (Nov. 4, 2024), https://www.immuta.com/blog/what-is-data-de-identification/ [https://perma.cc/LA7K-F7VS]. Recipients are essentially entities or individuals who derive some benefit from the data and may, in turn, provide some benefit to another through its use. *See id.*

[54] *See id.*; *Guidance Regarding Methods for De-Identification*, *supra* note 15.

[55] *Guidance Regarding Methods for De-Identification*, *supra* note 15.

[56] *Id.*; *Re-Identification*, NAT'L INST. OF STANDARDS AND TECH., https://csrc.nist.gov/glossary/term/re_identification [https://perma.cc/3F4H-QPXX] (last visited Jan. 27, 2025).

[57] *Guidance Regarding Methods for De-Identification*, *supra* note 15.

straightforward way for covered entities to determine if information is adequately de-identified.[58] It involves the removal of eighteen identifiers to ensure that the information cannot be traced back to an individual.[59] Importantly, the covered entity or business associate cannot harbor actual knowledge (not to suggest a flaw in the Safe Harbor method, but rather to acknowledge the legal standard it incorporates) that the data can be re-identified.[60] However, even if a covered entity is aware of methods to re-identify de-identified information—either on its own or when combined with other data—merely knowing about these methods does not mean the covered entity has "actual knowledge."[61] The Office for Civil Rights (OCR) does not expect HIPAA-covered entities to assume that all potential recipients of de-identified data have re-identification capabilities.[62] If the covered entity does have actual knowledge, that would mean the data is not de-identified, and therefore cannot be shared.[63]

Thus, when a patient provides PHI to a covered entity, that entity has the ability to de-identify the data in accordance with either method and subsequently license, share, or sell it to data brokers or other interested parties.[64] Once in the data ecosystem, the information can continue to circulate, passing through multiple entities, including data brokers.[65] Data brokers can then analyze the data using advanced analytics, integrate it with other datasets through linkage techniques, and package it into detailed profiles for resale.[66] The movement of de-identified data from its origin at a covered entity to its final destination at a data broker is far from a simple, linear process. Instead,

[58] COMM. ON STRATEGIES FOR RESPONSIBLE SHARING OF CLINICAL TRIAL DATA, BD. ON HEALTH SCIS. POL'Y & INST. OF MED., SHARING CLINICAL TRIAL DATA: MAXIMIZING BENEFITS, MINIMIZING RISK 209 (2015), https://www.ncbi.nlm.nih.gov/books/NBK285994/ [https://perma.cc/T4GE-ENDQ].

[59] *Guidance Regarding Methods for De-Identification*, *supra* note 15.

[60] *Id.*

[61] *Id.*

[62] *Id.*

[63] *Id.*

[64] *See* Katharine Miller, *De-Identifying Medical Patient Data Doesn't Protect Our Privacy*, STAN. UNIV. (July 19, 2021), https://hai.stanford.edu/news/de-identifying-medical-patient-data-doesnt-protect-our-privacy [https://perma.cc/S7XH-LGYL].

[65] *See id.*

[66] *Data Brokers: Key Players in the Data Selling Ecosystem*, *supra* note 14; *see* Gadotti et al., *supra* note 15, at 5 ("Recently, the Federal Trade Commission initiated a lawsuit against the data broker Kochava over concerns for the sale of location data that could be used to identify women who visit abortion clinics.").

it follows a complex and often opaque path, making oversight and accountability challenging.[67]

C. Differences Between De-identified, Pseudonymous and Anonymous Data

While de-identifying data is a prevalent form of data privacy, it is not the only approach. There are three general ways to protecting privacy with three levels of risk utility: (1) de-identified data, (2) pseudonymous data, and (3) anonymous data.[68] At one end, de-identified data presents high utility but also a high re-identification risk.[69] Pseudonymous data falls in the middle, with its level of risk dependent on various factors.[70] Anonymized data on the other end of the spectrum carries the lowest risk but also the lowest utility.[71] While the terms of de-identified and anonymized data are most often used interchangeably, each, including pseudonymous data, carries distinct implications regarding the potential for re-identification.[72]

The first approach to preserving privacy, de-identification, is a process by which data is stripped of direct identifiers (a piece of data that allows the user to be identified, e.g., Social Security numbers), so that it no longer provides the ability, to a degree, to identify an individual.[73] Once data has been de-identified, HIPAA places no restrictions on its use or disclosure, which means the data can be shared freely.[74] However, de-identification does not eliminate the risk of re-identification, particularly through linkage attacks, which combine de-identified datasets with other non-de-identified data sets (e.g.,

[67] *See Data Brokers: Key Players in the Data Selling Ecosystem*, *supra* note 14; Tom Kemp, *A Closer Look at Data Brokers' Sources of Data*, MEDIUM (July 6, 2023), https://tomkemp00.medium.com/a-closer-look-at-data-brokers-sources-of-data-ded248f8d760 [https://perma.cc/96D9-W75R].

[68] Simson S. Garfinkel, NAT'L INST. OF STANDARDS AND TECH. INTERNAL REPORT 8053: DE-IDENTIFICATION OF PERSONAL INFORMATION 43 (2015), http://dx.doi.org/10.6028/NIST.IR.8053 [https://perma.cc/7DEG-MGS3].

[69] Gadotti et al., *supra* note 15, at 5.

[70] *See* Garfinkel, *supra* note 68, at 17.

[71] Amitai Richman, *Advantages and Disadvantages of Anonymized Data*, K2VIEW (Sept. 11, 2023), https://www.k2view.com/blog/anonymized-data/ [https://perma.cc/VDZ7-7DWV].

[72] Raphael Chevrier, Vasiliki Foufi, Christophe Gaudet-Blavignac, Arnaud Robert & Christian Lovis, *Use and Understanding of Anonymization and De-Identification in the Biomedical Literature: Scoping Review*, 21 J. MED. INTERNET RSCH. 1, 9 (2019); Gadotti et al., *supra* note 15, at 5.

[73] 45 C.F.R. § 164.514(a); *see* discussion *supra* Section II.B on expert determination and safe harbor methods.

[74] *See Guidance Regarding Methods for De-Identification*, *supra* note 15.

location data) from data brokers.[75] A well-documented example of this occurred in the mid-1990s, when Massachusetts released hospital records summarizing every state employee's hospital visits. [76] Then-Governor William Weld assured the public that the data had been properly protected, as fields containing direct identifiers—such as name, address, and Social Security number—had been removed. [77] However, the dataset still contained indirect identifiers.[78] Dr. Latanya Sweeney, a graduate student at the time, obtained the dataset and cross-referenced it with publicly available voter registration records.[79] By using just three pieces of indirect information (zip code, birth date, and gender), she successfully re-identified the governor's medical history, including his diagnoses and prescriptions.[80] Linkage attacks like this remain a major concern, as the volume of publicly available and commercially aggregated data has expanded significantly, making it easier to match individuals across datasets. [81] The growing sophistication of advanced analytics further amplifies re-identification risks, challenging the assumption that de-identified data risks are low.[82]

The second approach, pseudonymization, is a product of a de-identification technique that "removes both the association with a data subject and adds an association, a code or pseudonym, between a particular set of characteristics relating to the data subject and one or more pseudonyms," thus allowing data to be linked to the same individual across datasets without immediately revealing their identity.[83] For example, removing the first and last name from a profile and giving it a randomized number to identify the profile instead.[84] While pseudonymization is used in both health and non-health

[75] *See* Justin Sherman, *People Search Data Brokers, Stalking, and 'Publicly Available Information' Carve-Outs*, LAWFARE (Oct. 30, 2023) [hereinafter *People Search Data Brokers*], https://www.lawfaremedia.org/article/people-search-data-brokers-stalking-and-publicly-available-information-carve-outs [https://perma.cc/VWG4-4732]; Garfinkel, *supra* note 68, at 18.

[76] *People Search Data Brokers*, *supra* note 75.

[77] *Id.*

[78] *See id.*

[79] *Id.*

[80] Garfinkel, *supra* note 68, at 18.

[81] *See Data Broker Market Overview*, *supra* note 20.

[82] *See* Donald Farmer & Katie Terrell Hanna, *Advanced Analytics*, TECHTARGET (Nov. 2023), https://www.techtarget.com/searchbusinessanalytics/definition/advanced-analytics [https://perma.cc/F3PL-Z34P]; *see also* Anne Trafton, *Study Finds the Risks of Sharing Health Care Data Are Low*, MIT NEWS (Oct. 6, 2022), https://news.mit.edu/2022/patient-data-risks-low-1006 [https://perma.cc/WX3H-3HMT].

[83] Garfinkel, *supra* note 68, at 43.

[84] *See Pseudonymization*, IMPERVA, https://www.imperva.com/learn/data-security/pseudonymization/ [https://perma.cc/RU4F-QU7P] (last visited May 9, 2025).

contexts, its regulatory treatment varies depending on the governing state privacy law framework.[85]

In the healthcare context, HIPAA allows covered entities to include a pseudonym (or code) within a de-identified dataset, provided that the pseudonym is not derived from or related to an individual's personal information and is not capable of being translated to identify the individual.[86] Additionally, HIPAA requires that the covered entity does not use or disclose the pseudonym in a way that would enable re-identification outside of the covered entity itself.[87] Because only the covered entity retains the ability to re-identify individuals, and because the pseudonym itself does not directly link back to the data subject under HIPAA's standards, the data remains classified as de-identified.[88] If the covered entity later re-identifies the dataset, it would fall back under HIPAA's scope and be considered PHI once again.[89]

These theoretical risks become particularly salient in today's data ecosystem, where the combination of de-identified data with external datasets enables powerful re-identification techniques like linkage attacks.[90] The success of a linkage attack depends on multiple factors, including the availability of external reference datasets and the specificity of indirect identifiers.[91] This reflects a broader reality that data brokers possess extensive information about individuals and can, through linkage, form or further contribute to detailed profiles of individuals.[92] These profiles are created through direct data collection, inferred characteristics, or by purchasing datasets from other data

[85] *See, e.g.*, 2023 Tenn. Pub. Acts 408; VA. CODE ANN. § 59.1-577, 1-578 (West 2023).

[86] 45 C.F.R. § 164.514(c); *HIPAA Privacy Regulations: Other Requirements Relating to Uses and Disclosures of Protected Health Information: De-Identification of Protected Health Information - § 164.514(a)*, BRICKER GRAYDON [hereinafter *HIPAA Privacy Regulations: Other Requirements*], https://www.brickergraydon.com/insights/resources/key/HIPAA-Privacy-Regulations-Other-Requirements-Relating-to-Uses-and-Disclosures-of-Protected-Health-Information-De-Identification-of-Protected-Health-Information-164-514-a [https://perma.cc/3URN-6QF3] (last visited Feb. 1, 2025).

[87] *HIPAA Privacy Regulations: Other Requirements*, *supra* note 86.

[88] *See* 45 C.F.R. § 164.514(c).

[89] *De-Identification of PHI in Accordance with the HIPAA Privacy Rule*, UNIV. OF PA. SCH. OF NURSING [hereinafter *De-Identification of PHI*], https://www.nursing.upenn.edu/live/files/907-de-identification-of-phi-in-accordance-with-the#:~:text=Re%2Didentification,-The%20implementation%20specifications&text=If%20a%20covered%20entity%20or,meet%20the%20definition%20of%20PHI [https://perma.cc/W5SU-JBAT] (last visited Feb. 1, 2025).

[90] *See* Garfinkel, *supra* note 68, at 17–18; Sherman, *supra* note 5.

[91] *See id.* at 19. This Note does not address the statistical variations of linkage attacks. *See generally* Gadotti et al., *supra* note 15.

[92] *Data Brokers: Key Players in the Data Selling Ecosystem*, *supra* note 14; *see Data Broker Market Overview*, *supra* note 20.

brokers.[93] When a de-identified dataset is introduced into this ecosystem, a data broker can use its existing knowledge to link individuals back to their de-identified records, effectively re-identifying them.[94] Once this occurs, the data broker has an even more complete profile, which can then be sold to marketers, advertisers, or other entities interested in highly targeted consumer data.[95]

A well-known example of the risks associated with pseudonymization and linkage attacks is the Netflix Prize dataset, released in 2006 as part of a machine learning competition.[96] Netflix removed direct identifiers and assigned a unique, consistent pseudonym to each user.[97] However, researchers demonstrated how users could still be re-identified.[98] By cross-referencing the Netflix dataset with publicly available IMDB reviews, which contained movie ratings and timestamps, they were able to match individuals between the two datasets.[99] Even though Netflix had stripped away names, patterns of behavior, such as viewing history and rating timestamps, were enough to infer identities.[100]

In the context of healthcare, this raises an important consideration. Although a covered entity is legally restricted under HIPAA from disclosing the re-identification key, data brokers operate under far fewer constraints.[101] It is not a far leap to assume that even if covered entities comply with HIPAA, data brokers could still use external data sources to re-identify individuals in de-identified datasets

[93] *See generally* Justin Sherman, *Data Brokers and Threats to Government Employees*, LAWFARE (Oct. 22, 2024, 1:45 PM), https://www.lawfaremedia.org/article/data-brokers-and-threats-to-government-employees [https://perma.cc/STQ5-3YU5].

[94] Tim Starks, *Proposed Data Broker Regulations Draw Industry Pushback on Anonymized Data Exceptions, Bulk Thresholds*, CYBERSCOOP (Apr. 22, 2024), https://cyberscoop.com/proposed-data-broker-regulations-draw-industry-pushback-on-anonymized-data-exceptions-bulk-thresholds/ [https://perma.cc/H2N4-CCEY].

[95] *See* Tom Kemp, *How Data Brokers Profile, Segment, and Score Us*, MEDIUM (July 26, 2023), https://tomkemp00.medium.com/how-data-brokers-profile-segment-and-score-us-5144af9a465 [https://perma.cc/9NUP-6KUA].

[96] Nico Otezia, *Data Privacy—The Netflix Prize Competition*, MEDIUM (July 2, 2022), https://medium.com/@EmiLabsTech/data-privacy-the-netflix-prize-competition-84330d01cc34 [https://perma.cc/DR8S-ZRPT].

[97] *Id.*

[98] *Id.*

[99] *Id.*

[100] *Id.*

[101] *See* Emile Ayoub & Elizabeth Goitein, *Closing the Data Broker Loophole*, BRENNAN CTR. FOR JUST. (Feb. 13, 2024), https://www.brennancenter.org/our-work/research-reports/closing-data-broker-loophole [https://perma.cc/3P7A-64EY].

that contain a pseudonym.[102] If this type of re-identification is not already happening, it is likely only a matter of time before it does.

The third approach to preserving privacy, anonymization, is where anonymized data is "data from which [a person] cannot be identified by the recipient of the information." [103] Anonymization provides a unique benefit that neither de-identified nor pseudonymous data offers: the anonymized data cannot be traced back to the individual.[104] Essentially, it is immune to privacy attacks.[105] Thus, it is in many ways the ultimate goal for preserving consumer privacy.[106] From a business' perspective, it is also beneficial because it can be used and shared without consent from the individuals whose data produced the anonymized version.[107] To achieve anonymity, there must be a balance between privacy and utility favoring privacy more so than de-identified or pseudonymous.[108] However, anonymous data has the potential to aid in quelling re-identification concerns present in the above-mentioned data types.[109]

D. Section 5 of The Federal Trade Commission (FTC) Act

The lack of comprehensive federal privacy legislation in the US has left significant regulatory gaps in privacy protection, particularly regarding data brokers, whose ability to aggregate and link de-identified and pseudonymized datasets creates substantial privacy risks.[110] In response, the FTC has attempted to fill this void by using its enforcement powers to police deceptive and unfair practices in the data economy.[111] Section 5 of the FTC Act prohibits "unfair or deceptive acts or practices in or affecting commerce."[112] Such broad authority allows

[102] *See, e.g.*, Otezia, *supra* note 96.

[103] Garfinkel, *supra* note 68, at 39.

[104] Gadotti et al., *supra* note 15, at 1.

[105] *See id.*

[106] *Id.*

[107] *Id.*

[108] *See id.*

[109] *Id.*

[110] Ayoub & Goitein, *supra* note 101.

[111] *See A Brief Overview of the Federal Trade Commission's Investigative, Law Enforcement, and Rulemaking Authority*, FED. TRADE COMM'N [hereinafter *Overview of the FTC's Authority*], https://www.ftc.gov/about-ftc/mission/enforcement-authority [https://perma.cc/2UDG-B39L] (May 2021).

[112] *Id.*

the FTC to regulate businesses, including data brokers that engage in practices harmful to consumers.[113]

A practice is deceptive if a business makes misleading claims or omits key information in a way that misleads consumers.[114] A practice is unfair if it causes substantial harm to consumers that is not outweighed by benefits and cannot be reasonably avoided.[115] The FTC can then take enforcement action against data brokers if they misrepresent their data anonymization practices,[116] fail to disclose how they use or sell consumer data, or collect and monetize sensitive information without consent.[117] Data brokers may be at particular risk for violating section 5 when they re-identify de-identified datasets or sell personal information under misleading claims of anonymity.[118] While FTC action offers some form of accountability, it is largely reactive, case-by-case, and depends on fitting modern data risks into older standards designed for more traditional forms of consumer harm. Although the FTC has recently taken a more active role in policing data broker practices under section 5 authority, such enforcement is inherently limited and politically contingent. If the FTC step back from such efforts, the need for comprehensive federal privacy legislation becomes even more urgent, especially in providing consistent protections across states.

E. State Consumer Privacy Laws, Tennessee Definitions

The fragmented approach to data privacy in the United States has left a regulatory gap, prompting states to develop consumer privacy

[113] *See* *Privacy and Security Enforcement*, FED. TRADE COMM'N, https://www.ftc.gov/news-events/topics/protecting-consumer-privacy-security/privacy-security-enforcement [https://perma.cc/K76T-W2B9] (last visited Feb. 16, 2025).

[114] *Overview of the FTC's Authority*, *supra* note 111.

[115] *Id.*

[116] Data brokers often claim that their data is "anonymized." *See* Ayoub & Goitein, *supra* note 101.

[117] *See* Lesley Fair, *What Goes on in the Shadows: FTC Action Against Data Broker Sheds Light on Unfair and Deceptive Sale of Consumer Location Data*, FED. TRADE COMM'N (Jan. 9, 2024), https://www.ftc.gov/business-guidance/blog/2024/01/what-goes-shadows-ftc-action-against-data-broker-sheds-light-unfair-deceptive-sale-consumer-location [https://perma.cc/W6CU-92BH]; *FTC Takes Action Against Mobilewalla for Collecting and Selling Sensitive Location Data*, FED. TRADE COMM'N (Dec. 3, 2024) [hereinafter *FTC Takes Action Against Mobilewalla*], https://www.ftc.gov/news-events/news/press-releases/2024/12/ftc-takes-action-against-mobilewalla-collecting-selling-sensitive-location-data [https://perma.cc/2SAF-F46E].

[118] *See* Ayoub & Goitein, *supra* note 101.

laws to address the growing concerns about data protection.[119] California was the first to pass such legislation, the California Consumer Privacy Act (CCPA), in 2018.[120] Since then, nineteen states, including Tennessee with its enactment of TIPA, have passed comprehensive privacy legislation.[121] State privacy laws vary in scope and focus. Some, like TIPA, are considered comprehensive consumer privacy laws because they regulate a broad range of personal data across industries.[122] Others are more targeted, such as the data broker laws in Texas and Vermont, which primarily impose data broker registration and disclosure requirements,[123] or health-specific laws like Washington's My Health My Data Act, which focuses exclusively on protecting health information.[124] Though their aims differ, they share a critical weakness: all exempt data de-identified under HIPAA, leaving such data vulnerable to re-identification by data brokers.[125]

TIPA provides a useful example of how most comprehensive data privacy laws are written.[126] For consumer rights (e.g., right to delete and right to correct) to apply under TIPA, the data must fall under the definition of personal information, defined as "information that identifies, relates to, or describes a particular consumer or is reasonably capable of being directly or indirectly associated or linked with, a particular consumer."[127] However, TIPA exempts de-identified data handled by HIPAA-covered entities and their business associates, as well as publicly available information, from its definition of personal information.[128] As a result, these categories of data are not subject to consumer right provisions, even though de-identified data may carry re-identification risks.[129]

[119] Brenna Goth, *Varied Data Privacy Laws Across States Raise Compliance Stakes*, BLOOMBERG L. (Oct. 11, 2023, 4:00 AM), https://news.bloomberglaw.com/in-house-counsel/varied-data-privacy-laws-across-states-raise-compliance-stakes [https://perma.cc/S5QX-V7TA].

[120] C Kibby, *US State Privacy Legislation Tracker*, IAPP (Mar. 4, 2025), https://iapp.org/resources/article/us-state-privacy-legislation-tracker/ [https://perma.cc/G5NY-6ZQL].

[121] *Id.*; 2023 Tenn. Pub. Acts 408.

[122] 2023 Tenn. Pub. Acts 408; *see* Goth, *supra* note 119.

[123] TEX. BUS. & COM. CODE §§ 509.005(a), 509.007(11) (2023); VT. STAT. ANN. tit. 9, § 2446.

[124] WASH. REV. CODE § 19.373.005(2).

[125] TEX. BUS. & COM. CODE §§ 509.005(a), 509.007(11) (2023); VT. STAT. ANN. tit. 9, § 2446; WASH. REV. CODE § 19.373.100(E)(viii) (2025).

[126] *See* 2023 Tenn. Pub. Acts 408.

[127] *Id.*

[128] *Id.* De-identified data is defined as "data that cannot reasonably be linked to an identified or identifiable natural person, or a device linked to that individual." *Id.*

[129] *See id.*

Consistent with its narrow scope, TIPA also differentiates between other forms of data.[130] While it does not explicitly address anonymized data, it introduces a related category: pseudonymous data.[131] Under TIPA, pseudonymous data is defined as:

> Personal information that cannot be attributed to a specific natural person without the use of additional information, so long as the additional information is kept separately and is subject to appropriate technical and organizational measures to ensure that the personal information is not attributed to an identified or identifiable natural person.[132]

However, by only defining pseudonymous data and leaving de-identified data unaddressed, TIPA reinforces the broader trend of under-regulating data that still carries re-identification risks.[133] Examining how various states approach exemptions and consumer rights reveals that de-identified data is consistently excluded from regulatory protections, despite the growing risk of re-identification.[134] Like comprehensive consumer privacy laws, Washington's My Health My Data Act also exempts de-identified data from its purview, reinforcing the broader regulatory trend.[135] Similarly, state data broker laws, such as those in Texas, also exclude de-identified data from regulation, meaning data brokers may freely sell or share such data.[136] While the Texas Data Broker Law prohibits data brokers from intentionally re-identifying de-identified data within a single dataset, it does not prevent them from purchasing separate datasets and cross-referencing them to infer identities.[137] Because the law narrowly defines re-identification as reversing de-identification in the original dataset, data brokers can still piece together individual profiles by linking de-identified data with external datasets, effectively bypassing the law's intent.[138]

[130] *Id.*

[131] *See* Lisa Pilgram, Thierry Meurers, Bradley Malin, Elke Schaeffner, Kai-Uwe Eckardt & Fabian Prasser, *The Costs of Anonymization: Case Study Using Clinical Data*, 26 J. MED. INTERNET RSCH. 1, 2 (2024), https://www.jmir.org/2024/1/e49445; 2023 Tenn. Pub. Acts 408.

[132] 2023 Tenn. Pub. Acts 408.

[133] *See id.*

[134] N.Y. S-929, 2025–2026 Legis. Session (N.Y. 2025).

[135] *See id.* ("Location or payment information that relates to an individual's physical or mental health . . . ."); Katelyn N. Ringrose, Amy C. Pimentel, Alexander H. Southwell & Sam Siegfried, *New York Assembly Passes Restrictive Health Information Privacy Act*, MCDERMOTT WILL & EMERY (Jan. 24, 2025), https://www.mwe.com/insights/new-york-passes-restrictive-health-information-privacy-act/ [https://perma.cc/CJ26-VAGJ].

[136] TEX. BUS. & COM. CODE § 509.008(11) (2023).

[137] *Id.* at § 509.002.

[138] *See id.*

Of note, New York's Health Information Privacy Act (NYHIPA), recently passed by the New York Assembly and awaiting Governor Kathy Hochul's signature, would still exempt de-identified data under HIPAA but separately define "regulated health information" as any reasonably linkable data related to an individual's physical or mental health, including location, payment data, and derived inferences.[139] This broad definition suggests an effort to capture more health-related data beyond what HIPAA covers, particularly in nontraditional contexts like consumer apps and digital services.[140] New York's framework departs from traditional assumptions about what types of data pose re-identification risks, expanding protections to encompass a wider range of reasonably linkable information that could reveal an individual's health status when combined with de-identified data.[141]

As illustrated by the state laws discussed above, the data privacy landscape in the United States is replete with inconsistent definitions and exemptions, reflecting a patchwork system of data protection that varies by state.[142] There seems to be a shift, looking to New York as evidence, toward recognizing other types of data that present re-identification risks,[143] an issue that the European Union has long sought to address through its stricter, more unified framework for data privacy, the General Data Protection Regulation (GDPR)[144]

1. The General Data Protection Regulation (GDPR)

GDPR is Europe's most comprehensive privacy and security law, safeguarding individuals within the European Union by regulating personal data collection, processing, and storage.[145] Data brokers may be subject to GDPR if they handle the personal data of EU citizens, meaning they may simultaneously be subject to both US and EU laws depending on the datasets processed.[146] Unlike TIPA and other US privacy laws,[147] GDPR takes a more expansive approach to the

[139] *See* N.Y. S-929 §1120(2); Malgorzata Poddębniak, *Everything You Need to Know About the New York Health Information Privacy Act*, PIWIK PRO (Feb. 19, 2025), https://piwik.pro/blog/new-york-health-information-privacy-act-nyhipa/ [https://perma.cc/YFK4-FVZW].

[140] *See id.*; 45 C.F.R. § 164.514(b)(2).

[141] *See* N.Y. S-929 §1120(2).

[142] Goth, *supra* note 119.

[143] *See* N.Y. S-929 §1120(2).

[144] *What is GDPR, the EU's New Data Protection Law?*, GDPR.EU, https://gdpr.eu/what-is-gdpr/ [https://perma.cc/K29F-NMJT] (last visited Nov. 1, 2024).

[145] *Id.*

[146] *See* Hannah Ruschemeier, *Data Brokers and European Digital Legislation*, 9 EUR. DATA PROT. L. REV. 27, 33 (2023).

[147] *See* 2023 Tenn. Pub. Acts 408.

definition of personal data.[148] Under Article 4(1), personal data includes any information relating to an identified or identifiable individual, further clarifying that data subjects are deemed identifiable if they can directly *or* indirectly be identified, which thus subjects the data to GDPR compliance.[149] GDPR does not explicitly recognize a category equivalent to the HIPAA notion of de-identified data. Instead, GDPR recognizes two binary categories: anonymized data,[150] data that is processed to ensure non-identifiability either by the data controller or by any other person, and pseudonymized data,[151] data that cannot be attributed to a specific individual without additional information that is kept separately and subject to technical and organizational safeguards.[152] Anonymized data is not subject to consumer protections under GDPR, unlike pseudonymized data, which is data that can "directly or indirectly" identify an individual.[153]

Although GDPR does not explicitly define de-identified data, its broad definition of personal data nonetheless captures what this Note would describe as de-identified data: data with downstream re-identification risks.[154] GDPR recognizes such risks in its treatment of pseudonymized data, and as a result, imposes clear obligations on entities processing such data.[155] By contrast, current US laws, federal

[148] *See* Regulation (EU) 2016/679, of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation), art. 4(1), 2016 O.J. (L 119) 1 [hereinafter GDPR]; 2023 Tenn. Pub. Acts 408.

[149] GDPR, *supra* note 148; *see also* Emily M. Weitzenboeck, Pierre Lison Malgorzata Cyndecka & Malcolm Langford, *The GDPR and Unstructured Data: Is Anonymization Possible?*, 12 INT'L. DATA PRIV. LAW 184, 189 (2022).

[150] GDPR, *supra* note 148, at recital 26.

[151] "The processing of personal data in such a way that the data can no longer be attributed to a specific data subject without the use of additional information, as long as such additional information is kept separately and subject to technical and organizational measures to ensure non-attribution to an identified or identifiable individual." *Id.* at art. 4(5).

[152] *Id.*

[153] *Id.* at recital 26.

[154] *See id.* at art. 4(1); *but see* Gadotti et al., *supra* note 15, at 2 ("De-identification in [US privacy laws] plays a role similar to anonymization in the GDPR—strongly reducing the legal obligations that apply to the processing of personal data.").

[155] GDPR, *supra* note 148, at art. 25(1); *but see* Case T-557/20, Single Resol. Bd. (SRB) v. Eur. Data Prot. Supervisor (EDPS), 2023 ECLI:EU:T:2023:219, ¶¶ 101–05 (Apr. 26, 2023) (holding that pseudonymized data is not considered personal data if the recipient lacks the means to re-identify individuals). This approach mirrors the US HIPAA model, which does not impose further privacy protections once health data is de-identified, as long as there is no "actual knowledge" that the recipient can re-identify the data. *See Guidance Regarding Methods for De-Identification*, *supra* note 15. Similarly, in SRB, the Court focused on whether the recipient had the ability to re-identify, rather than the broader risks of re-identification through external data sources. *SRB*, 2023 ECLI:EU:T:2023:219, at ¶¶ 101–05.

and state, offer no consistent protections for data with similar re-identification risks, allowing data brokers to operate with minimal oversight.[156] Such inconsistencies raise downstream privacy concerns given data brokers' expansive data access and powerful capabilities to link such data.[157]

Part I of this Note has outlined the fragmented legal framework governing de-identified data in the United States, where both HIPAA and state-level privacy laws, including data broker statutes, exempt such data from meaningful regulation. Despite well-documented risks of re-identification, especially when de-identified data is combined with auxiliary datasets, these exemptions provide an avenue for data brokers to operate with minimal oversight. In contrast, GDPR offers a more cohesive and risk-aware model by treating potentially linkable data as subject to privacy protections. These shortcomings in the United States create significant regulatory gaps, gaps that Part II will explore in depth through a specific state privacy law.

II. REGULATORY GAPS AND RE-IDENTIFICATION RISKS BY DATA BROKERS

This Note will now conduct an in-depth discussion of TIPA, Tennessee's consumer privacy law, using it as a model to examine how exemptions for de-identified data in a consumer privacy law create loopholes that data brokers can exploit to handle and market health data. This discussion further explores the risks of re-identification from the point at which the de-identified data leaves a covered entity, and why enforcement efforts to halt the dissemination of such data to data brokers is largely ineffective at the state and federal level.

A. HIPAA's Treatment of De-identified Data and Its Consequences in TIPA's Regulatory Framework

The distinction between de-identified and pseudonymous data under TIPA creates a regulatory loophole for data brokers to exploit. Under HIPAA, de-identified data remains classified as such even when a covered entity assigns a pseudonym or re-identification code to an identifiable piece of information, provided the code is not derived from or related to the individual's information and cannot be translated to

[156] *See* discussion *supra* Section I.E.

[157] *See* discussion *supra* Section I.E.

reveal their identity.[158] Because only the covered entity can re-identify the data and the pseudonym does not directly link to the data subject under HIPAA, the data remains de-identified until re-identified, at which point it once again becomes PHI.[159] Thus, such data is not considered pseudonymous.[160]

Though the intent of the TIPA was not solely to regulate health data, but rather to encompass a broader scope of general data privacy, its framework helps explain why data brokers remain largely unregulated.[161] This Note uses TIPA as an example of one end of the regulatory spectrum, where de-identified data is exempt from key consumer protections and data brokers are not meaningfully regulated unless they meet narrow statutory definitions, to show that imposing distinctions between pseudonymous and de-identified data does little to curb their activities.[162]

Even if one were to argue that assigning a pseudonym to de-identified data makes it pseudonymous, such data would still be exempt under TIPA.[163] In Tennessee, like many other states, pseudonymous data is excluded from all consumer rights protections, if the controller of the data implements safeguards that prevent re-identification.[164] The assumption then is that businesses do implement safeguards, as it is in their best interest, effectively leaving consumers with no rights.[165] This limited approach to pseudonymous data is not unique to Tennessee. For example, Virginia and Connecticut also restrict consumer rights in this area, but in a different way: they only grant consumers the right to opt out of the processing of their personal data (including pseudonymous data) for purposes of, among other things, targeted advertisement.[166] Thus, while distinguishing between de-identified and pseudonymous data offers consumers little

[158] GDPR, *supra* note 148. *But see SRB*, ECLI:EU:T:2023:219, ¶¶ 94–105 (Apr. 26, 2023) (holding that pseudonymized data is not considered personal data if the recipient lacks the means to re-identify individuals).

[159] *See* 45 C.F.R. § 164.514(c); *De-Identification of PHI*, *supra* note 89.

[160] *See De-Identification of PHI*, *supra* note 89.

[161] *See* 2023 Tenn. Pub. Acts 408. Unlike Washington's My Health My Data Act, TIPA does not explicitly define its intent as health data privacy, meaning the law covers a much broader swath of data, such as financial information that the Washington law does not (at least within the My Health My Data Act). *See id.*; WASH. REV. CODE § 19.373.005(2) ("With chapter 191, Laws of 2023, the legislature intends to provide heightened protections for Washingtonian's health data.").

[162] *See* 2023 Tenn. Pub. Acts 408; Pell et al., *supra* note 21 (noting that focusing on the activity of data brokerage, not classifications based on the type of data broker or if they make enough money to qualify as a data broker is not the main concern).

[163] 2023 Tenn. Pub. Acts 408.

[164] *Id.*; Colo. Rev. Stat. §§ 6-1-1307 (2023).

[165] *See id.*; 2023 Tenn. Pub. Acts 408.

[166] VA. CODE ANN. § 59.1-577 (West 2023); Conn. Pub. Acts 22-15 § 4(a)(5) (2022).

protection under a law like TIPA, it becomes more meaningful in states like Virginia and Connecticut, where limited opt-out rights apply, albeit in very specific instances.[167]

From the perspective of a data broker, the regulatory distinction between de-identified and pseudonymous data is critical. It creates opportunities for them to sidestep regulation by framing their practices as falling outside of the law. Even if a data broker complies with the processing requirements set forth by TIPA, the data broker's argument may well be that of this Note: that the data broker is processing de-identified data, which is categorically excluded from consumer rights protections.[168] In a state like Virginia where consumers have the right to opt out of the processing of personal data, including pseudonymous data, for limited purposes like targeted advertising, this distinction becomes a strategic tool: by classifying data as de-identified rather than pseudonymous, data brokers can avoid triggering even those narrow opt-out rights.[169] The central issue, then, is that because TIPA exempts de-identified data from consumer protections, data brokers, with their powerful analytical tools and access to vast amounts of data can exploit the residual risks to re-identify individuals.[170]

1. Covered Entities Sell De-Identified Data, While States Focus on Definitions, Not Risks

HIPAA removes protections from data once it is de-identified, allowing it to be freely shared or sold without consumer consent.[171] However, its de-identification standards fail to account for how data brokers use advanced analytics and auxiliary datasets to re-identify individuals.[172] Despite this risk, state laws like TIPA reinforce HIPAA's flawed assumptions by exempting de-identified data from consumer protections, prioritizing definitions over enforcement.[173]

By contrast, New York takes a step in the right direction by acknowledging that risks to identifiable information extend beyond rigid classifications, recognizing the potential for re-identification when datasets are linked.[174] New York's recognition of these risks

[167] *See id.*

[168] *See* 45 C.F.R. § 164.514(c); *De-Identification of PHI*, *supra* note 89.

[169] VA. CODE ANN. § 59.1-577 (West 2023).

[170] Stalla-Bourdillon, *supra* note 53; *Data Brokers: Key Players in the Data Selling Ecosystem*, *supra* note 14.

[171] *Guidance Regarding Methods for De-Identification*, *supra* note 15.

[172] *See Data Brokers: Key Players in the Data Selling Ecosystem*, *supra* note 14.

[173] *See* 2023 Tenn. Pub. Acts 408 § 47-18-3204(a)(4).

[174] *See* NY S-929 §1120(2).

demonstrates that state laws can move past binary classifications and instead focus on the actual risks posed by modern data practices, suggesting a broader need to reorient privacy legislation around risk, not definition.[175]

The scale of the data broker industry and its reliance on vast amounts of personal information[176] suggests that data brokers' business models depend, in part, on circumventing the re-identification risks imagined by HIPAA by acquiring auxiliary datasets to perform linkage attacks through advanced analytics, thereby re-identifying individuals,.[177] One example that substantiates this assumption is Vanderbilt Health's privacy policy. It states:

> We may use and share your Medical Information to create information that doesn't reveal who you are (de-identified information) as federal privacy law allows. We may also share your Medical Information with a business associate to create de-identified information, which we or others may use. We may use or share this information for any lawful purpose. *This includes, but is not limited to, commercial purposes without your permission and may allow third parties to do the same.*[178]

This policy reflects exactly what HIPAA permits: de-identified data can be sold or shared freely. But in practice, the framework does not account for downstream risks posed by data brokers. The example of Vanderbilt Health demonstrates that covered entities can and do share de-identified data with third parties, including data brokers.

B. There Is No Incentive for Data Brokers to Utilize Privacy-Preserving Record Linkage

Privacy-Preserving Record Linkage (PPRL) is widely used in healthcare and public health research to securely integrate data while complying with HIPAA's de-identification standards.[179] PPRL, at its simplest, allows datasets to be linked with one another across

[175] *See id.*

[176] Dimitri Shelest, *What Is a Data Broker? 2024 Insights on how They Collect, Use, and Sell Your Data*, ONEREP (Feb. 6, 2025), https://onerep.com/blog/what-are-data-brokers-and-how-come-they-sell-your-info [https://perma.cc/Y24A-NGSG].

[177] *See* Ayoub & Goitein, *supra* note 101.

[178] VANDERBILT HEALTH, NOTICE OF PRIVACY PRACTICES 4 (2023), https://www.vumc.org/information-privacy-security/sites/default/files/public_files/NPP_Final_Dec 2023.pdf [https://perma.cc/EUV4-KDA6] (emphasis added).

[179] HLN CONSULTING, DATA LINKAGE AND IDENTITY MANAGEMENT – PRIVACY PROTECTING RECORD LINKAGE (PPRL) MEETING SUMMARY 1 (2023), https://www.cdcfoundation.org/CDCFoundationPPRLSummary.pdf?inline#:~:text=Using%20PPRL%2C%20public%20health%20can,case%20counts%20or%20immunization%20rates [https://perma.cc/W3HX-PFV8].

organizations without exposing personal identities.[180] This approach is particularly effective in medical research and for insurers, where insights can be gained while maintaining patient privacy.[181] For example, an insurance company might link de-identified medical claims, pharmacy records, and lab results to assess chronic disease risk within a covered population.[182] By applying proprietary risk models to the linked data, the insurer can refine underwriting decisions or develop targeted health interventions.[183]

However, even after data is linked using PPRL, it would still be considered de-identified under TIPA, meaning it can be shared and sold freely and therefore offers no practical advantage to data brokers that already freely engage in such activities.[184] Further, once PPRL-linked datasets (the de-identified data) are passed downstream, whether by a covered entity or business associate, they can be lawfully acquired by data brokers.[185] For example, HealthVerity, a data platform that performs PPRL, states in its privacy policy that it "may sell or disclose de-identified information to third parties," and when it does so, it "prohibits the third party from attempting to re-identify a person from the de-identified information."[186] But such contractual restrictions do not prevent data brokers from combining de-identified datasets with auxiliary sources to infer or reconstruct identity.[187] Thus, while PPRL may functionally preserve privacy, it does not prevent the downstream dissemination of de-identified data.[188] In fact, data brokers have little incentive to adopt PPRL at all, as they can acquire and link de-identified datasets using advanced analytics without restriction.[189] Because de-identified data is already exempt under TIPA, PPRL is not

[180] *What is Privacy-Preserving Record Linkage (PPRL) and Why Does It Matter?*, DATAVANT (Aug. 14, 2023), https://www.datavant.com/hipaa-privacy/privacy-preserving-record-linkage [https://www.datavant.com/hipaa-privacy/privacy-preserving-record-linkage].

[181] *Id.*

[182] *Synchronize Risk, Rebates and Rewards*, HEALTHVERITY, https://healthverity.com/insurance/ [https://perma.cc/R8MG-2W5V] (last visited May 13, 2025).

[183] *Id.*

[184] *See* 2023 Tenn. Pub. Acts 408 § 47-18-3204(a)(4); 45 C.F.R. § 164.514(a); *Data Brokers: Key Players in the Data Selling Ecosystem*, *supra* note 14.

[185] *See* 2023 Tenn. Pub. Acts 408 § 47-18-3204(a)(4); TEX. BUS. & COM. CODE § 509.001(11) (2023).

[186] *Privacy and Cookie Notice*, HEALTHVERITY, https://healthverity.com/privacy-policy/#:~:text= HealthVerity%20may%20sell%20or%20disclose%20de%2Didentified%20information%20to%20third%20parties [https://perma.cc/TJS5-WHR3] (last visited Feb. 1, 2025).

[187] *Id.*

[188] *See id.*

[189] *See* Reviglio, *supra* note 4, at 12 n.20; *Data Brokers: Key Players in the Data Selling Ecosystem*, *supra* note 14; 2023 Tenn. Pub. Acts 408 § 47-18-3204(a)(4).

only ineffective but also redundant, offering no meaningful legal or practical safeguard in this context.[190]

C. FTC's Role in Regulating Data Brokers

The FTC has taken on the bulk of enforcement actions against data brokers in recent years, though no specific case has directly addressed the handling of de-identified data.[191] The FTC has primarily scrutinized data brokers under section 5 of the FTC Act, targeting unfair and deceptive practices related to consumer privacy.[192] While the FTC has attempted to act, TIPA's legal framework makes no such effort because it flatly exempts de-identified data.

In January 2024, the FTC prohibited the data broker X-Mode Social from selling individuals' sensitive location data, citing concerns that such information could reveal visits to healthcare facilities, places of worship, and other private locations.[193] The FTC found that X-Mode misled users of third party applications about their ability to opt out of data collection, illustrating how data brokers often obscure the true privacy risks associated with their practices.[194]

Similarly, when a company claims that de-identified data is protected but fails to disclose its potential for re-identification, the company may be engaging in deceptive practices by omitting material information about privacy risks. [195] These practices may also be considered unfair if a company expose consumers to privacy harms that "cause or are likely to cause" substantial injury, such as intrusive marketing (such as sending targeted advertisements about sensitive health conditions based on location data), that consumers cannot reasonably avoid.[196] In this context, "cannot reasonably avoid" means that because consumers are not provided with "full information

[190] 2023 Tenn. Pub. Acts 408 § 47-18-3204(a)(4); *see What Is Privacy-Preserving Record Linkage (PPRL) and Why Does It Matter?*, *supra* note 180; *see* Anushka Vidanage, Thilina Ranbaduge, Peter Christen & Rainer Schnell, *A Taxonomy of Attacks on Privacy-Preserving Record Linkage*, 12 J. PRIV. & CONFIDENTIALITY 1, 2–3 (2022).

[191] *See, e.g.*, *FTC Takes Action Against Mobilewalla*, *supra* note 117; *FTC Order Prohibits Data Broker X-Mode Social and Outlogic from Selling Sensitive Location Data*, FED. TRADE COMM'N (Jan. 9, 2024) [hereinafter *FTC Prohibits Data Brokers Selling Sensitive Data*], https://www.ftc.gov/news-events/news/press-releases/2024/01/ftc-order-prohibits-data-broker-x-mode-social-outlogic-selling-sensitive-location-data [https://perma.cc/33HS-GMRJ].

[192] *E.g.*, *FTC Takes Action Against Mobilewalla*, *supra* note 117.

[193] *FTC Prohibits Data Brokers Selling Sensitive Data*, *supra* note 191.

[194] *Id.*

[195] *See* Ayoub & Goitein, *supra* note 101.

[196] *See Overview of the FTC's Authority*, *supra* note 111.

regarding the Compan[y]'s usage and purpose [of the data]" they are unable to consent to the data brokers' data-sharing practices.[197]

It may well be the case that companies' privacy policies acknowledge these risks while still enabling extensive data exchanges.[198] Athenahealth's privacy policy, for example, permits the disclosure of de-identified patient information to third parties such as service providers or academic researchers, but only when permitted by contractual obligations and HIPAA. [199] This language reflects an implicit acknowledgment that de-identified data, while legally exempt from privacy protections, still carries risks—particularly when shared beyond the originating entity.[200]

Yet, while the FTC has signaled increasing concern about data practices that exploit data broker practices, its case-by-case enforcement approach cannot substitute for clear statutory safeguards. TIPA, like HIPAA, explicitly exempts de-identified data from consumer protection, leaving similar privacy risks unaddressed at the state level.[201] In the absence of meaningful restrictions on how de-identified data can be used or combined, these exemptions risk enabling the same harmful practices the FTC is only beginning to police. [202] As data brokers' capabilities continue to grow with increased sophistication in advanced analytics (e.g., machine learning), HIPAA's de-identification methods, and by extension, TIPA's reliance on them, can no longer keep pace with the increased risk of re-identification, especially given covered entities make clear that they do sell de-identified data. [203] Without clearer regulatory intervention that restricts data brokers' ability to re-identify de-identified data, state laws will be ineffective in fulfilling their intended purpose of protecting consumer privacy.

[197] Kirk J. Nahra, Genesis Ruano, Amy Olivero & Ali A. Jessani, *Recent Enforcement Actions Signal FTC Focus on Protecting Location Data*, WILMERHALE (Feb. 9, 2024), https://www.wilmerhale.com/en/insights/blogs/wilmerhale-privacy-and-cybersecurity-law/20240209-recent-enforcement-actions-signal-ftc-focus-on-protecting-location-data [https://perma.cc/WBJ2-KWZX]; *see* Fair, *supra* note 117.

[198] *See* Hannah Burks, *How Healthcare Marketing Is Catching Up*, HEALTHVERITY (Jan. 27, 2020), https://blog.healthverity.com/linking-patient-data-to-digital-touch-points [https://perma.cc/7J36-VT3D].

[199] *Athenahealth Privacy Policy*, ATHENAHEALTH (May 9, 2025), https://www.athenahealth.com/privacy-rights?utm_source.

[200] *Id.*

[201] 2023 Tenn. Pub. Acts 408 § 47-18-3210(a)(13).

[202] *Id.*

[203] *See* VANDERBILT HEALTH, *supra* note 178.

III. TOWARD COMPREHENSIVE PRIVACY PROTECTIONS

To address the growing privacy risks posed by data brokers and the re-identification of de-identified health data, a comprehensive federal approach is necessary. This Part will focus on two key components of such a solution: regulatory reforms and technological advancements. First, federal regulators should take immediate steps to close the state regulatory gaps that allow the unregulated sharing and selling of de-identified data, rather than relying on a patchwork of state laws.[204] While the eventual adoption of a comprehensive federal privacy framework is necessary, lawmakers can begin by expanding the definition of data brokers to include any entity engaged in data monetization, impose strict oversight on the sale and sharing of de-identified health data, and restrict covered entities under HIPAA from selling such data unless for narrowly defined purposes.[205] Such measures would help to provide greater regulation of the data broker industry.

Second, the federal framework should promote synthetic data generation as a technological safeguard to reduce reliance on de-identified data and eliminate many of the risks associated with re-identification.[206] This Part will also examine how synthetic data usage can be integrated into privacy regulations, particularly through industry-wide validation standards that ensure its reliability for research and analytics.[207] By encouraging the adoption of synthetic data and incorporating it into existing legal frameworks such as HIPAA and GDPR, regulators can provide organizations with a privacy-preserving alternative while maintaining the benefits of data-driven innovation.[208]

[204] *See* discussion *infra* Part IV.

[205] Even the new data broker law, enacted by former President Biden, exempts de-identified data but does have some promising consumer privacy mechanisms like deletion requests. *President Biden Signs Law Limiting Data Broker Sales*, ELEC. PRIV. INFO. CTR (Apr. 24, 2024), https://epic.org/president-biden-signs-law-limiting-data-broker-sales/ [https://perma.cc/9PTS-7888].

[206] *See* Steven M. Bellovin, Preetam K. Dutta & Nathan Reitinger, *Privacy and Synthetic Datasets*, 22 STAN. TECH. L. REV. 1, 36, 48–49 (2019).

[207] *See, e.g.*, Matthew S. McCoy, Anita L. Allen, Katharina Kopp, Michelle M. Mello, D.J. Patil & Pilar Ossorio, *Ethical Responsibilities for Companies that Process Personal Data*, 23 AM. J. BIOETHICS 1, 15 (2023).

[208] *See* James Jordon, Lukasz Szpruch, Florimond Houssiau, Mirko Bottarelli, Giovanni Cherubin, Carsten Maple, Samuel N. Cohen & Adrian Weller, *Synthetic Data – What, Why and How?*, ROYAL SOC'Y 1, 5–6 (2022).

A. Regulate the Sale, Sharing, and Licensing of De-Identified Data

Regulating the sale, sharing, and licensing (collectively, data brokerage activities) of de-identified data requires a more targeted approach that directly addresses the regulatory gaps in state consumer privacy laws, as well as the lack of federal privacy legislation, exploited by data brokers.[209] One of the most significant weaknesses in state privacy laws is the broad exemption for de-identified data, which assumes that once direct identifiers are removed, the data poses no privacy risk.[210] However, modern data analytics will continue to improve linkage capabilities between de-identified and auxiliary datasets, allowing data brokers to continue aggregating, analyzing, and selling consumer profiles without meaningful oversight.[211] To address this gap, state privacy laws must be amended to regulate the downstream use of de-identified data. First, state privacy laws should remove existing exemptions for de-identified data, thereby ensuring that consumer protections extend to all forms of data, regardless of whether explicit identifiers have been removed under HIPAA. Second, states should implement clear restrictions on the sharing of de-identified data by covered entities under HIPAA, and require transparency (e.g., specifically with whom they share data) in data transactions so that regulators and consumers can understand how de-identified data is used.[212] Such a requirement would eliminate the ability of data brokers to avoid regulation by claiming they only handle de-identified data.[213]

The issue of re-identification is compounded by the narrow regulatory definition of data brokers.[214] Many of the comprehensive consumer privacy and existing data broker-specific laws, such as that in California, apply only to entities whose primary business is selling or licensing data.[215] This definition fails to capture the full scope of entities engaged in data monetization, which thus allows major technology companies and financial institutions, among others, that collect data directly from individuals (first-party data brokers) to escape regulation despite their significant role in the data brokerage

[209] Pell et al., *supra* note 21 (noting that focusing on the activity of data brokerage better identifies the overall concern of data brokers).

[210] *See* 2023 Tenn. Pub. Acts 408 § 47-18-3210(a)(13).

[211] *See* Gadotti et al., *supra* note 15, at 5.

[212] *See, e.g.*, McCoy et al., *supra* note 207.

[213] *See* discussion *supra* Section II.A.

[214] *See* Pell et al., *supra* note 21.

[215] VT. STAT. ANN. tit. 9, § 2430(4)(A).

ecosystem.[216] To be effective in protecting de-identified data from being linked to auxiliary datasets (e.g., geolocation), regulatory frameworks must expand the definition of data brokers to include any entity that sells, shares, or licenses consumer data, regardless of whether data brokerage is its primary business model or its relationship with consumers. [217] Current state privacy laws also impose arbitrary applicability thresholds, such as revenue-based classifications or the number of consumer records processed, that prevent many entities engaged in data processing from being classified as data brokers.[218] Removing these artificial distinctions would prevent data brokers from structuring their operations in ways that evade regulation.[219] Further, federal regulators should require data brokers to disclose their partnerships with third-party entities that facilitate re-identification, including advertisers and analytics firms.[220] Transparency regarding these relationships would help regulators track how data flows through the ecosystem and prevent companies from operating in the shadows.[221]

Beyond redefining data brokers and closing state-level gaps, a more effective approach would cut off the flow of de-identified data at its source. If the goal is to halt the movement of consumer information before it reaches data brokers and is further disseminated, restrictions must be placed on the brokerage of de-identified data by covered entities and other data processors (e.g., business associates of covered entities).[222] Covered entities under HIPAA should be prohibited from sharing de-identified data to third parties (including data brokers) unless it is for a narrow set of approved purposes, such as public health

[216] *See id.*; discussion *supra* Section I.A.

[217] *See* discussion *supra* Section I.A; Pell et al., *supra* note 21.

[218] *See, e.g.*, N.H. Rev. Stat. Ann. § 507-H:2.

> "I. This chapter applies to persons that conduct business in this state or persons that produce products or services that are targeted to residents of this state that during a one-year period:
>
> (a) Controlled or processed the personal data of not less than 35,000 unique consumers, excluding personal data controlled or processed solely for the purpose of completing a payment transaction; or
>
> (b) Controlled or processed the personal data of not less than 10,000 unique consumers and derived more than 25 percent of their gross revenue from the sale of personal data.
>
> II. The secretary of state shall notice and post a link to RSA 507-H on the secretary of state's website."

[219] *See* Pell et al., *supra* note 21.

[220] *See* McCoy et al., *supra* note 207, at 16.

[221] *See id.*

[222] *See CFPB Proposes Rule*, *supra* note 16. Since the proposed rule would effectively halt data brokerage activities for sensitive financial information, it stands to reason that similar restrictions could plausibly apply to de-identified data as well. *See id.*

research or epidemiological studies. This would prevent third parties from acquiring such data under the pretense of de-identification and later re-identifying individuals. Further, restrictions must also extend to auxiliary datasets that enable re-identification, such as geolocation data and consumer purchase history.[223] These datasets are often combined with de-identified data to reconstruct identifiable profiles.[224] By addressing the entire data supply chain, rather than just the activities of data brokers, regulations would eliminate the conditions that make re-identification possible.[225]

Given the current regulatory loopholes at the state level and the re-identification risks associated with de-identified data, synthetic data presents a viable alternative that allows for research and analytics while eliminating the possibility of re-identification.[226]

B. Synthetic Data Generation as a Broad-Reaching Solution

Synthetic data is emerging as a transformative solution to address privacy challenges while maintaining data utility.[227] Unlike de-identified data, synthetic data is not derived directly from real individuals.[228] Instead, it is generated algorithmically to replicate the statistical properties and patterns of real-world datasets without including any actual personal identifiers, a distinction that makes synthetic data uniquely suited for privacy-sensitive applications, as it almost eliminates the risk of re-identification.[229] Increasingly, synthetic data is being adopted by organizations across different industries.[230] Data brokers may acquire these synthetic datasets from such organizations in a similar manner to how they acquire de-identified

[223] *See* Gadotti et al., *supra* note 15, at 5 ("Recently, the Federal Trade Commission initiated a lawsuit against the data broker Kochava over concerns for the sale of location data that could be used to identify women who visit abortion clinics.").

[224] *Data Broxkers: Key Players in the Data Selling Ecosystem*, *supra* note 14.

[225] *See* Pell et al., *supra* note 21.

[226] *See* discussion *supra* Section I.C.

[227] Jordon et al., *supra* note 208, at 1.

[228] Aldren Gonzales, Guruprabha Guruswamy & Scott R. Smith, *Synthetic Data in Health Care: A Narrative Review*, 2 PLOS DIGIT. HEALTH 1, 2–3 (2023); Ali, *supra* note 17.

[229] Gonzales et al., *supra* note 228 ("Because synthetic data can be composed purely or mixed with 'fake' data, it is harder to re-identify the records.").

[230] Carolina Trindade, Luís Antunes, Tânia Carvalho & Nuno Moniz, *Synthetic Data Outliers: Navigating Identity Disclosure*, ARXIV (June 4, 2024), https://arxiv.org/html/2406.02736v1.

data.[231] Tools like Synthetic Data Vault (SDV)[232] demonstrate the capabilities of synthetic data generation for various fields, including healthcare.[233] These tools enable organizations to harness data for advanced analytics, research, and innovation without exposing individuals to privacy vulnerabilities.[234] For instance, SubSalt, a company that develops synthetic data solutions using advanced machine learning and generative AI to enable secure, compliant data across analytics and research, provides a clear example of how synthetic data can be implemented at scale.[235] SubSalt's framework ensures that synthetic datasets preserve the statistical integrity of real data while eliminating re-identification risks.[236] By collaborating with expert determination partners, SubSalt applies rigorous tests modeled after HIPAA's Expert Determination method, to ensure that synthetic data meets HIPAA's "very small risk" threshold for re-identification.[237] These tests simulate real-world attacks, such as attempts to link synthetic data with external datasets, to verify that the generated data cannot be traced back to individuals.[238] If a dataset fails to meet these privacy thresholds, it is blocked from release to the consumer (the person or entity seeking use of the dataset), ensuring that only truly anonymized data is shared.[239]

At a federal level, synthetic data could be integrated into HIPAA regulations to require its use as part of the de-identification process

[231] *See* Sherman, *supra* note 24.

[232] The Synthetic Data Vault (SDV) is an open-source ecosystem of libraries developed at MIT that enables users to generate synthetic data across various data modalities. *Overview*, SYNTHETIC DATA VAULT (Mar. 28, 2023), https://sdv.dev/SDV/#:~:text=The%20 Synthetic%20Data%20Vault%20(SDV,properties%20as%20the%20original%20dataset [https://perma.cc/ACZ4-JKZW]. By leveraging machine learning models, SDV creates synthetic datasets that maintain the statistical properties and relationships of real-world data, facilitating tasks such as software testing, machine learning model training, and data analysis without compromising privacy. *The Synthetic Data Vault*, DATACEBO, https://sdv.dev/ [https://perma.cc/EH6L-LBEY] (last visited Dec. 31, 2024); Elise Devaux, *Synthetic Data Tools: Open Source or Commercial? A Guide to Building vs. Buying*, MEDIUM (Sept. 26, 2022), https://medium.com/statice/synthetic-data-tools-open-source-or-commercial-a-guide-to-building-vs-buying-580ddeee30e8 [https://perma.cc/CLN8-3XTM].

[233] Gonzales et al., *supra* note 228.

[234] *Id.*

[235] *The Query Engine for Regulated Data*, SUBSALT, https://www.getSubsalt.com/ [https://perma.cc/845P-4ST5] (last visited Dec. 31, 2024).

[236] *Id.*

[237] *Id.*

[238] *How It Works*, SUBSALT, https://www.getSubsalt.com/how-it-works [https://perma.cc/M6VG-F68Y] (last visited Dec. 31, 2024).

[239] *Id.*

before any health data is shared.[240] If HIPAA were updated to mandate synthetic data generation as an additional safeguard, it would create a national standard that enhances privacy while preserving the usability of health data for research, analytics, and technological innovation.[241] Rather than replacing existing de-identification standards, synthetic data could be integrated into the expert determination pathway, or required as an added step before data is shared externally. Covered entities and their business associates could be obligated to generate synthetic versions of identifiable health data prior to selling or sharing the data, providing stronger privacy protection at the point of disclosure. While not a perfect solution, it offers a more effective approach than current de-identification methods, which leave residual privacy risks.[242] Such a requirement would also provide a future-proof compliance framework, ensuring organizations remain aligned with evolving privacy laws because synthetic data might not be considered "personal data."[243] This is significant because under GDPR, only data that relates to an identified or identifiable individual is subject to compliance.[244] If organizations are able to generate fully synthetic datasets, where every record is entirely artificial, those datasets could qualify as anonymous data under GDPR, thereby reducing legal risk while accelerating innovation.[245] Thus, for organizations, fully synthetic datasets could offer greater ease in cross-border data sharing without as significant of privacy concerns.[246] However, much of whether a dataset could even qualify to be fully synthetic as opposed to "partially synthetic" would be a technical consideration and possible limitation.[247]

[240] *See The Query Engine for Regulated Data*, *supra* note 235; *Guidance Regarding Methods for De-Identification*, *supra* note 15.

[241] *See* Khaled El Emam, Lucy Mosquera & Jason Bass, *Evaluating Identity Disclosure Risk in Fully Synthetic Health Data: Model Development and Validation*, 22 J. MED. INTERNET RSCH. 2 (2020).

[242] *See* Reviglio, *supra* note 4.

[243] *See* Ana Beduschi, *Synthetic Data Protection: Towards a Paradigm Change in Data Regulation?*, 11 SAGE J. 1, 4 (2024). There is a lot of ongoing research as to the question of whether synthetic data is considered "personal data," especially under GDPR standards. *See id.* ("That is because laws such as the GDPR only apply to the processing of personal data (Article 4-1 GDPR). Nonetheless, it remains unclear what level of re-identification risk would be sufficient to trigger their application in the context of synthetic data processing.").

[244] *Id.* at 2–3.

[245] *Id.*

[246] *Id.*

[247] *Id.*

Yet, synthetic data generation is also an effective solution in curbing the privacy risks created by data brokers.[248] Unlike de-identified data, which can be re-identified through linkage with auxiliary datasets, synthetic data does not correspond to any real individual and therefore typically cannot be linked back to specific identities.[249] Synthetic data would effectively close the current regulatory gaps that allow data brokers to exploit downstream data-sharing practices, namely the exemptions for de-identified data.[250] Because those exemptions are unlikely to disappear under the current legal framework, mandating synthetic data generation offers a way to work with existing state and federal structures while providing an extra layer of protection.[251] By requiring synthetic data as a safeguard, covered entities and their business associates can limit the amount of usable information entering the data broker ecosystem, thereby curbing the risks of linkage attacks by data brokers.[252]

However, synthetic data is subject to a somewhat similar subset of re-identification risks known as "identity disclosure."[253] Identity disclosure refers to the risk of revealing an individual's personal information from a dataset, potentially leading to re-identification.[254] Synthetic data must accurately replicate the relationships and patterns found in real-world data, particularly for applications in sensitive domains like healthcare, where data quality can directly impact outcomes.[255] Without proper governance and validation standards, synthetic data could inadvertently leak information about the original

[248] *See* Morgan Guillaudeux, Olivia Rousseau, Julien Petot, Zineb Bennis, Charles-Axel Dein, Thomas Goronflot, Nicolas Vince, Sophie Limou, Matilde Karakachoff, Matthieu Wargny & Pierre-Antoine Gourraudet, *Patient-Centric Synthetic Data Generation, No Reason to Risk Re-Identification in Biomedical Data Analysis*, 6 NPJ DIGIT. MED. 1, 2 (2023).

[249] Ali, *supra* note 17.

[250] Michael Cairo, *Synthetic Data and GDPR Compliance: How Artificial Intelligence Might Resolve the Privacy-Utility Tradeoff*, 28 U. FLA. J. TECH. L. & POL'Y 71, 112–13 (2024) ("By incorporating truly anonymous, privacy-compliant, synthetic data into the BigTech business model, companies like Google and Facebook could continue to operate in their current, highly successful fashion while resolving the challenges presented by the Privacy-Utility Tradeoff by protecting their users' privacy while continuing to profit off of mass data collection.").

[251] *See id.*

[252] Ziqi Zhang, Chao Yan & Bradley A. Malin, *Membership Inference Attacks Against Synthetic Health Data*, 125 J. BIOMEDICAL INFORMATICS 1, 1 (2022).

[253] El Emam et al., *supra* note 241.

[254] *Id.*

[255] David Talby, *The Dangers of Using Synthetic Patient Data to Build Healthcare AI Models*, FORBES (May 26, 2023), https://www.forbes.com/councils/forbestechcouncil/2023/05/26/the-dangers-of-using-synthetic-patient-data-to-build-healthcare-ai-models/ [https://perma.cc/5WMR-VUDP].

datasets.[256] Addressing these challenges requires the establishment of industry-wide standards for evaluating the privacy and utility of synthetic data.[257] Regulators should define clear benchmarks for synthetic data quality, mandate transparency around its generation methods, and integrate synthetic data frameworks into existing privacy laws like HIPAA, GDPR, and TIPA.[258] By embedding synthetic data practices into regulatory and corporate policies, policymakers and businesses alike can create a data ecosystem that fosters innovation while safeguarding privacy, ensuring that the benefits of data-driven technologies can be realized without compromising individual rights.[259]

Such recommendations seek to provide a dual approach to the privacy risks created by data brokers: regulatory reforms that restrict the flow and use of de-identified data, and the integration of synthetic data as a technical safeguard. While neither pathway is without its limitations, each addresses a different point in the data lifecycle. This layered strategy provides a workable path forward under existing state and federal law, while also laying foundation for more nuanced federal privacy protections.

IV. CONCLUSION

Current state privacy frameworks, while intended to foster privacy protections, fall short in addressing the sophisticated re-identification techniques employed by data brokers.[260] HIPAA's de-identification standards are effective within their intended scope at the time of data sharing.[261] However, the downstream risk of re-identification by data brokers who leverage advanced analytics to link seemingly disparate data sets presents a regulatory gap.[262] To mitigate these risks, the adoption of synthetic data and the enactment of comprehensive federal privacy legislation offer viable paths forward,

[256] *Id.*

[257] Anmol Arora, Siegfried Karl Wagner, Robin Carpenter, Rajesh Jena & Pearse A. Keane, *The Urgent Need to Accelerate Synthetic Data Privacy Frameworks for Medical Research*, 7 LANCET DIGIT. HEALTH 1, 2 (2024).

[258] *See id.*; McCoy et al., *supra* note 207.

[259] *See* Brian Eastwood, *What is Synthetic Data – and How Can It Help You Competitively?*, MIT MGMT. (Jan. 23, 2023), https://mitsloan.mit.edu/ideas-made-to-matter/what-synthetic-data-and-how-can-it-help-you-competitively [https://perma.cc/UY6Q-WRHS].

[260] *E.g.*, 2023 Tenn. Pub. Acts 408.

[261] *See* Deven McGraw & Kenneth D. Mandl, *Privacy Protections to Encourage Use of Health-Relevant Digital Data in a Learning Health System*, 4 NPJ DIGIT. MED. 1, 2 (2021).

[262] *See* Reviglio*, supra note 4.*

ensuring privacy in an age of unprecedented data collection and sharing.[263]

Hannah Moore[*]

[263] *See* Michal S. Gal & Orla Lynskey, *Synthetic Data: Legal Implications of the Data-Generation Revolution*, 109 IOWA L. REV. 1087, 1093 (2024).

[*] The author would like to thank Professor David Hoffman, Dr. Bradley Malin, and Ben Winokur, for their help and edits throughout this process. Additional thanks to the Vanderbilt Journal of Entertainment and Technology Law for this opportunity and the thoughtful comments and discussions.