4-2024

Researcher Access to Social Media Data: Lessons from Clinical Trial Data Sharing

Christopher J. Morten
Columbia Law School, cjm2002@columbia.edu

Gabriel Nicholas
NYU School of Law Information Law Institute

Salomé Viljoen
Michigan Law School

RESEARCHER ACCESS TO SOCIAL MEDIA DATA: LESSONS FROM CLINICAL TRIAL DATA SHARING

Christopher J. Morten,[†] *Gabriel Nicholas*[††] *& Salomé Viljoen*[†††]

ABSTRACT

For years, social media companies have sparred with lawmakers over how much independent access to platform data they should provide researchers. Sharing data with researchers allows the public to better understand the risks and harms associated with social media, including areas such as misinformation, child safety, and political polarization. Yet researcher access is controversial. Privacy advocates and companies raise the potential privacy threats of researchers using such data irresponsibly. In addition, social media companies raise concerns over trade secrecy: the data these companies hold and the algorithms powered by that data are secretive sources of competitive advantage. This Article shows that one way to navigate this difficult strait is by drawing on lessons from the successful governance program that has emerged to regulate the sharing of clinical trial data. Like social media data, clinical trial data implicates both individual privacy and trade secrecy concerns. Nonetheless, clinical trial data's governance regime was gradually legislated, regulated, and brokered into existence, managing the interests of industry, academia, and other stakeholders. The result is a functionally successful (albeit imperfect) clinical trial data-sharing ecosystem. Part II sketches the status quo of researchers' access to social media data and provides a novel taxonomy of the problems that arise under this regime. Part III reviews the legal structures governing sharing of clinical trial data and traces the history of scandals, investigations, industry protest, and legislative response that gave rise to the mix of mandated sharing and experimental programs we have today. Part IV applies lessons from clinical trial data sharing to social media

---

DOI: https://doi.org/10.15779/Z38QF8JK81

 For their helpful input, we thank Ashraf Ahmed, Julie Cohen, Talia Gillis, Matthew Herder, Margot Kaminski, Aniket Kesari, Amy Kapczynski, Clarisa Long, Peter Lurie, Nicholson Price, David Pozen, Reshma Ramachandran, Joseph S. Ross, Jason Schultz, Saurabh Vishnubhakat, Ari Waldman, Felix Wu, Deborah Zarin, and commentators at the Privacy Law Scholars Conference, Cornell Legal Theory Workshop, Columbia Junior Scholars Workshop, NYU Law Information Law Institute, Cardozo Intellectual Property and Information Law Colloquium, and the Future of Law in Technology and Governance Workshop at University of Pittsburgh. We thank Xingni (Cindy) Chen, Kasey Clarke, Davis Gonsalves-DeDobbelaere, Jackson Guilford, Angela Kang, Stephanie Lim, Rosella LoChirco, Wisdom Onwuchekwa-Banogu, and Julia Zhu for invaluable research assistance. We also thank the editors of the Berkeley Technology Law Journal, including Gulnur Bekmukhanbetova, Han Bae, Will Kasper, Alex Le, Elizabeth Oh, Bani Sapra, Caressa Tsai, Yuhan Wu, and Nicole Zeinstra, for their expert editing help. All errors are our own.

† Associate Clinical Professor of Law, Columbia Law School.

†† Resident Research Fellow, Center for Democracy & Technology; Fellow, NYU School of Law Information Law Institute.

††† Assistant Professor of Law, Michigan Law School.

data and charts a strategic course forward. Three primary lessons emerge: first, the benefits of research on otherwise secret data are cascading and unpredictable; second, law without institutions to implement the law is insufficient; and, third, data access regimes must be tailored to the different sorts of data they make available.

TABLE OF CONTENTS

I. INTRODUCTION 112
II. THE STATE OF SOCIAL MEDIA DATA SHARING 122

A. HOW RESEARCHERS USE SOCIAL MEDIA DATA 122
B. CURRENT RESEARCHER ACCESS TO SOCIAL MEDIA DATA 126
C. WHAT HAPPENS WHEN RESEARCHERS TRY TO USE THIS ARCHITECTURE? 130
1. Social Science One (SS1) 130
2. NYU Ad Observatory 134
D. A TAXONOMY OF PROBLEMS WITH RESEARCHER ACCESS TO SOCIAL MEDIA DATA 137
1. Poor Research Quality 137
a) Limited by Data Access Arrangements 137
b) Unstable Data Access 138
c) Decontextualized Data Production 138
d) Streetlight Effect 139
e) Denominator Problem 139
2. Unrealized Research 140
a) Inability to Evaluate Social Media Claims 140
b) Unequal Access Leads to Less Diverse Research 140
c) Inability to Discover Unexpected Effects 140
d) Slow Responses to Sudden Problems 141
E. THE LEGAL LANDSCAPE OF DATA SHARING 141
1. What Made Things This Way? 142
a) Trade Secrecy (and Other Entitlement-Like Claims) 143
b) Privacy 145
2. Navigating a Path Forward Between Privacy and Trade Secrecy 148

III. CLINICAL TRIAL DATA SHARING: MANDATE AND EXPERIMENTS 149

A. WHAT IS CLINICAL TRIAL DATA, AND WHY DOES IT MATTER? 149
1. Clinical Trial Data Defined 149
a) Individual Patient-Level Data (IPD) 151
b) Summary Data 151
c) Metadata 152
2. The Value of Clinical Trial Data and Clinical Trial Data Sharing 152

3. The Dangers of Clinical Trial Data Sharing.....156
B. "DARK AGES" OF CLINICAL TRIAL SECRECY: LITTLE RESEARCHER ACCESS, UNREALIZED BENEFITS, AND HARM TO PATIENTS .157
C. LEGISLATING TODAY'S CLINICAL TRIAL DATA SHARING MANDATE 166
1. Mandatory Publication of Approval Packages.....166
2. Mandatory Submission and Publication of Clinical Trial Data to ClinicalTrials.gov.....169
D. IMPLEMENTATION OF THE CLINICAL TRIAL DATA SHARING MANDATE AND EXPERIMENTATION WITH RESEARCHER ACCESS TO MORE SENSITIVE DATA 174
1. Key Institutional Governors of the Clinical Trial Data Sharing Mandate: FDA and NIH.....175
a) FDA and NIH Curate Data.....176
b) FDA and NIH Fund Data-Sharing Initiatives and Research Itself.....177
c) FDA and NIH Regulate and Enforce the Data Sharing Mandate.....178
2. Pioneering Researcher Access to More Sensitive Data.....181
a) Sharing IPD.....182
i) NIH BioLINCC.....183
ii) Yale Open Data Access Project (YODA).....185
b) Sharing Metadata That Contains Alleged Trade Secrets.186
E. CLINICAL TRIAL DATA IN ACTION: A RECAP.....189

IV. TOWARD A SOCIAL MEDIA DATA SHARING MANDATE 190

A. CASCADING (AND UNPREDICTABLE) BENEFITS OF BASIC RESEARCH 191
B. EMPOWERING REGULATORS 192
1. Independent, Preferably Public, Funding.....193
2. Control Over Standards and Terms of Access and Use.....194
3. Meaningful Regulatory Enforcement 195
C. TREATING DIFFERENT DATA DIFFERENTLY.....195
1. Summary Data.....198
2. Metadata 199
3. Individual Data.....200

V. CONCLUSION..... 201

I. INTRODUCTION

In 2018, researchers at Harvard University announced that they had entered into a landmark voluntary partnership with Facebook called Social Science One (SS1) to gather and share data on the inner workings of the social media goliath. The announcement was met with great fanfare. Researchers had been clamoring for data access in order to better understand the dynamics of social media and its effects on everything from elections to teenage mental health to free speech online. Today, however, this grand experiment in voluntary social media data sharing is remembered as a fiasco. Facebook delivered only a fraction of the data it had promised; technical "fixes" made by the company to protect user privacy rendered certain data useless for research; and funders, academics, and civil society partners all eventually withdrew from the project.[1]

Two years after SS1, researchers at New York University's (NYU) Ad Observatory announced that they were taking a different approach to studying Facebook: conducting large-scale research, with or without the company's consent. The Ad Observatory focused on understanding political advertising on Facebook and tracked electoral races across the country. Ad Observatory researchers developed a browser extension, externally audited for security and privacy, that scraped ad data from Facebook and contributed it to an NYU-run database. Months later, Facebook suspended the Ad Observatory researchers' access to Facebook. Facebook's stated justification was to "protect people's privacy."[2]

These two abbreviated anecdotes illuminate a few things about the current state of researchers' access to social media. First, they highlight that significant numbers of researchers in academia and civil society actively want to research social media and will go to great lengths to do so. Second, they show that independent researchers lack sufficient access to various forms of social media data, including content data about what users see, moderation data about how platforms such as Facebook promote and censor content, and distribution data about what kinds of users see what kinds of content. Third, they show that when platforms themselves wield absolute control over which researchers get access to data (and how much, and on what terms), platforms can thwart critical research and shape the literature that emerges by selectively providing access to data.

As we explain in this Article, we need research on social media to flourish if we, as a social-media-obsessed world, are to flourish. For example,

1. *Infra* Section II.B.
2. *Id.*

understanding how content is shared and amplified on social media is essential to understanding how right-wing populism, xenophobia, and conspiratorial misinformation about COVID-19 have attracted large and growing online followings.[3] Understanding social media is also essential to understanding ourselves—how our psyches and societies are reshaped by our screentime and social media's new norms. Understanding social media is essential, too, to understanding social media platforms, some of the 21st century's richest and most powerful companies—how they forestall competition and regulation,[4] how they expand data collection in increasingly elaborate and far-reaching schemes of "informational capitalism" (or, perhaps, "surveillance capitalism"),[5] and more.[6]

---

3. ELIZABETH HANSEN SHAPIRO, MICHAEL SUGARMAN, FERNANDO BERMEJO & ETHAN ZUCKERMAN, NEW APPROACHES TO PLATFORM DATA RESEARCH (2021); CAITLIN VOGUS, IMPROVING RESEARCHER ACCESS TO DIGITAL DATA: A WORKSHOP REPORT 19 (2022), https://cdt.org/wp-content/uploads/2022/08/2022-08-15-FX-RAtD-workshop-report-final-int.pdf; *see also* Julia Angwin, *The Gatekeepers of Knowledge Don't Want Us to See What They Know*, N.Y. TIMES (July 14, 2023) ("To truly hold the platforms accountable, we must support the journalists who are on the front lines of chronicling how despots, trolls, spies, marketers and hate mobs are weaponizing tech platforms or being enabled by them.").

4. KRISTINA KARLSSON, NEW RULES FOR BIG TECH: A CONVERSATION FOR CHANGE 1 (2018) ("Facebook has continued to expand its market power and adapt to trends in the space by acquiring potential competitors, such as Instagram and WhatsApp. Antitrust regulators have failed to understand how these platforms are nascent competitors and thus waved through a series of mergers that greatly diminished consumer choice of social media platforms.").

5. JULIE E. COHEN, BETWEEN TRUTH AND POWER: THE LEGAL CONSTRUCTIONS OF INFORMATIONAL CAPITALISM (2019) (describing "informational capitalism" as an economic system in which information production and information processing are dominant modes of producing and capturing value); SHOSHANA ZUBOFF, THE AGE OF SURVEILLANCE CAPITALISM (2019) (describing "surveillance capitalism" as an economic system in which users' data is used to make predictions about users, control their behavior, and so extract value); Amy Kapczynski, *The Law Of Informational Capitalism*, 129 YALE L.J. 1460, 1466 (2020); Nathaniel Persily & Joshua A. Tucker, *Conclusion: The Challenges And Opportunities For Social Media Research*, *in* SOCIAL MEDIA AND DEMOCRACY: THE STATE OF THE FIELD, PROSPECTS FOR REFORM 313, 313 (Nathaniel Persily & Joshua A.Tucker eds., 2020).

6. Irene V. Pasquetto, Briony Swire-Thompson, Michelle A. Amazeen, Fabrício Benevenuto, Nadia M. Brashier, Robert M. Bond, Lia C. Bozarth, Ceren Budak, Ullrich K. H. Ecker, Lisa K. Fazio, Emilio Ferrara, Andrew J. Flanagin, Alessandro Flammini, Deen Freelon, Nir Grinberg, Ralph Hertwig, Kathleen Hall Jamieson, Kenneth Joseph, Jason J. Jones, R. Kelly Garrett, Daniel Kreiss, Shannon McGregor, Jasmine McNealy, Drew Margolin, Alice Marwick, FiIippo Menczer, Miriam J. Metzger, Seungahn Nah, Stephan Lewandowsky, Philipp Lorenz-Spreen, Pablo Ortellado, Gordon Pennycook, Ethan Porter, David G. Rand, Ronald E. Robertson, Francesca Tripodi, Soroush Vosoughi, Chris Vargo, Onur Varol, Brian E. Weeks, John Wihbey, Thomas J. Wood & Kai-Cheng Yang, *Tackling Misinformation: What Researchers Could Do with Social Media Data*, HARV. KENNEDY SCHOOL (HKS)

Yet researchers' access to data remains controversial. Independent privacy advocates raise concerns over the sensitivity of social media data held by companies and the potential threats of researchers using such data irresponsibly.[7] Social media companies themselves increasingly deploy (or perhaps "weaponize") arguments about individual privacy to justify intense secrecy.[8] These companies wield privacy arguments at both the doctrinal and theoretical levels, arguing that researcher access (1) would violate various extant laws, such as the European Union's General Data Protection Regulation (GDPR), and (2) is normatively undesirable because it would expose individuals who use social media to a raft of harms that outweigh the research's foreseeable benefits.[9] In addition, the same social media companies raise separate but equally serious concerns over intellectual property. Again, these companies raise commercial secrecy objections at both the doctrinal level and the theoretical level, asserting that researcher access would (1) violate state and federal trade secrecy law, and (2) be normatively undesirable because it would encourage "free riding" by competitors and thereby erode crucial "incentives to innovate."[10]

In industry's telling, and in much popular discourse, privacy and incentives to innovate have become a kind of "Scylla and Charybdis" of sharing social media data—two obstacles that any data-sharing effort must navigate to

---

MISINFORMATION REV. 1, 8 (2020) (on social media platforms' role in propagation of misinformation).

7. SHAPIRO ET AL., *supra* note 3, at 45 ("The rules and regulations around user privacy, combined with the political force of privacy advocates, are by far the biggest barrier to platform companies' ability and willingness to share data with researchers."); VOGUS, *supra* note 3, at 33 ("Properly balancing competing interests, such as the risks to user privacy, may require policymakers to take incremental steps to improve researchers' access to data, and to carefully assess whether those steps are serving the public interest.").

8. For a broad analysis, see Rory Van Loo, *Privacy Pretexts*, 108 CORNELL L. REV. 1 (2022). For a specific example, see generally AMY O'HARA & JODI NELSON, EVALUATION OF THE SOCIAL SCIENCE ONE—SOCIAL SCIENCE RESEARCH COUNCIL—FACEBOOK PARTNERSHIP (2020) (explaining how Facebook concluded it could not provide previously-promised data access to researchers because of concerns over user privacy).

9. *Id.*; *see also* Matias Vermeulen, *The Keys to the Kingdom* (July 27, 2021), https://knightcolumbia.org/content/the-keys-to-the-kingdom (analyzing whether GDPR creates barriers to researcher access).

10. FACEBOOK, COMMENTS TO THE FEDERAL TRADE COMMISSION ON DATA PORTABILITY 13 (2020) (arguing that portability of and access to "all observed and inferred data could also result in a different sort of burden: the disclosure of trade secret or other proprietary information developed by a business to enhance or differentiate its services. Enabling people to port that kind of information could reduce incentives for businesses to develop it in the first place"); VOGUS, *supra* note 3, at 25 ("[A]ccess to non-public data raises greater risks of invading users' privacy and revealing trade secrets or security measures used by hosts.").

succeed.[11] Social media companies cast this two-headed trap as so fearsome that it may ultimately doom even the cleverest efforts. Some regulators and legislators have nonetheless persisted in proposing and enacting new laws to expand researcher access to social media data,[12] but they face stiff headwinds. Concerns over privacy and incentives to innovate have chilled nascent efforts toward real transparency and accountability of social media.[13]

The key question that this Article addresses is this: Does a regulatory pathway exist to achieve meaningful researcher access to social media data while protecting privacy and incentives to innovate?

This is an urgent question, and we are far from the first to write on it. Daphne Keller;[14] Aline Iramina, Maayan Perel & Niva Elkin-Koren;[15] Rebekah Tromble;[16] and the Working Group established by the European Digital Media Observatory[17] are among those who have offered important views on this question. The European Union is already moving to mandate researcher access to social media platform data.[18] Its Digital Services Act, among other initiatives, requires qualifying platforms to grant access to certain

11. *E.g.*, Paddy Leerssen, *Platform Research Access in Article 31 of the Digital Services Act*, VERFASSUNGSBLOG (Sept. 7, 2021), https://verfassungsblog.de/power-dsa-dma-14/. The twin obstacles of privacy and incentives to innovate are discussed in greater detail in *infra* Part II.

12. *See generally* VOGUS, *supra* note 3 (discussing U.S. legislative proposals to guarantee researcher access to social media data); Alex Engler, *Platform Data Access Is a Lynchpin of the EU's Digital Services Act*, BROOKINGS INST. (Jan. 15, 2021), https://www.brookings.edu/blog/techtank/2021/01/15/platform-data-access-is-a-lynchpin-of-the-eus-digital-services-act/ (presenting researcher access provisions of EU's Digital Services Act).

13. *See* SHAPIRO ET AL., *supra* note 3; VOGUS, *supra* note 3.

14. Daphne Keller, *Delegated Regulation on data access provided for in the Digital Services Act—Comment of Daphne Keller*, EUR. COMM'N (May 22, 2023), https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/13817-Delegated-Regulation-on-data-access-provided-for-in-the-Digital-Services-Act/F3422727_en; Daphne Keller & Max Levy, *What's the Best Path Forward for Platform Transparency Regulation*, LAWFARE (July 11, 2022), https://www.lawfaremedia.org/article/getting-transparency-right.

15. Aline Iramina, Maayan Perel (Filmar) & Niva Elkin-Koren, *Paving the Way for the Right to Research Platform Data* (June 19, 2023), https://ssrn.com/abstract=4484052.

16. Rebekah Tromble, *Where Have All the Data Gone? A Critical Reflection on Academic Digital Research in the Post-API Age*, 7 SOC. MEDIA + SOC'Y 1 (2021).

17. EUROPEAN DIGITAL MEDIA OBSERVATORY WORKING GROUP, REPORT OF THE EUROPEAN DIGITAL MEDIA OBSERVATORY'S WORKING GROUP ON PLATFORM-TO-RESEARCHER DATA ACCESS (2022), https://edmo.eu/wp-content/uploads/2022/02/Report-of-the-European-Digital-Media-Observatorys-Working-Group-on-Platform-to-Researcher-Data-Access-2022.pdf.

18. Iramina et al., *supra* note 15.

requested information to vetted researchers, although the processes for doing so have not yet been finalized.[19]

We think the answer is yes—a regulatory pathway *does* exist to achieve meaningful researcher access to social media data while protecting privacy and incentives to innovate. While the Digital Services Act's vetted researcher access mandate is a valuable source of insight and inspiration, we choose to make a complementary case focusing and drawing on U.S. law to argue that researcher access can be achieved here in the United States—because, indeed, in other technology industries, it has already. We do not have to look only to "pro-regulatory" Europe for comparative lessons on the potential virtues of regulation: our own regulatory history and landscape offers such lessons, too.[20]

The main contribution of this Article is comparative. It imports hard-won lessons from other fields of technology—pharmaceuticals[21] and medical devices—to enrich the current debate over researcher access to social media data.[22] The complexity of these technologies rivals that of social media—as does the power of their industries and lobbies, especially in the United States. And yet in pharma and medical devices, we have successfully established mechanisms for broad sharing of what would otherwise be secret industry data.[23] Along the way, these fields successfully navigated a similarly narrow strait between potential harms to individual privacy and harms to incentives to innovate.

---

19. Regulation on a Single Market for Digital Services (Digital Services Act), 2022 O.J. (L 277) 1, 27 ("This Regulation therefore provides a framework for compelling access to data from very large online platforms and very large online search engines to vetted researchers affiliated to a research organisation within the meaning of Article 2 of Directive (EU) 2019/790, which may include, for the purpose of this Regulation, civil society organisations that are conducting scientific research with the primary goal of supporting their public interest mission."). For an explainer of researcher access and the processes ahead, see John Albert, *A Guide to the EU's New Rules for Researcher Access to Platform Data*, ALGORITHM WATCH (Dec. 7, 2022), https://algorithmwatch.org/en/dsa-data-access-explained/.

20. This point is not meant to undercut the significance of the Digital Services Act for non-EU researchers who will likely, under the delegated acts, gain access to hitherto unavailable social media platform data.

21. Throughout this Article, for concision, we generally use the terms "pharmaceutical" and "drug" broadly to describe both small-molecule drug products and biologic drug products. This broad usage is admittedly inexact but consistent with the common practice of the Food & Drug Administration (FDA) and others. *See*, *e.g.*, *Drugs@FDA Glossary*, FDA, https://www.accessdata.fda.gov/scripts/cder/daf/index.cfm?event=glossary.page (last visited Dec. 27, 2023) (defining "Drug" to include biological products).

22. Small portions of a preliminary version of the ideas in this Article were published in a 2022 white paper, GABRIEL NICHOLAS & DHANARAJ THAKUR, LEARNING TO SHARE: LESSONS ON DATA-SHARING FROM BEYOND SOCIAL MEDIA (2022).

23. *See infra* Part III.

In this Article, we focus on one specific kind of data generated by pharmaceutical and medical device companies: clinical trial data. Clinical trials are research studies on human volunteers that answer questions about the safety and efficacy of different health interventions, such as drugs, vaccines, and devices. They are the "gold standard" of evidence-based medicine. They are expensive to conduct, and their data is enormously valuable to doctors' care for patients, regulatory approval, businesses' decision-making and marketing, and scientific research.

Until the 1990s and 2000s, the pharmaceutical and medical device industries could and did keep clinical trial results proprietary. The result was a comparative dark age of information, with drug companies "cherry-picking" only their most favorable data for publication in the medical literature, and falsely marketing unsafe and ineffective products as wonder drugs. A series of high-profile scandals ensued, which involved companies that hid unfavorable data from independent researchers and the broader public, leading to widespread patient harm. These scandals ultimately provoked landmark federal legislation in 2007 that, for the first time, mandated that industry share certain clinical trial data at an across-the-board baseline level. Today, independent researchers around the world use this data to double check the industries' claims and the work of the industries' central regulator, the Food & Drug Administration (FDA), identify unsafe and ineffective products, and advance science.

Before the 2007 clinical trial data-sharing mandate, the pharmaceutical and medical device industries fought it by advancing privacy and incentives-to-innovate arguments similar to those that social media companies deploy today.[24] For example, the largest pharmaceutical lobby warned that mandatory clinical trial data sharing would "fail to protect adequately trade secrets and confidential commercial information," and therefore "harm the public health by discouraging the very innovation necessary to bring new medical advances to the market."[25] And like most social media data, much clinical trial data implicates acute privacy concerns, as individuals' detailed medical statuses are encoded in the data, including many statuses that expose people to discrimination and exploitation.[26]

24. *Supra* Section III.C.

25. Letter from William W. Chin, Executive Vice President, and Jeffrey K. Francer, Vice President & Senior Counsel Scientific & Regulatory Affairs, PhRMA, to Jerry Moore, NIH Regulations Officer, National Institutes of Health (Mar. 25, 2015) (on file with the National Institutes of Health).

26. *Supra* Section III.A.

Yet in the years since Congress legislated the clinical trial data-sharing mandate, no real harm to privacy or to incentives has occurred, even as independent research on that data has unlocked new uses and social benefits. If anything, the trend in clinical trial data sharing today is to push further, expanding researcher access to the most sensitive kinds of data, especially individual patient-level data (IPD) and methodological protocols that reveal exactly how companies conduct their trials and generate and interpret their own data.[27] As we show below, there are important proof-of-concept data-sharing initiatives led by academic centers and by administrative agencies in the United States and Canada that demonstrate even the most highly sensitive data can, under the right conditions, be shared responsibly with researchers.

We recognize that the parallels between social media data and clinical trial data are inexact. Clinical trial data sets are more standardized and far smaller than that of social media platforms. The data subjects in clinical trials are volunteers, enrolled pursuant to elaborate and independently vetted processes of informed consent, while the quality of informed consent for data collection from users of social media is widely perceived as laughable.[28] Some individuals' social media data is intensively sensitive in ways that even the most detailed medical data is not; social media data may reveal, for example, users' political affiliations and organizing activities, romantic preferences, travel histories, and more. The variety and profundity of harms that flow from discriminatory and other unwanted uses of social media data can therefore be even greater than the harms that flow from unwanted uses of medical data. Furthermore, social media and medical products implicate very different tradeoffs. Medical products are generally seen as innovations vital for society; social media innovations, such as algorithms targeting ads or recommending content, for example, are increasingly seen as socially deleterious.[29] Clinical trial and social media data access systems both need to manage tradeoffs between protection of trade secrecy and utility to researchers, but where they draw those lines will be very different.

Yet as we endeavor to show in this Article, the benefits of sharing are likely to be broadly similar. Indeed, we argue that important parallels do exist and that the history of clinical trial data sharing therefore holds important lessons for social media data sharing.[30] We focus on clinical trial data not because this

27. *Supra* Section III.D.

28. *See* Ari Ezra Waldman, *Privacy, Notice, and Design*, 21 STAN. TECH. L. REV. 74 (2018).

29. *See generally* MARIANA MAZZUCATO, THE VALUE OF EVERYTHING: MAKING AND TAKING IN THE GLOBAL ECONOMY (2018).

30. Social media companies sometimes insist that their technologies are unprecedented and sui generis, and thus cannot be regulated like technologies past; a rich literature shows that's false. *See, e.g.*, MARIANA MAZZUCATO, THE ENTREPRENEURIAL STATE (1st ed. 2013)

data is, as a technical matter, most similar to social media data, but because the technical, institutional, and legal structures that *govern* clinical trial data sharing are particularly mature, tested, and successful, as we show below. In future work, we and other scholars may draw other instructive lessons from efforts to share other kinds of medical data, such as electronic medical record data.[31]

In this Article, we offer three primary lessons for those studying, advocating, and legislating social media data sharing: first, the benefits of research on otherwise secret data are cascading and unpredictable; second, law without institutions to implement the law is insufficient; and third, different kinds of data must be treated differently.[32]

The history of clinical trial data sharing shows that effective researcher access and use of industry data is impossible without powerful independent institutions that can serve as counterweights to extraordinarily powerful industries. Such counterweight institutions, whether public agencies, private independent institutions, or both, could serve as "regulators" of the social media industry. To support research, these regulators may serve many roles:

(technology and pharmaceutical companies arguing they deserve regulatory exceptions); Rebecca Haw Allensworth, *Antitrust's High-Tech Exceptionalism*, 130 YALE L.J. F. 588 (2021) (detailing how courts granted tech companies special exceptions to antitrust rules due to "views about digital markets in the early 2000s—that they were uniquely dynamic, innovative, and competitive" that are not only false, but have also prevented competition in the tech sector); Yaël Eisenstat & Nils Gilman, *The Myth of Tech Exceptionalism*, NOEMA MAGAZINE (Feb. 10, 2022), https://www.noemamag.com/the-myth-of-tech-exceptionalism/ (detailing how big tech companies use the narrative of innovation to ward off regulation); Richard Waters, *Tech's Self-Declared Exceptionalism is Coming to an End*, FIN. TIMES (Sept. 19, 2019), https://www.ft.com/content/1cf9ac56-da5d-11e9-8f9b-77216ebe1f17; *see generally* LOUIS HYMAN, TEMP: THE REAL STORY OF WHAT HAPPENED TO YOUR SALARY, BENEFITS, AND JOB SECURITY (2019) (detailing the historical roots of gig work in outsourcing innovations of the 1960s and 1970s). The belief in new technologies' revolutionary status is closely linked to cults of genius that arise around technology company founders. Luke Savage, *Elon Musk is Destroying the Myths of Silicon Valley in Front of Our Very Eyes*, JACOBIN (Nov. 27, 2022), https://jacobin.com/2022/11/elon-musk-twitter-silicon-valley-myth?mc_cid=aa8219b840&mc_eid=f0c834022c ("The main ingredient in this futurist cocktail is typically said to be a rare breed of exceptional individuals who rise to the top through a combination of eccentric genius and personal grit.").

31. For more on efforts to share electronic medical record data, see, e.g., SHARONA HOFFMAN, ELECTRONIC HEALTH RECORDS AND MEDICAL BIG DATA (thorough survey of the state of electronic health record sharing as of 2016); *European Health Union: A European Health Data Space for People and Science*, EUR. COMM'N (May 3, 2022), https://ec.europa.eu/commission/presscorner/detail/en/ip_22_2711 (describing the European Health Data Space initiative). For a brief analysis of parallels between sharing such data and sharing social media data, see Naomi Shiffman, *Tools for Platform Research: Lessons from the Medical Research Industry*, TECH POL'Y PRESS (Apr. 26, 2023), https://techpolicy.press/tools-for-platform-research-lessons-from-the-medical-research-industry/.

32. *See infra* Part II, especially Section II.C through Section II.E.

they monitor and enforce industries' compliance with data sharing laws; collect, standardize, curate, steward, and share data; govern researchers' access and use of data; explain to researchers and the broader public how to use data; and sometimes fund worthy research. These institutions need not be public (though most are in the world of clinical trial data sharing); they can be academic or non-governmental organizations. But they do need to be functionally independent from industry; pharmaceutical industry-funded clinical trial data sharing initiatives failed to spark useful research and to check the industry's worst excesses.

The history of clinical trial data sharing also shows that different kinds of data should be treated differently. Perhaps the point is self-evident, but it is also vital. Today federal legislation mandates sharing of certain clinical trial data—so-called "summary data" characterizing broad trends, as well as certain "metadata" on how data is generated—on a public website accessible from anywhere in the world. This kind of blunt mandatory disclosure works well for data of high value to researchers and for which sharing poses low risk. For more sensitive data—individual participant data (IPD), which can easily be reidentified, or certain trial protocols that reveal industries' innovative and confidential scientific methods—blunt disclosure to the general public is inappropriate. Instead, more sensitive data tends to be shared only with trusted researchers subject to a raft of constraints on access and use.

Before we turn to the body of the Article, a word on the Article's limitations—on what this Article is and is not. First, we intend this Article as a primarily descriptive, positivist account of how law and technology currently work. Much of the description and analysis of clinical trial data sharing (and sharing of other kinds of medical data) is in the medical and scientific literature rather than the law review literature, and thus has received comparatively little attention from legal scholars, activists, and other researchers focused on social media.[33] We see value in building a bridge between distinct literatures and distinct readerships.

Second, we recognize and decline to address, in this Article, a large set of important theoretical and doctrinal questions attached to the value of social media data sharing. For example, what is the fundamental value of social media? Is the collection of social data ethical and desirable in the first place? What theory (or theories) of privacy should inform laws governing social media? Under existing doctrine, does any form of social media data qualify for

33. *But see* Naomi Shiffman, *Tools for Platform Research: Lessons from the Medical Research Industry*, TECH POL'Y PRESS (Apr. 26, 2023), https://techpolicy.press/tools-for-platform-research-lessons-from-the-medical-research-industry/ (drawing explicit parallels between sharing privacy-sensitive medical data and sharing social media data).

trade secrecy protection, or other forms of intellectual property protection? Should it, from a public policy perspective? The three of us have grappled with some of these questions in other work,[34] and will continue to, but we put these questions aside for this Article.

Third, this Article largely accepts the social media industry's professed concerns over privacy and incentives to innovate. There are, of course, compelling reasons to be skeptical.[35] But here we endeavor to show that it is possible to take the social media industry's concerns seriously and overcome them. This Article argues that legislators and regulators concerned with protecting privacy and intellectual property rights in sensitive privately held data can nonetheless devise rules and institutions to share that data with independent researchers responsibly. This is, at the very least, precisely what has happened with the pharmaceutical and medical device industries.

The Article proceeds as follows. Part II provides a legal and technical description of the current state of researchers' access to social media data and presents a novel taxonomy of its problems. It also describes the law and normative arguments that created and perpetuate today's status quo, with focus on trade secrecy and privacy in the United States. Part III lays out relevant lessons from clinical trial data, explaining what clinical trial data is, how it compares to social media data, and how regulatory and voluntary efforts managed to responsibly share even the most sensitive personal and trade secret data with independent researchers. Part III also gives the history of these efforts, describing first the "dark ages" of clinical trial data secrecy, when the pharmaceutical companies that created and exploited this data wielded near-total control over access to it, and then how the industry emerged from these dark ages after Congress passed data sharing requirements and invested in countervailing public and nonprofit institutions. Part IV applies the clinical trial data sharing framework's legal and institutional lessons to social media data and charts a strategic course forward toward responsible and effective social media data sharing. As noted above, one key lesson is the need to empower public or nonprofit institutions capable of confronting the powerful social media industry. Another is the value of treating different kinds of data differently. In particular, clinical trial data's tripartite distinction of individual data, summary data, and metadata promotes distinct governance structures

34. *See* Christopher J. Morten, *Publicizing Corporate Secrets*, 171 U. PENN. L. REV. 1319 (2023); Gabriel Nicholas, *Taking It with You: Platform Barriers to Entry and the Limits of Data Portability*, 27 MICH. TECH. L. REV. 263 (2021); *see generally* Salome Viljoen, *A Relational Theory of Data Governance*, 131 YALE L.J. 573 (2021).

35. *See* Van Loo, *Privacy Pretexts*, *supra* note 8; *see also* Yafit Lev-Aretz & Katherine J. Strandburg, *Privacy Regulation and Innovation Policy*, 22 YALE J.L. & TECH. 256 (2020).

that maximize researcher utility while minimizing risks to data subjects and incentives to innovate. Part V briefly concludes with a discussion of proposed legislation.

II. THE STATE OF SOCIAL MEDIA DATA SHARING

Social media companies have a wide range of approaches they can take to sharing data with researchers. This Part offers a snapshot of the status quo of how sharing occurs currently and the legal and technical arrangements that support that sharing. It also lays out the primary legal challenges to addressing the problem of researcher access to data that animate the rest of the Article.

A. HOW RESEARCHERS USE SOCIAL MEDIA DATA

Researchers are interested in all sorts of social media data for all sorts of reasons. Many seek to better understand the dynamics and external effects of social media ecosystems. Social and computer science researchers use platform data to better understand widespread popular problems such as the spread of mis- and dis-information,[36] the effects of algorithmic speech systems,[37] online

36. *See* Miriam J. Metzger, Andrew J. Flanagin, Paul Mena, Shan Jiang & Christo Wilson, *From Dark to Light: The Many Shades of Sharing Misinformation Online*, 9 MEDIA & COMMC'N 134, 135 (2021); AOIFE GALLAGHER, MACKENZIE HART & CIARÁN O'CONNOR, ILL ADVICE: A CASE STUDY IN FACEBOOK'S FAILURE TO TACKLE COVID-19 DISINFORMATION 8 (2021); Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini & Filippo Menczer, *The Spread of Low-Credibility Content by Social Bots*, 9 NATURE COMMC'NS 1 (2018).

37. *E.g.*, Andrew Guess, Kevin Aslett, Richard Bonneau, Jonathan Nagler & Joshua A. Tucker, *Cracking Open the News Feed: Exploring What U.S. Facebook Users See and Share with Large-Scale Platform Data*, 1 J. QUANTITATIVE DESCRIPTION: DIGIT. MEDIA 1, 10–11 (2021); Cody Buntain, Richard Bonneau, Jonathan Nagler & Joshua A. Tucker, *YouTube Recommendations and Effects on Sharing Across Online Social Platforms*, 5 PROCS. ACM ON HUM.-COMPUT. INTERACTION 1 (2021); MARC FADDOUL, GUILLAUME CHASLOT & HANY FARID, A LONGITUDINAL ANALYSIS OF YOUTUBE'S PROMOTION OF CONSPIRACY VIDEOS (2020).

extremism,[38] child welfare,[39] free speech online,[40] and online discourse around elections and other democratic processes.[41]

Some smaller scale work may not require researchers to have access to more or different data than is available to ordinary users. For instance, sociological research that focuses on small online communities can be done without special access to data, so long as researchers can embed themselves within those communities.[42] Larger scale and more macro-level research, however, requires access to more data than any one regular user has access to through non-automated means. For example, researchers looking to understand public views of gender-based violence on X, née Twitter (referred to from here as "Twitter"), need access to hundreds of thousands or millions

38. *E.g.*, Homa Housseinmardi, Amir Ghasemian, Aaron Clauset, Markus Mobius, David M. Rothschild & Duncan J. Watts, *Examining the Consumption of Radical Content on YouTube*, 118 PROCS. NAT'L ACAD. SCIS. 1 (2021); WEI WEI, KENNETH JOSEPH, HUAN LIU & KATHLEEN M. CARLEY, THE FRAGILITY OF TWITTER SOCIAL NETWORKS AGAINST SUSPENDED USERS 9 (Jian Pei et al. eds., 2015); Yannick Veilleux-Lepage & Emil Archambault, *Mapping Transnational Extremist Networks: An Exploratory Study of the Soldiers of Odin's Facebook Network, Using Integrated Social Network Analysis*, 13 PERSPS. ON TERRORISM 21 (2019).

39. *E.g.*, MUHAMMAD SHAHROZ NADEEM, PRIVACY VERIFICATION OF PHOTODNA BASED ON MACHINE LEARNING IN SECURITY AND PRIVACY FOR BIG DATA, CLOUD COMPUTING AND APPLICATIONS, 263–64 (Wei Ren et al. eds, 2019); Adrian Ulges, Christian Schulze, Damian Borth & Armin Stahl, *Pornography Detection in Video Benefits (a lot) from a Multi-Modal Approach*, *in* AMVA 12: PROCEEDINGS OF THE 2012 ACM INTERNATIONAL WORKSHOP ON AUDIO AND MULTIMEDIA METHODS FOR LARGE-SCALE VIDEO ANALYSIS 21 (2012).

40. *E.g.*, Michael D. Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Alessandro Flammini, Filippo Menczer, *Political Polarization on Twitter*, 5 PROCS. INT'L AAAI CONF. ON WEB & SOC. MEDIA 89, 90 (2011); ERWAN LE MERRER, BENOÎT MORGAN & GILLES TRÉDAN, SETTING THE RECORD STRAIGHTER ON SHADOW BANNING (2021).

41. *E.g.*, DEEN FREELON, CHARLTON D. MCILWAIN & MEREDITH CLARK, BEYOND THE HASHTAGS: #FERGUSON, #BLACKLIVESMATTER, AND THE ONLINE STRUGGLE FOR OFFLINE JUSTICE (2016); CENTER FOR DEMOCRACY AND TECHNOLOGY, AN UNREPRESENTATIVE DEMOCRACY: HOW DISINFORMATION AND ONLINE ABUSE HINDER WOMEN OF COLOR POLITICAL CANDIDATES IN THE UNITED STATES (Dhanaraj Thakur & DeVan L. Hankers eds., 2022); Orestis Papakyriakopoulos, Christelle Tessono, Arvind Narayanan & Mihir Kshirsagar, *How Algorithms Shape the Distribution of Political Advertising: Case Studies of Facebook, Google, and TikTok*, AIES '22: PROCS. 2022 AAAI/ACM CONF. ON AI, ETHICS, & SOC'Y (2022).

42. *E.g.*, Carolina Are, *The Shadowban Cycle: an autoethnography of pole dancing, nudity and censorship on Instagram*, 22 FEMINIST MEDIA STUDS. 2002 (2022); Ysabel Gerrard, *Beyond the Hashtag: Circumventing Content Moderation on Social Media*, 20 NEW MEDIA & SOC'Y 4492, 4497 (2018); Julia R. DeCook & Jennifer Forestal, *Of Humans, Machines, and Extremism: The Role of Platforms in Facilitating Undemocratic Cognition*, 67 AM. BEHAV. SCIENTIST 629 (2023).

of posts to be able to discern recurring behaviors and rhetorical patterns.[43] Researchers who attempt to reverse engineer or uncover patterns in recommendation algorithms require particularly large volumes of detailed data to produce significant results, since any one user's recommendations only reflects their own tastes, not the system as a whole.[44]

Researchers are also interested in accessing social media data in order to confirm or refute otherwise unverifiable claims made by companies, particularly about changes in their practices. *The Markup* used data collected from its Citizen Browser to reveal that Facebook had not stopped recommending anti-vaccine groups as it claimed it had.[45] In April 2022, researchers used data collected from Russian TikTok to show that TikTok had not had as complete of a ban of Russian pro-war propaganda as it had claimed.[46] Researchers have also used data to show when social media services have made good on their promises to improve. For example, researchers used data scraped from YouTube to confirm that it had reduced the prevalence of conspiratorial content in its recommendation algorithms.[47]

Giving researchers access to social media data can confirm theoretical problems on social media or uncover new problems not previously known to exist. The now-famous "filter bubble" phenomenon, for example, was able to be confirmed by researchers with access to data donated by social media users.[48] Work from Jonas Kaiser and Adrian Rauchfleisch studying YouTube's

43. *E.g.*, HEMANT PUROHIT, TANVI BANERJEE, ANDREW HAMPTON, VALERIE L. SHALIN, NAYANESH BHANDUTIA & AMIT P. SHETH, GENDER-BASED VIOLENCE IN 140 CHARACTERS OR FEWER: A #BIGDATA CASE STUDY OF TWITTER (2015) (using 14 million posts); Aparup Khatua, Erik Cambria & Apalak Khatua, *Sounds of Silence Breakers: Exploring Sexual Violence on Twitter*, 2018 IEEE/ACM INT'L CONFERENCE ON ADVANCES SOC. NETWORKS ANALYSIS & MINING (ASONAM) 397, 397 (2018) (using 700,000 posts); CENTER FOR DEMOCRACY AND TECHNOLOGY, *supra* note 37 (using over 100,000 posts).

44. *See, e.g.*, Matthew Hindman, Nathaniel Lubin & Trevor Davis, *Facebook has a Superuser-Supremacy Problem*, ATLANTIC: FACEBOOK PAPERS (Feb. 10, 2022), https://www.theatlantic.com/technology/archive/2022/02/facebook-hate-speech-misinformation-superusers/621617/.

45. Corin Faife & Dara Kerr, *Facebook Said it Would Stop Recommending Anti-Vaccine Groups. It Didn't*, MARKUP: CITIZEN BROWSER (May 20, 2021), https://themarkup.org/citizen-browser/2021/05/20/facebook-said-it-would-stop-recommending-anti-vaccine-groups-it-didnt.

46. MARC FADDOUL, SALVATORE ROMANO, ILIR RAMA, NATALIE KERBY & GIULIA GIORGI, TRACKING EXPOSED SPECIAL REPORT: CONTENT RESTRICTIONS ON TIKTOK IN RUSSIA FOLLOWING THE UKRAINIAN WAR 4 (2022), https://tracking.exposed/pdf/tiktok-russia-12april2022.pdf.

47. FADDOUL ET AL., *supra* note 37.

48. Seth Flaxman, Sharad Goel & Justin M. Rao, *Filter Bubbles, Echo Chambers, and Online News Consumption*, 80 PUB. OP. Q. 298, 312 (2016); Colin Lecher & Leon Yin, *One Year After the Capitol Riot, Americans Still See Two Very Different Facebooks*, MARKUP: CITIZEN BROWSER

recommendation algorithm in Brazil found that users could go down rabbit holes of videos of sexually suggestive videos of children.[49] The Stanford Internet Observatory used data from Mastodon to discover a large decentralized distribution network of human- and computer-generated child sexual abuse material.[50]

Some areas of social media research require access beyond what is available on the internet publicly. For example, most research related to personalization requires information on real people's profiles, activities, and recommendations, which, if not public, can only be obtained through donation by the users or the platform itself. Though more challenging from a privacy perspective, this research is still critically important. For instance, research using data donated from Facebook users found that the platform drastically overcounted some and undercounted other political ads, including tens of thousands of ads that ran during its "moratorium" on political ads around the U.S. 2020 elections, raising questions about the company's ability to effectively enforce its own policies.[51]

Researchers that study topics beyond social media may also be interested in data from platforms. Linguists, for example, use social media to understand emerging subject areas such as how emojis are used and how people from different generations speak online.[52] Machine learning researchers use labeled image data and unlabeled text data from social media to train generative AI models.[53] Hundreds of scientific articles have sought to use social media posts

(Jan. 6, 2022, 10:30 AM), https://themarkup.org/citizen-browser/2022/01/06/one-year-after-the-capitol-riot-americans-still-see-two-very-different-facebooks; Michael Wolfowicz, David Weisburd & Badi Hasisi, *Examining the Interactive Effects of the Filter Bubble and the Echo Chamber on Radicalization*, 19 J. EXPERIMENTAL CRIMINOLOGY 119, 124, 129 (2023).

49. Jonas Kaiser & Adrian Rauchfleisch, *The Implications of Venturing Down the Rabbit Hole*, 8 INTERNET POL'Y REV. 1 (2019).

50. DAVID THIEL & RENÉE DIRESTA, CHILD SAFETY ON FEDERATED SOCIAL MEDIA (2023), https://purl.stanford.edu/vb515nd6874.

51. VICTOR LE POCHAT, LAURA EDELSON, TOM VAN GOETHEM, WOUTER JOOSEN, DAMON MCCOY & TOBIAS LAUINGER, AN AUDIT OF FACEBOOK'S POLITICAL AD POLICY ENFORCEMENT 13 (2022), https://cybersecurityfordemocracy.org/audit-facebook-political-ad-policy-enforcement.

52. GRETCHEN MCCULLOCH, BECAUSE INTERNET: UNDERSTANDING THE NEW RULES OF LANGUAGE (2020).

53. Mehtab Khan & Alex Hanna, *The Subjects and Stages of AI Dataset Development: A Framework for Dataset Accountability*, 19 OHIO ST. TECH. L. J. 172, 174 (2023); Mike Isaac, *Reddit Wants to Get Paid for Helping to Teach Big A.I. Systems*, N.Y. TIMES (Apr. 18, 2023), https://www.nytimes.com/2023/04/18/technology/reddit-ai-openai-google.html.

to detect mental illness.[54] And at least while it was publicly available, the U.S. Geological Survey used Twitter data to track earthquakes, which in some cases has been shown to work even better than a Richter scale.[55] It is easy to imagine many other use cases of social data: ornithologists accessing photos of birds on Instagram, social scientists accessing relational information to predict gun violence, and so on.

Social media companies themselves of course stand to gain a lot of value from the data generated by their own services, and many have business models that entirely depend on such data.[56] Companies can use their data to target advertisements, increase the amount of time users spend on a service, or sell it to data brokers and other actors that can monetize the data. For instance, when Reddit began charging for its API in 2023, the company claimed it was because Google and OpenAI were using their data to train large language models, although critics argued it was also for them to wrestle control over their advertising revenue from third-party apps.[57] Companies can also use data from their platforms to better understand how users use their services, and use that information to improve the user experience or the safety and integrity of their communities. Many legal scholars have written about the market benefits of requiring social media companies to make certain data available to competitors,[58] but those efforts have different normative values from providing researchers with data—facilitation of markets as opposed to the generation of knowledge—and entail very different governance decisions outside the scope of this Article.

B. CURRENT RESEARCHER ACCESS TO SOCIAL MEDIA DATA

Social media companies vary widely in what data they share with researchers and how they make it available. Many platforms make little to no

54. Tianlin Zhang, Annika M. Schoene, Shaoxiong Ji & Sophia Ananiadou, *Natural Language Processing Applied to Mental Illness Detection: A Narrative Review*, 5 NPJ DIGIT. MED. 46, 5 (2022).

55. *How the USGS uses Twitter Data to Track Earthquakes*, TWITTER: BLOG (Oct. 7, 2015), https://blog.twitter.com/en_us/a/2015/usgs-twitter-data-earthquake-detection.

56. Amy Kapczynski, *The Law of Informational Capitalism*, 129 YALE L.J. 1460, 1469 (2020); COHEN, *supra* note 5.

57. Isaac, *supra* note 53. Despite charging for API access, ChatGPT is still likely trained on Reddit. *See ChatGPT's Web Browser Could Deflate Reddit's API Pricing*, SAM IS TOAST (June 17, 2023), https://samstoast.substack.com/p/chatgpt-can-already-circumvent-reddits.

58. Oscar Borgogno & Giuseppe Colangelo, *Data Sharing and Interoperability: Fostering Innovation and Competition Through APIs*, 35 COMPUT. L. & SEC. REV. 105314, 105314 (2019); Gabriel Nicholas, *Taking It with You: Platform Barriers to Entry and the Limits of Data Portability*, 27 MICH. TECH. L. REV. 263, 272 (2021); Chris Riley, *Unpacking Interoperability in Competition*, 5 J. CYBER POL'Y 94, 94 (2020).

data available to researchers, including private messaging apps such as WhatsApp, Telegram, and iMessage; team chat apps such as Slack and Discord; semi-private social networks such as Snapchat; and public social networks such as LinkedIn and Pinterest. There are also large public social networks such as YouTube and TikTok that, as of this writing, make some data available to researchers—and recently increased that amount due to recent regulatory efforts—but still not enough or under too restrictive agreements to be adopted by researchers en masse.[59]

Other large public-facing platforms offer data access but only under certain conditions. Most platforms at least have their data protected under terms of service, but some have additional restrictions they impose upon researchers in exchange for access to more data. Meta, for instance, allows approved academics and independent researchers to access data sets about election ads and URL shares on Facebook.[60] However, those researchers are required to sign a data agreement that, among other things, limits their ability to share data with third party reviewers, prevents them from using Facebook data in conjunction with other data, and allows Meta to review any published material ahead of time for "any Confidential Information or any Personal Data that may be included or revealed in those materials and which need to be removed prior to publication or disclosure."[61]

There are two primary ways companies make data available to researchers: static public datasets and application programming interfaces (APIs). Static data datasets allow companies to share a snapshot of the data on their platform, but since they are not dynamic, they can go out of date. APIs, on the other hand, allow live, up-to-date access to data hosted on a platform. They are more expensive for companies to build, maintain, and operate, but unlike static data sets, they allow companies to retain extensive control over who can access what data and how much.

Social media companies have not shied away from severely limiting data access through APIs, even to researchers. Before Twitter raised the cost of its

59. *See YouTube Researcher Program*, YOUTUBE, https://research.youtube/ (last visited Nov. 23, 2023); Vanessa Pappas, *Strengthening our Commitment to Transparency*, TIKTOK: NEWSROOM (July 27, 2022), https://newsroom.tiktok.com/en-us/strengthening-our-commitment-to-transparency; Emma Lurie, *Comparing Platform Research API Requirements*, TECH POL'Y PRESS (Mar. 22, 2023), https://techpolicy.press/comparing-platform-research-api-requirements/.

60. *Academic Resources: Meta Data for Independent Research*, META, https://research.facebook.com/data/ (last visited Nov. 23, 2023).

61. *Research Data Agreement*, SOC. SCI. ONE 2–3, 8, https://socialscience.one/files/socialscienceone/files/fort_non-monetary_rda_with_public_institution_and_developer_terms.pdf (last accessed Jan. 21, 2024).

API from free to $42,000 per month (a move many see as Elon Musk thumbing his nose at researchers),[62] Twitter offered researchers with university affiliations an Academic API, which allowed them to access Twitter's full archive of historical tweets and perform more refined searches.[63] However, it limited researchers to accessing ten million tweets per month, or the equivalent of about one fiftieth of all tweets sent *per day*.[64] YouTube's API is far more limited: by default, it allows researchers to make 100 search requests or 10,000 video information requests per day.[65] While some of these numbers sound large, they constitute a very small fraction of the activity that happens on these platforms.[66] Researchers complain that these limitations significantly stifle or prevent research.[67]

With APIs, platforms can also change the data they make available or revoke data access to individuals as they see fit. Facebook and Twitter, for example, both drastically reduced what and how much data users, including researchers, could access shortly after news of the Cambridge Analytica scandal broke.[68] Researchers with informal and ad hoc arrangements with

62. Get it? 420? *See* Chris Stokel-Walker, *Twitter's $42,000-per-Month API Prices Out Nearly Everyone*, WIRED (Mar. 10, 2023), https://www.wired.com/story/twitter-data-api-prices-out-nearly-everyone/.

63. Suhem Parack, *Introducing the New Academic Research Product Track*, TWITTER: DEVS. (Jan. 2021), https://twittercommunity.com/t/introducing-the-new-academic-research-product-track/148632/1.

64. *Id.*

65. *YouTube Data API Overview*, YOUTUBE, https://developers.google.com/youtube/v3/getting-started#calculating-quota-usage (last modified Nov. 11, 2022); Researchers can apply to increase their quota. *See How It Works*, YOUTUBE, https://research.youtube/how-it-works/ (last visited Nov. 23, 2023).

66. Twitter publishes 500 million tweets per day. *See* Claire Beveridge, *33 Twitter Stats that Matter to Marketers in 2023*, HOOTSUITE: BLOG (Mar. 16, 2022), https://blog.hootsuite.com/twitter-statistics/. YouTube likely has more than 500 hours of video uploaded per minute. *See Hours of Video Uploaded to YouTube Every Minute as of February 2020*, STATISTA (Feb. 2020), https://www.statista.com/statistics/259477/hours-of-video-uploaded-to-youtube-every-minute/.

67. VOGUS, *supra* note 3 ("[I]f researchers do not know what data a host collects and maintains, they do not know what data to ask the host for. This lack of knowledge, some researchers said, limits the research questions that they ask, because they do not know whether certain platforms may have data that would allow them to answer different kinds of questions."); Nathaniel Persily & Joshua A. Tucker, *How to Fix Social Media? Start with Independent Research.*, BROOKINGS (Dec. 1, 2021), https://www.brookings.edu/research/how-to-fix-social-media-start-with-independent-research/.

68. *Graph API*, META: DEVS., https://developers.facebook.com/docs/graph-api/changelog/versions (last visited Nov. 23, 2023); *Previewing Changes to the User and Mentions Timeline API Endpoints*, TWITTER: DEV. PLATFORM BLOG (Mar. 19, 2019), https://blog.twitter.com/developer/en_us/topics/tools/2019/previewing-changes-to-the-user-and-mentions-timeline-api-endpoints.

specific companies to access data may be particularly vulnerable to losing data access without warning.[69] Social media companies can also censure specific researchers for using data in ways they deem improper, as will be discussed further in Section II.B.2 with the case of NYU Ad Observatory.

Finally, social media companies can withdraw support for their data sharing tools or remove them entirely. Twitter and Reddit have both recently been in the news for starting to charge extremely high prices for their once-free APIs.[70] More quietly, Meta appears to be slowly sunsetting CrowdTangle, a popular social media monitoring tool acquired by Facebook in 2016.[71] CrowdTangle is a particularly popular tool with researchers for studying COVID misinformation,[72] election misinformation,[73] and online hate[74] in a wide range of languages. Recently however, Meta has reduced support for the product, allowing it to become buggy and less usable, and has plans to shut it down entirely.[75] Critics argue that Meta is deprecating CrowdTangle because it has contributed to negative press about the company.[76]

---

69. VOGUS, *supra* note 3.

70. *See* Isaac, *supra* note 53; Stokel-Walker, *supra* note 62.

71. Casey Newton, *Facebook Buys CrowdTangle, the Tool Publishers Use to Win the Internet*, VERGE (Nov. 11, 2016), https://www.theverge.com/2016/11/11/13594338/facebook-acquires-crowdtangle.

72. James W. Salazar, Jennifer D. Claytor, Anand R. Habib, Vinay Guduguntla & Rita F. Redberg, *Spread of Misinformation About Face Masks and COVID-19 by Automated Software on Facebook*, 181 JAMA INTERNAL MED. 1251, 1251 (2021); Aimei Yang, Jieun Shin, Alvin Zhou, Ke M. Huang-Isherwood, Eugene Lee, Chuqing Dong, Hye Min Kim, Yafei Zhang, Jingyi Sun, Yiqi Li, Yuanfeixue Nan, Lichen Zhen & Wenlin Liu, *The Battleground of COVID-19 Vaccine Misinformation on Facebook: Fact Checkers vs. Misinformation Spreaders*, 2 HARV. KENNEDY SCH. MISINFORMATION REV. 1, 11 (2021).

73. *E.g.*, Fabio Giglietto, Nicola Righetti, Luca Rossi & Giada Marino, *It Takes a Village to Manipulate the Media: Coordinated Link Sharing Behavior During 2018 and 2019 Italian Elections*, 23 INFO., COMMC'N & SOC'Y 867, 874 (2020); Zeve Sanderson, Megan A. Brown, Richard Bonneau, Jonathan Nagler & Joshua A. Tucker, *Twitter Flagged Donald Trump's Tweets with Election Misinformation: They Continued to Spread Both on and off the Platform*, 2 HARV. KENNEDY SCH. MISINFORMATION REV. 1, 14 (2021).

74. AVAAZ, MEGAPHONE FOR HATE: DISINFORMATION AND HATE SPEECH ON FACEBOOK DURING ASSAM'S CITIZENSHIP COUNT 15 (2019); Sandra Miranda, Fábio Malini, Branco Di Fatima & Jorge Cruz, *I Love to Hate!: The Racist Hate Speech in Social Media*, 9 PROCS. 9TH EUR. CONF. ON SOC. MEDIA 137, 139 (2022).

75. Davey Alba, *Meta Pulls Support for Tool Used to Keep Misinformation in Check*, BLOOMBERG (June 23, 2022), https://www.bloomberg.com/news/articles/2022-06-23/meta-pulls-support-for-tool-used-to-keep-misinformation-in-check?leadSource=uverify%20wall.

76. Kevin Roose, *Inside Facebook's Data Wars*, N.Y. TIMES (Oct. 4, 2021), https://www.nytimes.com/2021/07/14/technology/facebook-data.html; John Albert, *Facebook's Gutting of CrowdTangle: A Step Backward for Platform Transparency*, ALGORITHM WATCH (Aug. 3, 2022), https://algorithmwatch.org/en/crowdtangle-platform-transparency/.

Platformed-sanctioned methods, however, are not the only ways for researchers to be able to access social media data. Researchers can appeal directly to users themselves to give permission to read their data, usually either through authenticating a third-party application (aka a "Sign in with __" button) or through installing a browser extension that scrapes websites on their behalf. These platform-unsanctioned methods can pose additional risks for users because bad actors can use elevated permissions to exfiltrate data. Researchers who build these tools are also at risk of violating a platform's Terms of Service, if not the Computer Fraud and Abuse Act.[77]

However, unsanctioned methods allow for research that could not be otherwise possible under platform sanctioned methods, including research a platform may try to preclude since it could reflect unfavorably on the platform.[78]

C. WHAT HAPPENS WHEN RESEARCHERS TRY TO USE THIS ARCHITECTURE?

Two public controversies showcase the deficiencies and barriers of the current state of social media data access: Social Science One and the New York University Ad Observatory.[79] In the first case, researchers tried to work within the platform's data sharing architecture but ran into shortcomings and had no way to negotiate the additional access they needed, despite being well connected and resourced. In the second, researchers tried to work outside the platform's data sharing architecture, but the platform rejected them, despite their research being safe, secure, socially beneficial, and impossible to do within the company's platform-sanctioned methods.

1. Social Science One (SS1)

On March 17, 2018, *The New York Times* and *The Observer* revealed that the conservative political consulting firm Cambridge Analytica had harvested private information from more than fifty million Facebook profiles and used that data to influence elections around the world.[80] Facebook was already at

77. Sara R. Benson, *Social Media Researchers and Terms of Service: Are We Complying with the Law*, 47 AIPLA Q.J. 191 (2019). Twitter also sued researchers at the Center for Countering Digital Hate under the CFAA. *See* Bryan Pietsch, *Twitter, now X, sues group that researched hate speech on platform*, WASH. POST (Aug. 1, 2023).

78. SHAPIRO ET AL., *supra* note 3, at 14.

79. The use of Facebook in both examples is not meant to be a specific criticism of Facebook's practices. Facebook arguably shares *more* data than many other social media companies do, and therefore has more opportunities for illustrative failures. *See infra* Section I.C.1.

80. Matthew Rosenberg, Nicholas Confessore & Carole Cadwalladr, *How Trump Consultants Exploited the Facebook Data of Millions*, N.Y. TIMES (Mar. 17, 2018), https://

the center of controversy for its role in the 2016 United States presidential election, Brexit, and the spreading of Russian-influenced propaganda, but Cambridge Analytica turned a gradual public relations crisis into an acute one.

Facebook higher ups soon after began to look for new ways to support independent research to help avoid future election interference, and honed in on one method proposed by Harvard social scientists Gary King and Nate Persily.[81] King and Persily argued that researchers inside social media companies had access to data but no credibility or independence, while researchers outside the companies had the inverse. To resolve this, they proposed giving some academics access to a company's data but having them sign NDAs and preventing them from publishing. Those academics on the inside could then help decide what data is important and how to share it with third-party researchers in a privacy-preserving way.[82]

Facebook quickly put the proposal into practice. About three weeks after the Cambridge Analytica leak (and one day before Zuckerberg was slated to testify before the Senate), Facebook announced a new initiative to allow academics independent access to Facebook data.[83] King and Persily established SS1 as the organization that would operate within Facebook, and they brought on the Social Science Research Council (SSRC) to manage external researchers, who would apply for access to the data they made available. King and Persily raised ten million dollars for the initiative from an ideologically diverse group of seven foundations.[84]

In July 2018, SS1 announced the data set Facebook would release: every URL that had ever been shared publicly on Facebook between January 1, 2017 and June 11, 2018, along with information about who shared it, how often it

---

www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html; Carole Cadwalladr & Emma Graham-Harrison, *Revealed: 50 Million Facebook Profiles Harvested for Cambridge Analytica in Major Data Breach*, OBSERVER (Mar. 17, 2018), https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election.

81. O'HARA & NELSON, *supra* note 8, at 3–4.

82. Gary King & Nathaniel Persily, *A New Model for Industry-Academic Partnerships*, 53 PS: POL. SCI. & POLITICS 703, 703 (2020).

83. *Mark Zuckerberg*, FACEBOOK (Apr. 9, 2018), https://www.facebook.com/zuck/posts/10104797374385071; Elliot Schrange & David Ginsberg, *Facebook Launches New Initiative to Help Scholars Assess Social Media's Impact on Elections*, META (Apr. 9, 2018), https://about.fb.com/news/2018/04/new-elections-initiative/.

84. The list of foundations includes The William and Flora Hewlett Foundation, The Charles Koch Foundation, The John S. and James L. Knight Foundation, Laura and John Arnold Foundation, the Alfred P. Sloan Foundation, The Democracy Fund, and Omidyar Network. O'HARA & NELSON, *supra* note 8, at 7–8.

was shared, and how many people saw it.[85] SSRC and SS1 put out a request for proposals for research projects and granted $50,000 to each project along with access to the URL share dataset.[86]

However, the endeavor faced legal and political headwinds. Europe's General Data Protection Regulation (GDPR) took effect in May 2018 and California passed the California Consumer Privacy Act a month later, introducing new legal complexities. The Electronic Privacy Information Center (EPIC) also sent an open letter to SS1 claiming that the project complied with neither GDPR's personal data protection requirements nor Facebook's 2011 consent decree from the Federal Trade Commission to obtain user consent before sharing data.[87]

The project also faced technical headwinds. Facebook needed to comb through a huge volume of data to create the URL shares dataset. Facebook had over two billion active users, the URL shares dataset was initially calculated to include sixty billion public posts, and preparing just the shares and interaction metrics required processing more than fifty terabytes per day.[88] Many researchers would likely not have the computing resources to ingest this much data. Simultaneously, Facebook tried to respond to privacy concerns by implementing differential privacy, a statistical method that adds noise to a dataset to make individuals less identifiable, while still maintaining certain core patterns in the data.[89] In 2018, differential privacy was still relatively new and

85. Solomon Messing, Bogdan State, Chaya Nayak, Gary King & Nathaniel Persily, *Facebook URL Shares: Codebook*, HARVARD DATAVERSE 1 (July 11, 2018), https://doi.org/10.7910/DVN/EIAACS/PMQG9X ("URLs are included if shared by at least 20 unique accounts, and shared publicly at least once."). By the time the data set launched, it was expanded to go through February 19, 2019, but would only include URLs shared more than 100 times. Gary King & Nathaniel Persily, *Unprecedented Facebook URLs Dataset Now Available for Academic Research through Social Science One*, SOC. SCI. ONE BLOG (Feb. 13, 2020), https://socialscience.one/blog/unprecedented-facebook-urls-dataset-now-available-research-through-social-science-one.

86. *Social Science One Public Launch*, SOC. SCI. ONE BLOG (July 11, 2018), https://socialscience.one/blog/social-science-one-public-launch; O'HARA & NELSON, *supra* note 8, at 10.

87. Letter from Marc Rotenberg, Christine Bannan, Sunny Kang, Sam Lester, Electronic Privacy Information Center to Gary King & Nathaniel Persily, Soc. Sci. One (July 12, 2018), https://epic.org/wp-content/uploads/privacy/facebook/EPIC-ltr-SocialScienceOne-July-2018.pdf.

88. Josh Constantine, *Facebook Now Has 2 Billion Monthly Users . . . And Responsibility*, TECH CRUNCH (June 27, 2017), https://techcrunch.com/2017/06/27/facebook-2-billion-users/; O'HARA & NELSON, *supra* note 8, at 17.

89. Cynthia Dwork, Frank McSherry, Kobbi Nissim & Adam Smith, *Calibrating Noise to Sensitivity in Private Data Analysis*, *in* 3876 THEORY OF CRYPTOGRAPHY CONF. 2006, LECTURE NOTES IN COMPUT. SCI. 265, 265 (Shai Halevi & Tal Rabin eds., 2006); O'HARA & NELSON, *supra* note 8, at 17.

Facebook engineers underwent lots of trial and error to make it work at such a scale.[90]

The technical and legal challenges plagued the project with delays and eventually led to its collapse. SS1 and SSRC believed that Facebook would be able to provide the URL shares data by fall 2018, but they gave no information until January 2019, when they admitted to further delay. SSRC announced the first research grant winners in April 2019, which included more than sixty researchers from thirty academic institutions in eleven countries, but Facebook still had no URL shares data.[91] When Facebook did finally share data, it was a "light" version of the dataset, which excluded demographic and exposure data. This meant researchers could not study who and how many people different posts reached, likely hampering research on such topics as mis- and disinformation. At the end of SS1's year-long funding period, all seven funders sent a joint letter to SSRC announcing that they would discontinue funding. As they explained:

> It now seems clear that the technical and legal complexities associated with making proprietary data available to independent scholars are greater than any of the parties originally understood, and Facebook has as a result been unable to deliver all the data initially anticipated.[92]

Facebook continued the project on its own, and the full URL shares dataset was finally made available to researchers in February 2020. However, statistical analysis from King and others suggest that the differential privacy methods Facebook used added significant statistical bias.[93] In 2021, Facebook also revealed that the data accidentally excluded URLs shared by any U.S. user without detectable political leanings, about half of all US Facebook users.[94]

A 2019 post-mortem released by the Hewlett Foundation offered multiple interpretations of the events of SS1. One is that funders, SS1, and SSRC put

90. O'HARA & NELSON, *supra* note 8, at 17.

91. Elliot Schrage & Chaya Nayak, *First Grants Announced for Independent Research on Social Media's Impact on Democracy Using Facebook Data*, META (Apr. 29, 2019), https://about.fb.com/news/2019/04/election-research-grants/.

92. Letter from Funders Supporting Independent Scholarly Access to Facebook Data to The Social Science Research Council (Aug. 27, 2019), https://ssrc-static.s3.amazonaws.com/sdi/resources/SMDRG_funder_letter_august_2019.pdf.

93. Georgina Evans & Gary King, *Statistically Valid Inferences from Differentially Private Data Releases, with Application to Facebook URLs Dataset*, POL. ANALYSIS 1, 1 (2022), https://gking.harvard.edu/dpdw; Simon Hegelich, *Facebook Needs to Share More with Researchers*, NATURE (Mar. 24, 2020), https://www.nature.com/articles/d41586-020-00828-5.

94. Craig Timberg, *Facebook Made Big Mistake in Data It Provided to Researchers, Undermining Academic Work*, WASH. POST (Sept. 10, 2021), https://www.washingtonpost.com/technology/2021/09/10/facebook-error-data-social-scientists/.

the cart before the horse: "investing in research was premature given the uncertainty of data access."[95] Another is that SS1 was unable to motivate Facebook to share data.[96] Both may be right.

All in all, SS1 is widely seen as a failure, or as Persily put it to the press, "I'm happy to be quoted saying this: This was the most frustrating thing I've been involved in, in my life."[97] Persily later stated that the demise of SS1 "demonstrates why we need government regulation to force social media companies to develop secure data sharing programs with outside independent researchers."[98]

2. NYU Ad Observatory

Laura Edelson and Damon McCoy of the NYU Cybersecurity for Democracy group started the NYU Ad Observatory on September 15, 2020.[99] The Observatory was meant to increase political ad transparency on social media ahead of the 2020 elections, and let researchers independently search for and analyze political ads by state, races, targeting criteria, funding sources, money spent, and messaging. The Observatory quickly saw adoption, particularly from journalists reporting on federal and local elections, including in Florida, Kentucky, Missouri, and Utah.[100]

Data for the NYU Ad Observatory came from a mix of platform sanctioned and unsanctioned sources. It used reports provided by Facebook such as the Facebook API, CrowdTangle, and Ad Library reports, as well as

95. This is harder to verify since communications between SS1 and Facebook were under NDA. O'HARA & NELSON, *supra* note 8, at 18.

96. *Id.*

97. Issie Lapowsky, *Why Facebook's Data-Sharing Project Ballooned Into A 2-Year Debacle*, PROTOCOL (Feb. 13, 2020), https://www.protocol.com/facebook-data-sharing-researchers.

98. Timberg, *supra* note 94.

99. NYU TANDON SCHOOL OF ENGINEERING, NEW TOOL TO ANALYZE POLITICAL ADVERTISING ON FACEBOOK REVEALS MASSIVE DISCREPANCIES IN PARTY SPENDING ON PRESIDENTIAL CONTEST (2020), https://engineering.nyu.edu/sites/default/files/2020-09/NYU-Ad-Observatory.pdf.

100. Christine Stapleton, *How Much Have Local Congressional Candidates Spent on Facebook Ads?*, PALM BEACH POST (Sept. 18, 2020), https://www.palmbeachpost.com/story/news/politics/2020/09/18/how-much-local-congressional-candidates-spending-facebook-ads/3492178001/; Craig Silverman & Ryan Mac, *Facebook Promised to Label Political Ads, but Ads for Biden, the Daily Wire, and Interest Groups are Slipping Through*, BUZZFEED NEWS (Oct. 22, 2020), https://www.buzzfeednews.com/article/craigsilverman/facebook-biden-election-ads; Tessa Weinberg, *Social Media Ads Another Battleground to Reach Voters in Missouri Governor's Race*, MISSOURI INDEP. (Oct. 20, 2020), https://missouriindependent.com/2020/10/20/social-media-ads-another-battleground-to-reach-voters-in-missouri-governors-race/; Brittany Glas, *KSL Investigates: Who is Behind the Millions in Facebook Political Ads Targeting Utah?*, KSL.COM (Oct. 9, 2020), https://www.ksl.com/article/50028514/ksl-investigates-who-is-behind-the-millions-in-facebook-political-ads-targeting-utah.

an unsanctioned browser extension called the Ad Observer that users could install to scrape ad data from the Facebook website to donate to the Observatory. The Ad Observer is an open-source tool that underwent independent reviews of its code and privacy practices to ensure it adequately obtained user consent and collected only the data it needed.[101] Edelson claimed that they could not depend solely on data Facebook made available—particularly Facebook's Ad Library—because it had many reporting inconsistencies and thousands of missing ads.[102]

In late October 2020, Facebook sent a cease and desist letter to Edelson and McCoy, demanding NYU Cybersecurity for Democracy shut down its Ad Observer plug-in and delete any data collected from it. Civil society groups lashed back: more than fifty signed onto a letter from Mozilla demanding Facebook withdraw the cease and desist.[103] The Knight First Amendment Institute provided legal representation for Edelson and McCoy.[104] Little was heard from the case for the next several months while negotiations between Facebook and NYU Cybersecurity for Democracy continued behind closed doors.

On August 3, 2021, negotiations broke down and Facebook suspended Edelson, McCoy, and others' Facebook accounts, thereby cutting off their access to Facebook's sanctioned tools, the API, Ad Library, and CrowdTangle. Facebook had cut off other ad transparency tools in the past, including ones from ProPublica, Mozilla, and Who Targets Me, but they largely did this by updating their own website in a way that broke those tools, not by suspending

101. JASON CHUANG, AD OBSERVER PRIVACY PROPERTIES & DATA COLLECTION 1, MOZILLA BUGZILLA, https://bug1676407.bmoattachments.org/attachment.cgi?id=9187255 (last visited Jan. 30, 2024). Specifically, that info was from the "Why am I seeing this ad?" box of each ad a user saw. *Id.* at 2.

102. Jeremy B. Merrill, *How Facebook's Ad System Lets Companies Talk Out of Both Sides of Their Mouths*, MARKUP (Apr. 13, 2021), https://themarkup.org/citizen-browser/2021/04/13/how-facebooks-ad-system-lets-companies-talk-out-of-both-sides-of-their-mouths; Laura Edelson, *Audit of Facebook Ad Transparency Finds Missed Political Ads*, MEDIUM (Oct. 22, 2020), https://medium.com/online-political-transparency-project/audit-of-facebook-ad-transparency-finds-missed-political-ads-603f95027cc6.

103. Letter from Mozilla to Mark Zuckerberg, CEO of Facebook, Dear Facebook: Withdraw Your Cease & Desist to NYU (Oct. 28, 2020), https://foundation.mozilla.org/en/blog/dear-mr-zuckerberg/.

104. *Researchers, Knight Institute Condemn Facebook Effort to Squelch Research on Disinformation*, KNIGHT FIRST AM. INST. COLOM. U. (Oct. 23, 2020), https://knightcolumbia.org/content/researchers-knight-institute-condemn-facebook-effort-to-squelch-research-on-disinformation.

researchers' accounts.[105] In a blog post titled, "Research Cannot Be the Justification for Compromising People's Privacy," Facebook claimed that they "took these actions to stop unauthorized scraping and protect people's privacy in line with our privacy program under the FTC Order," and offered the Ad Library as an alternative.[106]

There was immediate public outrage from academics, civil society, journalists, and lawmakers.[107] Edelson published an opinion piece in *The New York Times* a week after the incident arguing against Facebook's justifications blocking their work.[108] Edelson testified before Congress at the end of September, where she argued that to use the Ad Library, researchers were required to "sign an agreement that limits how they use and share the data, which significantly hampers meaningful publication of any research findings, as the dataset that would be necessary for other researchers to reproduce any findings cannot be publicly shared."[109] Edelson also argued that many ads were missing from the Ad Library and that others were intentionally mislabeled as non-political by bad actors.[110] FTC Acting Director of the Bureau of Consumer Protection Samuel Levine soon sent a letter clarifying that the NYU Ad Observer did not break Facebook's consent decree:

> Had you honored your commitment to contact us in advance, we would have pointed out that the consent decree does not bar Facebook from creating exceptions for good-faith research in the public interest. Indeed, the FTC supports efforts to shed light on

105. Jeremy B. Merrill & Ariana Tobin, *Facebook Moves to Block Ad Transparency Tools—Including Ours*, PROPUBLICA (Jan. 28, 2019), https://www.propublica.org/article/facebook-blocks-ad-transparency-tools.

106. Mike Clark, *Research Cannot be the Justification for Compromising People's Privacy*, META (Aug. 3, 2021), https://about.fb.com/news/2021/08/research-cannot-be-the-justification-for-compromising-peoples-privacy/.

107. *Standing with Laura Edelson in Support of Tech Industry Accountability Research*, EDELSON-SOLIDARITY (Aug. 6, 2021), https://edelson-solidarity.neocities.org/; Lisa Macpherson, *Observe and Report: Facebook Versus NYU Ad Observatory Proves the Need for Policy Interventions*, PUB. KNOWLEDGE (Aug. 11, 2021), https://publicknowledge.org/observe-and-report-facebook-versus-nyu-ad-observatory-proves-the-need-for-policy-interventions/; Taylor Hatmaker, *Facebook Cuts Off NYU Researcher Access, Prompting Rebuke from Lawmakers*, TECHCRUNCH (Aug. 4, 2021), https://techcrunch.com/2021/08/04/facebook-ad-observatory-nyu-researchers/.

108. Laura Edelson & Damon McCoy, *We Research Misinformation on Facebook. It Just Disabled Our Accounts.*, N.Y. TIMES (Aug. 10, 2021), https://www.nytimes.com/2021/08/10/opinion/facebook-misinformation.html.

109. Testimony of Laura Edelson, NYU Cybersecurity for Democracy, Before the Subcomm. on Investigations & Oversight of the H. Comm. on Sci., Space, and Tech., 117th Cong. 2 (2021), https://docs.house.gov/meetings/SY/SY21/20210928/114064/HHRG-117-SY21-Wstate-EdelsonL-20210928.pdf.

110. Laura Edelson, Tobias Lauinger & Damon McCoy, *A Security Analysis of the Facebook Ad Library*, 2020 IEEE SYMP. ON SEC. & PRIV. (SP) 661, 667 (2020).

> opaque business practices, especially around surveillance-based advertising. While it is not our role to resolve individual disputes between Facebook and third parties, we hope that the company is not invoking privacy—much less the FTC consent order—as a pretext to advance other aims.[111]

Despite being cut off from some data, NYU Cybersecurity for Democracy was able to release a new version of the Ad Observatory ahead of the 2022 elections.[112] Facebook (now Meta) has not shared whether or not they have reinstated any of the researchers' accounts as of this writing, but the company has expanded their own Ad Library to include more in-depth targeting information about political ads. However, researchers continue to argue that Ad Library misses several political ads since those running the ads do not identify them as political.

D. A TAXONOMY OF PROBLEMS WITH RESEARCHER ACCESS TO SOCIAL MEDIA DATA

We identify two broad categories of problems that currently afflict social media data sharing. The first is *poor research quality*; existing approaches to giving researchers access to data negatively impact the quality and utility of research that gets produced. The second is *unrealized research*; some socially beneficial types of research cannot be done at all with the data currently made available.

1. Poor Research Quality

a) Limited by Data Access Arrangements

Platforms sometimes require researchers to sign burdensome contracts in order to gain access to data, as the NYU Ad Observatory argued Facebook has done.[113] Platforms can also impose large technical burdens, like how TikTok requires researchers using its research API to refresh results "at least every fifteen (15) days, and delete data that is not available from the TikTok Research API at the time of each refresh."[114] Even without requiring pre-publication approval, a platform has unilateral power over the data it makes

111. Samuel Levine, *Letter from Acting Director of the Bureau of Consumer Protection Samuel Levine to Facebook*, FED. TRADE COMM'N: CONSUMER BLOG (Aug. 5, 2021), https://www.ftc.gov/blog-posts/2021/08/letter-acting-director-bureau-consumer-protection-samuel-levine-facebook.

112. *New, Enhanced Adobservatory.Org Provides Transparency & Insights on Digital Political Spending*, NYU TANDON (Aug. 3, 2022), https://medium.com/cybersecurity-for-democracy/new-enhanced-adobservatory-org-provides-transparency-insights-on-digital-political-spending-784f87a12006.

113. *See supra* Section II.B.2.

114. *TikTok Research API Services Terms of Service*, TIKTOK § I.3.e, https://www.tiktok.com/legal/page/global/terms-of-service-research-api/en.

available and may be able to pressure researchers to suppress results that reflect on it negatively. This is particularly acute when companies provide ad hoc access to individual researchers, or when researchers receive direct funding from companies. The inability to share data further makes research results less robust and more difficult to publish since it is unreproducible and unverifiable.

b) Unstable Data Access

Platforms regularly change which data they make available to researchers and under what terms, often with little warning. Shortly after Musk acquired Twitter, for instance, the service very suddenly raised the cost of its API from free to $42,000 a month, making it inaccessible to nearly all academic researchers and jeopardizing hundreds of in-progress research projects.[115] Data access can change also because new threats to privacy and security are uncovered, as happened with SS1 and the Facebook Graph API in the wake of Cambridge Analytica.[116] The possibility of data access changing precludes entire research methodologies, such as longitudinal research, and threatens in-progress projects.

c) Decontextualized Data Production

Platforms often share only limited information about how they generate the data they share and how it has been filtered. Without understanding the provenance of data from platforms' tools, researchers often cannot know or predict how their data is skewed. This problem is not just theoretical. Studies show that tweets from Twitter's livestream API, which shares 1% of all live traffic, are not randomly sampled.[117] Often, platform-permissioned tools are not designed with research in mind, so they can be missing basic

115. Justin Calma, *Twitter Just Closed the Book On Academic Research*, VERGE (May 31, 2023), https://www.theverge.com/2023/5/31/23739084/twitter-elon-musk-api-policy-chilling-academic-research; *Letter: Twitter's New API Plans Will Devastate Public Interest Research*, COALITION FOR INDEP. TECH. RSCH. (Apr. 3, 2023), https://independenttechresearch.org/letter-twitters-new-api-plans-will-devastate-public-interest-research/. Even those who pay for it claim it doesn't work. *See* Matt Binder, *Twitter's API Keeps Breaking, Even For Developers Paying $42,000*, MASHABLE (June 29, 2023), https://mashable.com/article/twitter-api-elon-musk-developer-issues-apps.

116. *See Graph API Reference*, META FOR DEVELOPERS, https://developers. https://developers.facebook.com/docs/graph-api/changelog/version3.0#gapi-90 (last visited Aug. 8. 2023).

117. Fred Morstatter, Jürgen Pfeffer & Huan Liu, *When is it Biased? Assessing the Representativeness of Twitter's Streaming API*, WWW '14 COMPANION: PROCS. 23RD INT'L CONF. ON WORLD WIDE WEB 555 (2014).

information.[118] Even when these tools are designed for researchers, opacity around the processes in which they are built can lead to huge oversights that even the platforms themselves miss, as occurred with SS1.[119]

d) Streetlight Effect

The streetlight effect is a type of bias wherein people only search for something where it is easiest to look, just as someone who lost their keys outside at night might only look where there are streetlights.[120] A similar effect plays out in social media research: researchers often study the platforms where they can access the most data, not necessarily the ones most relevant to the effect they are trying to study.[121] Entire domains of research can end up centralizing around non-representative data sources, as some argue occurred with Twitter.[122]

The streetlight effect also creates perverse incentives for companies not to share data. Companies that provide data may end up receiving more scrutiny and criticism from researchers. They may not even experience the public relations benefits of openness because they may be publicly criticized, as frequently and harshly as companies that share no data at all, for sharing insufficient data or in ways that make it difficult to use.[123]

e) Denominator Problem

The denominator problem is when researchers are unable to use the volume of overall activity on a platform to contextualize their findings.[124] For instance, imagine that a researcher found five thousand tweets in Hindi over a week-long period of time that promote ethnic violence against Muslims. Without certain baseline information, such as the total number of tweets per week, tweets in Hindi per week, or total active users versus active Hindi-speaking users, that researcher will not know whether their five thousand tweets should be considered a lot or a little.

---

118. Tromble, *supra* note 16 ("[T]he non-randomness of data captured via [Twitter's] APIs means that, even in the best of times, many Twitter studies have drawn conclusions based on substantially biased inferences.").

119. *See supra* Section II.B.1.

120. DAVID H. FREEDMAN, WRONG: WHY EXPERTS* KEEP FAILING US—AND HOW TO KNOW WHEN NOT TO TRUST THEM (2010).

121. *E.g.*, Tromble, *supra* note 16; Michael Zimmer & Nicholas Proferes, *A Topology of Twitter Research: Disciplines, Methods, and Ethics*, 66 ASLIB J. INFO. MGMT. 250 (2014).

122. Nicolas Kayser-Bril, *Under the Twitter Streetlight: How Data Scarcity Distorts Research*, ALGORITHM WATCH, https://algorithmwatch.org/en/data-access-researchers-left-on-read/.

123. SHAPIRO ET AL., *supra* note 3, at 24–26.

124. *Id.* at 46.

2. Unrealized Research

a) Inability to Evaluate Social Media Claims

Social media companies frequently roll out changes to their systems. Sometimes, these changes are publicly announced and are meant to address controversies or harms uncovered by research.[125] However, without access to adequate data, researchers are unable to evaluate the effectiveness of these interventions, or whether they have been rolled out at all. Claims related to opaque technical systems, such as recommendation algorithms and content moderation practices, are nearly impossible to evaluate, making it difficult for the public to distinguish between public relations puffery and meaningful changes.

b) Unequal Access Leads to Less Diverse Research

Researchers with personal connections to large social media companies are more easily able to gain access to data through both informal and formal means. Well-connected researchers are more likely to convince companies to share data in ad hoc ways for one-off projects.[126] They are also better able to defend their unsanctioned access since they may have powerful allies, such as when the Knight First Amendment Institute offered legal defense to NYU Cybersecurity for Democracy for its Ad Observatory.[127] Less resourced and connected researchers may not even have the budget to purchase the computing power necessary to do certain research.

This unmeritocratic approach to doling out access to data may lead to worse outcomes. The best-connected researchers are not necessarily the ones who come up with the best research questions or plans of execution. Underrepresented researchers may bring unique insights and approaches that more well-connected researchers do not.

c) Inability to Discover Unexpected Effects

Social media companies share non-public data with researchers in some areas more than others. Meta, for example, offers more information about political advertisements than it does non-political advertisements, in part

125. *E.g.*, Vanessa Pappas & Kudzi Chikumbu, *A Message to Our Black Community*, TikTok Newsroom (June 1, 2020); Vijaya Gadde & Kavyon Beykpour, *Setting the Record Straight on Shadow Banning*, X BLOG (July 26, 2018); Mark Zuckerberg, *A Blueprint for Content Governance and Enforcement*, X BLOG (Nov. 15, 2018).

126. *E.g.*, Nancy Scola, *Facebook's Next Project: American Inequality*, POLITICO (Feb. 19, 2018), https://www.politico.com/story/2018/02/19/facebook-inequality-stanford-417093 (reporting that Facebook shared data with Stanford economist Raj Chetty, "a favorite among tech elites," but not with other researchers or the broader public).

127. *Researchers, Knight Institute Condemn Facebook Effort*, *supra* note 104.

because researchers have appealed to democratic values to gain such access.[128] By limiting access to other data not deemed as important, however, platforms may prevent researchers from discovering new, unexpected effects of different technological architectures, user interfaces, and policy designs. A change in the way a social network displays advertisements, for instance, could drastically increase how often users fall for cryptocurrency fraud. This effect would be unexpected and important, but impossible for researchers to discover for a number of reasons: researchers do not have access to data regarding how the company rolled out the change to advertisements (e.g., A/B test data), which content gets flagged as cryptocurrency fraud, which ads can be categorized as cryptocurrency ads, or how much engagement those ads receive. Companies are disincentivized from finding or sharing with the public new negative social impacts of their services.

d) Slow Responses to Sudden Problems

Sudden social, economic, and political upheavals often play out on social media. Fast evolving and paradigm shifting events such as COVID-19, the January 6th attacks, and the Russian attack on Ukraine are both reflected on and affected by the online information ecosystem.[129] Researcher organizations that use platform data access mechanisms to run social media monitoring programs, including the Stanford Internet Observatory and the Global Disinformation Lab at UT Austin, may be uniquely poised to give platforms the information they need to act quickly. Sharing timely data with external researchers, such as watchdog organizations and journalists, could help companies and the public better understand what is happening on platforms, and in turn, improve responses to such upheavals. Platforms, however, do not have policies to allow emergency access to data, even if it may be useful for all parties.

E. THE LEGAL LANDSCAPE OF DATA SHARING

This Section takes a step back to consider the state of social media data sharing from a legal point of view.

128. *See* Paddy Leerssen, Tom Dobber, Natali Helberger & Claes de Vreese, *News from the Ad Archive: How Journalists Use the Facebook Ad Library to Hold Online Advertising Accountable*, 26 INFO. COMMC'N & SOC'Y 1381, 1383 (2021).

129. *See, e.g.*, Mia Sato, *Ukrainian Influencers Bring the Frontlines to TikTok*, VERGE (Mar. 16, 2022), https://www.theverge.com/c/22971491/ukraine-tiktok-influencers-russian-invasion; Cathleen O'Grady, *In the Line of Fire*, 375 SCI. 1338 (2022).

1. What Made Things This Way?

As the case studies above highlight, the barriers to data access are not only technical, but also legal. Subject to a few narrow exceptions outlined in state and federal privacy laws, social media data is subject to private ordering: once data subjects have consented to their data being collected, companies enjoy broad discretion to determine who gains access to social media data and on what terms such access is granted.

Companies assert both legal rights and legal duties to control and manage access to proprietary data. Technically, there is no recognized legal property right in data per se, despite enduring debate over recognizing one.[130] Instead, companies rely on two kinds of legal claims to approximate full-throated entitlement rights over data access and control: rights to limit access to data to protect commercial secrets and competitive advantage, and obligations companies owe data subjects to limit access to data, which may arise under companies' terms of service or privacy laws. Together, these two kinds of legal claims allow companies to justify broad, contractually governed discretion over how researchers gain access to data.

Both trade secrecy and privacy claims generally arise out of underlying contractual legal relationships that structure companies' claims to and obligations regarding social media data. Two kinds of contractual relationships govern, to a large degree, how social media data is collected, processed, and used. First, terms of service govern collection and the relationship between companies and data subjects, and second, data use agreements govern data access and the relationship between companies and researchers.

Companies have been able to constrain access to data in the contractual realm because of their success at invoking underlying privacy and trade secrecy rationales—rationales that companies use as obstacles to increased public oversight and control over researcher access. Thus, we focus on privacy and trade secrecy because these are the doctrinal obstacles and normative justifications that platforms invoke in public statements against researcher access. To retrieve affirmative public rights of researcher access from the realm of private contractual ordering requires us to address these privacy and trade secrecy claims.

130. *See generally* James Grimmelmann & Christina Mulligan, *Data Property*, 72 AM. U. L. REV. 829 (2023) (arguing for personal property-like rights in personal data); Michael C. Pollack, *Taking Data*, 86 U. CHI. L. REV 77 (2019) (arguing for property-like rights in personal data against government use of data); Jorge L. Contreras, *The False Promise of Health Data Ownership*, 94 N.Y.U. L. REV 624 (2019) (detailing the challenges and risks of recognizing personal property claims to health data).

a) Trade Secrecy (and Other Entitlement-Like Claims)

First, companies make trade secrecy claims to protect their commercial interests in data acquired from users and used to develop their products.[131] As Tait Graves and Sonia Katyal have written (in a broad survey of recent trends in trade secrecy law), "companies are increasingly exploiting [gaps in trade secrecy doctrine] to assert trade secret rights in a growing range of nontraditional contexts."[132] Under now-dominant definitions of a trade secret, information qualifies for trade secret protection if it (1) is generally not known to others in the same industry; (2) is not readily ascertainable from the use of limited time and effort; (3) has actual or potential independent economic value to competitors; and (4) is reasonably guarded as secret.[133] This broad definition permits companies to claim—often without substantiation—proprietary rights over a sweeping range of information.[134] Once a claim of trade secrecy is made, companies wield the claim to withhold the information from researchers and even regulators.[135] These companies argue that disclosure of the secret information—even to these noncommercial audiences—will inevitably lead to some leaks to competitors, encouraging free riding and thereby eroding crucial incentives to innovate.[136]

Tech platforms have a track record of making such trade secrecy claims. For instance, in its 2020 comments to the FTC on data portability, Facebook alleged that data such as granular use logs, non-human understandable data, and data stored in formats that rely on proprietary technology "make clear that

131. Frederick Mostert & Alex Urbelis, *Social Media Platforms Must Abandon Algorithmic Secrecy*, FIN. TIMES (June 16, 2021), https://www.ft.com/content/39d69f80-5266-4e22-965f-efbc19d2e776 (noting the obstacles trade secret law creates for accountability and transparency); King & Persily, *supra* note 82 ("[P]rogress in data sharing for social good will occur only if all incentives are aligned—if individual privacy is protected, company trade secrets and related proprietary information are respected, and the standards and independence of the scientific process are secured.").

132. Charles Tait Graves & Sonia K. Katyal, *From Trade Secrecy to Seclusion*, 109 GEO. L.J. 1337, 1351 (2021).

133. U.T.S.A. § 1(4); 18 U.S.C. § 1839(3).

134. Graves & Katyal, *supra* note 132, at 1352–68; *see also* Deepa Varadarajan, *Business Secrecy Expansion and FOIA*, 68 UCLA L. REV. 462 (2021); Morten, *Publicizing Corporate Secrets*, *supra* note 34; Amy Kapczynski, *The Public History of Trade Secrets*, 55 UC DAVIS L. REV. 1367 (2022); Christopher J. Morten & Amy Kapczynski, *The Big Data Regulator, Rebooted: Why and How the FDA Can and Should Disclose Confidential Data on Prescription Drugs and Vaccines*, 109 CALIF. L. REV. 493 (2021).

135. Graves & Katyal, *supra* note 132, at 1353–54; *see also* Elizabeth A. Rowe, *Striking a Balance: When Should Trade-Secret Law Shield Disclosures to the Government?*, 96 IOWA L. REV. 791 (2011) (describing the phenomenon of regulated companies withholding alleged trade secret information from regulators).

136. Rowe, *supra* note 135, at 793–94.

including all observed and inferred data could also result in a different sort of burden: the disclosure of trade secret or other proprietary information developed by a business to enhance or differentiate its services. Enabling people to port that kind of information could reduce incentives for businesses to develop it in the first place."[137] In 2021, Facebook withheld internal research on the impact of its platforms on youth mental health from senators, stating that "its internal research is proprietary and 'kept confidential to promote frank and open dialogue and brainstorming internally.'"[138] In 2023, the Information Technology Industry Council (ITI), issued a statement expressing concern over the European Union Data Act's data sharing provisions. ITI, which includes Google, Meta, Microsoft, and Snap as members, argued the law should be amended to permit companies to "refus[e] to share data in specific circumstances where disclosure of trade secrets would be likely to cause serious damage to the data holder."[139] A bit further afield, Uber Eats and two other food delivery platforms challenged a New York City municipal ordinance requiring platforms share customer data with the underlying restaurant fulfilling an order. All three platforms asserted that the law constitutes a violation of their trade secrecy rights under the Second Circuit standard.[140]

Companies also use other entitlement-like claims to limit extra-contractual researcher access. For instance, despite recent cases limiting the application of such laws to certain forms of research, many companies still include language in their terms of service indicating that activity that violates their terms will be referred to law enforcement for prosecution under the Computer Fraud and Abuse Act (CFAA).[141] Copyright enforcement has similarly endowed

137. FACEBOOK, *supra* note 10.

138. Georgia Wells, Jeff Horwitz & Deepa Seetharaman, *Facebook Knows Instagram is Toxic for Teen Girls, Company Documents Show*, WALL ST. J. (Sept. 14, 2021), https://www.wsj.com/articles/facebook-knows-instagram-is-toxic-for-teen-girls-company-documents-show-11631620739.

139. *Global Tech Association ITI Raises Concerns Ahead of Crucial Week for EU Data Act Adoption*, INFO. TECH. INDUSTRY COUNCIL (Mar. 13, 2023), https://www.itic.org/news-events/news-releases/global-tech-association-iti-raises-concerns-ahead-of-crucial-week-for-eu-data-act-adoption.

140. *See, e.g.*, Complaint, Portier v. City of New York, No. 21-cv-10347, 2021 WL 5758964 (S.D.N.Y. 2021) ("The personal data that users have entrusted to Uber Eats constitute trade secrets, which required significant investment and expenditure to accumulate. The Ordinance plainly interferes with Uber Eats' exclusive and economic use of those trade secrets."); *see also* DoorDash v. City of New York, No. 21-cv-7695 (S.D.N.Y. filed Sept. 15, 2021), Grubhub v. City of New York, No. 21-cv-10602 (S.D.N.Y. filed Dec. 10, 2021).

141. Nat Meysenburg, *Cybersecurity Research Should Not Be A Crime*, NEW AM. (Nov. 18, 2021), https://d1y8sb8igg2f8e.cloudfront.net/documents/Research_Exemptions_One-Pager.pdf; *see also* Sandvig v. Barr 451 F. Supp. 3d 73 (D.C. Cir. 2020) (concluding that the CFAA does not criminalize mere terms-of-service violations on consumer websites, and

platforms with legal rights to control and manage access. For instance, the Digital Millennium Copyright Act (DMCA) not only establishes a takedown regime for unauthorized content, but also includes prohibitions against circumventing technical access protections, knowingly and improperly obtaining valuable trade secrets, and distributing technologies that facilitate circumvention.[142] The practical upshot of the DMCA, particularly the provision against trafficking in circumvention technologies themselves, is that platforms enjoy strong rights over access control protocols.[143]

b) Privacy

Second, companies assert that the privacy obligations they owe consumers (either via the contractual promises they make to data subjects or due to privacy regulations with which they must comply) are reasons to deny researcher access.[144] These concerns, while sometimes used as pretexts by companies to protect the value of walled-off data assets, are not always levied in bad faith or without merit. Users have legitimate privacy interests in the data at issue in researcher access; protecting this legitimate interest makes researcher access a legally and ethically tricky problem.[145] Indeed, researchers

---

therefore that the Plaintiffs' proposed research plans were not criminal activity under the CFAA); Facebook, Inc. v. Power Ventures, Inc., 844 F.3d 1058, 1065–69 (9th Cir. 2016) (holding a third-party platform civilly liable under the CFAA for accessing Facebook users' data).

142. To be clear, the DMCA does not directly apply to social media data (which is not as a general matter copyrightable), but it has featured significantly as a background law governing the relationship between online platforms and external researchers of those platforms, and depending on the research in question, may be implicated in a given form of social media research.

143. COHEN, *supra* note 5, at 126.

144. While not all information privacy laws apply to social media data, laws like the Children's Online Privacy and Protection Act (COPPA) and the Fair Credit Reporting Act (FCRA) either explicitly or arguably extend to social media activity. Children's Online Privacy and Protection Act of 1998 (COPPA) (codified at 15 U.S.C. § 6501–06 (2018); Fair Credit Reporting Act (FCRA) (1970) (codified at 15 U.S.C. § 1681 (2018). In addition to federal laws, several states have passed prominent privacy laws that impose additional obligations on social media platforms. This includes both specific uses of data such as Illinois' Biometric Information Privacy Act (BIPA), and omnibus laws like California's Consumer Privacy Act (CCPA). California Consumer Privacy Act (CCPA), Cal. Civ. Code §§ 1798.100(a)(1), 1798.110(3); 1798.135. Biometric Information Privacy Act (BIPA), Pub. Act 095-994 (codified at 740 Ill. Comp. Stat. 14/1 (2008)); Family Educational Rights and Privacy Act (FERPA) (1974) (codified at 20 U.S.C. § 1232g (2018)).

145. The argument that researcher access is normatively good for user privacy is orthogonal to the argument of this Article. That said, there are compelling reasons to think that well-designed researcher access mechanisms for social media data may have salutary effects on the overall privacy of social media users. This view is suggested by the FTC's favorable response to NYU's Ad Observatory and other research that seeks to "shed light on

themselves have recognized that proposals to increase access to social data pose privacy risks to platform users.[146]

Some information privacy laws affirmatively grant data subjects additional rights and impose additional duties on platforms. For example, the California Consumer Privacy Act (CCPA) grants data subjects rights to request information about what data is being collected about them and whether any of their personal data is being sold or disclosed to third parties.[147] It also grants data subjects the right to opt out of the sale of their personal information.[148] The Children's Online Privacy Protection Act (COPPA) imposes additional obligations on platforms regarding data collected from children under thirteen years of age.[149] To comply, platforms must post comprehensive policies regarding their practices for such data and obtain verified parental consent prior to any data collection, among other requirements. Although COPPA does not prohibit children under the age of thirteen from sharing their data with platforms, many social media platforms prohibit children under age thirteen from using their services due to the costs and risks associated with violating COPPA.[150] These contractual terms—and several federal privacy laws, including COPPA—are in turn regulated by the Federal Trade Commission Act's § 5 authority and state consumer protection laws.[151]

---

opaque business practices," in the wake of Meta's efforts to use obligations under its 2012 FTC consent decree as a justification to shut down that research. *See supra* Section II.C. Such cases, where two sides of a dispute both marshal privacy arguments to advance their claims (in this case, companies and social media researchers), present an instance of what David Pozen calls a 'privacy-privacy tradeoff.' *See* David E. Pozen, *Privacy-Privacy Tradeoffs*, 83 U. CHI. L. REV. 221 (2016).

146. Daphne Keller, *User Privacy vs. Platform Transparency: The Conflicts are Real and We Need to Talk About Them*, CTR. FOR INTERNET & SOC'Y (Apr. 6, 2022), https://cyberlaw.stanford.edu/blog/2022/04/user-privacy-vs-platform-transparency-conflicts-are-real-and-we-need-talk-about-them-0; *see also* David E. Pozen, *Privacy-Privacy Tradeoffs*, 83 U. CHI. L. REV. 221 (2016).

147. California Consumer Privacy Act of 2018 (CCPA), Cal. Civ. Code §§ 1798.100(a)(1), 1798.110(3); 1798.135.

148. *See id. But see* Salome Viljoen, *The Promise and Pitfalls of California's Consumer Privacy Act*, DIGITAL LIFE INITIATIVE: CRITICAL REFLECTIONS (Apr. 11, 2020), https://www.dli.tech.cornell.edu/post/the-promise-and-pitfalls-of-the-california-consumer-privacy-act (canvassing the law's deficiencies).

149. 15 U.S.C. §§ 6501–06.

150. *Id.* COPPA applies both to services that are "directed to children" under 13, such as children's online games, and those that knowingly collect personal information from people under 13. Platforms look to avoid charges of "actual knowledge" under COPPA by requiring users to input a birthdate on their registration page, and disallowing any user that responds with a year that suggests they are under 13.

151. *See* 15 U.S.C. § 45(a)(1) (2018) (prohibiting "unfair or deceptive acts or practices in or affecting commerce"). All states have incorporated similar consumer protection clauses into

As the case studies above highlight, the legal barriers erected by privacy obligations to researcher access (as well as the perceived legal risks accompanying these barriers) are significant. In the case of SS1, the growing legal complexities around compliance with the GDPR and CCPA were key contributors to the consortium's failure. In the case of the NYU Ad Observatory, Facebook invoked privacy duties—its supposed obligations under its FTC consent decree, and its obligations to users under their terms of service—to cut off researcher access.

These cases also demonstrate additional complexities when it comes to assessing the merit of privacy claims. On the one hand, social media companies may invoke privacy obligations in bad faith to withhold data that makes them look bad.[152] In the case of the NYU Ad Observatory, for example, Facebook's attempt to use its FTC consent decree to block access to data was undermined by the FTC itself.[153] The agency clarified that it welcomed and encouraged greater researcher access to platform data.

On the other hand, companies also underinvest in privacy, and sharing data with researchers can raise legitimate privacy risks. Perhaps the most infamous example here is the Cambridge Analytica scandal, which nominally involved data harvested for a research project. SS1 sits somewhere in between this example and the NYU Ad Observatory example. Researchers and Facebook became mired in concerns over what SS1 would mean for Facebook's obligations under significant, new data protection laws. Some viewed Facebook's privacy concerns as pretextual; the company used exaggerated estimates of the perceived legal risk of new laws to wriggle out of obligations it no longer wanted to fulfill. However, Facebook was not alone in its assessment of risk. Credible third-party groups, including EPIC, clearly thought that SS1 raised genuine privacy concerns.[154]

---

their civil codes, and state attorney general offices use their enforcement authority under such statutes and myriad other state privacy laws to regulate consumer digital terms and services. Danielle Keats Citron, *The Privacy Policymaking of State Attorneys General*, 92 NOTRE DAME L. REV. 747, 754 (2016). State attorneys general have set up specialized units or departments to bring digital privacy-related enforcement actions. *See, e.g.*, *Bureau of Internet and Technology*, N.Y. ATT'Y GEN.'S OFF., https://ag.ny.gov/bureau/internet-bureau; Privacy Unit (last visited Nov. 23, 2023), *Privacy and Data Security*, CAL. ATTY' GEN.'S OFF., https://oag.ca.gov/privacy. For a list of state privacy laws, see *Privacy Laws by State*, ELEC. PRIV. INFO. CTR. (EPIC), https://epic.org/privacy/consumer/states.html (last visited Nov. 23, 2023).

152. Van Loo, *supra* note 8.

153. *See supra* Section II.C.

154. Letter from Marc Rotenberg, EPIC President, Christine Bannan, EPIC Administrative Law and Policy Fellow, Sunny Kang, EPIC International Consumer Council, and Sam Lester, EPIC Consumer Privacy Fellow, to Gary King and Nathaniel Persily, ELEC.

Regardless of whether companies raise privacy concerns in good or bad faith, courts and would-be legislators must consider the merit of such claims.[155] On this count, the privacy concerns of data sharing clearly present a challenge to unfettered researcher access, and they require good faith engagement.

2. Navigating a Path Forward Between Privacy and Trade Secrecy

Alongside the strong legal claims of companies over social media data is the conspicuous *absence* of rights to access for other entities. Users themselves have some individual rights over their data, but researchers and even government agencies have limited countervailing legal rights over data to supersede those of companies.[156] This is notable, given that absolute rights of any kind are rare in law, particularly with respect to intangible goods, and that government claims that limit or supersede private (commercial) claims of right in the course of ordinary socioeconomic legislation were once more common.[157]

The lack of public rights in social media is also extraordinary given the magnitude of the public interests at stake. Social media companies are some of the largest companies in the world. They exert significant influence on the public sphere, affecting how billions of people around the world interact with one another and with the news of the day. These spaces are key to self, social, and political formation. They generate billions, if not trillions, of dollars of revenue. And yet very little is known about how they actually work.

As this Article endeavors to show, we must overcome the legal barriers to researcher access imposed by trade secrecy and privacy claims to examine how these platforms work. Or, more accurately (and more humbly), we must find ways to navigate safely past these barriers. For this, we now turn to the lessons

---

PRIV. INFO. CTR. (July 12, 2018), https://epic.org/wp-content/uploads/privacy/facebook/EPIC-ltr-SocialScienceOne-July-2018.pdf.

155. Salome Viljoen, *Privacy Puzzles* (draft on file with author).

156. The obvious exception here is access for purposes of criminal investigations, which are subject to warrants. Several scholars have argued that the law permits, even requires, expanding countervailing public rights over privately held data. *See* Mary D. Fan, *The Public's Right to Benefit from Privately Held Consumer Big Data*, 96 N.Y.U. L. REV. 1438 (2021) (arguing for a public right to benefit from privately held consumer data such as social media data); Aziz Huq, *The Public Trust in Data*, 110 GEO. L. J. 333 (2021); Salome Viljoen, *A Relational Theory of Data Governance*, 131 YALE L.J. 573 (2021) (developing an account of data interests that accrue at the population-level and must be governed via public rights and institutions).

157. Amy Kapczynski, *The Lochnerized First Amendment and the FDA: Toward a More Democratic Political Economy*, 118 COLUM. L. REV. 179, 179–80 (2018), Jedediah Purdy, *Beyond the Bosses' Constitution: The First Amendment and Class Entrenchment*, 118 COLUM. L. REV. 2161, 2174 (2018); Amy Kapczynski, *The Public History of Trade Secrets*, 55 U.C. DAVIS L. REV. 1367, 1429–36 (2022); Christopher J. Morten, *Publicizing Corporate Secrets*, 171 U. PENN. L. REV. 1319, 1340–47 (2023).

of another powerful industry where researchers have been granted access to valuable and sensitive commercial data: pharmaceutical and medical device companies' clinical trials.

III. CLINICAL TRIAL DATA SHARING: MANDATE AND EXPERIMENTS

What is clinical trial data? What is the clinical trial data sharing mandate, and why might it matter for governance of social media data? What mechanisms have emerged for responsible sharing of even the most sensitive components of clinical trial data? This Part answers these questions.

In this Part, Section III.A introduces clinical trial data. Section III.B provides historical context for the clinical trial data sharing mandate that emerged in the United States in the 21st century. Section III.C then describes the 2007 legislation—the Food & Drug Administration Amendments Act (FDAAA)—that forms the foundation of that mandate. The law works, albeit imperfectly, and it has unlocked benefits for researchers, patients, and the broader public. Section III.D then describes the institutions that implement FDAAA and other laws that govern researcher access to clinical trial data. Section III.D also shows that some institutions that share clinical trial data have been able to achieve deeper sorts of data sharing with researchers. These relationships have made the most sensitive components of trial data—individual patient data (IPD) and detailed trial methodologies that may implicate companies' trade secrets—accessible to researchers. Section III.E distills key features.

Today researchers have meaningful access to much of the very same data that companies rely on for their research and development (R&D), regulatory approvals, and profits. So far, at least, clinical trial data sharing also capably protects the interests of the people who create this data by volunteering for clinical trials.

A. WHAT IS CLINICAL TRIAL DATA, AND WHY DOES IT MATTER?

1. Clinical Trial Data Defined

Clinical trials are research studies on human volunteers. Clinical trials answer questions about different health interventions, such as surgeries, drugs, vaccines, knee replacements, and changes in exercise or diet.

The highest quality clinical trials are randomized and controlled. Human subjects are assigned at random to different "groups" within the trial; one of the groups is a "control group" that receives a standard intervention, a placebo, or no intervention at all. By comparing outcomes in the treatment and control

group, the safety, efficacy, and other properties of the intervention under study can be measured. Randomized controlled trials are the most important means of testing whether a particular intervention is safe and effective—the "gold standard" of evidence-based medicine.[158]

Clinical trials are traditionally categorized into one of four "phases." "Phase 1" trials are the first trials conducted on a new intervention. Small and cautious, they are primarily used to evaluate safety. "Phase 2" trials are larger and longer; they gather more safety information and begin to explore the intervention's efficacy. "Phase 3" trials are still larger; they weigh benefits and harms and examine rare adverse events in a larger population. "Phase 4" trials are done after an intervention is already on the market and in wide use, to study longer-term safety and effectiveness, new uses in new patient populations, and other outstanding questions.[159]

Clinical trials generate lots of data, especially large Phase 3 and Phase 4 trials. One 1999 estimate concluded that a typical Phase 3 clinical trial design with 2,000 patients studied for twelve months could "generate up to 3 million data points."[160] Thousands of clinical trials are conducted every year, making the total quantity of trial data enormous.

There are numerous components of clinical trial data, each with its own properties, utility, and sensitivities. Before proceeding further, we provide a brief taxonomy of clinical trial data. Clinical trial data contains three distinct components: (1) individual patient-level data (IPD), (2) summary data, and (3) metadata. [161] Together, these three components constitute the body of information collectively referred to as clinical trial data.

158. *Clinical Research: Benefits, Risks, and Safety*, NAT'L INST. ON AGING, NAT'L INST. OF HEALTH, https://www.nia.nih.gov/health/clinical-trials-and-studies/clinical-research-benefits-risks-and-safety.

159. *Step 3: Clinical Research, Clinical Research Phase Studies*, FDA, https://www.fda.gov/patients/drug-development-process/step-3-clinical-research#Clinical_Research_Phase_Studies [https://perma.cc/5V65-MRQG].

160. ROUNDTABLE ON RESEARCH AND DEVELOPMENT OF DRUGS, BIOLOGICS, AND MEDICAL DEVICES, INSTITUTE OF MEDICINE, ASSURING DATA QUALITY AND VALIDITY IN CLINICAL TRIALS FOR REGULATORY DECISION MAKING: WORKSHOP REPORT 45 (Jonathan R. Davis et al. eds., 1999).

161. COMMITTEE ON STRATEGIES FOR RESPONSIBLE SHARING OF CLINICAL TRIAL DATA, INSTITUTE OF MEDICINE, SHARING CLINICAL TRIAL DATA: MAXIMIZING BENEFITS, MINIMIZING RISK 7 (2015).

a) Individual Patient-Level Data (IPD)

The first and perhaps most obvious component of clinical trial data is individual patient-level data (IPD).[162] IPD is the "raw" data collected on individual patients. Among other things, it reveals the precise health statuses of different patients—the testing, care, and diagnoses they receive; the side effects and other "adverse events" they experience; and so on.[163] Expert users of data, such as academics and the FDA's regulatory scientists, may be most interested in IPD, but other users, such as journalists and patient groups, may find it difficult to use and understand.

IPD is the most sensitive data component from a patient privacy perspective. It is the rich, detailed personally identifying information (PII) of the clinical trial world, as it links specific health status information with specific individuals.[164] IPD can be de-identified by redacting obvious identifiers such as name, birth year, and zip code,[165] but it remains IPD after de-identification as it continues to characterize the health status of individual people rather than larger groups. Thus, even after de-identification, IPD remains at risk of re-identification and subsequent effects on individual patients.

b) Summary Data

The second data component of clinical trial data is summary data, also known as aggregate data. As the name suggests, this data does not reveal the health status of individual people but instead reveals something about *groups* of people—e.g., the treatment and control arms of a clinical trial, or demographic subgroups of patients in the trial (such as patients over age sixty-five). Some summary data includes explanations and simple "takeaways" digestible to non-expert readers, such as high-level conclusions about a drug's safety and efficacy (or lack thereof) in a group of people.

---

162. In the context of clinical trials, the phrase individual *participant* data is used synonymously with individual patient-level data.

163. For richer description of the "structure" of IPD in clinical trials, see Deborah A. Zarin & Tony Tse, *Sharing Individual Participant Data (IPD) within the Context of the Trial Reporting System (TRS)*, 13 PLOS MED 1, e1001946 (2016).

164. Patients in trials are typically assigned a code number or other anonymous identifier and are not identified by name. But detailed demographic data such as age, gender, weight, height, race, and zip code is often included in "anonymous" IPD, making "anonymized" IPD identifiable. Katherine Tucker, Janice Branson, Maria Dilleen, Sally Hollis, Paul Loughlin, Mark J. Nixon & Zoë Williams, *Protecting patient privacy when sharing patient-level data from clinical trials*, 16 BMC MED. RSCH. METHODOLOGY 77 (2016). IPD is a form of protected health information (PHI); PHI is the term of art used in the HIPAA Privacy Rule. 45 C.F.R. § 164.514 (2021).

165. *See, e.g.*, 45 C.F.R. §§ 160.103, 164.514 (defining "identifiable health information" and "protected health information").

Summary data may span multiple trials. The FDA, for example, synthesizes IPD from multiple trials to produce summary data useful to patients and doctors.[166]

The term "summary" clinical trial data suggests brevity, but some important summary data runs long. Standard summary clinical study reports (CSRs) can run many thousands of pages and provide expert readers with a wealth of information.[167]

c) Metadata

The third component of clinical trial data is metadata. Metadata is data about the other data components. It describes how, exactly, IPD and/or summary data is generated, recorded, analyzed, and presented. Analysis of metadata alongside IPD and summary data can confirm that IPD and summary data are trustworthy—and reveal and discourage manipulation and mistakes.[168]

In the context of clinical trials, the term "metadata" commonly refers to specific standardized documents and data elements: the clinical trial protocol, the statistical analysis plan (SAP), and any analytic code used in connection with the SAP. Together, these resources provide a trial's precise methodology: what questions the trial was intended to answer; what patients were included in and excluded from the trial; what patient "outcomes" it measured (such as tumor size or cholesterol levels); how those measurements were taken and processed; and more.

2. The Value of Clinical Trial Data and Clinical Trial Data Sharing

Clinical trial data is terrifically valuable and expensive to generate. Even a simple trial costs millions of dollars to run; larger, longer Phase 3 trials typically cost tens of millions of dollars.[169] The costs are worth incurring because the

166. For example, the FDA publishes simple "Medication Guides" that explain to patients how to make safe use of certain relatively risky drugs. 21 C.F.R. § 208.24 (2022). The FDA also publishes more detailed "approval" packages, described *infra* Section III.C.1, that likewise synthesize findings from multiple trials.

167. Joshua M. Sharfstein James Dabney Miller, Anna L. Davis, Joseph S. Ross, Margaret E. McCarthy, Brian Smith, Anam Chaudhry, G. Caleb Alexander & Aaron S. Kesselheim, *Blueprint for Transparency at the U.S. Food and Drug Administration: Recommendations to Advance the Development of Safe and Effective Medical Products*, 45 J.L. MED. & ETHICS 7, 17 (2017).

168. *See* John P.A. Ioannidis, Arthur L Caplan & Rafael Dal-Ré, *Outcome Reporting Bias in Clinical Trials: Why Monitoring Matters*, 2017 BMJ 356 (2017) (describing value of comparing published trial results against trial protocols).

169. Linda Martin, Melissa Hutchens, Conrad Hawkins & Alaina Radnov, *How Much Do Clinical Trials Cost?*, 16 NATURE REVS. DRUG DISCOVERY 381 (2017); Thomas J. Moore, James Heyward, Gerard Anderson & G. Caleb Alexander, *Variation in the Estimated Costs of Pivotal*

data generated is scientifically and commercially valuable. Drug, vaccine, device, and other for-profit companies around the world spend tens of billions of dollars on clinical trials to guide their research, to support marketing efforts, and to generate sufficient data to earn approval from the FDA and other regulators around the world.[170]

As with social media data, the stakeholders in clinical trial data are numerous. Key stakeholders include patients themselves; doctors, nurses, and other providers whose care is shaped by trial results; hospitals, clinics, and other organizations that employ the providers (and are liable for many of their actions); innovative companies that develop and sell new drugs, devices, and vaccines; generic and biosimilar companies that seek to sell similar products at lower prices; scientific researchers in academia, government, and nonprofit nongovernmental organizations who do basic research; government regulators who conduct, referee, and pay for research; journalists, academics, and civil society researchers who watchdog those regulators and the healthcare system as a whole; and the public at large, who pay for the regulators and pay a fortune for healthcare.[171]

All these stakeholders are important, but for purposes of this Article, we focus on researchers and research uses of clinical trial data that provide benefits to the broader public. In 2015, a landmark report from the Institute of Medicine (now known as the National Academy of Medicine) characterized the benefits of IPD sharing as follows:[172]

Clinical Benefit Trials Supporting the US Approval of New Therapeutic Agents, 2015–2017: A Cross-Sectional Study, 10 BMJ OPEN 1 (2020). Some industry-funded estimates based on industry-provided data put the average cost of a Phase 3 trial over $200 million. *See, e.g.*, CONG. BUDGET OFF., RESEARCH AND DEVELOPMENT IN THE PHARMACEUTICAL INDUSTRY (2021), https://www.cbo.gov/publication/57126#footnote-069-backlink (citing Joseph A. DiMasi Henry G. Grabowski & Ronald W. Hansenal, *Innovation in the Pharmaceutical Industry: New Estimates of R&D Costs*, 47 J. HEALTH ECON. 20, 24–25).

170. GRAND VIEW RESEARCH, CLINICAL TRIALS MARKET SIZE, SHARE & TRENDS ANALYSIS REPORT BY PHASE (PHASE I, PHASE II, PHASE III, PHASE IV), BY STUDY DESIGN, BY INDICATION (PAIN MANAGEMENT, ONCOLOGY, CNS CONDITION, DIABETES, OBESITY), BY REGION, AND SEGMENT FORECASTS, 2023–2030, https://www.grandviewresearch.com/industry-analysis/global-clinical-trials-market.

171. *National Health Expenditure Data: Historical*, CTRS. FOR MEDICARE & MEDIAID SERVS., https://www.cms.gov/research-statistics-data-and-systems/statistics-trends-and-reports/nationalhealthexpenddata/nationalhealthaccountshistorical (last modified Sept. 6, 2023) (providing statistics on U.S. health spending, and reporting that U.S. healthcare spending reached $4.3 trillion in 2021).

172. INSTITUTE OF MEDICINE, *supra* note 161, at 32 (citations omitted); *see also* NATIONAL ACADEMIES OF SCIENCES., ENG'RS & MED., REFLECTIONS ON SHARING CLINICAL TRIAL DATA: CHALLENGES AND A WAY FORWARD (2020).

> From the perspective of society as a whole, sharing of data from clinical trials could provide a more comprehensive picture of the benefits and risks of an intervention and allow health care professionals and patients to make more informed decisions about clinical care. Moreover, sharing clinical trial data could potentially lead to enhanced efficiency and safety of the clinical research process by, for example, reducing unnecessary duplication of effort and the costs of future studies, reducing exposure of participants in future trials to avoidable harms identified through the data sharing, and providing a deeper knowledge base for regulatory decisions.
>
> In the long run, sharing clinical trial data could potentially improve public health and patient outcomes, reduce the incidence of adverse effects from therapies, and decrease expenditures for medical interventions that are ineffective or less effective than alternatives. In addition, data sharing could open up opportunities for exploratory research that might lead to new hypotheses about the mechanisms of disease, more effective therapies, or alternative uses of existing or abandoned therapies that could then be tested in additional research.

In the following Sections, we show in more detail how independent researchers have used access to IPD and other clinical trial data to interrogate manufacturers' claims about their products and help protect the public from unsafe, ineffective, or exaggerated products (think Ad Observatory, but for drugs). For now, one vivid example of the value of clinical trial data sharing: the antidepressant paroxetine ("Paxil").

Paroxetine was never approved for use in children but became popular with providers, who wrote over two million prescriptions for children per year in the early 2000s on the basis of a 2001 medical journal article. The drug's manufacturer, GlaxoSmithKline, funded and disseminated the article,[173] which claimed that the medicine was "generally well tolerated and effective" in young patients.[174] In fact, paroxetine caused suicidal thinking and suicide in many children.[175] In 2003 and 2004, after widespread anecdotal reports of teen suicides caused by paroxetine, FDA scientists reanalyzed earlier-submitted clinical trial data and concluded that the drug causes increased risk of suicide and suicidal ideation.[176] This led to stricter prescribing rules and a wave of

173. Joanna Le Noury, John M. Nardo, David Healy, Jon Jureidini, Melissa Raven, Catalin Tufanaru & Elia Abi-Jaoude, *Restoring Study 329: Efficacy and Harms of Paroxetine and Imipramine in Treatment of Major Depression in Adolescence,* 351 BMJ 1 (2015).

174. *Id.*

175. *Id.*

176. Tarek A. Hammad, *Review and Evaluation of Clinical Data*, U.S. FOOD & DRUG ADMIN. 45 (Aug. 16, 2004), https://web.archive.org/web/20080625161255/https://www.fda.gov/

litigation against GlaxoSmithKline. GlaxoSmithKline ultimately pled guilty to fraud,[177] and paroxetine is no longer widely prescribed to children.

In the 2010s, independent academic researchers eventually convinced GlaxoSmithKline to share more comprehensive data from the trial described in the 2001 article. They found that the trial data had shown the risks all along and that GlaxoSmithKline had misrepresented the data.[178] The researchers concluded that the affair "illustrates the necessity of making primary trial data and protocols available to increase the rigor of the evidence base."[179] Had GlaxoSmithKline's data been shared with independent researchers in 2001, they might have raised the alarm then, and years of harm might have been averted.

Independent research conducted with clinical trial data is not limited to investigation of questions of safety and efficacy, vital as those questions obviously are. Independent research also helps private and public payers allocate resources better. For example, the nonprofit organization Institute for Clinical and Economic Review (ICER) uses trial data and other medical data to undertake detailed analyses of the cost-effectiveness of various medical interventions, including everything from comparison of all FDA-approved multiple sclerosis drugs[180] to service dogs as treatment for post-traumatic stress disorder.[181] Meta-analysis of pooled clinical trial data established that the blockbuster influenza drug oseltamivir ("Tamiflu") is only modestly effective and that massive stockpiling was a poor use of billions of dollars of public money.[182]

OHRMS/DOCKETS/ac/04/briefing/2004-4065b1-10-TAB08-Hammads-Review.pdf ("No individual trial showed a statistically significant signal for suicidality. However, many had a RR of 2 or more and some of the overall estimates, across various trial groupings, were statistically significant.").

177. Katie Thomas & Michael S. Schmidt, *Glaxo Agrees to Pay $3 Billion in Fraud Settlement*, N.Y. TIMES (July 2, 2012), https://www.nytimes.com/2012/07/03/business/glaxosmithkline-agrees-to-pay-3-billion-in-fraud-settlement.html.

178. *See* Joanna Le Noury, John M. Nardo, David Healy, Jon Jureidini, Melissa Raven, Catalin Tufanaru & Elia Abi-Jaoude., *Restoring Study 329: Efficacy and Harms of Paroxetine and Imipramine* at 2.*in Treatment of Major Depression in Adolescence*, 2015 BMJ 351; *see also* Deborah A. Zarin & Tony Tse, *Sharing Individual Participant Data (IPD) within the Context of the Trial Reporting System (TRS)*, 13 PLOS MED e1001946, 4–5 (2016).

179. Noury et al., *supra* note 178, at 1.

180. *Multiple Sclerosis: CIS, RRMS, and SPMS*, INST. FOR CLINICAL & ECON. REV., https://icer.org/assessment/multiple-sclerosis-2023/ (last updated Feb. 21, 2023).

181. *PTSD: Service Dogs*, INST. FOR CLINICAL & ECON. REV., https://icer.org/assessment/ptsd-service-dogs-2021/ (last updated Jan. 31, 2022).

182. Peter Doshi, Tom Jefferson & Chris Del Mar, *The Imperative to Share Clinical Study Reports: Recommendations from the Tamiflu Experience*, 9 PLOS MED 1 (2012).

3. The Dangers of Clinical Trial Data Sharing

Of course, sharing clinical trial data with researchers has risks, too. There are legitimate and strong countervailing interests that often militate against sharing. The two predominant interests here are patients' privacy (especially as to IPD) and innovative companies' competitive interests.[183] The latter are often articulated in terms of "incentives to innovate" and "protection from free-riders," or framed in terms of specific intellectual property doctrines, such as trade secrecy.

Others' work has thoroughly analyzed both these important interests, in the context of clinical trial data, in the context of healthcare more broadly, and in the context of valuable data writ large.[184] In the Sections that follow, we will show specific instances of such arguments being raised by the pharmaceutical

183. *See, e.g.*, Morten & Kapczynski, *The Big Data Regulator, Rebooted*, *supra* note 134, at 531; *FDA Commissioner Scott Gottlieb, M.D., On New Steps FDA is Taking to Enhance Transparency of Clinical Trial Information to Support Innovation and Scientific Inquiry Related to New Drugs*, U.S. FOOD & DRUG ADMIN. (Jan. 16, 2018) (identifying (1) patient privacy and (2) trade secrecy and the related concept of "confidential commercial information" as justifications for caution in sharing clinical trial data); Memorandum in Support of Pfizer Motion to Intervene at 3, Pub. Health & Med. Pro. for Transparency v. Food and Drug Admin., No. 4:21-CV-01058-P (N.D. TX Jan. 21, 2022) (expressing Pfizer's view that its clinical trial data and related data on its COVID-19 vaccine contain "personal privacy information of individuals who participated in clinical trials and confidential business and trade secret information of Pfizer"); Letter to Jerry Moore, *supra* note 25; 79 Fed. Reg. 69,566 (Nov. 21, 2014); Docket No. NIH-0003 (Mar. 23, 2015) at 2 (arguing that benefits of clinical trial data sharing "must be pursued in a manner that protects other important public health goals such as maintaining patient privacy and protecting incentives for innovative medical research.").

184. For scholarship on privacy and intellectual property (IP) as pro-secrecy interests in clinical trial data specifically, see Erika Lietzan, *A New Framework for Assessing Clinical Data Transparency Initiatives*, 18 MARQ. INTELL. PROP. L. REV. 33 (2014); INSTITUTE OF MEDICINE, *supra* note 161; Morten & Kapczynski, *supra* note 134. For scholarship on privacy and IP interests in broader medical data, see Sharona Hoffman, *Citizen Science: The Law and Ethics of Public Access to Medical Big Data*, 30 BERKELEY TECH. L.J. 1741 (2015); Timo Minssen & Justin Pierce, *Big Data and Intellectual Property Rights in the Health and Life Sciences*, *in* BIG DATA, HEALTH LAW, AND BIOETHICS 307 (I. Glenn Cohen et al. eds., 2018); Elizabeth Rowe, *Sharing Data*, 104 IOWA L. REV. 287 (2018). Much has been written on privacy and IP interests in personal data, and on methods of generating and using that data. *See, e.g.*, HELEN NISSENBAUM, PRIVACY IN CONTEXT: TECHNOLOGY, POLICY, AND THE INTEGRITY OF SOCIAL LIFE (2010); COHEN, *supra* note 5. Governmental bodies like the FDA and the European Medicines Agency have also sometimes taken the position that some clinical trial data—metadata especially—may constitute valid, protected trade secrets. *See, e.g.*, *External Guidance on the Implementation of the European Medicines Agency Policy on the Publication of Clinical Data for Medicinal Products for Human Use*, EUR. MEDS. AGENCY 49–52 (Oct. 15, 2018), https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/external-guidance-implementation-european-medicines-agency-policy-publication-clinical-data_en-3.pdf.

or medical device industry and then accommodated or rebutted by the legislators and governors of the clinical trial data-sharing mandate.

Note here that the parallels with social media data are strong. Just as platform companies have invoked patient privacy and innovation to limit sharing their data with researchers, so too have large, incumbent companies that hold and profit from clinical trial data.[185] For example, in 2015, shortly after the National Institutes of Health (NIH) proposed a new rule mandating expanded sharing of certain summary and metadata from clinical trials with researchers and the broader public,[186] the Pharmaceutical Research and Manufacturers of America (PhRMA) association warned, ominously, that "the rule does not adequately protect the process of medical research innovation. Failure to protect adequately trade secrets and confidential commercial information would harm public health by discouraging the very innovation necessary to bring new medical advances to the market."[187] NIH responded that PhRMA's concerns were overblown and that NIH's rule struck an appropriate balance.[188] Since NIH's rule went into effect in 2017, NIH has proven correct—as the next two Sections show.

B. "DARK AGES" OF CLINICAL TRIAL SECRECY: LITTLE RESEARCHER ACCESS, UNREALIZED BENEFITS, AND HARM TO PATIENTS

This Section explains how today's clinical trial data sharing mandate emerged out of comparative "dark ages" of data secrecy, contestation, and unnecessary human suffering.

Consider the United States in 1960. There was then no explicit law governing researcher access to clinical trial data and other kinds of medical research data. In addition, more rudimentary information technology meant that data was more difficult to share and use.

Because no law mandated researcher access, drug companies, medical device manufacturers, universities, and other entities that conducted clinical trials were free to disseminate or withhold data as they saw fit. They massaged data, such as by publishing selective data in medical journals that painted their

185. For a broad, independent view of the pharmaceutical industry's claims of trade secrecy in clinical trial data, see W. Nicholson Price II & Timo Minssen, *Will Clinical Trial Data Disclosure Reduce Incentives to Develop New Uses of Drugs?*, 33 NATURE BIOTECHNOLOGY 685 (2015).

186. *See infra* Sections III.B & III.C.1. NIH did not propose, and has not proposed, mandatory sharing of IPD from the same broad swath of clinical trials.

187. Letter to Jerry Moore, *supra* note 25, at 2.

188. *See, e.g.*, 81 Fed. Reg. 64982, 64968, 64995.

products in the best possible light.[189] The medical literature was thus incomplete and manipulated.

In fact, as of 1960, drug companies sometimes withheld clinical trial data not just from researchers but from the FDA itself. An infamous example: In 1960 and 1961, one FDA scientist, Frances Kelsey, grew concerned over a lack of safety data to the FDA on the drug thalidomide, even as the drug had been approved and entered widespread use in Europe and Australia. Kelsey came to suspect that the drug's manufacturer, the William S. Merrell Company, was withholding safety data from the FDA[190] and requested this missing data from the company.[191] Kelsey's insistence on receiving the data parallels the FTC's recent insistence that social media companies turn over certain data pursuant to a past consent decree.[192] Kelsey's lengthy review of thalidomide prevented widespread use in the United States. By late 1961, reports of thousands of horrifying birth defects and fetal deaths caused by the drug in other countries led to its withdrawal from pharmacies worldwide. Kelsey was justifiably hailed as a hero for protecting Americans from its harms.

The thalidomide catastrophe, and a broader "full disclosure movement" that coalesced in drug regulation in the wake of other, smaller drug scandals,[193] prompted Congress to enact the first important federal clinical trial data sharing legislation: the 1962 Kefauver-Harris Amendments to the Food, Drug & Cosmetics Act. This legislation mandated, for the first time, that drug companies submit clinical trial data to the FDA as a condition of market approval, and it gave the FDA legal authority to dictate exactly how that data was packaged and presented to the agency. If companies didn't comply, the

189. *See* DANIEL CARPENTER, REPUTATION AND POWER (2010); PATRICK RADDEN KEEFE, EMPIRE OF PAIN: THE SECRET HISTORY OF THE SACKLER DYNASTY (2021); *see also* MILTON M. SILVERMAN & PHILIP R. LEE, PILLS PROFITS, AND POLITICS 105 (1974) (explaining that the new FDA Commissioner took office in 1966 and criticized the practice of "conscious withholding of unfavorable animal or clinical data" from the FDA).

190. These were case reports on peripheral neuropathy held by the company. CARPENTER, *supra* note 189, at 221.

191. *See* Stephen Phillips, *How a Courageous Physician-Scientist Saved the U.S. From Birth-Defects Catastrophe*, UCHICAGO MED. (Mar. 9, 2020), https://www.uchicagomedicine.org/forefront/biological-sciences-articles/courageous-physician-scientist-saved-the-us-from-a-birth-defects-catastrophe; CARPENTER, *supra* note 189, at 221.

192. *E.g.*, Agreement Containing Consent Order, *In re* Facebook, Inc., File No. 092 3184, Part IV (Fed. Trade Comm. Dec. 5, 2021), https://www.ftc.gov/sites/default/files/documents/cases/2011/11/111129facebookagree.pdf.

193. CARPENTER, *supra* note 189, at 237.

FDA could keep products off the U.S. market. The FDA became the world's largest reservoir of clinical trial data, which it remains today.[194]

But the Kefauver-Harris Amendments did not guarantee researcher or public access to that data. The "full disclosure" in "full disclosure movement" meant full disclosure *to the FDA*, not to independent researchers. Against a statutory blank canvas, the FDA had no legal obligation to disclose any of the clinical trial data in its possession to the broader public.[195] Through the 1960s, the FDA's choice was to keep most of this data confidential; its expert reviewers worked mostly in secret. Independent researchers outside the FDA typically learned the results of clinical trials from the medical literature, where industry continued to cherry-pick the data it wanted to share.

At least as early as 1969, some FDA officials expressed a desire to change this state of affairs and make all clinical trial data held by the agency public once the product in question had been approved for sale.[196] Tentative, inconsistent efforts to do so through discretionary agency action proved unsuccessful, in part because they were undone by a rotating cast of more industry-friendly, pro-secrecy FDA commissioners, and in part because for years, the FDA was threatened with legal challenge by the powerful pharmaceutical industry.[197]

---

194. U.S. FOOD & DRUG ADMIN., DRIVING BIOMEDICAL INNOVATION: INITIATIVES TO IMPROVE PRODUCTS FOR PATIENTS 22 (2011), https://www.ipqpubs.com/wp-content/uploads/2012/02/FDA-Driving-Biomedical-Innovation.pdf.

195. The Freedom of Information Act (FOIA), first enacted in 1966, might seem to offer researchers a vehicle to demand and obtain clinical trial data held by the FDA, as, on its face, FOIA empowers any member of the public to demand most documents held by almost every federal agency, including the FDA. *See* Margaret B. Kwoka, *FOIA, Inc.*, 65 DUKE L.J. 1361 (2016); Morten & Kapczynski, *supra* note 134. Yet in practice, FOIA has proved to be of modest utility. In 1974, under a secrecy-friendly, Nixon-appointed Commissioner, the FDA first promulgated regulations promising to keep essentially all industry-submitted clinical trial data secret from FOIA requesters, and these regulations remain on the books today. 21 C.F.R. § 4.61, promulgated in 1974, since recodified to 21 C.F.R. § 20.61; FDA INFORMATION DISCLOSURE MANUAL (1999), https://www.governmentattic.org/6docs/FDA-InfoDisc Manual_1999.pdf. For deeper analysis, see Rebecca S. Eisenberg, *The Role of the FDA in Innovation Policy*, 13 MICH. TELECOMM. & TECH. L. REV. 345, 381 (2007); Lietzan, *supra* note 184, at 51–53. Some skilled and determined researchers have succeeded in using FOIA to obtain clinical trial data, but only with great effort. *See* CARPENTER, *supra* note 189, at 381; Charles Seife, *FDA Documents Reveal Depths of Internal Rancor Over Drug's Approval Process*, UNDARK (Aug. 2, 2017), https://undark.org/2017/08/02/fda-eteplirsen-janet-woodcock/.

196. *See* Silverman & Lee, *supra* note 189, at 241 (recounting that the then-FDA Commissioner, appointed in 1969, "urged . . . that the results of all animal and human trials and similar clinical data should be made public").

197. Robert M. Halperin, *FDA Disclosure of Safety and Effectiveness Data: A Legal and Policy Analysis*, 1979 DUKE L.J. 286, 294 (1979); Thomas O. McGarity & Sidney A. Shapiro, *The*

From the 1970s to the 1990s, there remained no coherent statutory regime guaranteeing researcher access to clinical trial data, even as researchers clamored for access. In 1978, a bill that would have mandated disclosure of summary data, metadata, and IPD, called the Drug Regulation Reform Act (DRRA), was defeated in Congress.[198] The bill failed to pass despite support from the Center for Law and Social Policy, the Environmental Defense Fund, and Public Citizen.[199] In 1980, McGarity and Shapiro published an article in the *Harvard Law Review* criticizing the FDA's then-skimpy disclosure of industry-generated clinical trial data in the agency's possession;[200] this practice contrasted with the EPA's much richer data disclosure of testing data on pesticides[201] and the FDA's own richer data disclosure on food additives.[202] The 1984 Hatch-Waxman Act was, in early drafts of the legislation, to have included a DRRA-like provision that would have required the FDA to publish volumes of clinical trial data when product applications were approved or denied.[203] The pharmaceutical industry's lobby watered down the statutory language, arguing that mandatory disclosure would undermine patient privacy and its trade secrecy interests.[204] At the same time, the FDA Commissioner testified in Congress on the alleged benefits of data secrecy and urged construction of the watered-down statutory language in ways that perpetuated the secretive status quo.[205] In the late 1990s, the FDA began voluntary,

---

Trade Secret Status of Health and Safety Testing Information: Reforming Agency Disclosure Policies, 93 HARV. L. REV. 837, 837–38 (1980).

198. William W. Vodra, *The Drug Regulation Reform Act of 1978: Putting Some Economic Issues into Different Contexts*, 1 MANAGERIAL & DECISION ECON. 184 (1980).

199. Drug Regulation Reform Act of 1978: Hearings on S. 27755 Before the Subcomm. on Health & Sci. Rsch. of the Comm. On Hum. Res., 95th Cong. 2 (1978) at 625–26 (CLSP), 645–46 (EDF), 668 (Public Citizen).

200. McGarity & Shapiro, *supra* note 197, at 867. Around the same time, as Matthew Herder has documented, there were also some brief but important flashes of discretionary data transparency from the FDA, as when the agency disclosed secret data on the safety and efficacy of sulfinpyrazone (Anturane) to combat misleading messaging by the drug's manufacturer. *See* Matthew Herder, *Reviving the FDA's Authority to Publicly Explain Why New Drugs Are Approved or Rejected*, 178 JAMA INTERNAL MED. 1013 (2018).

201. *Id.* at 873.

202. *See id.* at 872.

203. *See* James T. O'Reilly, *Knowledge Is Power: Legislative Control of Drug Industry Trade Secrets*, 54 U. CIN. L. REV. 1, 16–17 (1985); Jane A. Fisher, *Disclosure of Safety and Effectiveness Data Under the Drug Price Competition and Patent Term Restoration Act*, 41 FOOD DRUG COSM. L. J. 268 (1986).

204. O'Reilly, *supra* note 203; Fisher, *supra* note 203.

205. Then-FDA Commissioner Frank Young intervened during the negotiation and passage of the Hatch-Waxman Act in 1984 to express the view that the statutory text of Act did and should not expand the agency's obligation to disclose safety and efficacy data, despite statutory language mandating that "[s]afety and efficacy data" "be made available to the public, upon request," under various circumstances. O'Reilly, *supra* note 203, at 20–21; Fisher, *supra*

discretionary disclosure of some summary data and metadata from clinical trials, but shared this data only after product approval for a subset of approved products[206] and on a leisurely timeline.[207]

The pharmaceutical industry largely thwarted researcher access into the 2000s.[208] For example, in 2000, David Willman of *The Los Angeles Times* reported a meticulous, Pulitzer-Prize-winning series of articles[209] on seven drugs that had been withdrawn between 1993 and 2000 for causing death and other serious side effects, revealing weaknesses in the FDA's drug approval process and in the pharmaceutical industry's ethics.[210] Willman remarked on the difficulty of his investigation and the FDA's then-still-prevalent culture of data secrecy. For example, data from one important clinical trial showing deaths in kidney transplant patients taking the immunosuppressive drug tacrolimus ("Prograf") had been disclosed to the FDA but not made readily available to outside researchers; per Willman, "the only way for doctors or patients to find that data is to search the medical literature or seek the FDA's review documents" through FOIA.[211] Similarly, in 2004, Barry Meier of *The New York Times* reported that medical researchers seeking to investigate the safety of antidepressants "could get only pieces of" relevant trial data, as "drug companies refused to turn over data . . . even though these researchers had

note 203, at 283–84 (describing letter from Commissioner Young asserting that the FDA would construe the Act to permit the FDA keep data secret if the data "have commercial value as confidential business information").

206. INSTITUTE OF MEDICINE, THE FUTURE OF DRUG SAFETY: PROMOTING AND PROTECTING THE HEALTH OF THE PUBLIC INSTITUTE OF MEDICINE 142–43 (Alina Baciu et al. eds., 2006); Zarin & Tse, *supra* note 178.

207. *See* Marion F. Gruber, *US FDA Review and Regulation of Preventive Vaccines for Infectious Disease Indications: Impact of the FDA Amendments Act 2007*, 10 EXPERT REV. VACCINES 1011, 1018 (2011) (observing that prior to 2007, "only limited documentation had to be sent forward for redaction and posting immediately upon product approval, with supportive documentation to be provided in the following months"); Marian S. McDonagh, Kim Peterson, Howard Balshem & Mark Helfand, *US Food and Drug Administration Documents Can Provide Unpublished Evidence Relevant to Systematic Reviews*, 66 J. CLINICAL EPIDEMIOLOGY 1071, 1078 (2013); McGarity & Shapiro, *supra* note 197, at 867.

208. *See* Shankar Vedantam, *Antidepressant Makers Withhold Data*, NBC NEWS (Jan. 28, 2004, 8:59 PM), https://www.nbcnews.com/id/wbna4091562 (documenting unmet public demand for clinical trial data as of 2004); Barry Meier, *Contracts Keep Drug Research Out of Reach*, N.Y. TIMES (Nov. 29, 2004), https://www.nytimes.com/2004/11/29/business/contracts-keep-drug-research-out-of-reach.html (documenting industry resistance to legislation mandating clinical trial data sharing).

209. CARPENTER, *supra* note 189, at 735.

210. David Willman, *How a New Policy Led to Seven Deadly Drugs*, L.A. TIMES (Dec. 20, 2000), https://www.latimes.com/nation/la-122001fda-story.html.

211. *Id.*

helped come up with it."[212] Meier added that companies blocked researchers from "shar[ing] their own data with colleagues who had not worked" on a particular trial, siloing researchers from one another.[213] In 2006, two representatives of the prominent nonprofit Public Citizen, Peter Lurie and Allison Zieve, summarized the lamentable state of affairs: "Those committed to the free exchange of scientific information have long complained about various restrictions on access to [the FDA's] pharmaceutical data and the resultant restrictions on open discourse."[214]

During this time, there was some voluntary sharing of data by drug and device manufacturers. As noted above, these companies selectively published data in medical literature. Some companies went further and made databases of certain clinical trial data and other data (e.g., genetic data) available to academic and other researchers. Companies that shared more were praised for "transparency," but this transparency was selective and subject to some of the same "pathologies" of voluntary sharing of social media data identified in Part II—decontextualization and streetlight effects especially. (For example, Merck, a company that received praise in the 1990s for voluntary sharing of some kinds of data,[215] was later shown to have hidden other data on the safety of rofecoxib ("Vioxx") that contributed to the deaths of tens of thousands of people.[216]) As Deborah Zarin and Tony Tse stated in 2007, there were twelve "pharmaceutical industry-sponsored clinical trial databases," but they were "generally not reviewed by experts external to the company." An independent investigation "found that when conclusions were listed in these databases, they tended to be more favorable for the company's product than those found in published articles or FDA reviews of the same trials."[217]

Perhaps not coincidentally, the 1990s and 2000s were marked by a series of increasingly high-profile scandals involving drug companies that hid unfavorable clinical trial data from independent researchers and the broader public, leading to widespread harm to patients. Two of the highest-profile scandals involved the drugs paroxetine ("Paxil") and rofecoxib ("Vioxx").

212. Barry Meier, *Contracts Keep Drug Research Out of Reach*, N.Y. TIMES (Nov. 29, 2004), https://www.nytimes.com/2004/11/29/business/contracts-keep-drug-research-out-of-reach.html.

213. *Id.*

214. Peter Lurie & Allison Zieve, *Sometimes the Silence Can Be Like the Thunder: Access to Pharmaceutical Data at the FDA*, 69 L. & CONTEMP. PROBS. 85 (2006).

215. Eliot Marshall, *Ethics in Science: Is Data-Hoarding Slowing the Assault on Pathogens?*, 275 SCI. 777 (1997); Eliot Marshall, *HGS Opens its Databanks-for a Price*, 266 SCI. 25 (1994).

216. *Infra* notes 219–224 and surrounding text.

217. Zarin & Tse, *supra* note 178, at 2115–18.

The basic details of the paroxetine scandal are summarized above. GlaxoSmithKline gathered evidence that its drug fueled tens of thousands of teen suicides, then intentionally hid that evidence from the public.[218] The paroxetine scandal gripped the public consciousness and helped spur Congress to action. When the FDA decided to warn doctors and parents to stop giving paroxetine to children in 2003, the story made headline news.[219] Media not only covered paroxetine's contributions to a spike in teen suicides but also big pharma's culture of data secrecy. A 2004 *New York Times Magazine* story observed "public outrage at revelations that a number of pharmaceutical companies had deliberately withheld damning information about [antidepressants including Paxil]—specifically, data from clinical trials that suggested that these drugs were both more dangerous and less effective for adolescents than millions of consumers had been led to believe."[220]

The rofecoxib ("Vioxx") scandal was perhaps even more shocking. Rofecoxib, a painkiller, was approved by the FDA in 1999 and quickly became a blockbuster, earning Merck billions of dollars.[221] Then, in 2004, Merck abruptly removed the drug from the market, with encouragement from the FDA and other drug regulators because—Merck admitted—it caused heart attacks, strokes, and heart failures.[222] Merck held, internally, clinical trial data establishing these deadly side effects but did not disclose the data to independent researchers or the broader public. Merck moved to withdraw the drug only because a courageous FDA scientist with access to the data, David Graham, double-checked the agency's analysis and raised concerns, first with the agency and then with the U.S. Senate and the broader public.[223] The relevant trial data was first made available to independent researchers only

218. *See supra* Section II.A.2; *see also* Benedict Carey, *Antidepressant Paxil Is Unsafe for Teenagers, New Analysis Says*, N.Y. TIMES (Sept. 16, 2015), https://www.nytimes.com/2015/09/17/health/antidepressant-paxil-is-unsafe-for-teenagers-new-analysis-says.html?ref=health&_r=0; David Dobbs, *The Human Cost of a Misleading Drug-Safety Study*, ATLANTIC (Sept. 18, 2015), https://www.theatlantic.com/health/archive/2015/09/paxil-safety-bmj-depression-suicide/406105/.

219. *See, e.g.*, Associated Press, *Paxil Not for Kids, FDA Warns*, L.A. TIMES (June 20, 2003), https://www.latimes.com/archives/la-xpm-2003-jun-20-na-paxil20-story.html.

220. Jonathan Mahler, *The Antidepressant Dilemma*, N.Y. TIMES MAG. (Nov. 21, 2004), https://www.nytimes.com/2004/11/21/magazine/the-antidepressant-dilemma.html

221. Harlan M. Krumholz, *What Have We Learnt From Vioxx?*, 334 BRIT. MED. J. 120, 122 (2007).

222. Joseph S. Ross, David Madigan, Kevin P. Hill, David S. Egilman, Yongfei Wang & Harlan M. Krumholz., *Pooled Analysis of Rofecoxib Placebo-Controlled Clinical Trial Data: Lessons for Postmarket Pharmaceutical Safety Surveillance*, 169 ARCHIVES INTERNAL MED. 1976, 1976–77 (2009).

223. Matthew Herper, *Face of the Year: David Graham*, FORBES (Dec. 13, 2004), https://www.forbes.com/2004/12/13/cx_mh_1213faceoftheyear.html?sh=ed7b36d6d576.

years later, through litigation.[224] These researchers quickly proved that signals of these risks were present in data held by Merck and the FDA nearly *3.5 years* before the drug was withdrawn from the market.[225] Had independent researchers gotten access to the data sooner, they could have caught the problem and averted at least 39,000 deaths.[226] Prominent scientists pointed to Vioxx as evidence that clinical trial data should be "stored on an academic site, analysed by non-company investigators, and eventually made accessible to the public for scrutiny."[227] *The New York Times* covered the Vioxx scandal at length, publishing stories on the FDA's promises of greater clinical trial data sharing[228] and the pharmaceutical industry's unreliable commitments to transparency.[229]

In short, Vioxx and Paxil were "Cambridge Analytica moments" for the pharmaceutical industry. GlaxoSmithKline's efforts to downplay safety problems with a different drug, rosiglitazone ("Avandia"), constituted a third such moment, prompting more Congressional hearings and calls for reform.[230] Pharmacia's manipulation of data on another blockbuster painkiller drug, celecoxib ("Celebrex"), arguably created yet a fourth.[231] Clinical trial data

224. *See* Aaron S. Kesselheim & Jerry Avorn, *The Role of Litigation in Defining Drug Risks*, 297 JAMA 308, 309 (2007).

225. Ross et al., *supra* note 222, at 1979.

226. David J. Graham, David Campen, Rita Hui, Michele Spence, Craig Cheetham, Gerald Levy, Stanford Shoor & Wayne A. Ray, *Risk of Acute Myocardial Infarction and Sudden Cardiac Death in Patients Treated with Cyclo-Oxygenase 2 Selective and Non-Selective Non-Steroidal Anti-Inflammatory Drugs: Nested Case-Control Study*, 365 LANCET 475, 480 (2005); *see also* Carolyn Abraham, *Vioxx Took Deadly Toll: Study*, GLOBE & MAIL (Jan. 25, 2005), https://www.theglobeandmail.com/life/vioxx-took-deadly-toll-study/article1113848.

227. Krumholz, *supra* note 221, at 334.

228. Gardiner Harris, *F.D.A. Moves Toward More Openness with the Public*, N.Y. TIMES (Feb. 20, 2005), https://www.nytimes.com/2005/02/20/us/health/fda-moves-toward-more-openness-with-the-public.html.

229. Alex Berenson, *Despite Vow, Drug Makers Still Withhold Data*, N.Y. TIMES (May 31, 2005), https://www.nytimes.com/2005/05/31/business/despite-vow-drug-makers-still-withhold-data.html.

230. Joanne Silberner, *FDA Criticized for Diabetes Drug Avandia*, NPR (May 22, 2007), https://www.npr.org/2007/05/22/10318764/fda-criticized-for-diabetes-drug-avandia; *Senators Reveal Efforts by the FDA to Suppress Scientific Dissent and Downplay Safety Concerns*, U.S. SENATE COMM. ON FIN. (July 24, 2007), https://www.finance.senate.gov/chairmans-news/senators-reveal-effort-by-the-fda-to-suppress-scientific-dissent-and-downplay-safety-concerns.

231. Pharmacia published a misleading study in the medical literature suggesting that celecoxib helps guard against ulcers while withholding from the public more complete data showing the drug actually does not. The scandal was reported by *The Washington Post*, *The New York Times*, and other major newspapers. *See, e.g.*, Susan Okie *Missing Data on Celebrex*, WASH. POST (Aug. 5, 2001), https://www.washingtonpost.com/archive/politics/2001/08/05/missing-data-on-celebrex/59d3748b-6683-4ca7-8890-d711aad07241/; Melody Petersen, *Study Finding Celebrex Safer Was Flawed, Journal Says*, N.Y. TIMES (June 1, 2002), https://

secrecy had become a matter of national attention. Resulting public outrage[232] prompted Congress to revisit the possibility of legislation mandating data sharing by pharmaceutical and medical device companies and resulted in breakthrough federal legislation that forms the foundation of today's data sharing mandate.

The pharmaceutical and medical device industries fought data-sharing legislation from the start. As Galbraith details,[233]

> Not surprisingly, the pharmaceutical industry's trade group did not support the FACT Act [proposed federal legislation that would mandate sharing of clinical trial data]. Originally, the Pharmaceutical Research and Manufacturers of America (PhRMA) asserted that a results reporting requirement was unnecessary.[202] However, in January of 2005, faced with pressure from lawmakers, the medical community, and the public, the four largest pharmaceutical trade groups in the world, including PhRMA, released a joint statement on the disclosure of clinical trial information.[203] While the group members pledged to release a nominal amount of information regarding ongoing trials, they did not commit to submitting the data to a comprehensive, government-sponsored registry.[204] Instead, the provisions left open the possibility of publishing the information on individual, company-sponsored websites that could contain internal rules that might not be publicly disclosed and consequently may differ from one site to the next . . . . Furthermore, with regard to completed trials, the pharmaceutical manufacturers agreed only to make public "summary results" of the studies and, additionally, asserted such disclosure "must maintain protections for . . . intellectual property and contract rights."

Just as Facebook and other social media platform companies claim today, pharmaceutical companies in the 2000s argued that laws mandating data sharing would compromise their trade secrets and the privacy of individual data subjects.[234] Congress enacted mandate legislation anyway.[235] When NIH then proposed the rule implementing the legislation, the pharmaceutical lobby

---

www.nytimes.com/2002/06/01/us/study-finding-celebrex-safer-was-flawed-journal-says.html.

232. *See, e.g.*, CARPENTER, *supra* note 189, at 588; *Drug Safety: Improvement Needed in FDA's Postmarket Decision-making and Oversight Process*, GAO-06-402, U.S. GOV'T ACCOUNTABILITY OFF. (Mar. 31, 2006), https://www.gao.gov/products/gao-06-402.

233. Christine D. Galbraith, *Dying to Know: A Demand for Genuine Public Access to Clinical Trial Results Data*, 78 MISS. L. J. 705, 738–39 (2009).

234. *Id.* at 752, 764 (citing, *inter alia*, Shankar Vedantam, *Antidepressant Makers Withhold Data on Children*, WASH. POST (Jan. 29, 2004), at A1); Joel Lexchin, *The Secret Things Belong Unto the Lord Our God: Secrecy in the Pharmaceutical Arena*, 26 MED. & L. 417 (2007).

235. *Infra* Section III.C.

again sang the same tune, warning that "the rule does not adequately protect the process of medical research innovation. Failure to protect adequately trade secrets and confidential commercial information would harm public health by discouraging the very innovation necessary to bring new medical advances to the market."[236]

As we describe in the next Section, the pharmaceutical lobby's concerns proved unfounded. NIH and other stewards of sensitive and previously secret clinical trial data have proven capable of collecting it from industry and sharing it with researchers without compromising patient privacy or incentives to innovate. The legislation that the pharmaceutical lobby resisted now forms the cornerstone of today's clinical trial data sharing mandate, pushing the industry out of the dark ages.

C. LEGISLATING TODAY'S CLINICAL TRIAL DATA SHARING MANDATE

The story of today's clinical trial data sharing mandate begins with the legal system: first legislation, and then regulation to implement and extend legislation. As in many other contexts, public law provided a necessary counterweight to private power. Public law mandated that drug and device companies make clinical trial data available to researchers and empowered federal regulators, such as the FDA and the NIH, to enforce compliance and govern that data.

The single most important piece of American law in the clinical trial data sharing mandate is the Food and Drug Administration Amendments Act (FDAAA), enacted in 2007. FDAAA was described by the then-FDA commissioner as "massive legislation" informed by a "spirit of transparency."[237] A key achievement of FDAAA was to mandate universal disclosure of summary and metadata from clinical trials (though not IPD).

FDAAA achieved much broader researcher access to clinical trial data in two ways: (1) mandatory publication by the FDA of "approval packages" that contain clinical trial data (and more); and (2) mandatory submission of clinical trial data to NIH, for validation and posting by NIH on a public website, ClinicalTrials.gov. We discuss each in turn.

1. Mandatory Publication of Approval Packages

FDAAA mandates that every time the FDA approves a new drug or vaccine, the agency must publish an "approval package"[238] providing summary

236. Letter to Jerry Moore, *supra* note 25, at 2.

237. Andrew C. von Eschenbach, *The FDA Amendments Act: Reauthorization of the FDA*, 63 FOOD & DRUGS L.J. 579, 581 (2008)

238. Also known as an "action package."

data and metadata from all the clinical trials on which it relied for approval.[239] The approval package provides a summary of both the drug manufacturer's data and the FDA's independent analysis.[240] The FDA must publish the approval package within thirty days of approval.[241] In effect, this provision of FDAAA obliges the FDA to share some of its vast reservoir of data with the public.

Today the FDA publishes approval packages as a matter of routine practice, on a website it calls "Drugs@FDA."[242] These packages fuel important research.[243] For example, a 2013 review article observed that "FDA documents contain unpublished evidence that can be highly useful in resolving

239. 21 U.S.C. § 355(l). The FDA had done this previously, since at least the late 1990s, but only discretionarily, and more slowly and less consistently. *See* Gruber, *supra* note 207, at 1017–18. Note that this disclosure mandate is limited to new molecular entities and biological products; newly approved products that do not contain any previously unapproved active moiety or active ingredient (such as newly approved reformulations of existing drugs) are exempt from the statute's disclosure mandate. 21 U.S.C. § 355(l)(2)(A).

240. U.S. FOOD & DRUGS ADMIN, *NDAs/BLAs/Efficacy Supplements: Action Packages and Taking Regulatory Actions*, MAPP 6020.8 Rev. 1, 13 (June 13, 2016), https://www.fda.gov/media/72739/download (specifying that action packages consist of "a compilation of (1) FDA-generated documents related to review of an NDA or efficacy supplement (i.e., from submission to final action), (2) documents (e.g., meeting minutes, pharmacology reviews) pertaining to the format and content of the application generated during drug development (investigational new drug [IND]), and (3) labeling submitted by the applicant").

241. *Id.*

242. *Drugs@FDA*, U.S. FOOD & DRUGS ADMIN., https://www.accessdata.fda.gov/scripts/cder/daf/index.cfm (last visited Nov. 12, 2023). Drugs@FDA is not quite comprehensive of all drugs. *See Drugs@FDA Frequently Asked Questions*, U.S. FOOD & DRUG ADMIN., https://www.accessdata.fda.gov/scripts/cder/daf/index.cfm?event=faq.page# contains ("What products are not in Drugs@FDA?"). FDA maintains separate but very similar databases for vaccines and other biological products. *See*, *e.g.*, *Vaccines Licensed for Use in the United States*, U.S. FOOD & DRUG ADMIN., https://www.fda.gov/vaccines-blood-biologics/vaccines/vaccines-licensed-use-united-states; *Approved Cellular and Gene Therapy Products*, U.S. FOOD & DRUG ADMIN., https://www.fda.gov/vaccines-blood-biologics/cellular-gene-therapy-products/approved-cellular-and-gene-therapy-products.

243. Letter from Peter Doshi et al. to the FDA (Aug. 23, 2019), http://freepdfhosting.com/19eabf06a7.pdf (identifying uses to which researchers put approval packages: systematic reviews and meta-analyses of medical products, and improving methods; researching regulatory, publication, and drug approval processes; comparing regulatory review times and outcomes across jurisdictions; developing consumer and professional decision-making tools and case studies of particular drug approval decisions; and evaluating the impact of federal policy); *see also* Erick H. Turner, *How to Access and Process FDA Drug Approval Packages for Use in Research*, 347 BRIT. MED. J. 1 (2013); Aviv Ladanie, Hannah Ewald, Benjamin Kasenda & Lars G. Hemkens, *How to Use FDA Drug Approval Documents for Evidence Synthesis*, 362 BMJ 1, 1 (2018). *But see* Matthew Herder, Christopher J. Morten & Peter Doshi, *Integrated Drug Reviews at the US Food and Drug Administration—Legal Concerns and Knowledge Lost*, 180 JAMA INTERN MED, 629 629–30 (2020) (criticizing recent the FDA's move to less information-rich approval packages).

publication bias and selective outcome and analysis reporting, identifying important harms, and filling gaps in knowledge about understudied subpopulations, outcomes, and comparisons."[244] In effect, approval packages equip independent researchers to overcome structural problems that afflict independent research, including the problem of decontextualized data production (by giving researchers more objective context, including the FDA's own analysis) and the streetlight effect (by giving researchers access to the FDA's data, rather than simply to a cherry-picked subset that drug manufacturers choose to publish in the medical literature).

To show the value of the FDA's approval packages to independent researchers and the broader public, a few concrete examples: In 2014, independent researchers used an approval package to detect and publicize errors in clinical trial data reporting by the drug company Roche on its anti-influenza drug oseltamivir ("Tamiflu"). [245] In the same year, different researchers used an approval package to establish that the anti-inflammatory drug roflumilast ("Daxas") provides net benefits to patients with severe chronic obstructive pulmonary disease (COPD), but not patients with milder disease, reshaping prescribing habits.[246] In similar ways, independent academic and nonprofit researchers have used approval package data in combination with other data (from the medical literature and other sources) to conduct research on the diabetes drug rosiglitazone ("Avandia"), [247] the painkiller valdecoxib ("Bextra"),[248] and cosmetic injections of botulinum toxin (better known under the brand name Botox).[249]

244. McDonagh et al., *supra* note 207, at 1072.

245. Tom Jefferson, Mark Jones, Peter Doshi, Elizabeth Spencer, Igho Onakpoya & Carl J. Heneghan, *Oseltamivir for Influenza in Adults and Children: Systematic Review of Clinical Study Reports and Summary of Regulatory Comments*, 348 BMJ 1, 7 (2014).

246. Tsung Yu, Kevin Fain, Cynthia M. Boyd, Sonal Singh, Carlos O. Weiss, Tianjing Li, Ravi Varadhan & Milo A. Puhan, *Benefits and Harms of Roflumilast in Moderate to Severe COPD*, 69 THORAX 616, 622 (2014).

247. Joshua D Wallach, Kun Wang, Audrey D. Zhang, Deanna Cheng, Holly K. Grossetta Nardini, Haiqun Lin, Michael B. Bracken, Mayur Desai, Harlan M. Krumholz & Joseph S. Ross, *Updating Insights into Rosiglitazone and Cardiovascular Risk Through Shared Data: Individual Patient and Summary Level Meta-Analyses*, 368 BMJ 1 (2020). This paper was corrected in 2021. 373 BMJ n1302 (2021).

248. Sidney Wolfe, *Public Citizen to Call on FDA to Ban Celebrex and Bextra*, PUB. CITIZEN (Dec. 17, 2004), https://www.citizen.org/news/public-citizen-to-call-on-fda-to-ban-celebrex-and-bextra/.

249. *Petition Requesting Regulatory Action Concerning the Spread of Botulinum Toxin (Botox, Myobloc) to Other Parts of the Body*, PUB. CITIZEN (Jan. 23, 2008), https://www.citizen.org/article/petition-requesting-regulatory-action-concerning-the-spread-of-botulinum-toxin-botox-myobloc-to-other-parts-of-the-body/; Peter Lurie, *Statement: FDA Grants Public Citizen*

The FDA's data transparency has benefits for the agency's public credibility, as well. In November 2020, at a moment when the American public's trust in the FDA had been damaged by interference in its COVID-19 vaccine review process from then-President Trump and his political appointees,[250] the agency was able to restore some trust in the agency and in the vaccines themselves by committing to publish complete approval packages even as the agency was short-cutting other steps of the standard vaccine approval process in the emergency setting of a global pandemic. [251] Independent researchers dissected these approval packages once published and, by and large, confirmed COVID vaccines' safety and efficacy, and the wisdom of the FDA's decision to hurry them into patients' arms.[252]

2. Mandatory Submission and Publication of Clinical Trial Data to ClinicalTrials.gov

A separate provision of FDAAA mandates that an even broader set of summary data and metadata must be shared with researchers via an independent means: ClinicalTrials.gov, a free and publicly accessible website administered by the NIH.[253] Regardless of whether a particular drug, vaccine, or device is approved or unapproved by the FDA, the results of Phase 2, 3, or 4 trials studying the drug or device in the United States must, by law, be published on ClinicalTrials.gov. [254] FDAAA's ClinicalTrials.gov mandate requires that the results of clinical trials be individually submitted to NIH by the companies, universities, and other entities ("responsible parties" per the statute) that run them.

Petition on Botox, PUB. CITIZEN (Apr. 30, 2009), https://www.citizen.org/article/statement-fda-grants-public-citizen-petition-on-botox/.

250. *See, e.g.*, Alec Tyson, Courtney Johnson & Cary Funk, *U.S. Public Now Divided Over Whether to Get COVID-19 Vaccine*, PEW RSCH. CTR. (Sept. 17, 2020), https://www.pewresearch.org/science/2020/09/17/u-s-public-now-divided-over-whether-to-get-covid-19-vaccine/.

251. Stephen M. Hahn, Commissioner of Food and Drugs, *COVID-19 Update: FDA's Ongoing Commitment to Transparency for COVID-19 EUAs*, U.S. FOOD & DRUGS ADMIN. (Nov. 17, 2020), https://www.fda.gov/news-events/press-announcements/covid-19-update-fdas-ongoing-commitment-transparency-covid-19-euas.

252. *See, e.g.*, Hilda Bastian, *The FDA Really Did Have to Take This Long*, ATLANTIC (Aug. 23, 2021), https://www.theatlantic.com/science/archive/2021/08/fda-pfizer-vaccine-full-approval/619870/ (observing that "early, publicly available data have now been thoroughly scrutinized").

253. 42 U.S.C. § 282(j).

254. Deborah A. Zarin, Kevin M. Fain, Heather D. Dobbins, Tony Tse, & Rebecca J. Williams, *10-Year Update on Study Results Submitted to ClinicalTrials.gov,* 381 NEW ENG. J. MED. 1966 (2019).

FDAAA is detailed and exacting. It specifies the precise summary data and metadata that responsible parties must submit to ClinicalTrials.gov and thereby disclose, data element by data element.[255] When FDAAA was being debated and implemented, many entities that conduct clinical trials protested that the statute's and subsequent rule's data elements were overly detailed, overly rigid, or unreasonably different from the idiosyncratic ways in which they formatted their own data.[256] However, the consistent, predictable format of summary data provided on ClinicalTrials.gov has helped independent researchers understand and use its data.

The mandatory metadata-sharing provisions of FDAAA merit attention too, as they likewise help independent researchers contextualize trial results and perform useful research. FDAAA requires responsible parties to share detailed metadata: "[t]he full protocol or such information on the protocol for the trial as may be necessary to help to evaluate the results of the trial."[257] NIH has elaborated on this statutory provision with a rule specifying that responsible parties must also share their statistical analysis plans.[258] This mandatory sharing of metadata makes the summary data richer for researchers, and permits researchers to root out errors and manipulation.

FDAAA's mandatory metadata-sharing requirement was fought by the pharmaceutical and medical device industries. As NIH observed when it promulgated the rule that implemented this provision of FDAAA, multiple commentators from relevant industries alleged that requiring disclosure of trial protocols would violate privacy and intellectual property interests: "Some asserted that protocols contain personally identifiable information, proprietary information, or other information that, if publicly disclosed, could be damaging to business interests."[259] The largest biotech industry lobbying group, the Biotechnology Innovation Organization (BIO), argued that NIH's commitment to sharing protocols (and summary data, too) "may undermine

255. 42 U.S.C. §§ 282(j)(3)(C), (D); *see also* 42 C.F.R. § 11.48.

256. *See, e.g.*, Clinical Trials Registration and Results Information Submission, 81 Fed. Reg., *supra* note 188, at 64,982, 65,006 ("While the Agency appreciates that accepting a variety of submission formats . . . may be less burdensome for responsible parties, [FDAAA] requires the final rule to establish a standard format for the submission of clinical trial information. This standard format will, in turn, facilitate search and comparison of entries in the registry data bank, as is also required under the statute.").

257. 42 U.S.C. § 282 (j)(3)(D)(iii)(III). A trial's protocol is "[t]he written description of a clinical study. It includes the study's objectives, design, and methods. It may also include relevant scientific background and statistical information." *Protocol*, CLINICALTRIALS.GOV, https://clinicaltrials.gov/ct2/about-studies/glossary (last visited Nov. 11, 2023).

258. 42 C.F.R. § 11.48(a)(5).

259. Clinical Trials Registration and Results Information Submission, 81 Fed. Reg., *supra* note 188, at 64,982, 65,000.

incentives to innovate by forcing premature disclosure of proprietary information." [260] The largest medical device industry lobbying group, AdvaMed, echoed BIO and went further, threatening litigation over NIH's interference with its alleged trade secrets:

> [NIH's] disclosure of "trade secret and confidential commercial information" would constitute a taking in violation of the Fifth Amendment, AdvaMed stated. The device lobby group also asserted the disclosure of proprietary, confidential clinical trial data for products not approved would chill interest in developing new and innovative devices.[261]

NIH proceeded anyway. However, in a concession to industry, NIH allows companies to redact portions of their trial protocols that they consider trade secrets before posting them to ClinicalTrials.gov,[262] "so long as the redaction does not include any specific information that is otherwise required to be submitted under" the law.[263]

NIH held the line on summary data and, through rulemaking, extended FDAAA's disclosure mandate to reach experimental products not yet approved by the FDA.[264] NIH does not permit companies to redact any portion of their summary data, even if they fear competitors' use of the information.[265]

Industry's threats of litigation proved hollow. NIH has never been sued by industry over its implementation of FDAAA. Nor have the FDA or the U.S. Department of Health and Human Services (HHS). The pharmaceutical and

260. Erin Durkin, *Califf, Biden Task Force Tout NIH Rule Requiring Failed Trial Data be Posted*, 22 INSIDEHEALTHPOLICY.COM'S FDA WEEK 11 (2016).

261. *Id.*

262. Clinical Trials Registration and Results Information Submission, 81 Fed. Reg., *supra* note 188, at 64,982, 65,000 ("[I]f there is a case in which a responsible party believes that a protocol does contain trade secret and/or confidential commercial information, the responsible party may redact that information, so long as the redaction does not include any specific information that is otherwise required to be submitted under this rule.").

263. 42 C.F.R. § 11.48(a)(5); Clinical Trials Registration and Results Information Submission, 81 Fed. Reg., *supra* note 188, at 64,982, 65,000.

264. Clinical Trials Registration and Results Information Submission, 81 Fed. Reg., *supra* note 188, at 64,982, 64,986.

265. *Id.* at 64,982, 64,996 ("A few commenters suggested that if the proposal is adopted, only a limited number of primary or key secondary outcomes prior to regulatory approval should be required to be submitted, or the final rule should allow the submission of redacted results information, especially when the product has not been approved, licensed, or cleared by FDA. The Agency disagrees; we believe that results information submission for all pre-specified primary and secondary outcomes, as required in the statute, is necessary to serve the public interest in having access to full and complete information.").

medical device industries have stopped criticizing FDAAA and quietly begun complying with its mandates.

To be sure, compliance with FDAAA's ClinicalTrials.gov reporting rules is less than perfect: Independent analysis by "FDAAA Trials Tracker," a project of the Bennett Institute for Applied Data Science at Oxford University, suggests that only about 78% of trials with a legal obligation to comply with reporting rules have done so.[266] In addition, many trials that do report are late; in 2021, independent experts estimated that fewer than 50% of covered trials report results on time.[267] But this data sharing is meaningful, as much of this data is unavailable elsewhere. NIH's ClinicalTrials.gov has become the world's largest publicly accessible database of clinical trial data.[268]

And ClinicalTrials.gov has proven the value of the clinical trial data sharing mandate. Since assuming its modern form in 2017,[269] ClinicalTrials.gov's vault of data has been used in a wide range of socially beneficial research. For example, a 2014 study compared data reported on ClinicalTrials.gov with data reported in medical literature and found that "nearly all had at least 1 discrepancy in the cohort, intervention, or results reported between the two sources."[270] This study underscored ongoing errors in and manipulation of medical literature (where data reporting is less standardized and, in some journals, less scrutinized than ClinicalTrials.gov). Researchers used ClinicalTrials.gov—primarily the metadata reported pursuant to FDAAA—to critique the proliferation of many small, relatively low-quality trials of COVID therapeutics in 2020 and early 2021.[271] Such critique helped to prompt the U.S.

266. *Who's Sharing Their Clinical Trial Results?*, FDAAA TRIALS TRACKER, https://fdaaa.trialstracker.net/ (last updated Oct. 25, 2023).

267. Nicholas J. DeVito & Ben Goldacre, *Evaluation of Compliance with Legal Requirements Under the FDA Amendments Act of 2007 for Timely Registration of Clinical Trials, Data Verification, Delayed Reporting, and Trial Document Submission*, 18 JAMA INTERNAL MED. 1128 (2021).

268. Guodong Liu, Gang Chen, Lawrence I. Sinoway & Arthur Berg, *Assessing the Impact of the NIH CTSA Program on Institutionally Sponsored Clinical Trials*, 6 CLINICAL & TRANSLATIONAL SCI. 196 (2013).

269. It was in 2017 that NIH's Final Rule defining regulated entities' precise responsibilities finally went into effect. *See FDAAA 801 and the Final Rule*, CLINICALTRIALS.GOV, https://clinicaltrials.gov/ct2/manage-recs/fdaaa (last updated Jan. 2022).

270. Jessica E. Becker, Harlan M. Krumholz, Gal Ben-Josef, & Joseph S. Ross, *Reporting of Results in ClinicalTrials.gov and High-Impact Journals*, 311 JAMA 1063, 1064 (2014).

271. Krishna Pundi, Alexander C. Perino, & Robert A. Harrington, *Characteristics and Strength of Evidence of COVID-19 Studies Registered on ClinicalTrials.gov*, 180 JAMA INTERNAL MED. 1398 (2020); Paul P. Glasziou, Sharon Sanders & Tammy Hoffmann, *Waste in Covid-19 Research*, 369 BMJ 1 (2020); Deborah A. Zarin & Stephen Rosenfeld, *Lack of Harmonization of Coronavirus Disease Ordinal Scales*, 18 CLINICAL TRIALS 263 (2020).

government to promise better coordination of government-funded trials.[272] Deborah Zarin, Director of ClinicalTrials.gov from 2005–2018, wrote in 2022,[273]

> [The ClinicalTrials.gov] database has been in existence since 2008, and has been continually updated and improved during that time. Thousands of responsible parties have used it to submit over 51,000 sets of results. Research has shown that about half of these—results for about 25,000 trials—are not available in the published literature, making ClinicalTrials.gov the unique public source of this information.

Research into safety, efficacy, and the accuracy of companies' claims often complements the work of government regulators. Independent research critiques and ultimately reinforces the credibility and reliability of government regulators such as the FDA. This sort of research not only informs the public, but also actively checks and reshapes the regulatory process. For example, independent analysis of drug safety by the nonprofit organization Public Citizen, using data from ClinicalTrials.gov, Drugs@FDA, and other sources, helped convince the FDA to remove at least twenty-three dangerous drugs from the U.S. market, as of 2019.[274] Independent analysis of the clinical trial data that supported approval of Purdue Pharma's addictive oxycodone product, Oxycontin, and other opioid painkillers by drug regulators worldwide has underscored the paucity of evidence on addiction that regulators initially demanded, and has helped shape a present-day consensus that regulators must more carefully scrutinize new drugs for addictive potential.[275] In the past two years, independent analysis of the results of COVID-19 vaccines clinical trials has consistently corroborated the FDA's conclusion that the vaccines are safe,

272. *See, e.g.*, *Accelerating COVID-19 Therapeutic Interventions and Vaccines (ACTIV): Overview*, NAT'L INST. HEALTH, https://www.nih.gov/research-training/medical-research-initiatives/activ (last visited Nov. 11, 2023).

273. Ed Silverman, *'A Blind Eye': NIH Fails to Ensure Clinical Trial Results are Reported, and Still Funds Researchers Who don't File Results*, STAT (Aug. 17, 2022), https://www.statnews.com/pharmalot/2022/08/17/nih-clinical-trials-transparency-fda/.

274. Public Citizen & Center for Science in the Public Interest et al. as Amici Curiae Supporting Respondents, Food Marketing Institute v. Argus Leader Media, 139 S. Ct. 2356 (2019) (No. 18-481), 2018 WL 7890208, 18, https://www.citizen.org/wp-content/uploads/food_market_institute_v_argus_leader.pdf.

275. *See* James Heyward, Thomas J. Moore, Jennifer Chen, Kristin Meek, Peter Lurie & G. Caleb Alexander, *Key Evidence Supporting Prescription Opioids Approved by the U.S. Food and Drug Administration, 1997 to 2018,* 173 ANNALS INTERNAL MED. 956 (2020); Jessica Pappin, Itai Bavli & Matthew Herder, *On What Basis Did Health Canada Approve OxyContin in 1996? A Retrospective Analysis of Regulatory Data*, 19 CLINICAL TRIALS 584, 585 (2022).

and helped to counter some of the hesitance and misinformation that have surrounded the vaccines.[276]

The ClinicalTrials.gov database is free and accessible all over the world.[277] As such, it reduces longstanding inequities in access to trial data[278] and has catalyzed research not just in the United States but around the world. Some of the research conducted with ClinicalTrials.gov is conducted by researchers outside the United States.[279] Data from ClinicalTrials.gov has also been used to study the extent of research conducted in Global North-South collaboration.[280]

D. IMPLEMENTATION OF THE CLINICAL TRIAL DATA SHARING MANDATE AND EXPERIMENTATION WITH RESEARCHER ACCESS TO MORE SENSITIVE DATA

This Section elaborates on FDAAA's data-sharing mandate in two important regards.

First, this Section elaborates on implementation: How, exactly, does clinical trial data sharing *work*? For example, who enforces compliance with data-sharing mandates, and how? Because this Section focuses on implementation, it necessarily focuses on institutions. These institutions perform a number of important roles in the clinical trial data sharing ecosystem: they request or mandate submission of clinical trial data by industry, academia, and other sectors that perform clinical trial research; verify clinical trial data and hold it securely; mediate access to it; oversee uses by

276. *See, e.g.*, Steven K. Korang, Elena von Rohden, Areti Angeliki Veroniki, Giok Ong, Owen Ngalamika, Faiza Siddiqui, Sophie Juul, Emil Eik Nielsen, Joshua Buron Feinberg, Johanne Juul Petersen, Christian Legart, Afoke Kokogho, Mathias Maagaard, Sarah Klingenberg, Lehana Thabane, Ariel Bardach, Agustín Ciapponi, Allan Randrup Thomsen, Janus C. Jakobsen & Christian Gluud, *Vaccines to prevent COVID-19: A Living Systematic Review with Trial Sequential Analysis and Network Meta-Analysis of Randomized Clinical Trials*, 17 PLOS ONE 1, 2 (2022); Kushal T. Kadakia, *Leveraging Open Science to Accelerate Research*, 384 NEW ENG. J. MED. 1, 3 (2021); Bastian, *supra* note 252.

277. Drugs@FDA is too.

278. Satyen Shenoy, *From Bench to the Public: Open Access*, 31 MED. WRITING 6, 6 (2022) ("Paywalls and subscription fees are neither new nor unheard of in scientific publishing. However, for long, these practices have been a hindrance to dissemination of research findings, especially to the scientific and medical community in the global south, due to non-affordability.").

279. *See, e.g.*, Glasziou, *supra* note 271 (analysis of ClinicalTrials.gov data by researchers in Australia).

280. Hesborn Wao, Yan Wang, & Melvin A. Wao, *Factors associated with North-South Research Collaboration Focusing on HIV/AIDS: Lessons from ClinicalTrials.gov*, 18 AIDS RSCH. & THERAPY 1 (2021).

researchers; and monitor and enforce compliance with the laws that govern each of these steps.

Second, this Section describes how some institutions have begun pioneering giving researchers access to more sensitive data. As traced in Section III.C, FDAAA's clinical trial data sharing mandate is limited to high-reward, low-risk data: summary data and some metadata. The mandate does not reach IPD—the most sensitive data, from a privacy perspective—nor does it reach all information industry describes as its trade secrets. Yet, as we show, some institutions have pioneered mechanisms for sharing this data responsibly.

A key theme is that *institutional* governance of medical data sharing is vital to the success of *legal* governance of the same. Law on paper is only modestly effective without associated institutions to implement, elaborate, and enforce that law. It is institutions—people—that get things done.

1. *Key Institutional Governors of the Clinical Trial Data Sharing Mandate: FDA and NIH*

FDAAA's results-sharing mandate did not effectuate itself; FDAAA requires two federal agencies, the FDA and NIH, to implement the legislation's data-sharing mandate, and govern access to and use of clinical trial data.

The FDA, NIH, and other federal scientific agencies play a variety of important roles in managing not just clinical trial data but a wealth of other scientific and technical data. As Contreras observed, "the state's role in fostering innovation and scientific advancement is often analyzed in terms of incentives that the state may offer to private actors" such as tax credits, IP protections, direct grants, and provision of infrastructure.[281] Yet Contreras convincingly argues that this view is incomplete, at least in the fields of medicine and biotechnology. In the United States, the medical "innovation system" depends on the U.S. government not just as incentive-setter but as a central *actor* in the "information economy," managing data flows:

> The state plays a number of well-understood roles with respect to the planning, provisioning, and maintenance of publicly owned infrastructure resources such as highways, prisons, and public utilities. Likewise, the state is often involved in the oversight, regulation, and operation of private and public-private infrastructural resources such as airports and telecommunications networks. Why then should the same types of complementary and

281. Jorge Contreras, *Leviathan in the Commons: Biomedical Data and the State*, *in* GOVERNING MED. KNOWLEDGE COMMONS 19–20 (Katherine J. Strandburg et al. eds., 2017).

> overlapping relationships not arise with respect to data resources that form an integral part of the research infrastructure?[282]

Contreras maps nine distinct roles that U.S. government agencies play in the governance of medical data, writ large: (1) creator, (2) funder, (3) convenor, (4) collaborator, (5) endorser, (6) curator, (7) regulator, (8) enforcer, and (9) consumer.[283]

In the world of clinical trial data sharing, the FDA and NIH play all nine roles, but in this Section, we focus on four overlapping roles we consider particularly important to the success of clinical trial data sharing: curator, funder, regulator, and enforcer.

a) FDA and NIH Curate Data

Institutions curate data by aggregating, hosting, and explaining data for other stakeholders to access and use. FDAAA mandates that NIH and the FDA play these curatorial roles: NIH with ClinicalTrials.gov, and FDA with Drugs@FDA.[284]

NIH's National Library of Medicine (NLM) aggregates and hosts the massive ClinicalTrials.gov database. NLM also actively safeguards the quality, accuracy, and usability of each submission of clinical trial data.[285] NLM conducts an extensive quality control process to ensure that data is submitted to ClinicalTrials.gov completely and in the correct format.[286] NLM also maintains an elaborate "customer support" site and helpline for staff at universities, drug companies, and other institutions who encounter problems when preparing and submitting data to the database.[287] In this way, NLM protects the credibility and usability of the database.

NLM has curated not just data *submission* but data *use* by researchers and the general public; it maintains an extensive Glossary and FAQ page to guide researchers through searching and interpreting the database.[288] NLM has also

282. *Id.* at 25.

283. *Id.* at 22–24.

284. 42 U.S.C. § 282(j)(3) (NIH); 21 U.S.C. § 355(l) (FDA).

285. *See* Contreras, *supra* note 281, at 38.

286. Rebecca Williams, *ClinicalTrials.gov Webinar: Updated Quality Control and Posting Procedures*, NAT'L INST. HEALTH (Oct. 15, 2019), https://www.nlm.nih.gov/oet/ed/ct/30_day_post.html.

287. *Submit Studies to ClinicalTrials.gov PRS*, CLINICALTRIALS.GOV, https://clinicaltrials.gov/ct2/manage-recs/submit-study (last visited Nov. 11, 2023).

288. *Frequently Asked Questions*, CLINICALTRIALS.GOV, https://clinicaltrials.gov/ct2/manage-recs/faq (last visited Nov. 11, 2023).

published research guides in the medical literature, detailing how to make effective use of ClinicalTrials.gov.[289]

The FDA similarly aggregates, hosts, and explains the data it publishes on its own Drugs@FDA website in the form of the approval packages required by FDAAA. The FDA does not simply republish industry-submitted trial data, but also independently reviews the data and provides its own written critique and summary.[290] Like NLM, the FDA maintains a glossary[291] and FAQ[292] to help researchers use Drugs@FDA.

b) FDA and NIH Fund Data-Sharing Initiatives and Research Itself

The FDA and NIH serve separate roles as funders. They fund private initiatives to steward and share data, and they fund academic researchers who make socially beneficial uses of data. This role flows from law; Congress's appropriations bills earmark public money to the agencies for these very purposes. This role, too, explains the success of the clinical trial data sharing mandate.

NIH is the world's largest medical research grant-maker,[293] and some of the billions disbursed go to researchers who use ClinicalTrials.gov in their research.[294] The FDA has formed multi-year partnerships with Johns Hopkins, Stanford, the University of Maryland, the Mayo Clinic, and Yale to study

289. *See, e.g.*, Tony Tse, Kevin M. Fain & Deborah A. Zarin, *How to Avoid Common Problems when Using ClinicalTrials.gov in Research: 10 Issues to Consider*, 361 BMJ 1 (2018).

290. *Drugs@FDA: FDA-Approved Drugs*, FDA, https://www.accessdata.fda.gov/scripts/cder/daf/ (last visited Nov. 11, 2023); Herder, Morten & Doshi, *supra* note 243.

291. *Drugs@FDA Glossary of Terms*, FDA, https://www.fda.gov/drugs/drug-approvals-and-databases/drugsfda-glossary-terms (last visited Nov. 11, 2023).

292. *Drugs@FDA Frequently Asked Questions*, FDA, https://www.accessdata.fda.gov/scripts/cder/daf/index.cfm?event=faq.page (last visited Nov. 11, 2023).

293. W. Nicholson Price II, *Grants*, 34 BERKELEY TECH. L.J. 1, 4 (2019).

294. *See, e.g.*, Richeek Pradhan, David C. Hoaglin, Matthew Cornell, Weisong Liu, Victoria Wang & Hong Yu, *Automatic Extraction of Quantitative Data From ClinicalTrials.gov to Conduct Meta-Analyses*, 105 J. CLIN. EPIDEMIOL. 92 (2019) (independent research funded by NIH to create an automated tool to extra data from ClinicalTrials.gov more quickly and easily); Joshua D. Wallach, John H. Krystal, Joseph S. Ross & Stephanie S. O'Malley, *Characteristics of Ongoing Clinical Trials for Alcohol Use Disorder Registered on ClinicalTrials.gov,* 77 JAMA PSYCHIATRY 1081 (2020) (independent research funded by the National Institute on Alcohol Abuse and Alcoholism (an NIH Institute) studying the quantity and quality of trials for alcohol use disorder registered on ClinicalTrials.gov); Sarah F. Ackley, Scott C. Zimmerman, Willa D. Brenowitz, Eric J. Tchetgen Tchetgen, Audra L. Gold, Jennifer J. Manly, Elizabeth Rose Mayeda, Teresa J. Filshtein, Melinda C. Power, Fanny M. Elahi, Adam M. Brickman, & M. Maria Glymour, *Effect of Reductions in Amyloid Levels on Cognitive Change in Randomized Trials: Instrumental Variable Meta-Analysis*, 372 BMJ 1, n.156 (2021) (NIH-funded independent meta-analysis of existing clinical trial data available on ClinicalTrials.gov and other sources to explore the link between beta-amyloid levels in the brain and cognitive function).

pharmaceutical and medical device regulation to scrutinize and improve the FDA's regulatory work. These government-academic initiatives are called Centers of Excellence in Regulatory Science and Innovation (CERSI).[295] Researchers funded by the FDA in this way have critiqued and improved the FDA's own work, e.g., by using FDA-published trial data and other data to question the use of "real-world evidence" in lieu of traditional clinical trials[296] and asking whether the FDA is sufficiently attentive to evidence of side effects gathered after drug approval.[297]

The FDA has also experimented with funding academic institutions, nonprofits, and patient groups to become data-sharing platforms themselves. That is, the FDA has sponsored private institutions to aggregate and share certain clinical trial data. These initiatives include the Rare Disease Cures Accelerator-Data and Analytics Platform (RDCA-DAP).[298] Indeed, some other emerging "private" medical data-sharing initiatives led by patients, academia, and/or industry are funded partly with public resources; they do not always emerge entirely "organically" without the hand of the state. One such example is the Yale Open Data Access (YODA) Project, discussed more below.

c) FDA and NIH Regulate and Enforce the Data Sharing Mandate

Finally, we consider the roles of NIH and the FDA as regulators and enforcers of FDAAA's clinical trial data sharing mandate. NIH and the FDA force the pharmaceutical and medical device industries to share otherwise proprietary clinical trial data, consistent with FDAAA's mandate. Congress gave FDAAA "teeth" by specifying draconian potential consequences for failing to submit clinical trial results to ClinicalTrials.gov, including fines of over $10,000 per day per missing trial and a "freeze" on any grant money

295. *Centers of Excellence in Regulatory Science and Innovation (CERSIs)*, FDA (Jan. 5, 2023), https://www.fda.gov/science-research/advancing-regulatory-science/centers-excellence-regulatory-science-and-innovation-cersis. For an example grant, see Joseph S. Ross, *Yale-Mayo Clinic FDA Center of Excellence in Regulatory Science and Innovation (CERSI)*, GRANTOME, https://grantome.com/grant/NIH/U01-FD005938-03/ (last visited Nov. 11, 2023).

296. Victoria L. Bartlett, Sanket S. Dhruva, Nilay D. Shah, Patrick Ryan & Joseph S. Ross, *Feasibility of Using Real-World Data to Replicate Clinical Trial Evidence*, 2 JAMA NETWORK OPEN: STAT. & RSCH. METHODS 1, 7 (2019).

297. Meera M. Dhodapkar, Xiaoting Shi, Reshma Ramachandran, Evan M. Chen, Joshua D. Wallach, & Joseph S. Ross, *Characterization and Corroboration of Safety Signals Identified from the US Food and Drug Administration Adverse Event Reporting System, 2008-19: Cross Sectional Study*, 379 BMJ 1, 8 (2022).

298. *Funded by FDA, C-Path and Nord to Launch Rare Disease Data Analytics Platform*, NAT'L ORG. FOR RARE DISORDERS (Aug. 7, 2019), https://rarediseases.org/funded-by-fda-c-path-and-nord-to-launch-rare-disease-data-analytics-platform/.

disbursed by NIH, the FDA, and other constituent agencies of HHS.[299] FDAAA also requires the FDA to name and shame responsible parties out of compliance with FDAAA's reporting rules, via public "Notices of Noncompliance" on a FDA-managed website crosslinked to ClinicalTrials.gov.[300]

The FDA and NIH have performed poorly in their role as enforcers. Since FDAAA's enactment, the FDA's enforcement efforts have been almost laughably minimal: just five Notices of Noncompliance issued and zero fines imposed, despite thousands of trials out of compliance (among tens of thousands of trials with results required under FDAAA).[301] It was only in 2022 that NIH began sending letters threatening to withhold grant money from grantees out of compliance with FDAAA's data sharing mandate.[302] NIH and the FDA have been criticized from many sides for not doing more enforcement, including by researchers seeking access to missing data,[303] civil

299. 42 U.S.C. § 282(j).

300. *Id.*; *ClinicalTrials.gov—Notices of Noncompliance and Civil Money Penalty Actions*, FDA, https://www.fda.gov/science-research/fdas-role-clinicaltrialsgov-information/clinicaltrialsgov-notices-noncompliance-and-civil-money-penalty-actions (last visited Mar. 17, 2024).

301. *Id.*; *see also* Reshma Ramachandran, Christopher J. Morten & Joseph S. Ross, *Strengthening the FDA's Enforcement of ClinicalTrials.gov Reporting Requirements*, 326 JAMA 2131 (2021).

302. Ed Silverman, *After Years of Lax Oversight, the NIH is Starting to Contact Institutions About Unreported Clinical Trial Results*, STAT: PHARMALOT (Nov. 7, 2022), https://www.statnews.com/pharmalot/2022/11/07/nih-clinical-trials-transparency-fda-2/.

303. Nicholas J. DeVito & Ben Goldacre, *Evaluation of Compliance With Legal Requirements Under the FDA Amendments Act of 2007 for Timely Registration of Clinical Trials, Data Verification, Delayed Reporting, and Trial Document Submission*, 181 JAMA INTERNAL MED. 1128, 1130 (2021).

society groups,[304] journalists,[305] a former director of ClinicalTrials.gov,[306] HHS's Office of Inspector General,[307] and one of us.[308]

Yet even the FDA and NIH's meager enforcement has contributed to a significant increase in data-sharing compliance rates. Since 2020, when the FDA first promised to begin issuing Notices of Noncompliance and threatened fines,[309] the percentage of applicable clinical trial results reported to the database rose from approximately 60–65%[310] to about 75–80%.[311] Even light-touch enforcement prompts compliance. A 2021 analysis showed that when the FDA simply sent a few dozen short letters to responsible parties, stating that the agency had reason to believe their trials might be out of compliance with FDAAA's data reporting rules, more than 90% of recipients provided the missing data with a median response time of just a few weeks.[312]

And the present, C-grade state of enforcement and compliance with ClinicalTrials.gov's reporting mandate is nonetheless sufficient to unlock enormous benefits.[313] As former ClinicalTrials.gov Director Zarin wrote in 2022, there are approximately 25,000 trial results reported on ClinicalTrials.gov

304. *See, e.g.*, *Clinical Trials Transparency Campaign*, UNIVS. ALLIED FOR ESSENTIAL MEDS., https://www.uaem.org/transparency_campaign (last visited Nov. 11, 2023); *Clinical trial transparency at US universities*, TRANSPARIMED (Mar. 25, 2019), https://www.transparimed.org/_files/ugd/01f35d_8c22b87eda8e44ac83cf76642de94053.pdf?index=true (criticism from UAEM and TranspariMED).

305. Charles Piller, *FDA and NIH Let Clinical Trial Sponsors Keep Results Secret and Break the Law*, SCI. (Jan. 13, 2020), https://www.science.org/content/article/fda-and-nih-let-clinical-trial-sponsors-keep-results-secret-and-break-law.

306. Silverman, *supra* note 273; *Drug Researchers Refuse to Follow the Law. The Government Isn't Stopping Them*, SCI. FRIDAY (Jan. 24, 2020), https://www.sciencefriday.com/segments/clinical-trial-reporting-government/.

307. *The National Institutes of Health did Not Ensure that All Clinical Trial Results were Reported in Accordance with Federal Requirements*, DEP'T HEALTH & HUMAN SERVS. OFF. INSPECTOR GEN. (Aug. 12, 2022), https://oig.hhs.gov/oas/reports/region6/62107000.asp.

308. Ramachandran, *supra* note 301; Christopher Morten, Peter G. Lurie & Charles Seife, *Lost opportunities from FDA, NIH inaction when sponsors fail to report clinical trial results*, STAT (Apr. 13, 2020), https://www.statnews.com/2020/04/13/lost-opportunities-clinical-trial-results-unreported-lost-opportunities/.

309. Civil Money Penalties Relating to the ClinicalTrials.gov Data Bank; Guidance for Responsible Parties, Submitters of Certain Applications and Submissions to FDA, and FDA Staff, 85 Fed. Reg. 50028 (Aug. 17, 2020), https://www.fda.gov/regulatory-information/search-fda-guidance-documents/civil-money-penalties-relating-clinicaltrialsgov-data-bank.

310. Nicholas J. DeVito, Seb Bacon & Ben Goldacre, *Compliance with Legal Requirement to Report Clinical Trial Results on ClinicalTrials.gov: A Cohort Study*, 395 LANCET 361, 365 (2020); Piller, *supra* note 305.

311. *Who's Sharing Their Clinical Trial Results?*, FDAAA TRIALS TRACKER, https://fdaaa.trialstracker.net/ (last visited Nov. 11, 2023).

312. Ramachandran, *supra* note 301, at 2132.

313. *Supra* Section II.C.2.

that are unreported in the medical literature, and thus presumably accessible to researchers nowhere but ClinicalTrials.gov.[314]

Why such meager enforcement from the FDA and NIH? One major reason is that FDAAA imposed new regulatory obligations on both agencies without allocating new funding.[315] Both the FDA and NIH have many other obligations, and neither agency had strong incentives to dedicate personnel and attention to ClinicalTrials.gov. In addition, HHS's choice to divide enforcement responsibilities between the two agencies[316] rather than vesting responsibility entirely with one has made it easier for each agency to point to the other as the laggard.

2. Pioneering Researcher Access to More Sensitive Data

The entire clinical trial data sharing mandate described above requires sharing of just two components of clinical trial data: summary data and metadata. To recap, FDAAA mandates that summary data be disclosed without redaction.[317] It mandates that metadata be disclosed as well,[318] though NIH rules permits companies (and other trial sponsors) to redact information in trial protocols deemed a trade secret or confidential commercial information.[319] This means that FDAAA's clinical trial data sharing mandate does not reach IPD, the most detailed and most sensitive trial data.[320] The mandate also does not reach some metadata in trial protocols that companies deem trade secrets.

Yet some institutions that share clinical trial data have pioneered ways to share sensitive information with independent researchers. These efforts show it is possible to navigate treacherous hazards to privacy and trade secrecy with careful institutional and legal design.

314. Ed Silverman, *'A Blind Eye': NIH Fails to Ensure Clinical Trial Results are Reported, and Still Funds Researchers who don't File Results*, STAT: PHARMALOT (Aug. 17, 2022), https://www.statnews.com/pharmalot/2022/08/17/nih-clinical-trials-transparency-fda/.

315. *See* Ramachandran, *supra* note 301, at 2132.

316. Office of the Commissioner of Food and Drugs; Delegation of Authority, 77 Fed. Reg. 59196 (Sept. 26, 2012).

317. 42 U.S.C. § 282(j)(3)(C).

318. *Id.* § 282(j)(3)(D)(iii)(III) (mandating disclosure of the trial protocol or "or such information on the protocol for the trial as may be necessary to help to evaluate the results of the trial").

319. 42 C.F.R. § 11.48(a)(5).

320. *See supra* Section II.A.1.

a) Sharing IPD

Sharing raw clinical trial data that describes, in detail, the health statuses of individual patients—IPD—poses profound risks to patient privacy.[321] As the Institute of Medicine put it in 2015, "privacy concerns have been stated as a key obstacle to making these data available."[322] Yet some kinds of research depend on IPD and cannot be done without it. For example, only researchers with access to IPD and the trial's complete methodology can conduct reanalysis to confirm the correctness of the trial sponsor's conclusions.

Numerous institutions now share IPD with researchers, and do so responsibly.[323] Some of these databases are public—e.g., NIH's Biologic Specimen and Data Repositories Information Coordinating Center (BioLINCC). Other databases are nonprofit and academic—e.g., the Yale Open Data Access Project (YODA). Others are industry-run.[324]

We describe these two IPD-sharing databases here. We do not attempt a comprehensive survey of IPD-sharing initiatives but instead present these as proofs-of-concept. Key features permit them to share sensitive data with researchers while protecting the data's integrity and the interests of the data subjects.

As we trace below, a constant of these databases is that they are *not* universally accessible; they do not publish data for use by any and all comers. Instead, they discriminate among prospective users and provide access only to researchers deemed sufficiently responsible.

Further, the institutions that manage these databases use legal and/or technological constraints to limit researchers' access to and use of the data, reducing the risk of harmful uses. Researchers' access may be "tiered"; different kinds of researchers obtain different levels of access to different

321. *See supra* Section III.A.3.

322. *Sharing Clinical Trial Data: Maximizing Benefits*, *supra* note 161.

323. One driver of the recent uptick in IPD sharing has been prestigious medical journals, which have encouraged researchers who seek to report the results of clinical trials in those journals to commit to sharing deidentified IPD. *See* Darren Taichman, Peush Sahni, Anja Pinborg, Larry Peiperl, Christine Laine, Astrid James, Sung-Tae Hong, Abraham Haileamlak, Laragh Golloghy, Fiona Godlee, Frank A. Frizelle & Fernando Florenzano, *Data Sharing Statements for Clinical Trials: A Requirement of the International Committee of Medical Journal Editors*, 376 NEW ENGL. J. MED. 2277 (2017).

324. *See Our Mission*, CLINICALSTUDYDATAREQUEST.COM, https://clinicalstudydatarequest.com/Default.aspx## (last visited Nov. 11, 2023); *Convener, Collaborator, Catalyst in the Fight Against Cancer*, PROJECT DATA SPHERE, https://www.projectdatasphere.org/ (last visited Nov. 11, 2023). Michael J. Pencina, *Supporting Open Access to Clinical Trial Data for Researchers: The Duke Clinical Research Institute—Bristol-Myers Squibb Supporting Open Access to Researchers Initiative*, 172 AM. HEART J. 64, 67(2016).

components or kinds of data. All this underscores the vital role of *institutions* in clinical trial data sharing; these databases require active stewardship.

i) NIH BioLINCC

In addition to the enormous ClinicalTrials.gov database, NIH curates and controls smaller databases of clinical trial data. A notable one is BioLINCC, a database that contains sensitive IPD from clinical trials in cardiovascular, pulmonary, and hematological diseases.[325] BioLINCC has been in operation since the 2000s.[326] NIH created and administers the center, but much of the information contained in BioLINCC's databases is contributed not by NIH itself but by nongovernmental entities, including drug and device companies.[327] NIH requires these entities to submit data to BioLINCC as a condition of accepting NIH funding for their research. This straightforward quid pro quo leverages NIH's separate role as funder.

Because BioLINCC data typically contains IPD, NIH shares data conditionally, limiting access and use. BioLINCC requires would-be researchers to submit data use applications, which document the intended uses of specific data sets (prospective researchers' "Research Plan"), data security practices, and commitments. NIH discriminates among users; NIH provides commercial users access only to a subset of BioLINCC's data and provides no access at all to would-be researchers that do not submit a credible Research Plan.[328]

NIH then enforces researchers' compliance with their Research Plans through contract. NIH imposes a data use agreement on every researcher who obtains access to IPD from BioLINCC. The data use agreement governs transfer, maintenance, and use of protected data. The agreement imposes

325. *BioLINCC Resource Overview*, NIH, https://biolincc.nhlbi.nih.gov/resource_overview/. In additional to clinical trial data, BioLINCC also shares with researchers other non-clinical trial medical data and biospecimens. *Id.*

326. *The BioLINCC Handbook: A Guide to the NHLBI Biologic Specimen and Data Repositories*, NAT'L HEART, LUNG, & BLOOD INST. 1 (2021), https://biolincc.nhlbi.nih.gov/media/guidelines/handbook.pdf. BioLINCC also shares physical samples of materials useful in biomedical research.

327. *See Guidelines for Preparing Clinical Study Data Sets for Submission to the NHLBI Data Repository*, NAT'L HEART, LUNG, & BLOOD INST., https://www.nhlbi.nih.gov/grants-and-training/policies-and-guidelines/guidelines-for-preparing-clinical-study-data-sets-for-submission-to-the-nhlbi-data-repository (instructions to non-BHLBI investigators running NHLBI-funded studies).

328. *The BioLINCC Handbook*, *supra* note 326, at 8 ("[F]or studies with commercial use data restrictions, investigators requesting data for commercial use would be eligible to receive only the subset of the overall dataset that was provided by subjects who consented to commercial research.").

constraints on researchers, both positive (incentivizing users to do beneficial things) and negative (disincentivizing users from doing harmful things). BioLINCC's current standard agreement includes all the following:[329]

Provisions that prohibit . . .

- commercial uses of data;
- further sharing of data; and
- reidentification of or contact with any patient whose IPD is in the data set.

Provisions that require . . .

- appropriate data security safeguards;
- regular updates to NIH on the status of research;
- notification to NIH in the event of data breach;
- notification to NIH and the FDA in the event the data user identifies in the data an ongoing risk to public health and safety;
- dissemination of any findings to the public, e.g., by publication in the peer-reviewed medical or scientific literature; and
- destruction of data when research is complete.

Data use agreements can specify penalties in the event a researcher breaches the agreement. These penalties can be financial or non-financial. BioLINCC's data use agreement does not contemplate financial penalties but does promise to ban breachers from any future access to data.[330]

BioLINCC's information-sharing program has succeeded. Hundreds of requesters have sought and received access to thousands of data sets, leading to dozens of high-profile scientific and medical publications in cardiology, infectious disease, and other fields of medical research.[331] Over 250 articles

329. Sean A. Coady, George A. Mensah, Elizabeth L. Wagner, Miriam E. Goldfarb, Denise M. Hitchcock & Carol A. Giffen, *Use of the National Heart, Lung, and Blood Institute Data Repository*, 376 NEW ENG. J. MED. 1849 (2017) (describing data use agreements used by NIH's BioLINCC); *see also How Can Covered Entities Use and Disclose Protected Health Information for Research and Comply with the Privacy Rule?*, NAT'L INST. HEALTH, https://privacyruleandresearch.nih.gov/pr_08.asp (last visited Nov. 11, 2023) (NIH publication describing data use agreement).

330. *The BioLINCC Handbook*, *supra* note 326, at 20 ("[F]ailure to adhere to the terms of the RMDA will be taken into consideration with respect to any future requests for data and/or biospecimens from the NHLBI repositories.").

331. Joseph S. Ross, Jessica D. Ritchie, Emily Finn, Nihar R. Desai, Richard L. Lehman, Harlan M. Krumholz, & Cary P. Gross, *Data Sharing Through an NIH Central Database* Repository: A Cross-Sectional Survey of BioLINCC Users, 6 BMJ OPEN (2016); Carol A. Giffen, Leslie E. Caroll, John T. Adams, Sean P. Brennan, Sean A. Coady & Elizabeth L. Wagner, *Providing Contemporary Access to Historical Biospecimen Collections: Development of the NHLBI Biologic Specimen and Data Repository Information Coordinating Center (BioLINCC)*, 13 BIOPRESERVATION & BIOBANKING 271 (2015).

were published based on BioLINCC data accessed between January 2000 and May 2016.[332] In practice, NIH's scrutiny and data use agreements seem to work. No researcher misuse of BioLINCC data covered by a data use agreement has been reported in the years of BioLINCC's existence.

ii) Yale Open Data Access Project (YODA)

Another prominent institution with a track record of successfully sharing IPD is YODA, a nonprofit academic data center that holds complete data sets (including IPD) on over 400 trials.[333]

YODA is not the only non-governmental, not-for-profit institution that shares IPD. Two additional examples are Vivli and Project Data Sphere.[334]

YODA operates similarly to NIH's BioLINCC. Like BioLINCC, YODA holds data on its own servers, gatekeeps requests for access to data, and enforces compliance with its own rules for data sharing and use. To get YODA data, researchers must establish that they have a credible research plan and proper security measures in place.[335] YODA refuses some applicants, especially when those applicants seek access to sensitive IPD. In difficult cases, YODA uses a peer-review-like process: it solicits reviews from two independent scientists to help decide whether to approve or deny applications.[336] Like BioLINCC, YODA imposes data use agreements on all researchers who get access to the data.

YODA has convinced major medical technology companies—including Medtronic and Johnson & Johnson—to share, voluntarily, complete clinical trial data sets that would otherwise be proprietary. These companies benefit in various ways from contributing data to YODA, including a "halo effect" of good publicity and early access to scientific insights contributed by the researchers who use their data.[337] The companies that contribute data to

332. *See* Coady et al., *supra* note 329, at 1849.

333. *Our Mission*, YALE UNIVERSITY OPEN DATA ACCESS (YODA) PROJECT, https://yoda.yale.edu/ (last visited Jan. 31, 2023).

334. *About Vivli: Overview*, VIVLI, https://vivli.org/about/overview/ (last visited Jan. 31, 2023); *About Project Data Sphere*, VIVLI, https://www.projectdatasphere.org/about (last visited Jan. 31, 2023).

335. *Frequently Asked Questions (FAQs)*, YALE U. OPEN DATA ACCESS (YODA) PROJECT, https://yoda.yale.edu/about/frequently-asked-questions-faqs#Data%20Request%20Review%20Process (last visited Jan. 31, 2023).

336. *Id.*

337. Researchers are required, under the terms of the YODA DUA, to share insights with the company that contributed the trial data under study even before they publish their findings for the world. *Procedures to Guide External Investigator Access to Clinical Trial Data*, YALE U. OPEN DATA ACCESS (YODA) PROJECT, https://yoda.yale.edu/sites/default/files/files/

YODA reserve their own rights to bring breach-of-contract claims against researchers who breach YODA's data use agreements.

Though rather small, YODA has been a success thus far: between 2014 and 2018, Johnson & Johnson voluntarily shared data from 200 clinical trials through YODA, generating at least a dozen new scientific publications,[338] including analyses of the safety of ulcerative colitis treatments[339] and the efficacy of schizophrenia drugs (which critiqued exaggerated claims made in the medical literature).[340] All this occurred without evidence of privacy violations, breaches of the data use agreements, or harmful use of data by Johnson & Johnson's competitors.[341]

YODA operates on a mixture of grants provided by industry (Medtronic and Johnson & Johnson), philanthropy, and government. The FDA and the Centers for Medicare & Medicaid Services (CMS) have both funded YODA, showing the role public money and institutions can play in nurturing private governors of data.[342]

b) Sharing Metadata That Contains Alleged Trade Secrets

In this Section III.D.2.b, we turn to an institution that has pioneered responsible sharing of (purported) trade secret data with researchers: Health Canada, Canada's central drug regulator.

Since 2019, Health Canada has shared rich data sets from clinical trials of agency-approved products, under a program called Public Release of Clinical

YODA%20Project%20Data%20Release%20Procedures%20February%202019.pdf (last visited Nov. 11, 2023).

338. Joseph S. Ross, Joanne Waldstreicher, Stephen Bamford, Jesse A. Berlin, Karla Childers, Nihar R. Desai, Ginger Gamble, Cary P. Gross, Richard Kuntz, Richard Lehman, Peter Lins, Sandra A. Morris, Jessica D. Ritchie, Harlan M. Kumholz, *Overview and Experience of the YODA Project with Clinical Trial Data Sharing After 5 Years*, 5 SCI. DATA 1, 8–9 (Nov. 27, 2018).

339. *See* David Cheng, Kelly C. Cushing, Tianxi Cai, Ashwin N. Ananthakrishnan, *Safety and Efficacy of Tumor Necrosis Factor Antagonists in Older Patients with Ulcerative Colitis: Patient-Level Pooled Analysis of Data from Randomized Trials*, 19 CLINICAL GASTROENTEROLOGY & HEPATOLOGY 939, 944 (2021).

340. Alexander Hodkinson, Carl Heneghan, Kamal R. Mahtani, Evangelos Kontopantelis & Maria Panagioti, *Benefits and Harms of Risperidone and Paliperidone for Treatment of Patients with Schizophrenia or Bipolar Disorder: A Meta-Analysis Involving Individual Participant Data and Clinical Study Reports*, 19 BMC MED. 1, 6–8 (2021).

341. Ross, *supra* note 338.

342. Joseph S. Ross, *Sharing Data Through the Yale University Open Data Access (YODA) Project: Early Experience*, YOUTUBE (Oct. 11, 2017), https://www.youtube.com/watch?v=E2ex74Zn7I0.

Information (PRCI).[343] The data shared through PRCI is generated and compiled not by Health Canada but by the drug and device companies who submit it when seeking approval. In effect, PRCI works similarly to the FDA's Drugs@FDA database but is simultaneously deeper (providing more detailed summary data and metadata) and narrower (covering fewer products). As of March 2024, data on over 600 distinct drugs and devices, from dozens of companies, had been posted to PRCI.[344]

Academic researchers have used PRCI data to analyze and communicate the safety and efficacy of important medical products, constituting an important check on and complement to the work of Health Canada, the FDA, and other national regulators. For example, an academic group recently used PRCI data to show that extended-release oxycodone hydrochloride ("Oxycontin") was approved in the 1990s by Health Canada, the FDA, and other national regulators without any evaluation of the risks of misuse and addiction.[345]

The clinical trial data shared by Health Canada through PRCI implicates both patient privacy and trade secrecy. To protect these interests, Health Canada asks regulated entities to redact what it deems "confidential business information" (CBI)—essentially, trade secrets under U.S. law[346]—as well as information identifying individual trial participants before making data

343. *Clinical Information on Drugs and Health Products*, GOV'T CAN., https://www.canada.ca/en/health-canada/services/drug-health-product-review-approval/clinical-information-drugs-health-products.html (last updated Mar. 12, 2019). The European Union began and then suspended a similar program. For details, see Alexander C. Egilman, Amy Kapczynski, Margaret E. McCarthy, Anita T. Luxkaranayagam, Christopher J. Morten, Matthew Herder, Joshua D. Wallach & Joseph S. Ross, *Transparency of Regulatory Data Across the European Medicines Agency, Health Canada, and US Food and Drug Administration*, 49 J.L. MED. & ETHICS 456, 456–57, 459 (2021).

344. *Search for Clinical Information on Drugs and Medical Devices*, HEALTH CAN., https://clinical-information.canada.ca/search/ci-rc (last updated Mar. 17, 2024).

345. Jessie Pappin, Itai Bavli & Matthew Herder, *On What Basis Did Health Canada Approve OxyContin in 1996? A Retrospective Analysis of Regulatory Data*, 19 CLINICAL TRIALS 584, 584–85 (2022).

346. Health Canada's definition of CBI is nearly identical to the definition of "trade secret" that predominates in U.S. law: "business information[] that is not publicly available, in respect of which the person has taken measures that are reasonable in the circumstances to ensure that it remains not publicly available, and that has actual or potential economic value to the person or their competitors because it is not publicly available and its disclosure would result in a material financial loss to the person or a material financial gain to their competitors." *Guidance Document—Disclosure of Confidential Business Information Under Paragraph 21.1(3)(c) of the Food and Drugs Act*, GOV'T CAN., https://www.canada.ca/en/health-canada/services/drug-health-product-review-approval/request-disclosure-confidential-business-information/disclosure-confidential-business-information/guidance.html#a1.2 (last visited Nov. 11, 2023).

accessible to routine users of PRCI.[347] Users who wish to access and use these redacted data sets may do so with few restrictions, much like Drugs@FDA and ClinicalTrials.gov.

Yet Health Canada shares even more information with select researchers, including unredacted trade secrets. According to Paragraph 21.1(3)(c) of the Canadian Food and Drugs Act,[348] Health Canada will share trade secrets (CBI) on certain conditions. First, researchers must submit a data use application that proves their proposed use is noncommercial and relates to "protection or promotion of human health or the safety of the public."[349] Second, the application must also explain "[h]ow the results of the proposed project will be disseminated to the Canadian public."[350] Any researchers granted access must then sign data use agreements insisting "the specified CBI can be used only for the purposes of the proposed project and must be kept confidential using appropriate safeguards."[351] In the event a researcher detects a safety, efficacy, or quality problem in the data, Health Canada requests the researcher notify Health Canada as well as the public at large.[352]

In 2016, a medical researcher, Peter Doshi, used Paragraph 21.1(3)(c) to obtain detailed, previously secret data on the safety and efficacy of several medical products, including oseltamivir ("Tamiflu") and vaccines for human papillomavirus (HPV).[353] Doshi's access to this CBI—and his legal authority to disseminate analysis of it—was upheld by the Canadian Federal Court.[354]

347. *Guidance document on Public Release of Clinical Information: Profile Page*, HEALTH CAN., https://www.canada.ca/en/health-canada/services/drug-health-product-review-approval/profile-public-release-clinical-information-guidance.html (last updated Mar. 29, 2019).

348. *Disclosure of Confidential Business Information*, HEALTH CAN., https://www.canada.ca/en/health-canada/services/drug-health-product-review-approval/request-disclosure-confidential-business-information/disclosure-confidential-business-information.html (last updated Nov. 17, 2020).

349. *Guidance Document—Disclosure of Confidential Business Information Under Paragraph 21.1(3)(c)*, *supra* note 347.

350. *Id.*

351. *Id.*

352. *Id.* ("Recipients of disclosed information are expected to make the findings of their project with the disclosed information publicly available when the findings provide additional knowledge about the therapeutic product under study. If the recipient of disclosed information has made a determination that the safety, efficacy or quality of a product(s) may change as a result of the evaluation of the CBI then the results should be submitted to Health Canada.").

353. Trudo Lemmons, *Precedent Pushing Practice: Canadian Court Orders Release of Unpublished Clinical Trial Data*, BMJ OPINION (July 19, 2018), https://blogs.bmj.com/bmj/2018/07/19/precedent-pushing-practice-canadian-court-orders-release-of-unpublished-clinical-trial-data/.

354. Doshi v. Attorney General of Canada, [2018] F.C. 710 (Can. Ont.), https://www.canlii.org/en/ca/fct/doc/2018/2018fc710/2018fc710.pdf.

Doshi has not made inappropriate use of the data, and industry has not subsequently sued Health Canada to block similar disclosures.

E. CLINICAL TRIAL DATA IN ACTION: A RECAP

Perhaps the single most important lesson of Part III is that clinical trial data sharing *works*. Today's clinical trial data sharing mandate guarantees researchers meaningful access to components of clinical trial data that the R&D-driven pharmaceutical industry kept proprietary for decades. The mandate has fostered beneficial research that could not have occurred otherwise, some of which has challenged industries' overblown claims and improved the FDA's regulation. Indeed, the mandate seems to have contributed to a "new normal" of improved drug safety; in the years since FDAAA was enacted, we have not had scandals of unsafe products and manufacturer cover-ups on the level of Paxil or Vioxx.[355]

The pharmaceutical and medical device industries resisted clinical trial data sharing on the argument that sharing would harm privacy and incentives to innovate. But so far, clinical trial data sharing has capably protected those interests.

The clinical trial data sharing mandate emerged over years, not overnight, and remains a work in progress. Key to the mandate's qualified success are the institutions that give ongoing effect to its underlying law, especially FDAAA. Law cannot simply proscribe or prescribe behavior, nor can it reallocate power with the stroke of a pen. In our view, law must also create and nurture institutions to give law meaning and teeth. For the clinical trial data sharing mandate, the key institutions are the FDA and NIH, but they are surrounded by an array of other institutions, some private and some independent but government-funded.

Another key to the success of clinical trial data sharing, in our view, has been the recognition that different components of clinical trial data deserve different treatment. Clinical trial summary data and most metadata are low risk and high reward; they can be shared freely with users without restrictions on access and use. A small fraction of metadata may implicate trade secrecy, but such data can be shared carefully; data use agreements and other constraints preventing competitive use can protect innovative companies' first-mover advantages. Sharing IPD poses profound privacy risks, but IPD too can be

355. That is not to say that the pharmaceutical and medical device industries, or the FDA, have had a perfect track record since 2007. *See* Nicholas S. Downing, Nilay D. Shah, Jenerius A. Aminawung, Alison M. Pease, Jean-David Zeitoun, Harlan M. Krumholz, & Joseph S. Ross, *Postmarket Safety Events Among Novel Therapeutics Approved by the US Food and Drug Administration Between 2001 and 2010*, 317 JAMA 1854 (2017) (surveying safety problems).

shared responsibly with some users, subject to appropriate institutional and technical constraints.

IV. TOWARD A SOCIAL MEDIA DATA SHARING MANDATE

Part IV applies some of the primary lessons learned from clinical trial data sharing and charts a course toward responsible and effective social media data sharing. Section IV.A focuses on how the benefits of independent research cascade, emerge, and are unpredictable at the outset. Section IV.B focuses on the need for regulators. Here we use the term "regulators" to refer to both public and private entities that can impose accountability and exert countervailing power over social media companies by providing alternative forms of expertise, employment, and perspectives. Section IV.C, drawing from the concept of contextual integrity, transposes many of clinical trial data sharing's solutions for navigating the Scylla and Charybdis of trade secrecy and privacy. These solutions apply context-specific controls over social media data to treat contextually and normatively distinct kinds of data differently, using tiered access and a variety of constraints on data access and use tailored to the goals and needs of particular applications.

In our view, clinical trial data sharing's hybrid, "both and" approaches are successful. Various clinical trial data sharing initiatives deploy a mix of mandated sharing and voluntary arrangements, across data types of varying sensitivity, in order to balance the interests of commercial secrecy, individual privacy, and public benefits of research.

Clinical trial data sharing also shows that meaningful independent researcher access cannot be achieved without laws mandating that industry share more data. Clinical trial data's journey from the dark ages to today's robust ecosystem was made possible by the legal transformation of the rights in such data. What began as data governed almost exclusively by private ordering eventually incorporated public demands to constrain those interests and indexed a public right to quality research to provide accountability to a high-stakes sphere of life. Clinical trial data's iterative process of legislation and regulation to enact and build on that legislation was the legal foundation needed to build a robust data sharing ecosystem.

Finally, the example of clinical trial data shows the importance of ensuring that data access mandates do not operate as mere transparency requirements. Laws to grant researcher access must materially and legally empower regulators to avoid this pitfall.

Data access mandates that allow companies to retain either discretionary control over who is granted access or financial control over how the work of access is funded do more harm than good. At best, such proposals will

empower a subset of well-connected and resourced researchers through narrow interpretations of such rules. At worst, such proposals may weaken pressure to impose more substantive regulation over the digital economy.

We do not believe transparency alone can provide a sufficient solution to the larger issues surveyed above in the social media research ecosystem, as the case of SS1 amply demonstrates. As AI Now noted in its 2023 annual report, data access regulation alone is not enough to promote a stronger and more robust independent researcher ecosystem.[356]

A. CASCADING (AND UNPREDICTABLE) BENEFITS OF BASIC RESEARCH

One lesson of clinical trials is that the benefits of basic research are not always obvious before research begins. Benefits are instead cascading and unpredictable. Just because these benefits are not readily apparent at the time access is granted does not mean such benefits will not be significant. (And to be clear, in the case of social media data, many pressing societal benefits for researcher access are already readily apparent, as we have argued in Part II).

Basic research is infrastructural. It is the first step in the process of refining unknown unknowns into known unknowns or known knowns.[357] Basic research provides the scientific building blocks upon which many other forms of research and productive innovation rely. At the outset, the cascading, indirect benefits of basic research are near-impossible to predict because the stuff of value being built on or adapted for commercial use—a useful material or a surprising scientific breakthrough—is not even known to exist at the time.[358] It seems obvious to say, but discovery of the previously unknown is the point of basic research.

The value and unpredictability of discovery are important to emphasize when weighing the potential benefits of researcher access against claims of the risks to secrecy and privacy. Addressing direct, currently known needs are just one of the emergent beneficial properties of the new institutions that will be created to facilitate social media access.

As Part III showed, researchers' access to clinical trial data has led to many cascading benefits: illumination of harms that regulators missed, improved patient care and public health, higher quality trials, combating misinformation,

356. AI NOW, 2023 LANDSCAPE: CONFRONTING TECH POWER 41 (2023), https://ainowinstitute.org/wp-content/uploads/2023/04/AI-Now-2023-Landscape-Report-FINAL.pdf (calling access to data a "weak policy response" to the problem of independent research).

357. To riff on the old chestnut from Donald Rumsfeld. *See* David Pozen, *Deep Secrecy*, 62 STAN. L. REV. 257, 259 (2009).

358. A famous example is the birth of a booming plastics industry following the funding of the space program.

and more. Nonprofit and broadly accessible clinical trial databases, including ClinicalTrials.gov, Drugs@FDA, and BioLINCC expand and democratize access to scientific data.

Reliable and growing access to clinical trial data has also helped to create a cadre of independent researchers able to use that data. Grants from NIH and FDA have contributed to a corps of independent experts able to manage and use this data for public benefit. This material independence in turn has fostered a larger ecosystem of expertise and knowledge production that exists outside of—and largely independent of—the pharmaceutical and medical device industries.

Independent access to social media data, done right, can also empower a greater diversity of researchers with the tools to access this data, and thus conduct scientific research with this resource. Because researchers will no longer need to rely on individual, bespoke relationships with companies, or be willing to assume the legal risk of proceeding without such relationships in place, it is reasonable to assume that greater numbers of researchers from less well-resourced institutions will be able to gain access to social media data. The same goes for researchers that may be interested in U.S. social media data but reside outside of the United States—making this data available to qualified researchers opens up access to a global research community. Indeed, we have already seen a similar benefit of the European Union's recent efforts to grant researchers access to E.U. data; many U.S. researchers are extremely enthusiastic about the research potential of accessing E.U. data.[359]

Robust ecosystems of researcher data access take time to develop. They cannot be achieved in a day. Nevertheless, achieving a successful state of social media data access depends in part on the steps taken now. The cascading benefits of clinical trial data have taken years to realize and are still emerging. We are only at the very beginning of the process of implementing researcher access to social media data, and whether the process realizes its potential depends on the steps taken today.

B. EMPOWERING REGULATORS

To be successful, researcher access laws and policies must create and empower institutions, inside and outside government, with the funding, mandate, and expertise to manage the technical governance mechanisms of research data and to keep social media companies in compliance with existing law and accountable if they are not.

359. See discussion of the Digital Services Act's mandated access for vetted researchers in the Introduction, *supra*.

Legislation to require access and prescribe certain data practices is an important first step. But to produce real results, the experience of clinical trial data sharing suggests that laws also need to empower regulators to engage in the day-to-day work of both keeping social media companies compliant with data sharing requirements and managing the technical governance mechanisms of access.

Empowerment of such regulators means a few different things, and it can take a range of forms. Below we offer a menu of options, a mix of which have been successfully deployed in the clinical trial data setting. Given the early days of social media data sharing, we endorse experimentation, hybridization, and pluralism in approach among the options surveyed below. But the key lesson behind all these options is that social media platforms should not retain gatekeeping (or funding) authority over who is granted access to data, what studies are deemed fundable or feasible, or which results may be published.

1. Independent, Preferably Public, Funding

First, empowered regulators must have access to secure, reliable public funding. Currently, much of the funding (directly or indirectly) for researcher access to social media data is provided by companies themselves. This leaves researchers vulnerable to changes in market forces or company priorities.[360] It also produces a chilling effect on research considered overly critical. It is neither a sustainable model on which to build long-term access nor conducive to robust independent research.

As seen in Section III.D.2.a, public funding does not have to mean servers running under direct government control. Government agencies can and do fund several different institutional models of data curation and sharing. NIH directly funds, manages, and hosts its own databases, including ClinicalTrials.gov and BioLINCC. But the FDA and NIH also provide funding to private data stewards, including YODA. Recipients of public funding can be other public institutions (like public universities or research consortia), private academic or non-profit research institutions, or clusters of all the above (similar to CERSI).

Access mandates that both empower public and civil society institutions with independent funding and foster non-industry expertise in managing and providing access to such data can build these communities' material and intellectual capacity to do their work. Researcher access done right can thus play a key role in fostering the growth of meaningful regulators in the digital economy. As Part III shows, such institutions can play key roles in movement

360. *See, e.g.*, Calma, *supra* note 115.

and coalition building. Free from material dependency on the companies, independent technology research ecosystems can provide the intellectual and civic seeds of the broad political mobilization needed to transform how we develop and manage the digital infrastructures of social and public life.

2. Control Over Standards and Terms of Access and Use

Second, empowered regulators are those that have meaningful control over (1) standards and processes of data sharing and (2) researchers' data access and use. Control over the standards and processes of data sharing means regulators must curate and safeguard data by protecting its quality, accuracy, and useability. Control over researchers' access and use means just that. Control can be effectuated through technical means, contracts (data use agreements), guides and protocols for use, and more.

Regulators can exert control via a range of options that empower them in their relationships with both companies and researchers. At its most simple and direct, institutional control begins with laws that require companies to share certain data with regulators, as seen with ClinicalTrials.gov and in the Canadian example of trusted researcher access in Sections III.C.2 and III.D.2.b. We believe some degree of compulsory data sharing is required to foster successful, independent research. However, as Section III.D more broadly shows, voluntary forms of sharing can supplement mandatory forms, expand the universe of data made available to researchers, and build on their success. As Part III also shows (particularly in Section III.C.2) and as will be discussed below, when companies do not provide the data they are required to share, regulators should also be empowered to enforce sharing requirements.

Importantly, institutional control also means data stewards should be tasked with administering researcher access and use of data to ensure researchers comply with necessary controls and safeguards.

The destination of compelled data can be a government curator, as is the case with ClinicalTrials.gov. This approach is particularly promising for managing datasets on features shared across social media companies, like active users, volume of activity, distribution of that activity, language, and country of origin.

However, curators need not be government entities. In the United States, the FDA funded RDCA-DAP and YODA, two exemplary non-governmental data sharing platforms. Non-governmental options may be particularly attractive for data that is more sensitive to privacy concerns that militate against permitting government agencies the capacity to hold, see, or use such data. Regardless of whether institutions are public or private, they should be given the means to manage data responsibly. This means funding to keep

servers running and curatorial experts employed. This also means: legal rights to determine how data is to be shared from companies; rights to curate and assess data for quality; and rights to set the terms (and/or manage the process) of screening applicants for access via their own data use agreements. Curatorial institutions ought to have the rights to hold data on their own servers, serve as gatekeepers for access to data, and develop internal protocols for screening and evaluating researcher access proposals, including peer-review mechanisms for access to particularly sensitive data.

3. Meaningful Regulatory Enforcement

Part III also highlights the importance of meaningful enforcement of data sharing mandates to ensure compliance. The experience of ClinicalTrials.gov presented in Section III.C.1 suggests both that some enforcement is necessary and that even minimal enforcement through "naming and shaming" a handful of noncompliant entities can spur significant compliance.[361]

One condition of granting private entities data curation roles might be a requirement to regularly report noncompliance to the relevant public regulator. Public data stewards and regulators can be given the capacity to enforce compliance directly via mechanisms like naming and shaming, imposing fines, or a court-enforceable right of action to compel access, to name a few. If public stewards lack authority to enforce the law themselves, then they should at least be able to highlight non-compliance to the public and the relevant regulator.

The experience of clinical trial data sharing shows the modest but meaningful effectiveness of simple "naming and shaming" companies and other entities that withhold data from researchers despite a mandate to share. For instance, the FDAAA Trials Tracker, built by the Bennett Institute for Applied Data Science at Oxford, keeps track of which companies and clinical trials have shared their results as required under FDAAA.[362] For social media, regulation can help remove barriers to third-party development of similar accountability mechanisms.

C. TREATING DIFFERENT DATA DIFFERENTLY

Existing models of clinical trial data sharing show that it is possible to share data with researchers while also protecting data subjects from harm and preserving incentives to innovate. Clinical trial data sharing offers lessons

361. *See supra* Section III.D.1.c, on public institutional governors as regulators and enforcers.

362. As they say on their website, "The FDA are not publicly tracking compliance. So we are, here." *FDAAA Trials Tracker*, BENNET INST. FOR APPLIED DATA SCI., https://fdaaa.trialstracker.net/rankings/ (last visited Jan. 28, 2023).

about the design of both the technology and the law. In both domains, the clinical trial sector has developed data-sharing mechanisms that are specific, contextual, and allow researchers to access useful data while remaining independent.

Valid privacy and trade secrecy concerns should be treated with a scalpel, not a broadsword. In order to do this, data sharing mechanisms need to be tailored to the affordances of the data they offer and the risks posed by that data to data subjects, researchers, and platforms. This basic insight is not new. Scholars including Helen Nissenbaum, Dan Solove, and Neil Richards have argued for some time that theories and applications of information privacy need to be attentive to the contextually specific purposes and norms that both motivate and constrain information sharing.[363]

However, Section III.D.2 shows how the legal, institutional, and technological responses that structured the still-evolving clinical trial data governance regime paralleled—perhaps even prefigured—these theoretical developments in information privacy law. Different tiers and mechanisms of access for different kinds of clinical trial data, users, and uses gradually emerged in response to live policy considerations of how to balance the risks to commercial secrecy and privacy with the social benefits of access. In other words, the solutions that emerged in clinical trial sharing look quite similar to what information privacy theorists have long observed and recommended for digital personal information subject to privacy and other concerns. This Section, IV.C, transposes many of clinical trial data sharing's solutions for navigating the twin barriers of trade secrecy and privacy. In line with existing theories of privacy law, these apply context-specific controls over social media data to treat contextually and normatively distinct kinds of data differently.

To this end, we argue that, as an initial matter, social media should adopt clinical trial data's useful tripartite distinction of data types: individual data, summary data, and metadata. Social media companies tend to lump all these types of data together, raising the lowest common denominator of necessary protection. In other words, all data gets treated with the privacy and security sensitivity of individual data and the trade secrecy sensitivity of metadata, even though certain data—especially summary data—could easily be shared that does not raise those concerns.

When resisting sharing data with researchers, social media companies by and large focus on the promises and pitfalls associated with sharing data about individuals' social media activity. This is evident in their most common

363. *See* NISSENBAUM, *supra* note 184, at 129; DANIEL SOLOVE, UNDERSTANDING PRIVACY 187–89 (2008); NEIL RICHARDS, WHY PRIVACY MATTERS 22–34 (2022).

methods, such as APIs and static data sets, and large data sharing initiatives such as SS1. Yet the same companies provide little information on how these data are generated (metadata) or aggregate data on their users and their activity (summary data).

Without metadata, there are looming questions about the provenance and representativeness of data available to researchers. Without metadata, researchers must trust companies to have answered these questions in their own undocumented methodologies, despite evidence that some of these companies unreliable and unrepresentative data before.[364] For instance, Facebook's Ad Library comes in part from ads that the company's automatic detection algorithm flags as political.[365] However, Facebook does not offer any metadata on what classifiers it uses. Therefore, some entire topics may not be included in the library, and researchers would have no idea.

Without summary data, researchers face difficulty contextualizing their results (e.g., understanding relative effect size) and verifying the numbers they receive from companies. For instance, researchers did not know that nearly half of all data was missing from SS1, or that so many advertisements were mislabeled on Facebook's Ad Library (before the NYU Ad Observatory uncovered it) because it was not possible to see if the numbers made sense.

The minimal metadata and summary data that social media companies do currently provide to researchers lacks the requisite methodological clarity and specificity to be useful. Instagram, for instance, shares some information about how it ranks posts for users' feeds or explore pages, but the information provided is too general to be used in academic research.[366] The primary method companies use to share summary data is content moderation transparency reports, but these contain little information beyond how much content governments have requested be taken down and how often the platform complied.[367] Social media companies keep secret even basic platform usage information such as monthly active users and volume of uploads. For

364. *See supra* Section II.B.

365. *About the Meta Ad Library*, META BUS. HELP CTR., https://www.facebook.com/business/help/2405092116183307?id=288762101909005 (last visited Nov. 23, 2023).

366. *See generally* Adam Mosseri, *Shedding More Light on How Instagram Works*, INSTAGRAM BLOG (June 8, 2021), https://about.instagram.com/blog/announcements/shedding-more-light-on-how-instagram-works.

367. Caitlin Vogus & Emma Llansó, *Report—Making Transparency Meaningful: A Framework for Policymakers*, CTR. FOR DEMOCRACY & TECH. (Dec. 14, 2021), https://cdt.org/insights/report-making-transparency-meaningful-a-framework-for-policymakers/.

instance, the public learned that Instagram passed two billion monthly active users only when journalists leaked the information.[368]

Existing legal frameworks do little better. Proposed laws in the United States and passed laws in the European Union almost always focus on access to individual data, rather than summary data and metadata, and in turn, impose severe limitations to maintain privacy and trade secrecy. The Platform Accountability and Transparency Act, for instance, mostly focuses on sharing individual data with researchers, particularly high-profile users and content moderation actions taken against them. The Ad Transparency Act also focuses on individual ads instead of requiring companies to describe underlying ad targeting systems. And while the Digital Services Act in theory allows researchers to access all three types of data, this data is only available to certain vetted researchers.[369]

Below, we elaborate how not just individual data but summary and metadata on social media could be made available to researchers, and how access could be tailored to accommodate the privacy and trade secrecy considerations of each.

1. Summary Data

Summary data can be used by researchers to better understand who, how, and how many people use social media, while posing little trade secrecy or privacy risk. High level metrics (e.g., number of users, frequency of posts, or time spent on platform) broken down into certain categories (e.g., language or country of origin) can contextualize research and guide directions of future research. And if those categories are standardized, researchers can make comparisons across platforms. Summary data can also reveal self-sorted categories based on individual data, such as how many people use a given hashtag or remix a certain sound clip. For clinical trials, it took years of regulatory battles and clarification to get pharmaceutical and medical device companies to share summary data, but the resulting data sharing paradigm directly benefited the public, including by revealing discrepancies between

368. Salvador Rodriguez, *Instagram Surpasses 2 Billion Monthly Users While Powering Through a Year of Turmoil*, CNBC (Dec. 14, 2021), https://www.cnbc.com/2021/12/14/instagram-surpasses-2-billion-monthly-users.html.

369. Digital Services Act, OJ L 277, 27.10.2022, Article 31. The closest thing to summary data made available to the public is which platforms have enough E.U. users to be considered Very Large Online Platforms and which do not. *See* Digital Services Act, Article 33; *see also* John Albert, *A Guide to the EU's New Rules for Researcher Access to Platform Data*, ALGORITHM WATCH (Dec. 7, 2022), https://algorithmwatch.org/en/dsa-data-access-explained/.

published medical literature and real data and forcing unsafe products off the market.[370]

Summary data that reveals information about narrow subcategories of social media users may be useful to researchers, but greater specificity can raise privacy concerns. Social media companies can similarly offer broad categories of summary data publicly and narrower categories with increased privacy risk only to more vetted researchers.

Summary data sharing initiatives should not require companies to collect data they do not already gather or infer themselves.[371] But companies may have tools for approximating some of this information for their own internal research, which they can readily share with external researchers. For instance, Meta does not collect the race of its users, but it still evaluates the impact of different product changes on different racial groups; a methodology called Bayesian Improved Surname Geocoding makes a prediction about a user's race using their last name and zip code.[372] Meta could potentially give researchers access to this or similar methodologies, or the data they collect from them.

2. Metadata

Social media companies could provide metadata about data they generate internally and share externally with researchers, including how data has been scrubbed or filtered, which data may be missing or overrepresented, and how different systems work. This metadata poses fewer risks to privacy than individual data and variable risks to trade secrecy. These privacy and trade secrecy risks that can be placed on a sliding scale. "Riskier" data can be shared only with trusted researchers, shared subject to stringent data use agreements, and shared subject to technical constraints, such as limits on storage and retransmission of data. We see the clinical trial sector engage in some of this line drawing activity, particularly with NIH's ClinicalTrials.gov and Health Canada's PRCI.[373]

370. Clinical Trials Registration and Results Information Submission, 81 Fed. Reg., *supra* note 188, at 64,982, 65,006.

371. *E.g.*, *Twitter Infers Users Age Rather Than Always Collecting It. Geo, Gender, Language, and Age Targeting*, TWITTER BUS., https://business.twitter.com/en/help/campaign-setup/campaign-targeting/geo-gender-and-language-targeting.html (last visited Jan. 30, 2023).

372. Roy L. Austin, Jr., *Race Data Measurement and Meta's Commitment to Fair and Inclusive Products*, META NEWSROOM (Nov. 18, 2021), https://about.fb.com/news/2021/11/inclusive-products-through-race-data-measurement/.

373. *See* discussion *supra* Sections III.C.2 (ClinicalTrials.gov permits companies to redact metadata they consider trade secrets from public disclosure) & III.D.2.b (Health Canada will share trade secrets with researchers who promise confidentiality and high-value research).

Metadata does pose some real privacy risks. Metadata for social media encompasses a broader range of forms than metadata for clinical trials, and some social media metadata may reveal things about individual users, such as information on the users that initially posted content banned or restricted by a social media platform.

The trade secrecy risks posed to social media companies by sharing metadata likewise vary along on a sliding scale. Divulging methods of how summary data—i.e., statistics on hashtags—get generated is on the low-risk end of the spectrum, as is divulging the methods by which individual data gets produced and organized. Moderation and recommender systems pose greater risk to trade secrecy interests, as does information on systems for evaluating whether features should be rolled out. Metadata on how ad targeting systems work is perhaps still higher risk, as these ad targeting systems are currently social media platforms' main drivers of revenue. This sliding scale moves slowly from what is clearly data *about* data to what is data about how larger systems work. As such, it becomes harder to fit clearly into the category metadata and moves further from the factual parallelism of medical data.

We expect that controlled sharing of metadata from social media companies will yield real public benefits, broadly similar to those achieved by sharing metadata from clinical trials. With clinical trials, for instance, data sharing revealed limitations—even profound problems—with Tamiflu, Paxil, and Vioxx, but improved trust in certain COVID-19 vaccines. Similarly, social media metadata could be used to reveal the harms of some systems, but also to bolster public trust of others.

3. Individual Data

The concerns with individual data are a mirror of the concerns of those with summary data: they are not very likely to implicate trade secrecy concerns but can raise privacy concerns on a sliding scale from moderate to severe. And again, the tactic to manage this variance is to treat different data differently. Clinical trial data sharing initiatives do this very effectively. Clinical trial IPD is made available through tiered, tightly controlled access systems such as BioLINCC and YODA. The level of access provided to researchers and the sorts of research permitted depends on the data, the researchers, the intended research, and the associated privacy risks. More than two tiers of researcher access can exist, beyond one tier for "trusted researchers" and another for the broad public. The tailored access that YODA and BioLINCC provide useful models here.[374]

374. *Infra* Section II.D.

Some social media companies already tier data access, corroborating the notion that it can be done. For example, when Twitter offered its public facing API it had regular, enterprise, and academic versions.[375] Facebook has some data it shares publicly and other data it shares with those who sign an agreement, including now the data from SS1.[376]

The experience of clinical trial data sharing shows that platforms can share more individual data than they already do, and that the stewards of that data can be trusted actors outside of social media companies themselves. Social media companies could, through tiered access data sharing programs, share some of the most sensitive social media data with trusted researchers who commit to avoid harmful uses. This sensitive data includes complete lists of removed posts, individual ad targeting information, and inferred data. Some of the most sensitive social media data that poses the greatest privacy risks, such as personally identifiable information and direct messages, may remain off-limits to even the most trusted researchers.

V. CONCLUSION

Social media is in its data secrecy dark age, just as pharmaceuticals were in previous decades.[377] This Article has traced parallels between clinical trials' past and social media's present. For instance, both have witnessed high-profile crises caused by a lack of accountability and transparency: for clinical trials, Paxil's teen suicides and Vioxx's heart failures; for social media, Cambridge Analytica, the rise of online populism, and the degradation of truth in media and democracy. Just as intrepid health journalists in the 1990s and 2000s used the limited tools they had to shine a light on the shadowy pharmaceutical industry, so too have tech journalists and social media company whistleblowers bravely revealed some of the public consequences of surveillance capitalism and the attention economy. Pharmaceutical, medical device, and social media companies have all adopted similar tactics to appease or deflect popular demand for more information, including limited, cherry-picked "transparency" efforts.

In the past few years, a rash of new federal laws have been proposed that would mandate social media companies to share data with researchers—and, perhaps, bring in the light sufficient to end these dark ages. The Platform Accountability and Transparency Act, for instance, would empower the FTC

375. Adam Torres, *Enabling The Future of Academic Research with the Twitter API*, X DEVELOPER PLATFORM (Jan. 26, 2021), https://developer.twitter.com/en/blog/product-news/2021/enabling-the-future-of-academic-research-with-the-twitter-api.

376. *Infra* Section I.B.

377. *Infra* Section II.B.

to compel social media companies to share data with qualified researchers approved by the National Science Foundation.[378] The Social Media Data Act proposes requiring platforms to create in-depth ad libraries for academic researchers.[379] Other proposed U.S. laws such as the Kids Online Safety Act, the Digital Services Oversight and Safety Act, and the ACCESS Act could also allow researchers to access social media data in other ways.[380]

As of this writing, none of these proposals have become law. They remain the subject of intense debate, even controversy. Social media companies have fought them, just as pharmaceutical and medical device companies fought the legislation that mandates transparency of their clinical trial data. As if on cue, social media companies have invoked privacy and trade secrecy—this Article's "Scylla and Charybdis"—as doctrinal and normative reasons to oppose these proposals.[381]

This Article has argued it is possible for legislation and regulation to protect privacy and trade secrecy while simultaneously mandating and mediating researcher access to sensitive data. The precedent of clinical trial data sharing reveals both some pitfalls that await lawmakers seeking to create an effective social media data sharing mandate and some paths to avoid them. Even when clinical data sharing rules were enacted into law, it took years of rulemaking, enforcement, and public pressure to get pharmaceutical companies to actually share their data. And though those battles continue today, the fight has produced safer medical products. For those regulating social media in the United States, the history of sharing clinical trial data shows that merely requiring data access, as legislative proposals do now, is necessary but not sufficient: law also needs to empower regulatory institutions that can enforce those laws and tailor data sharing systems to narrowly manage the privacy and trade secrecy risks that accompany each data type.

In Part IV, we have done our best to distill useful lessons for governance of social media. Undoubtedly many readers will disagree that these are the right lessons. We hope, at very least, that the "thick" accounts of the need for researcher access to social media data and the history of clinical trial data sharing offered in Parts II and III inspire readers to make their own comparisons and derive their own lessons.

---

378. Platform Accountability and Transparency Act, S. 5339, 117th Cong., § 7 (2022).

379. Social Media DATA Act, H.R.3451, 117th Cong., § 2 (2022).

380. Kids Online Safety Act, S.3663, 117th Cong., § 7 (2022); Digital Services Oversight and Safety Act, H.R.6796, 117th Cong., §§ 6, 10 (2022); ACCESS Act of 2021, H.R.3849, 117th Cong., § 3 (2022).

381. *See supra* Section II.E.

Social media companies cast their industry as sui generis, one too complex and innovative for transparency regulation. But what is old is new again: social media is replaying some of the familiar beats of the sixty-plus-year battle for clinical trial data transparency. Social media is changing our world and our institutions in ways that we may not have sixty years to learn to counter. Researchers need better access to social media data to help us navigate this brave new world. We hope that lessons from the clinical trial precedent will help.