THE DATA-POOLING PROBLEM

Michael Mattioli[†]

ABSTRACT

American innovation policy as expressed through intellectual property law contains a curious gap: it encourages individual research investments, but does little to facilitate cooperation among inventors, which is often a necessary precondition for innovation. This Article provides an in–depth analysis of a policy problem that relates to this gap: increasingly, public and private innovation investments depend upon the willingness of private firms and institutions to cooperatively pool industrial, commercial, and scientific data. Data holders often have powerful disincentives to cooperate with one another, however. As a result, important research that the federal government has sought to encourage through intellectual property policy and through other targeted investments is being held back.

This Article addresses this issue by offering three contributions—one theoretical, one empirical, and one prescriptive. The theoretical contribution builds upon legal, economic, and public choice literature to explain why pooling data is relevant to innovation policy and why the level of data sharing in some settings may be suboptimal. This discussion offers a conceptual framework for scholars and policymakers to examine how data–pooling relates to innovation policy goals.

This leads to the second contribution: an ethnographic study of private efforts to pool data in an important field of research. This Article focuses on the field of cancer treatment because it is one of the most active areas where efforts to pool data have recently coalesced. Interviews with lawyers, executives, and scientists working at the vanguard of "Big Data" projects in the field of cancer treatment offer a detailed view of how, precisely, data–pooling problems can hinder technological progress. The study's most significant finding is that impediments to the pooling of patient treatment and clinical trial data are diverse, nuanced, and not reducible to collective action problems that are already well understood by legal scholars and economists, such as the free–rider dilemma.

DOI: https://dx.doi.org/10.15779/Z38R785P10

† Associate Professor of Law, Indiana University Maurer School of Law (Bloomington). I wish to express my thanks to Pamela Samuelson, Katherine Strandburg, Michael Madison, Brett Frischmann, Robert Merges, Arti Rai, Paul Ohm, Ted Sichelman, Mark Janis, Marshall Leaffer, Yvonne Cripps, Fred Cate, Gideon Parchomovsky, Nicholson Price, Brian Broughman, Jeff Stake, Daniel Cole, Jessica Eaglin, and Victor Quintanilla, all of whom offered valuable input. I also wish to thank organizers of the following events: The Second Thematic Conference on Knowledge Commons (NYU School of Law), The Creativity Without Law Conference (Case Western Reserve University School of Law), The Vincent and Elinor Ostrom Workshop (Indiana University). Finally, my thanks go to Indiana University's Office of the Vice Provost for Research for providing critical financial support.

These findings lead to the third key contribution: a set of targeted policy suggestions designed to facilitate data pooling through regulatory action, amendments to federal healthcare legislation, and tax incentives. These prescriptive measures are tailored to address the sharing of health–related data, but they capture an approach that could be applied in other settings where technological progress depends upon data–pooling. Ultimately, this Article argues for a vision of innovation policy in which cooperative exchanges of data are recognized as important preconditions for innovation that may require government support.

TABLE OF CONTENTS

I. INTRODUCTION

This Article explores a policy problem at the intersection of federal innovation policy[1] and "Big Data."[2] The problem's outlines are starkly simple: the U.S. government seeks to promote innovation primarily by

1. Legal scholars and economists commonly use the term "innovation policy" to refer to policy interventions designed to bring about technological advances. *See, e.g.*, Brett M. Frischmann & Mark A. Lemley, *Spillovers*, 107 COLUM. L. REV. 257 n.26 (2007) (discussing the role of pecuniary spillovers in innovation policy); Adam B. Jaffe, Josh Lerner & Scott Stern, *Introduction*, *in* 1 INNOVATION POLICY AND THE ECONOMY (2000) (describing innovation policy as encompassing "longstanding issues, such as the appropriate level and form of public support of research, as well as . . . intellectual property and the appropriate antitrust treatment of . . . industries where technology standards play a key role."); Clarisa Long, *Patent Signals*, 69 U. CHI. L. REV. 625, 675 (2002) (discussing the role of the patent system in innovation policy). *See generally* NICHOLAS S. VONORTAS ET AL., INNOVATION POLICY: A PRACTICAL INTRODUCTION (Nicholas S. Vonortas, Phoebe C. Rouge & Anwar Aridi eds., 2014).

2. "Big Data" describes the practice of drawing new and valuable insights from large datasets that typically hold little independent value. *See generally* VIKTOR MAYER-SCHÖNBERGER & KENNETH CUKIER, BIG DATA: A REVOLUTION THAT WILL TRANSFORM HOW WE LIVE, WORK, AND THINK 1–18 (2013) (discussing the Big Data phenomenon generally).

encouraging individual research investments, but cooperation among information holders is an important precondition for innovation in many settings.[3] An increasingly important form of such cooperation is the aggregation (i.e., pooling) of data held by firms, institutions, and individuals. This Article examines the forces that discourage data pooling in an important field of research—oncology treatment—and considers how the federal government might intervene to better promote innovation.

Scholars have long recognized that innovation can blossom from combinations of technological information, such as industrial knowledge and descriptions of inventions. In the 1980s, the economists Richard Nelson and Sidney Winter wrote, "[I]nnovation in the economic system—and indeed the creation or any sort of novelty in art, science, or practical life—consists to a substantial extent of a recombination of conceptual and physical materials that were previously in existence."[4] Nearly sixty years earlier, Joseph Schumpeter described the same idea when he defined innovation as the discovery of new relationships between previously combined conceptual components.[5] Today, scholars of industrial organization call this form of technological advancement, "recombinant innovation."[6]

Recently, it seems that a new kind of recombinant innovation has drawn widespread attention: Big Data. Rather than drawing upon combinations of technological information, Big Data draws upon combinations of factual information. Credit card receipts, Internet search histories, and patient

3. For a fascinating analysis of the importance of health insurers in creating new information that is valuable to research and innovation, see Rebecca S. Eisenberg & W. Nicholson Price II, *Promoting Healthcare Innovation on the Demand Side* (U. of Mich. Law, Public Law and Legal Theory Research Paper Series No. 503, 2016), http://ssrn.com/abstract=2766707. This intersectional or combinational type of innovation is distinct from "cumulative innovation" which scholars often describe through reference to Isaac Newton's "shoulders of giants" trope. *See infra* Section I.A; *see also* Oren Bar-Gill & Gideon Parchomovsky, *The Value of Giving Away Secrets*, 89 VA. L. REV. 1857, 1858 (2003) ("Cumulative innovation characterizes most industrial sectors."); Suzanne Scotchmer, *Standing on the Shoulders of Giants: Cumulative Research and the Patent Law*, 5 J. ECON. PERSP. 29, 37 (1991) ("Most innovators stand on the shoulders of giants, and never more so than in the current evolution of high technologies, where almost all technical progress builds on a foundation provided by earlier innovators.").

4. RICHARD R. NELSON & SIDNEY G. WINTER, AN EVOLUTIONARY THEORY OF ECONOMIC CHANGE 130 (1982).

5. *See* JOSEPH A. SCHUMPETER, BUSINESS CYCLES: A THEORETICAL, HISTORICAL, AND STATISTICAL ANALYSIS OF THE CAPITALIST PROCESS (1939).

6. ANDREW HARGADON, HOW BREAKTHROUGHS HAPPEN: THE SURPRISING TRUTH ABOUT HOW COMPANIES INNOVATE 31–52 (2003) (terming this characteristic "recombinant innovation").

treatment records from hospitals are standard examples of information that, in earlier times, was treated as digital detritus. Recently, however, computer scientists, statisticians, and engineers who specialize in data science have developed useful algorithms with the help of vast sets of such data.[7]

The value of Big Data algorithms lies in their predictive power. Just as weather forecasting techniques can determine the likelihood of a storm, Big Data algorithms may soon be able to predict the likelihood that a crime will occur on a particular street corner, the odds that a consumer will purchase a product after seeing an online advertisement, or whether a cancer patient will respond well to a particular treatment.[8] Importantly, experts believe that the most valuable applications of Big Data—"the real deal," as one expert interviewed for this Article called it—will be drawn from sets of data from different sources.[9] As one technology commentator recently put it, "when

7. *See, e.g.*, IAN AYRES, SUPER CRUNCHERS 88–111 (2007) (discussing the how the aggregation of patient records combined with analytics could improve healthcare); STEPHEN BAKER, THE NUMERATI 98–99 (2008) (describing the usefulness of gathering large amounts of data for analysis); MAYER-SCHÖNBERGER & CUKIER, *supra* note 2 (describing Big Data and its implications for science, industry, and society); ERIC SIEGEL, PREDICTIVE ANALYTICS: THE POWER TO PREDICT WHO WILL CLICK, BUY, LIE, OR DIE 203 (2013) (exploring Big Data's impact on society); PATRICK TUCKER, THE NAKED FUTURE: WHAT HAPPENS IN A WORLD THAT ANTICIPATES YOUR EVERY MOVE? 183–202 (2014) (discussing the use of data analytics to predict crime); BIG DATA IS NOT A MONOLITH (Cassidy R. Sugimoto, Hamid R. Ekbia & Michael Mattioli eds.) (2016) (presenting a multifaceted picture of Big Data, including discussions about its impact on different academic disciplines and industrial domains).

8. *See supra* note 7; *see also* Jessica M. Eaglin, *Constructing Recidivism Risk for Sentencing*, 67 EMORY L.J. (forthcoming 2017), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2821136 (discussing predictive data–based tools used in connection with criminal sentencing).

9. *See, e.g.*, U.S. Patent No. 7,444,655 (listing Microsoft as assignee on "Anonymous Aggregated Data Collection"). *See generally* MAYER-SCHÖNBERGER & CUKIER, *supra* note 2; Christine Borgman, *The Conundrum of Sharing Research Data*, 63 J. AM. SOC'Y FOR INFO. SCI. & TECH. 1059 (2012) (discussing the importance of data aggregation). For reasons explained in Part II, the ethnographic study at the heart of this Article focuses on efforts to pool cancer research data. A useful example of the power of aggregating and analyzing large sets of data in a different domain is "Google Photos," a service offered by Google Inc. By pooling and analyzing image data contained in photographs contributed by many users, Google engineers have developed algorithms that can recognize objects in photographs with unprecedented accuracy. *See* Joshua A.T. Fairfield, *Mixed Reality: How the Laws of Virtual Worlds Govern Everyday Life*, 27 BERKELEY TECH. L.J. 55, 97 (2012) (describing Google's "particularly aggressive" photo scraping and recognition capability).

we recombine . . . multiple datasets together, that sum . . . is worth more than its individual ingredients."[10]

Experts in this new field believe the best way to aggregate data for this purpose is through cooperative licensing—i.e., pooling. In simple terms, a central administrator collects data from multiple data holders, analyses the data, and then delivers helpful insights back to the contributors and possibly licenses the data to third–parties.[11] There is an intuitive appeal to this model: patent holders have long pooled related inventions to facilitate the production and development of new technologies.[12] This mode of cooperation has been helpful to patent licensors and licensees alike by dramatically reducing transaction costs, which are thought to be a chief impediment to patent licensing in settings where ownership of related patents is highly dispersed.[13] Without explicitly referring to patent pools, experts in some industries and fields of research have recently attempted to organize institutions similarly structured around the goal of collective data licensing.[14]

Any apparent similarities between pooling patents and pooling data might be superficial, however. For one thing, contributing data to a pool can entail steep up–front costs. It is costly to record, organize, and store vast amounts of data on an ongoing basis.[15] It is even costlier to ensure that data is accurate, that it is disclosed in a manner that comports with various laws and regulations (pertaining to privacy, for instance), that its provenance and

10. MAYER-SCHÖNBERGER & CUKIER, *supra* note 2, at 108; *see also* Borgman, *supra* note 9, at 1070 ("Indeed, the greatest advantages of data sharing may be in the combination of data from multiple sources, compared or 'mashed up' in innovative ways.").

11. The process is highly analogous to patent pools for standards–essential patents—patents which are required to implement an industry standard—where groups of patent holders pool patents under the supervision of a third party administrator. Kassandra Maldonado, *Breaching RAND and Reaching for Reasonable: Microsoft v. Motorola and Standard-Essential Patent Litigation*, 29 BERKELEY TECH. L.J. 419, 446 (2014).

12. *Id.* (describing effective use of patent pools to facilitate licensing standard essential patents).

13. Patent pools are federations of patent holders that reduce transaction costs by collectively licensing complementary patent rights under unified agreements. For a deep historical view of patent pooling, see Michael Mattioli, *Power and Governance in Patent Pools*, 27 HARV. J.L. & TECH. 421 (2014); FLOYD L. VAUGHAN, THE UNITED STATES PATENT SYSTEM (1956); OLIVER E. WILLIAMSON, THE ECONOMIC INSTITUTIONS OF CAPITALISM 20–22 (1985); Robert P. Merges, *Contracting into Liability Rules: Intellectual Property Rights & Collective Rights Organizations*, 84 CALIF. L. REV. 1293, 1393 (1996).

14. *See, e.g.*, *infra* Part III (presenting the results of an ethnographic study of oncology data pools).

15. Priscilla M. Regan, *Federal Security Breach Notifications: Politics and Approaches*, 24 BERKELEY TECH. L.J. 1103, 1108 (2009).

pedigree are adequately documented and disclosed, and so forth.[16] Patent holders also incur upfront costs, of course, in the form of research and patent prosecution expenses. From an ex ante perspective, however, the value of developing a patentable invention does not typically hinge upon eventual membership in a patent pool; rather, it turns on the underlying value of the invention itself. In contrast, as this Article explains, some types of data are useful *only* when aggregated. As a result, if a thriving pool does not already exist at the time the data is generated, weak incentives may exist for the relevant data holder to maintain the data and prepare it for pooling. The work of Nobelists in economics and political theory also suggests that data holders will be reluctant to pool their data due to the risk of free riders—i.e., third parties who obtain and benefit from the data without compensating the data holders.[17] Put simply, there are some compelling reasons to expect that efforts to pool useful data will flounder due to widespread "social dilemmas."[18]

Anecdotal evidence, while limited, supports these doubts. Christine Borgman, a leading computer science commentator, recently wrote that "[t]he 'dirty little secret' behind the promotion of data sharing is that not much sharing may be taking place."[19] Recent journalistic accounts of Big Data often explain the potential benefits of data pooling, but rarely cite examples of such cooperation in practice.[20] Instead, most real–world

16. *See, e.g.*, Michael Mattioli, *Disclosing Big Data*, 99 MINN. LAW REV. 525 (2014). The costs of sharing cancer research data are specifically outlined *infra* in Section III.C.; *see also* Part II, *infra* (discussing HIPAA).

17. This concern stems from a vast body of scholarship, both theoretical and empirical, concerning information generation and sharing. *See, e.g.*, Joel Mokyr, *The Commons of Knowledge: A Historical Perspective*, 4 ANN. PROC. WEALTH & WELL-BEING NATIONS 29 (2012), https://www.beloit.edu/upton/assets/4_MOKYRpages29_44.pdf. Leading intellectual property scholars addressed this concern in reference to scientific research data in the 1990s. *See infra* notes 119, 121 and accompanying text (discussing the work of Pamela Samuelson, J.H. Reichman, and Paul Uhlir). Under U.S. law, license agreements and trade secrets law offer data holders potential recourse against parties who directly misappropriate data. Data holders also can attempt to prevent data misappropriation through physical or electronic barriers, such as encryption. *See infra* Section II.A.

18. *See, e.g.*, KENNETH J. ARROW, SOCIAL CHOICE AND INDIVIDUAL VALUES 59 (1951) (discussing the defects of voting procedures); Robyn M. Dawes & David M. Messick, *Social Dilemmas*, 35 INT'L J. PSYCHOL. 111, 111 (2000) (defining social choice problems or social dilemmas as "situations in which each member of a group has a clear and unambiguous incentive to make a choice that—when made by all members—provides poorer outcomes for all than they would have received if none had made the choice.").

19. Borgman, *supra* note 9, at 1060 (identifying problems and limitations).

20. *See, e.g.*, James Glantz, *Is Big Data an Economic Dud?*, N.Y. TIMES (Aug. 17, 2013), http://www.nytimes.com/2013/08/18/sunday-review/is-big-data-an-economic-big-

examples of Big Data involve large, vertically integrated corporations looking inward and drawing insights from the data they already hold.[21] Highly publicized reports of a recent failed effort to pool health data in the United Kingdom seem to lend credence to these doubts.[22]

If these concerns are empirically supported, policymakers should be concerned.[23] The federal government has invested heavily in the future development of Big Data. In March 2012, the White House announced a federal initiative under which six agencies committed over $200 million in funds to advance the development of "tools and techniques needed to access, organize, and glean discoveries from huge volumes of data."[24] In a press release, the White House likened this initiative to earlier federal efforts that led to "advances in supercomputing and the creation of the

---

dud.html (discussing an emerging view held by economists that Big Data is not living up to its promise); Gary Marcus & Ernest Davis, *Eight (No, Nine!) Problems with Big Data*, N.Y. TIMES, Apr. 6, 2014 (exploring the limitations of Big Data and advocating a more tempered view of the phenomenon); Barnard Marr, *Where Big Data Projects Fail*, FORBES (Mar. 17, 2015, 12:28 PM), https://www.forbes.com/sites/bernardmarr/2015/03/17/where-big-data-projects-fail/ (arguing that many Big Data projects will ultimately fail to deliver); MATHIEU COLAS ET AL., CAPGEMINI CONSULTING, CRACKING THE DATA CONUNDRUM: HOW SUCCESSFUL COMPANIES MAKE BIG DATA OPERATIONAL (2014), https://www.capgemini-consulting.com/resource-file-access/resource/pdf/big_data_pov_03-02-15.pdf (reporting the results of a survey of 226 executives across several industries—e.g., retail, financial services, energy and utilities, and pharmaceuticals—in which only 27% of the organizations described their Big Data investments as "successful" and only 8% reported them as "very successful").

21. Authors frequently cite Target's analysis of its own customer records to predict when a customer's purchasing habits indicate they may be pregnant, Google's use of its own search records to predict the needs and wants of users, and Netflix's "mining" of its customers' viewing habits to make film recommendations and to develop its own programming. *See, e.g.*, David Carr, *Giving Viewers What They Want*, N.Y. TIMES (Feb. 24, 2013), http://www.nytimes.com/2013/02/25/business/media/for-house-of-cards-using-big-data-to-guarantee-its-popularity.html; Charles Duhigg, *How Companies Learn Your Secrets*, N.Y. TIMES MAG. (Feb. 16, 2012), http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html; GOOGLE, *Flu Trends: How Does This Work*, https://www.google.org/flutrends/about/ (last visited Aug. 12, 2017).

22. *See, e.g.*, *infra* notes 117–18 and accompanying text (discussing the failed UK National Health Service's National Programme for IT).

23. As explained in Part II, this problem can be understood as one of underuse caused by a proliferation of exclusionary rights—i.e., a "tragedy of the anticommons." The term is typically used in reference to patent rights, but here the term is adopted to describe the possibly similar phenomenon with respect to data. *See, e.g.*, David E. Adelman, *A Fallacy of the Commons in Biotech Patent Policy*, 20 BERKELEY TECH. L.J. 985, 993 (2005) (describing the "anticommons" theory).

24. Press Release, White House, Obama Administration Unveils "Big Data" Initiative: Announces $200 Million in New R&D Investments (Mar. 29, 2012), https://obamawhitehouse.archives.gov/the-press-office/2015/11/19/release-obama-administration-unveils-big-data-initiative-announces-200.

Internet."[25] The White House later commended dozens of early–stage data–sharing ventures that "answered the President's call" for cooperative partnerships designed to accelerate the development of Big Data. (Two of these initiatives are subjects of the study in Part III of this Article.)[26] In early 2016, President Obama tasked Vice President Joe Biden to head "The Cancer Moonshot"—a federal initiative to promote the development of cancer cures, in part, by providing an infrastructure for sharing genomic information.[27] If data held by non–public actors such as private hospitals, corporations, and individuals is not widely pooled, however, then the effectiveness of the government's targeted investments in this area may be limited.

Hindrances to private data pooling could also undermine the federal government's broader goals of promoting technological progress. A primary goal of the patent system, for instance, is to encourage research investments in a multitude of technological fields—including those fields where Big Data may be poised to spur innovation, such as oncology treatment.[28] Because recent Supreme Court jurisprudence has called the patentability of some kinds of algorithms into question, some types of Big Data algorithms may not be patentable.[29] Nevertheless, a data pool could open the door to an algorithm, which in turn could open many more doors to related patentable inventions regardless of the original algorithm's patentability.[30] In short, there is widespread agreement among experts in

25. *Id.*

26. Press Release, Executive Office of the President, "Data to Knowledge to Action" Event Highlights Innovative Collaborations to Benefit Americans (Nov. 12, 2013), https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/Data2Action%20Press%20Release.pdf.

27. Laurie McGinley, *Biden to Tackle Broad Range of Cancer Issues, Including Drug Prices, After Leaving White House*, WASH. POST (Jan. 4, 2017), https://www.washingtonpost.com/news/to-your-health/wp/2017/01/04/biden-to-tackle-cancer-drug-prices-as-part-of-post-white-house-moonshot-work/; *Genomic Data Commons Data Portal*, NAT'L CANCER INST., https://portal.gdc.cancer.gov/ (last visited Aug. 12, 2017).

28. For an important related discussion of the intersection of patent law and data, see Brenda M. Simon & Ted Sichelman, *Data-Generating Patents*, 111 NW. U. L. REV. 377 (2017).

29. *See* Alice Corp. Pty. Ltd. v. CLS Bank Int'l, 134 S. Ct. 2347 (2014) (indicating that a software method may not be patentable if it covers only an abstract idea, and if there is no additional "inventive concept" that applies to the underlying abstraction); *see also* Robert Merges, *Symposium: Go Ask* Alice *— What Can You Patent After* Alice v. CLS Bank*?*, SCOTUSBLOG (June 20, 2014), http://www.scotusblog.com/2014/06/symposium-go-ask-alice-what-can-you-patent-after-alice-v-cls-bank/ (discussing the implications of the *Alice* case on patentability).

30. *See infra* Part IV.

many fields that data pooling is an important precondition for technological progress. The pooling of data is, as a result, relevant to policies that seek to promote innovation.[31]

After explaining the policy relevance of data pooling in greater detail, this Article drills down to explore how it is affecting just one field of research: cancer treatment. This field was selected because it is one of the most active areas where efforts to pool data have recently coalesced.[32] Interviews with lawyers, executives, and scientists working at the vanguard of Big Data projects in the field of oncology offer a detailed view of how precisely data–pooling problems can hinder technological progress.

The interviews conducted for this study reveal that, contrary to conventional wisdom, the data–pooling problem as it affects cancer treatment is not reducible to either a free–rider dilemma or privacy concerns. Rather, the impediments are contextual: concerns over professional, competitive, and reputational standing, for instance, are powerfully discouraging the pooling of cancer treatment and research data. According to subjects interviewed, hospitals and other healthcare providers are reluctant to share data that might reflect poorly on the quality of service they provide; pharmaceutical companies, meanwhile, wish to closely guard data that could reduce the value of existing or future intellectual property, including patents and trade secrets; many academic researchers are loath to share data that could fuel publications; individual patients, meanwhile, may feel hesitant to share data that could expose them to various forms of discrimination.[33]

Drawing upon these findings and others, this Article considers whether the government should seek to promote technological progress by encouraging data pooling. The Article then presents a menu of possible interventions that address some of the impediments uncovered by the study. These suggestions are specific to the field of cancer treatment and research,

---

31. Relatedly, the Defend Trade Secrets Act of 2016 aims to "incentivize future innovation." S. REP. NO. 114-220, at 3 (2016), https://www.congress.gov/congressional-report/114th-congress/senate-report/220/1. Trade secret law offers a less compelling example of government policy that would be undermined by data–sharing problems, however: this is because trade secrecy may discourage the disclosure of data collection preparation methods, which is necessary for the useful exchange of data. I explored this topic in an earlier article. *See* Mattioli, *supra* note 13.

32. *See supra* notes 24–28.

33. A topic outside the scope of this study is whether and how technological norms and infrastructures impede data sharing. Such norms may include, for instance, mandatory registration or "login" procedures that offer little to no value to data holders and that, in the aggregate, impose high transaction costs on innovative data aggregation. Telephone Interview with Anonymous Subjects #7 and #8, Nat'l Insts. of Health (NIH) (Oct. 21, 2014).

but they capture a methodology that could be helpful in other settings as well. A broad goal of this Article is to encourage similar studies of data pooling in other industries and research settings.

Part II of this Article explains how the aggregation of information (including data) relates to federal innovation policy. This background discussion helps situate the Big Data phenomenon within a policy framework and explains the theoretical basis for expecting pooling efforts to run up against collective action problems. Part III presents an original ethnographic study of data pooling in the field of cancer research. As briefly noted above, discussions with individuals involved in several data pooling projects reveal that some of the most significant problems data pools face are more complex and nuanced than theory predicts. Part IV considers the appropriateness of a policy response to the problems uncovered by the study. The discussion then offers a set of policy proposals designed to address some of the impediments to pooling uncovered by the study. There is no "one–size–fits–all" solution, however. This closing discussion argues for a view of innovation policy in which cooperative data pooling is regarded as an important precondition for innovation that may sometimes require government intervention. Part V concludes.

II. A BIG DATA ANTICOMMONS?

Big Data is a promising new platform for innovation that often requires, as a prerequisite, the aggregation of data held by different firms, institutions, and individuals. There is reason to doubt that such aggregation will occur frequently or broadly enough to be meaningful, however, without support from the government.[34] Scholars from the fields of law, economics, public choice, and other disciplines have theorized that information of many kinds (including data) is ill–suited to widespread exchange.[35] The chief problems, these theorists believe, are that valuable information is often costly to prepare for exchange and highly subject to free–riding, as well as being risky to share when doing so might run afoul of privacy laws.[36] Anecdotal accounts from the front lines of Big Data seem to support this prediction.[37] Because barriers to data pooling stand to undermine federal innovation policy goals, policymakers should be concerned, and should explore potential corrective steps.

34. *See infra* Section II.B. (discussing barriers to naturally emerging data pools).
35. *See, e.g.*, *id.*
36. *Id.*
37. *Id.*

A. INFORMATION EXCHANGE AND INNOVATION POLICY

Combinations of technological information, such as industrial knowledge and descriptions of inventions, can fuel innovation.[38] Joseph Schumpeter described this phenomenon when he wrote that technological innovation involves "combin[ing] factors in a new way," or through "carrying out new combinations" of ideas.[39] The economic historian Abbott Payson Usher similarly called innovation "the constructive assimilation of preexisting elements into new syntheses, new patterns, or new configurations."[40] Innovation, Usher explained, "establishes relationships that did not previously exist."[41] The esteemed economists Richard R. Nelson and Sidney G. Winter echoed Usher and Schumpeter's views.[42] Nobelist Kenneth Arrow similarly viewed technological information held by different firms as "the major input" of inventive activity, "apart from the talent of the inventor."[43] As explained in Part I, experts in some fields call this type of technological advancement "recombinant innovation."[44]

---

38. Some of history's preeminent scientists and inventors have described the act of invention as a process of combining and repurposing existing information. In a 1908 essay, the famed mathematician Henri Poincaré explained that he discovered new mathematical relationships not by happening upon them at chance, but rather by deliberately combining well–known mathematical concepts ("entities" as he called them) until useful combinations revealed themselves. HENRI POINCARÉ, SCIENCE AND METHOD 50–51 (Francis Maitland trans., Courier Corp. 1914) ("What, in fact, is mathematical discovery? . . . Discovery consists precisely in not constructing useless combinations, but in constructing those that are useful, which are an infinitely small minority. Discovery is discernment, selection. . . . Among the combinations we choose, the most fruitful are often those which are formed of elements borrowed from widely separated domains."). Thus, although Poincaré's creative process was highly structured, it was not mechanical. "The real work . . . is not merely . . . manufacturing as many combinations as possible," but rather "in choosing between these combinations with a view to eliminating those that are useless." *Id.* at 57.

39. SCHUMPETER, *supra* note 5, at 84, 88.

40. ABBOTT PAYSON USHER, A HISTORY OF MECHANICAL INVENTIONS 11 (1929). Interestingly, Usher relied upon this definition to explain mechanical inventions as well as the creation of artistic works. *Id.*

41. *Id.*

42. SCHUMPETER, *supra* note 5, at 130.

43. Kenneth Arrow, *Economic Welfare and the Allocation of Resources for Invention*, *in* THE RATE AND DIRECTION OF INVENTIVE ACTIVITY: ECONOMIC AND SOCIAL FACTORS 618 (1962); *see also* Robert P. Merges, *Institutions for Intellectual Property Transactions: The Case of Patent Pools*, *in* EXPANDING THE BOUNDARIES OF INTELLECTUAL PROPERTY: INNOVATION POLICY FOR THE KNOWLEDGE SOCIETY 125 (Rochelle Dreyfuss et al. eds., 2000), https://www.law.berkeley.edu/files/pools.pdf ("Arrow set the stage for a new type of theory, one that recognized the need to assemble information and property rights from disparate sources in the process of bringing a product to market.").

44. *See, e.g.*, ANDREW HARGADON, SUSTAINABLE INNOVATION: BUILD YOUR COMPANY'S CAPACITY TO CHANGE THE WORLD 127–47 (2015).

It is helpful to note that recombinant innovation is distinct from "incremental innovation"—a form of technological advancement frequently discussed by intellectual property scholars.[45] Incremental innovation refers to a vertical process of improving upon existing technologies. Recombinant innovation, by contrast, involves the horizontal assembly of complementary technological information from different sources.[46] The invention of the mimeograph machine is a paradigmatic example. Thomas Edison developed the device by combining the idea of a printing press with that of a rapidly moving stylus mechanism used in automatic telegraph machines.[47] Edison did not invent these components nor did he improve upon them; rather, he combined them in a useful and complementary way.[48]

Although intellectual property scholars rarely use the term, the patent system encourages recombinant innovation. A central goal of the patent system is to encourage technological progress generally.[49] Even more specifically though, patent law's "obviousness bar" denies patent protection

45. *See, e.g.*, Rebecca S. Eisenberg, *Patents and the Progress of Science: Exclusive Rights and Experimental Use*, 56 U. CHI. L. REV. 1017, 1055 (1989) ("Scientists have been proclaiming their indebtedness to the research of their predecessors for centuries"); Mark A. Lemley, *The Economics of Improvement in Intellectual Property Law*, 75 TEX. L. REV. 989, 997 (1997) ("Rather, knowledge is cumulative—authors and inventors must necessarily build on what came before them."); Suzanne Scotchmer, *Standing on the Shoulders of Giants: Cumulative Research and the Patent Law*, 5 J. ECON. PERSP. 29, 29 (1991) ("Most innovators stand on the shoulders of giants, and never more so than in the current evolution of high technologies, where almost all technical progress builds on a foundation provided by earlier innovators."); Arti Kaur Rai, *Regulating Scientific Research: Intellectual Property Rights and the Norms of Science*, 94 NW. U. L. REV. 77, 90 (1999) ("The communal character of science is also manifested in a recognition by scientists of their dependence upon a cumulative cultural heritage.").

46. This is a somewhat stylized dichotomy: many inventions embody a mixture of incremental and recombinant innovation. Scholars have defined innovation broadly; for example, innovation has been defined as "the search for, and the discovery, development, improvement, and adoption of new processes, new products, and new organizational structures and procedures." Thomas M. Jorde & David J. Teece, *Innovation, Cooperation and Antitrust*, 4 BERKELEY TECH. L.J. 1, 5 (1989); *see also* Andrew Hargadon, *Brokering Knowledge: Linking Learning and Innovation*, 24 RES. ORG. BEHAV. 41, 44 (2002) (presenting case studies that illustrate this point).

47. JAN FAGERBERG ET AL., THE OXFORD HANDBOOK OF INNOVATION MANAGEMENT 171 (2014).

48. *See id.* Similar historical examples are legion. As Lee Fleming and Olav Sorenson have noted, "one might think of the automobile as a combination of the bicycle, the horse carriage, and the internal combustion engine. The steamship can be characterized as combining the boat with steam power." Lee Fleming & Olav Sorenson, *Technology as a Complex Adaptive System: Evidence From Patent Data*, 30 RES. POL'Y 1019, 1020 (2001).

49. Dan L. Burk, The Role of Patent Law in Knowledge Codification, 23 Berkeley Tech. L.J. 1009, 1009–10

to inventions that "would have been obvious . . . to a person having ordinary skill in the art to which the claimed invention pertains."[50] This rule denies patent protection to inventions that cover obvious combinations of prior art (technological information), while favoring inventors who combine prior art in ways that are unexpected or that were previously discouraged.[51] Patent law's obviousness bar does not establish a rule that *all* inventors must combine existing technological information creatively, of course, but it clearly rewards inventors who do.[52] In this way, it seems to reflect a policy judgment that the public stands to benefit when technological information is combined in unexpected ways.[53]

Saurabh Vishnubhakat and Arti Rai recently offered a valuable empirical view of this phenomenon:

> Because the USPTO assigns relevant USPC classifications to each patent, a patent's classes identify the distinct technologies that the inventor combined to produce the invention—and the combination identifies the particular interdisciplinarity at work in that instance of inventive activity. Historically, the rate at which

50. Laura G. Pedraza-Fariña, *Patent Law and the Sociology of Innovation*, 2013 WIS. L. REV. 813, 818 (2013); *see also id.* at 861 ("First, the obviousness inquiry should be structured so as to reward, and thus incentivize, those inventions that transport ideas, techniques, and problems across disciplinary boundaries, especially when vested interests are likely to delay or block fruitful intersections between communities of practice.").

51. 35 U.S.C. § 103 (2012) (imposing nonobviousness as a requirement for patentability).

52. Today, this body of law makes patent protection available (assuming other threshold requirements are met), to inventors who combine existing ideas and technological knowledge in unexpected ways, such as when there has been no prior teaching, suggestion, or motivation to do so. *See* Justin Lee, *How* KSR *Broadens (Without Lowering) the Evidentiary Standard of Nonobviousness*, 23 BERKELEY TECH. L.J. 15, 21 (2008).

53. Federal case law reflects the view that when fields of knowledge are particularly distinct, "the bringing together of knowledge held in widely diverse fields itself becomes invention." Johnson & Johnson v. W. L. Gore & Associates, Inc., 436 F. Supp. 704, 723 (D. Del. 1977). Obviousness doctrine has been shaped in many respects by adjudications of validity challenges asserted against patents that resulted from combinations of two or more prior art references—i.e., recombinant innovations. *See, e.g.*, Christelle K. Pride, *Misguided Panic and Missed Opportunity for Pharmaceutical Inventions: How Unexpected Results Eclipsed Reasonable Expectation of Success in* BMS v. Teva, 31 BERKELEY TECH. L.J. 587, 594 (2016) (considering implications of major nonobviousness decisions on future trajectory of case law); Tolga S. Gulmen, *Model Jury Instructions on Nonobviousness in the Wake of* KSR*: The Northern District of California's Approach*, 24 BERKELEY TECH. L.J. 99, 100–05 (2009) (providing a history of nonobviousness decisions arising from challenges to recombinant technology patents); Jeffrey A. Lefstin, *The Formal Structure of Patent Law and the Limits of Enablement*, 23 BERKELEY TECH. L.J. 1141, 1162 (2008) (analyzing major obviousness decisions dealing with recombination).

> new inventions have introduced new technological capabilities, representing new technological classes, has slowed considerably. Yet surprisingly, the rate at which new combinations of technological classes have emerged has systematically kept pace with the number of new patents.[54]

Interestingly, federal law is largely unconcerned with whether, as a preliminary matter, inventors have adequate access to sufficient technological information that can later be combined.[55] The reason for this sanguine attitude may simply be that such information has often been widely accessible. The ideas that Thomas Edison recombined into the mimeograph, for instance, had been described in published patents.[56] Technological information flows through many other channels as well, such as academic publications and trade journals.[57] Innovative firms can also acquire technological information by hiring individuals with expertise in diverse fields, and through corporate mergers and acquisitions.[58] A veritable subfield of economic and legal scholarship has developed around the study of how such channels of technological information influence innovation.[59]

---

54. Saurabh Vishnubhakat & Arti K. Rai, *When Biopharma Meets Software: Bioinformatics at the Patent Office*, 29 HARV. J.L. & TECH. 205, 239 (2015).

55. The National Cooperative Research and Production Act is one of the most significant pieces of legislation at the federal level designed to encourage cooperation among researchers. National Research and Production Act, 15 U.S.C. §§ 4301–06 (2000). This law and has not been a chief subject of interest among legal scholars focused on innovation policy, however.

56. *See* FAGERBERG ET AL., *supra* note 47, at 171; *see also*, KSR Int'l Co. v. Teleflex, 550 U.S. 398, 426–27 (2007) (framing obviousness analysis from the perspective of persons of ordinary skill in the art); *In re* Dembiczak, 175 F.3d 994, 998–1000 (Fed. Cir. 1999) (setting forth "suggestion, teaching, or motivation to combine" as a test for obviousness).

57. The history of airplane technology includes many examples of information exchanges yielding innovation. *See generally* Peter B. Meyer, *An Inventive Commons: Shared Sources of the Airplane and Its Industry*, *in* GOVERNING KNOWLEDGE COMMONS 341 (Brett M. Frischmann, Michael J. Madison & Katherine J. Strandburg eds., 2014).

58. The Ford Motor Company provides a helpful example of how recombinant innovation can develop through hiring. Ford developed important new methods of building automobiles by hiring engineers with complementary technical know–how in diverse and seemingly unrelated fields, including bicycle design and grain processing. DAVID HOUNSHELL, FROM THE AMERICAN SYSTEM TO MASS PRODUCTION, 1800-1932: THE DEVELOPMENT OF MANUFACTURING TECHNOLOGY IN THE UNITED STATES 217–226 (1985). Indeed, "[r]ather than chasing whole new ideas . . . Ford focused on recombining old ideas in new ways." *Id.* at 217; *see also* HARGADON, *supra* note 44, at 133.

59. *See generally* STEVEN JOHNSON, WHERE GOOD IDEAS COME FROM (2010); EVERETT M. ROGERS, DIFFUSION OF INNOVATIONS (2003) (exploring the conditions that often give rise to innovation and emphasizing the importance of conditions that facilitate connections between disparate sources of information); Katherine Strandburg et al., *Law*

Useful technological information can also be obtained through licensing, joint development projects, and similar private cooperative arrangements.[60] The organizational theorist David J. Teece has noted, "interactions among firms and institutions are important to the innovation process . . . Information exchange and cooperative relationships of various kinds lie at the heart of . . . tremendously innovative assemblage[s] of physical and human assets."[61] In her work examining the innovative activities of small firms, Maryann P. Feldman has similarly written, "interactions and cooperation among autonomous organizations commanding specialized complementary assets and sources of knowledge may be critical to innovative success."[62] Andrew Hargadon, who conducted rich ethnographic studies of corporate innovation yielded from cooperative information exchange, has consistently explained that because the social world is "fragmented into many small domains," innovation requires "bridging multiple domains and moving ideas from where they are known to where they are not."[63]

The foregoing discussion can be reduced to a few concise points: first, the aggregation of technological information is an important precondition for innovation; second, the federal government broadly seeks to encourage such innovation; third, the task of aggregating technological information for this purpose is largely left up to private actors. Sometimes, useful information is publicly disclosed (in a published patent or trade publication, for instance); other times, useful information can be obtained through private arrangements, including transactions. As the following Section explains, these concepts are central to a timely policy question.

B. THE DATA–POOLING QUESTION

The recent Big Data phenomenon represents, by all accounts, a new kind of recombinant innovation. Rather than relying upon combinations of technological information, data scientists who specialize in Big Data rely

---

and the Science of Networks: An Overview and an Application to the Patent Explosion, 21 BERKELEY TECH. L.J. 1293 (2006) (applying network science to examine flows of technological information).

60. Maryann P. Feldman has noted, however, that large, vertically integrated companies are "able to internalize innovative inputs and provide complementary assets to facilitate innovation." Maryann P. Feldman, *Knowledge Complementarity and Innovation*, 6 SMALL BUS. ECO. 363, 370 (1994).

61. David J. Teece, *Information Sharing, Innovation, and Antitrust*, 62 ANTITRUST L.J. 465, 470 (1994).

62. Feldman, *supra* note 60, at 363.

63. Hargadon, *supra* note 46, at 44.

upon combinations of factual information—i.e., data.[64] Investors and technologists have high hopes for this phenomenon. In a 2012 television interview, a widely–known American venture capitalist famously described Big Data as the "new oil."[65] Policymakers seem to have embraced this optimism. As explained in the Introduction, the federal government has committed over $200 million toward Big Data projects, and the Obama Administration made the promotion of Big Data an important component of federal innovation policy.[66] In his 2016 State of the Union Address, President Obama announced that Vice President Joseph Biden would head a new related initiative called the "Cancer Moonshot."[67] A major goal of this initiative is to encourage more robust and effective data sharing.[68]

The sources of data that can fuel Big Data practices are myriad, ranging from smartphones, to social networks, credit card purchase records, personal health devices, and more.[69] Practitioners in this new field seek to develop valuable algorithms by applying new statistical methods to large sets of such data. For example, a seemingly unrelated collection of web search records can reveal where influenza is most likely to next strike;[70] a pool of credit card purchase records can show that people who purchase small felt pads to protect their floors from furniture damage typically have

64. Contrary to what the term may appear to suggest, "Big Data" refers to a methodology, and not a particular type or quantity of data. *See* sources cited *supra* notes 2, 7. Some researchers call this practice "data-intensive science." THE FOURTH PARADIGM: DATA-INTENSIVE SCIENTIFIC DISCOVERY (Tony Hey, Stewart Tansley & Kristin Tolle eds., 2009) (coining the term "data-intensive science").

65. This quote has been widely attributed to the venture capitalist Ann L. Winblad. *See, e.g.*, Perry Rotella, *Is Data the New Oil?*, FORBES (Apr. 2, 2012, 11:09 AM), http://www.forbes.com/sites/perryrotella/2012/04/02/is-data-the-new-oil/.

66. *See, e.g.*, *supra* notes 2, 7 and accompanying text (various leading sources that describe the Big Data phenomenon).

67. Press Release, White House, Fact Sheet: Investing in the National Cancer Moonshot (Feb. 1, 2016), https://obamawhitehouse.archives.gov/the-press-office/2016/02/01/fact-sheet-investing-national-cancer-moonshot.

68. *See* TYLER JACKS, ELIZABETH JAFFEE & DINAH SINGER ET AL., CANCER MOONSHOT BLUE RIBBON PANEL REPORT 3 (2016), https://www.cancer.gov/research/key-initiatives/moonshot-cancer-initiative/blue-ribbon-panel/blue-ribbon-panel-report-2016.pdf (describing that a goal of the project is to "[c]reate a National Cancer Data Ecosystem to collect, share, and interconnect a broad array of large datasets so that researchers, clinicians, and patients will be able to both contribute and analyze data, facilitating discovery that will ultimately improve patient care and outcomes."). This document also explains the importance of obtaining diverse oncology data. *Id.* at 12.

69. *See supra* notes 20, 21 and accompanying text (reporting on a variety of situations where Big Data has been used, including types of data).

70. *See, e.g.*, Jeremy Ginsberg et al., *Detecting Influenza Epidemics Using Search Engine Query Data*, 457 NATURE 1012 (2009).

high credit scores;[71] seemingly ordinary shopping records can predict pregnancy with remarkable accuracy;[72] and the treatment records of one million cancer patients could offer a good prediction of how well future patients will fare under a particular course of treatment.[73] These and similar examples feature prominently in popular books and recent press accounts.[74]

As its name suggests, Big Data relies upon vast corpuses of factual information.[75] Experts have emphasized that the most powerful sets of data for this purpose are drawn from different sources. In 2012, for example, IBM collaborated with the carmaker Honda and the Pacific Gas and Electric Company in California to research the best times and locations for electric cars to be recharged.[76] By pooling historical data from the power grid with car–generated data and additional data from GPS receivers, IBM was able to develop an algorithm that determined the ideal locations for car recharging stations.[77] An expert interviewed for this Article opined that drawing insights from pooled datasets in this manner is the very essence of Big Data.[78] "The single data set has very little value," she explained, adding that "[t]he unique intellectual property is from the integration of data sets and the algorithms that you can generate on top of them. It's much more [valuable] if you combine, [say], air quality data and behavioral medicine

71. *See* Paul Ohm & Scott R. Peppet, *What if Everything Reveals Everything?*, *in* BIG DATA IS NOT A MONOLITH 45 (Cassidy R. Sugimoto et al. eds., 2016) (describing this example in greater detail); MAYER-SCHÖNBERGER & CUKIER, *supra* note 2.

72. *See* MAYER-SCHÖNBERGER & CUKIER, *supra* note 2, at 58.

73. *The Promise of Big Data for Cancer Patients and Practices*, COTA HEALTHCARE (Apr. 6, 2017), https://www.cotahealthcare.com/post/how-cancer-practices-and-patients-benefit-from-big-data

74. Algorithms derived from Big Data can have broad utility and some may meet the threshold requirements for patentability. Ognjen Zivojnovic, *Patentable Subject Matter After* Alice*—Distinguishing Narrow Software Patents from Overly Broad Business Method Patents*, 30 BERKELEY TECH. L.J. 807, 809 (2015). As one attorney for Microsoft commented in a telephone interview conducted for this Article, "There is valuable IP in the analytics of big data. Methods of analyzing the data." Telephone Interview with Anonymous Subject #66, Microsoft (Jan. 22, 2014).

75. *See* MAYER-SCHÖNBERGER & CUKIER, *supra* note 2, at 107 (discussing "Recombinant Data"); Erik Brynjolfsson & Andrew McAfee, *The Innovation Dilemma: Is America Stagnating?*, TALKING POINTS MEMO CAFE (Feb. 11, 2014, 6:01 AM), http://talkingpointsmemo.com/cafe/the-innovation-dilemma-is-america-stagnating ("[D]igital innovation is recombinant innovation in its purest form"); Ohm & Peppet, *supra* note 71, at 45 (drawing upon a definition coined by Microsoft that involves "seriously massive and often highly complex sets of information.").

76. *See* MAYER-SCHÖNBERGER & CUKIER, *supra* note 2, at 102–03.

77. *Id.*

78. Telephone Interview with Anonymous Subject #42 (Sept. 19, 2014).

data and understand the relationship and [discover] some unique combinatorial issue."[79]

This is where a potential problem arises. Much of the data that could serve as grist for the mill of Big Data inventions is generated and held privately.[80] As a result, the only practical way that many data scientists can obtain it is through licensing agreements. Theory suggests, however, that licensing data directly from multiple licensors on the scale necessary to draw useful insights from it could involve very high transaction costs.

A brief hypothetical can illustrate this problem.[81] Suppose a data scientist wishes to develop an algorithm that can predict the likely success of a particular medical treatment. This would–be inventor determines that developing the algorithm will first require her to identify patterns within large sets of patient treatment records. The data scientist must then conduct a search to learn which hospitals, academic research centers, or individuals hold this data.[82] This preliminary step alone could be costly and time–consuming. But let us assume the researcher forges ahead and finds that, say, twenty hospitals and five academic research centers hold the data she needs. The data scientist must then negotiate licensing agreements with each data holder. This step could significantly add to her costs. It could involve hiring a lawyer to draft agreements, holding meetings with each data holder, and other related costs. Added to these costs is the possibility that one or more of the data holders will simply refuse to cooperate or hold out for prohibitively high costs—i.e., the well–known "hold–out" or "hold–up" problem.[83] Assessing these transaction costs and risks, the would–be

79. *Id.*

80. *See, e.g.*, *infra* Part III.B. (discussing information holdout concerns in context of cancer research).

81. This hypothetical is modeled upon the cancer-research data landscape, the focus of the study in Part III of this Article.

82. *See, e.g.*, R.H. Coase, *The Problem of Social Cost*, 3 J.L. & ECON. 1, 15 (1960) ("In order to carry out a market transaction, it is necessary to discover who it is that one wishes to deal with, to inform people that one wishes to deal and on what terms, to conduct negotiations leading up to a bargain, to draw up the contracts, then undertake the inspection needed to make sure that the terms of the contract are being observed, and so on.").

83. *See, e.g.*, Carl Shapiro, *Navigating the Patent Thicket: Cross Licenses, Patent Pools, and Standard Setting*, 1 INNOVATION POL'Y & ECON. 119–50 (2001) (discussing patent holdout and license holdup problems); Benjamin C. Li, *The Global Convergence of FRAND Licensing Practices: Towards "Interoperable" Legal Standards*, 31 BERKELEY TECH. L.J. 429, 436–38 (2016).

inventor might decide to abandon her plans at the outset. As a result, the algorithm will never be developed.[84]

Intellectual property scholars will be quick to recognize this scenario. Manufactures and service providers face the same situation when they must license numerous complementary patent rights that apply to a single product.[85] Researchers could encounter a similar problem when attempting to license numerous "upstream" patents held by different companies in order to engage in an avenue of research that will yield a "downstream" innovation.[86] Leading commentators have argued that the transaction costs and holdout risks could appear so high in these situations that the would–be licensee (i.e., the manufacturer or the researcher) will decide to forego its plans, leading to a drop in commercialization and innovation. In a seminal 1998 article in *Science*, Rebecca Eisenberg and Michael Heller described this problem as one of underuse caused by a proliferation of exclusionary rights. They called this "The Tragedy of the Anticommons."[87] This term refers to "The Tragedy of the Commons"—the well understood inverse problem of *overuse* caused by a paucity of exclusionary rights.[88] Since the time of that publication, the existence and severity of this problem have been widely debated but the term has stuck.[89] Although data generally enjoys thin formal intellectual property protection and hence weaker exclusionary rights as compared to patents, the analogy still seems apt: trade secrecy, contracts, and practical measures such as encryption can discourage unlicensed uses of data to varying degrees.[90]

Patent holders have sometimes addressed these kinds of dilemmas by forming patent pools—institutions through which patent holders offer to license a collection of related complementary patents under a unified

84. The economist, political theorist, and Nobelist Kenneth Arrow once hinted at this possibility when he wrote that information is an important "input" to the process of innovation, but that aggregating information from multiple sources is likely fraught with transactional difficulties. Arrow, *supra* note 43, at 615.

85. Jeremy Mulder, *The Aftermath of* eBay*: Predicting When District Courts Will Grant Permanent Injunctions in Patent Cases*, 22 BERKELEY TECH. L.J. 67, 84 (2007) (describing this problem as a "patent thicket" and explaining how it inhibits innovation).

86. *Id.*

87. Michael A. Heller & Rebecca S. Eisenberg, *Can Patents Deter Innovation? The Anticommons in Biomedical Research*, 280 SCIENCE 698, 698–701 (1998).

88. *See id.*

89. Matthew D. Satchwell, *The Tao of Open Source: Minimum Action for Maximum Gain*, 20 BERKELEY TECH. L.J. 1757, 1763 (2005) (referencing the "now famous tragedy of the anticommons" as an established part of the intellectual property literature base).

90. Mattioli, *supra* note 13.

agreement at a standard rate.[91] A central administrator typically licenses the collected patent rights, collects royalties, and distributes those sums to the patent holders according to a formula agreed upon beforehand.[92] This approach dramatically reduces the transaction costs that would otherwise proliferate and stall the productive use of the patents.[93] Rather than needing to search for relevant patent holders and negotiate a series of licenses, prospective licensees (i.e., manufacturers, service providers, and researchers) can simply approach a single pool for a license offered at a standard rate.[94] In a recent article, Robert P. Merges and I demonstrated that patent pools routinely conserve vast transaction costs within technology markets.[95]

It may seem appealing, then, to think that pooling is a viable solution to the problem of aggregating data. A data pool structured similarly to a patent pool might facilitate the aggregation of related data and spur innovation. This idea seems to have resonated with executives, technologists, and scientists in a number of industries that are embracing Big Data.[96] As Part III of this Article explores, experts in the field of cancer research are attempting to assemble such data pools as of this writing.[97]

As attractive as data pooling might seem, this cooperative model might not address some of the most important costs and problems associated with large–scale data licensing. Data holders who wish to form a pool would likely incur high costs just to make their data usable to others.[98] Data holders must maintain ever–expanding sets of data on an ongoing basis; they must ensure that this data is stored in formats that permit future use; they must verify the accuracy of the data, and if possible, correct errors or ambiguities in it; and the law may require a data holder to alter its data in order to

91. *See generally* Shapiro, *supra* note 83; Merges, *supra* note 13, at 1293; Merges, *supra* note 43, at 123, 129–30, 132, 144; Mattioli, *supra* note 13, at 421.

92. *Id.*

93. *See* Robert P. Merges & Michael Mattioli, *Measuring the Costs and Benefits of Patent Pools*, 78 OHIO ST. L.J. (forthcoming 2017), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2759027 (estimating transaction cost savings).

94. *Id.*

95. *Id*.

96. *See, e.g.*, *supra* notes 20–28.

97. The examples studied in this Article are based upon this hope. *See also* MAYER-SCHÖNBERGER & CUKIER, *supra* note 2, at 147–48 ("More likely, we'll see the advent of new firms that pool data from many consumers, provide an easy way to license it, and automate the transactions.").

98. Jorge L. Contreras, *Data Sharing, Latency Variables, and Science Commons*, 25 BERKELEY TECH. L.J. 1601, 1165 (2010) (describing database creation costs in analogous context).

preserve the privacy of individuals, among other barriers.[99] As Arti Rai has written, settings where health data is pooled "raise major transaction cost challenges, particularly to the extent data holders cannot guarantee that de-identified or anonymized data is impervious to re-identification."[100] As Amitai Aviram and Avishalom Tor observed before the advent of Big Data, the cost of establishing institutions to aid information exchanges could be costly enough to discourage such cooperation.[101] Moreover, data holders may often need to maintain and share clear records of the foregoing steps—i.e., "metadata" reflecting the provenance and pedigree of the underlying data—which they have weak incentives to document and disclose.[102]

Apart from these upfront costs, a vast body of scholarship from different disciplines suggests that pooling data is inherently problematic due to free–rider problems.[103] Literature on this subject is, in fact, too large to

99. Mattioli, *supra* note 13.

100. Arti K. Rai, *Risk Regulation and Innovation: The Case of Rights-Encumbered Biomedical Data Silos*, 92 NOTRE DAME L. REV. 1641 (2017).

101. Avishalom Tor & Amitai Aviram, *Overcoming Impediments to Information Sharing*, 55 ALA. L. REV. 231 (2003). Standard–setting is an important counterexample, however. When companies within an industry jointly develop a new technological standard, they will often agree beforehand to abide by certain information–sharing rules. Mark A. Lemley & Carl Shapiro, *A Simple Approach to Setting Reasonable Royalties for Standard-Essential Patents*, 28 BERKELEY TECH. L.J. 1135, 1136–37 (2013). These rules are typically promulgated and enforced by standard-setting organizations—institutions composed of a variety of private and public actors. *Id.* The International Organization for Standardization ("ISO"), which oversees the development of the MPEG video standard, for instance, requires its members to disclose information about patents relevant to the joint enterprise and to relinquish any claims of copyright they might assert over such disclosures. Christopher S. Gibson, *Globalization and the Technology Standards Game: Balancing Concerns of Protectionism and Intellectual Property in International Standards*, 22 BERKELEY TECH. L.J. 1403, 1481 (2007). The ISO also serves as a platform for the sharing of technological information (e.g., know–how, ideas, etc.) that leads to the development of mature standards. This process does not always go smoothly, however: legal scholars have observed that firms involved in standard–setting projects are often hesitant to disclose information to other members that might advantage their competitors. *Id.* at 1436 (describing an examine involving Chinese unwillingness to disclose state–owned trade secrets).

102. As I noted in a previous publication, data holders have few incentives to disclose such information, whereas patent holders must disclose a great deal of detail as a condition of the patent application process. *See generally* Mattioli, *supra* note 13.

103. Mokyr, *supra* note 17. Some might argue problem is aggravated by the relatively thin intellectual property protection U.S. law affords factual information such as data. *See, e.g.*, Charles C. Huse, *Database Protection in Theory and Practice: Three Recent Cases*, 20 BERKELEY TECH. L.J. 23, 24 (2005). Facts typically fall below the "originality" threshold of copyright law and outside of the defined classes of patentable subject matter. *Id.* As a result, holders of most kinds of data typically have a limited set of tools at their disposal to prevent misappropriation. *See generally* Jacqueline Lipton, *Balancing Private*

adequately summarize here. For purposes of this discussion, however, the work of two key scholars, Kenneth Arrow and Elinor Ostrom, is helpful to note.[104] Arrow identified a few interrelated challenges on this point: First, potential buyers or licensees of information sometimes cannot ascertain its value without first examining it—a result that sometimes negates the need for exchange in the first place.[105] Second, an actor who sells information on the open market risks suffering a loss in value if the buyer duplicates and redistributes that information.[106] Third, the demand for information can often be low when potential licensees who wish to use it as an "input" to the invention process see its value as highly speculative.[107] In Arrow's view, these "restriction[s] on the transmittal of information will reduce the efficiency of inventive activity in general and will therefore reduce its quantity also."[108]

Elinor Ostrom's scholarship offers some equally important and relevant insights. Ostrom is perhaps known best for the vivid portraits she depicted of self-governing communities that manage shared resources, at times avoiding the so-called "Tragedy of the Commons." Her 2012 book UNDERSTANDING KNOWLEDGE AS A COMMONS was motivated by these concerns and provides a roadmap for scholars to examine knowledge-sharing ethnographically.[109] Although research questions related to knowledge sharing could be examined through many lenses, Ostrom's framework is designed to capture important factors that might otherwise go overlooked—how characteristics of the information being shared and the

Rights and Public Policies: Reconceptualizing Property in Databases, 18 BERKELEY TECH. L.J. 773 (2003). These tools include trade secrecy, contracts, and practical barriers such as encryption. *Id.* at 786–87. From the perspective of data holders, these options are weaker than intellectual property rights in a critical respect, however: they cannot be used to stop unwanted uses of data that have already been widely disclosed. M. Scott McBride, *Bioinformatics and Intellectual Property Protection*, 17 BERKELEY TECH. L.J. 1331, 1354–55 (2002). That is, trade secrets, contracts, and encryption do not provide data holders with a mechanism to enjoin widespread unlicensed downstream reproduction and use. *Id.* When data is widely disclosed, the genie is often simply out of the bottle. *Id.*

104. A third scholar whose work may shed useful light on the subject is Friedrich Hayek, whose examined this theme—which he referred to as "The Knowledge Problem"—quite extensively. *See generally* FRIEDRICH HAYEK, THE CONSTITUTION OF LIBERTY (1960).

105. *See* Dan L. Burk & Brett H. McDonnell, *The Goldilocks Hypothesis: Balancing Intellectual Property Rights at the Boundary of the Firm*, 2007 U. ILL. L. REV. 575, 585 (2007) (discussing this problem as it relates to trade secrecy).

106. Arrow, *supra* note 43.

107. *Id.*

108. *Id.*

109. *See generally* UNDERSTANDING KNOWLEDGE AS A COMMONS (Charlotte Hess & Elinor Ostrom eds., 2007).

broader cultural, legal, and economic context can influence sharing, for instance.[110] Ostrom and the many scholars she inspired showed through ethnographic studies that the problems of sharing technological information are often nuanced, and more complicated than theory alone suggests.[111] Katherine Strandburg, Michael Madison, and Brett Frischmann have advanced this area of study in important ways, including through the development of the "Knowledge Commons" research framework—an approach to studying information-sharing institutions inspired by Ostrom's work.[112]

Ultimately, the work of Arrow, Ostrom, and others indicates that pooling information can present a classic "social dilemma"—i.e., a situation "in which each member of a group has a clear and unambiguous incentive to make a choice that, when made by all members, provides poorer outcomes for all than they would have received if none had made the choice."[113] Based upon this, there is good reason to expect efforts to pool data will often be unsuccessful, thus impeding the development of Big Data innovations. Stated differently, data may not be pooled frequently enough or widely enough to spur meaningful technological advances.

Recent scholarly and press accounts seem to support these concerns. Economists have recently doubted that Big Data will ever carry economic or social benefits;[114] one distinguished academic commentator recently predicted that Big Data will be "far less important than the great innovations of the 19th and 20th centuries."[115] Press reports of data–sharing failures seem to support these doubts.[116] One notable example was the failed UK

110. *Id.*

111. *See generally id.* (exploring many possible conditions that could lead to failed information exchanges). The ethnographic study presented in Part III of this Article was conducted according to Ostrom's Institutional Analysis and Development ("IAD") methodological framework.

112. *See generally* GOVERNING KNOWLEDGE COMMONS (Brett M. Frischmann, Michael J. Madison & Katherine J. Strandburg eds., 2014); UNDERSTANDING KNOWLEDGE AS A COMMONS, *supra* note 109; ELINOR OSTROM, GOVERNING THE COMMONS: THE EVOLUTION OF INSTITUTIONS FOR COLLECTIVE ACTION (1990).

113. Dawes & Messick, *supra* note 18, at 111–16 (defining social choice problems or social dilemmas).

114. *See* Glantz, *supra* note 20.

115. *See* Marcus & Davis, *supra* note 20.

116. *See e.g.*, John Markoff, *Troves of Personal Data, Forbidden to Researchers*, N.Y. TIMES (May 21, 2012), http://www.nytimes.com/2012/05/22/science/big-data-troves-stay-forbidden-to-social-scientists.html (reporting on widely–held concerns by scientific researchers who believe that private data–holders are impeding research through their unwillingness to share data that is useful for Big Data projects); Bernardo A. Huberman, *Sociology of Science: Big Data Deserve a Bigger Audience*, 482 NATURE 308 (2012)

National Health Service's National Programme for IT.[117] The plan, which was to pool patient medical records under a single roof, collapsed in large part because data holders (mostly hospitals) had strong disincentives to share information widely.[118]

Legal and economic scholarship written before the age of Big Data further supports the prediction of a Big Data anticommons. Pamela Samuelson and J.H. Reichman warned in the 1990s that information–sharing failures threatened to slow the pace of federally–funded scientific research.[119] Scholars of management have documented how the same types of cooperative problems impede information sharing within corporations.[120] Paul Uhlir and J.H. Reichman have noted that the severity of the problem may increase when diverse institutions are concerned and suggest that antitrust law might further discourage the pooling of data:

> The evidence shows that such [database] pools are very difficult to form when the value of upstream research products defies easy measurement and the relevant players in a given industry have very different agendas, as would occur when federal agencies, academic institutions, and different types of private companies are all involved. Moreover, there are far greater risks that such pools lead to collusive, anti-competitive behavior, to the erection of formidable barriers to entry.[121]

("Many of the emerging 'big data' come from private sources that are inaccessible to other researchers."); Brett Hemenway & Bill Welser, *Cryptographers Could Prevent Satellite Collisions*, SCI. AM. (Feb. 1, 2015), https://www.scientificamerican.com/article/cryptographers-could-prevent-satellite-collisions/ (describing incident where U.S. and Russian satellites collided because both governments kept orbital data secret due); Austin Frakt, *Addiction Research and Care Collide With Federal Privacy Rules*, N.Y. TIMES (Apr. 27, 2015), https://www.nytimes.com/2015/04/28/upshot/federal-push-for-privacy-hampers-addiction-research-and-care.html (reporting how privacy regulations prompted insufficient disclosure of health records maintained by the federal government).

117. *See* Rajeev Syal, *Abandoned NHS IT System Has Cost £10bn So Far*, GUARDIAN (Sept. 17, 2013, 7:06 PM), http://www.theguardian.com/society/2013/sep/18/nhs-records-system-10bn.

118. *See* Lizzie Presser et al., *Care.data and Access to UK Health Records: Patient Privacy and Public Trust*, TECH. SCI. (Aug. 11, 2015), http://techscience.org/a/2015081103/.

119. J.H. Reichman & Pamela Samuelson, *Intellectual Property Rights in Data?*, 50 VAND. L. REV. 51, 56 (1997) (arguing that propertization of data would impede scientific progress and exchange).

120. Ulrike Cress et al., *Information Exchange With Shared Databases as a Social Dilemma*, 33 COMM. RES. 370 (2006) ("When group members exchange information via shared databases people are often reluctant to contribute information they possess.").

121. J.H. Reichman & Paul F. Uhlir, *A Contractually Reconstructed Research Commons for Scientific Data in A Highly Protectionist Intellectual Property Environment*,

Press accounts of Big Data innovations loosely align with these concerns. The most notable examples of Big Data innovations in practice have been developed by large, vertically–integrated firms that drew upon their own vast internal datasets, rather than through cooperative exchanges.[122]

Will data–pooling problems dampen innovation? Policymakers should be concerned by this possibility. As explained in Part II, barriers to data pooling could subvert recent federal policies designed to promote Big Data.[123] The patent system's goal of promoting technological progress is implicated by the success of data pooling as well. Recent Supreme Court jurisprudence has called the patentability of some kinds of software algorithms into question.[124] In particular, the decision in *Alice Corp. v. CLS Bank International* has cast doubts on the patentability of at least some kinds of patent claims pertaining to algorithms while providing only murky guidance on how such subject matter might qualify for patent protection.[125] But even so, experts believe Big Data has enormous potential to impact innovation. An unpatentable algorithm might open the door to new classes of patentable technologies and fields of research. The underlying policy concern thus seems clear: federal innovation policy seeks to promote technological advances that may not come about due to data–pooling problems.

Because of Big Data's immense and unrealized potential to stimulate and generate innovation, policymakers can benefit from understanding how successfully private actors are pooling useful data. To that end, the following Section presents an original ethnographic study of recent efforts to pool data in an important field of research.

---

66 L. & CONTEMP. PROBS. 315, 403–04 (2003). The "database pools" these authors posited would be similar to patent pools in some respects. *See id.*

122. *See supra* notes 21–27.

123. *See supra* Part II.

124. *See, e.g.*, Merges, *supra* note 29 (discussing the implications of the recent *Alice* decision on the patentability of software methods, including algorithms related to data).

125. Alice Corp. Pty. Ltd. v. CLS Bank Int'l, 134 S. Ct. 2347 (2014). For an example of such commentary, see, e.g., Merges, *supra* note 29.

III. A STUDY OF RECENT DATA–POOLING EFFORTS IN CANCER RESEARCH[126]

Is Big Data likely to develop into a thriving platform for innovation? According to experts in this new field, the answer to this question will hinge upon how effectively data held by different firms, institutions, and individuals can be aggregated. Here, this question is explored through an original ethnographic study of several burgeoning efforts to pool cancer treatment and research data.[127]

A. STUDY METHODOLOGY

The decision to focus this study on cancer treatment and research data was motivated primarily by the fact that cancer research is one of few fields where data holders have sought to organize pools in order to advance innovation.[128] As a result, the setting is ripe for investigation. Secondly, as mentioned in the Introduction, the federal government has publicly applauded efforts to pool cancer treatment and research data.[129]

It is important to note, however, that this study's focus on cancer treatment and research data necessarily limits the extent to which its findings can be generalized. This study does not aim to conclusively prove or disprove the existence of a widespread data–pooling problem. The goal here is more modest. This is an investigation designed to learn whether the data–pooling problem predicted by theory is playing out in an important field of research. A secondary goal of this article is to serve as a model for how data–pooling problems in other industry and research settings can be examined and addressed in the future.

The organizations examined in this study were all founded on a simple, hopeful thesis: if medical data (e.g., patient treatment data, drug trial observations, etc.) is pooled on a large scale, data scientists will be able to develop algorithms that predict how well future patents will respond to certain therapies. The goal is not to invent new treatments, but rather to develop new methods of prescribing existing treatments. CancerLinQ, a

126. A companion chapter to this study appears in a forthcoming book published by Cambridge University Press. That chapter covers aspects of the history of data in the field of medicine and the field of oncology data not presented here. *Cancer: From a Kingdom to a Commons*, *in* GOVERNING MEDICAL KNOWLEDGE COMMONS (eds. Katherine Strandburg, Brett Frischmann & Michael J. Madison, 2017).

127. This study received necessary Institutional Review Board approval from Indiana University.

128. These organizations were selected for study because they were reported on with the greatest frequency in the national press the time this study was conducted.

129. *See supra* notes 24–27.

project under the direction of the American Society of Clinical Oncology ("ASCO"), is an early–stage effort to pool cancer treatment data from hospitals and individual practices across the United States.[130] The White House has commended CancerLinQ as an exemplar of the potential for social good that Big Data holds.[131] Project Data Sphere, a second group studied, is a research joint venture that has drawn the participation of several large pharmaceutical companies, including Celgene, Pfizer, and Sanofi.[132] The project, which began in 2012, aims to improve cancer treatment by pooling clinical drug trial data.133 A third organization, CancerCommons, also aims to facilitate innovations in cancer care.[134] Unlike CancerLinQ and Project Data Sphere, however, CancerCommons seeks treatment data from individual patients.[135] A final initiative examined is the Data Alliance Collaborative ("DAC"). The DAC, which was organized by Premier Healthcare in 2013, seeks to pool Big Data methods and processes between member hospitals.[136] For instance, by pooling patient treatment records from different cities, members of the group have developed an algorithm that predicts the likelihood that future patients will be readmitted.[137]

Interview subjects were identified by a variety of means. Some were placed in contact with me by the data pools they worked with. Others were contacted directly because they had been quoted or otherwise mentioned in press accounts of the data pools studied. Some subjects provided introductions to additional interview subjects experienced with pooling health–related data. This second tier of interviewees, which included bioinformaticists at the National Institutes of Health, professors of medicine at leading research universities, and other prominent players in the field, provided context and outside perspectives on the institutions examined. All interview subjects were selected based upon their experience in the field

130. *ASCO CancerLinQ: Learning Intelligence Network for Quality*, AM. SOC'Y OF CLINICAL ONCOLOGY, http://www.cancerlinq.org (last visited Aug. 13, 2017).

131. Press Release, *supra* note 24, at 2.

132. Press Release, Project Data Sphere, CEO Roundtable on Cancer Launches the Project Data Sphere Initiative, A New Data Sharing and Analytic Platform for Cancer Patient Benefit (Apr. 8, 2014), https://www.projectdatasphere.org/projectdatasphere/html/PressRelease/LAUNCH.

133. *Id.*

134. *About Us,* CANCER COMMONS, https://www.cancercommons.org/about/ (last visited Aug. 10, 2015).

135. *Id.*

136. *See Shape the Future of Healthcare Data and Care Delivery Models*, PREMIER, INC., https://www.premierinc.com/transforming-healthcare/quality-improvement-in-healthcare/data-alliance-collaborative/ (last visited Aug. 12, 2017).

137. Telephone Interview with Anonymous Subject #17, member, Data Alliance Collaborative (Sept. 10, 2014).

(e.g., their credentials and work history), and their willingness to participate. All interviews lasted at least thirty minutes, were audio recorded with the permission of the interview subjects, and were later transcribed.[138] Follow–up interviews were conducted with nearly all subjects. Some interview subjects agreed to be identified by name, while others agreed to be quoted only if they were identified as anonymous subjects. This latter group of subjects have been identified in the footnotes anonymously with unique identifying numbers.

An important, and perhaps unavoidable, limitation of this study is that it includes only the insights of individuals who agreed to be interviewed. This means that the study may not capture the full range of opinions that exist on this topic—i.e., that the results presented here are biased in some way. Nonetheless, even if the picture presented here is incomplete, the insights offered are still helpful and relevant to the research question that motivated this study.

Methodologically, this study's ethnographic approach was inspired by the classical IAD framework developed by the Nobelist ethnographer Elinor Ostrom. Aspects of the study were also greatly influenced by the recent adaptation of Ostrom's IAD's framework by "Knowledge Commons" pioneers Katherine Strandburg, Michael Madison, and Brett Frischmann.[139] In practical terms, this entailed asking subjects a semi–structured set of interview questions that fell into the following categories: (i) background industry environment and context; (ii) the characteristics of the data or informational resources being shared, and skills needed to create or prepare these resources; (iii) the "default" status of the data to be shared, (iv) the identities of the firms or institutions participating; (v) the goals and objectives of the data–sharing project; (vi) any institutional governance mechanisms, such as information–sharing rules that members promise to abide by; and, (vii) technological infrastructure.[140]

Three themes emerged from these interviews, each of which is explored in the following subsections. The most significant finding is that the chief impediments to the pooling of cancer treatment data are not the free–rider problems that theory predicts. The pooling of cancer treatment and research data in the United States appears to be significantly impeded by concerns regarding professional, competitive, and reputational standing. The details

138. Several interview subjects participated in this study on the condition of anonymity. These individuals are identified in this article as "Anonymous Subject" followed by a unique identifying number.

139. *See generally* GOVERNING KNOWLEDGE COMMONS, *supra* note 112; OSTROM, *supra* note 112; UNDERSTANDING KNOWLEDGE AS A COMMONS, *supra* note 109.

140. *See* GOVERNING KNOWLEDGE COMMONS, *supra* note 112, at 20–21.

of these concerns are laid out in the sections that follow and are analyzed in Section IV. A second hindrance to data–pooling is the cost of preparing data for exchange, a topic briefly touched upon in Part II. A third hindrance is that potential data contributors are often doubtful that the benefits of sharing data with a pool will outweigh the necessary costs and risks. The possible reach of these findings, and whether they are worrisome enough to prompt policy action, are considered in Section IV of this Article.

Some interview subjects offered global comments that helpfully captured the challenges of pooling cancer treatment and research data. As the CEO of a data-pooling non-profit remarked, "everyone is behaving in an economically rational way because everyone has incentives to not share. Everyone is competing for everything—from patients, to grant dollars, to tenure, to vacations, to promotions."[141] A prominent professor of medicine involved in several oncology data–pooling projects echoed this view, commenting that those who hold valuable data have "every incentive not to share" and that "some of the most important data live at the places that have the least interest in sharing."[142]

B. COMPETITIVE CONCERNS

Subjects interviewed for this Article reported that many of the institutions that hold valuable cancer–related data are reluctant to share it widely because doing so might cost them reputational harm or a competitive edge. The precise nature of these concerns is contextual, varying from one data holder to another, but the overall result is the same: a decision not to cooperate.

Subjects reported that some hospitals and cancer care centers do not wish to disclose patient treatment data because doing so could publicize unfavorable information about the quality of care their institutions provide. One subject reported that a representative of a large medical institution told him, "the reason we don't want to share this data is that we are afraid people will use it to compare our outcomes with other institutions in an inappropriate way."[143] The subject made it clear that he was paraphrasing this remark but added that it was consistent with his own first–hand observations. Another subject described this phenomenon in more dramatic terms: "Hospitals reason, 'I'm going to get sued as soon as anyone can see how many people died from leaving sponges in bodies.' So they have a disincentive from a liability perspective to share."[144] Another subject

141. Telephone Interview with Anonymous Subject #57 (Aug. 29, 2014).
142. Telephone Interview with Anonymous Subject #42 (Sept. 19, 2014).
143. Telephone Interview with Anonymous Subject #57 (Aug. 29, 2014).
144. Telephone Interview with Anonymous Subject #44 (Sept. 11, 2014).

elaborated by explaining the kinds of valuable information patient treatment data can hold:

> I've seen hospitals reluctant to share [patient treatment] data because of market advantages. Often the information contains financial information. Overall, a lot of healthcare data also has trade secrets embedded in it. It's how you're planning your primary care network to basically act as a referral into your hospital network—that kind of stuff can be figured out by seeing where patients are coming in, and that kind of thing. I've seen a lot of reluctance [to share] that kind of information. And certainly reluctance on cost and charge information, because it can reveal [what the hospital] is getting paid for which procedures. Also, at the doctor level: Doctors are reticent to see patient information shared because it could expose the hospital to greater risks of malpractice. So, there's a lot of worry, and it's totally reasonable.[145]

Subjects also reported that concerns about patient privacy discourage some hospitals and cancer care centers from pooling data. Disclosing private patient data could lead to reputational harm, subjects explained, but also liability under privacy–related laws—most notably, the Health Insurance Portability and Accountability Act of 1996 (HIPAA).[146] "[T]here's this enormous liability risk of brokering access to de-identified data for research when you've been using it identified for quality improvement or for treatment and payment," explained one subject.[147] "That risk really blocks a lot of things."[148] Added another subject, "the big cancer centers, those that are not sharing data, will cite HIPAA … [I]t is a well-intentioned thing to protect patient privacy [but] it has really stymied medical research."[149] Interestingly, several experts suggested that HIPAA provides a plausible excuse for institutions that do not wish to share data for reasons unrelated to privacy, such as reputational concerns. This argument is "particularly hard to argue with," one subject stated.[150]

Subjects reported that pharmaceutical companies harbor similar concerns about reputational or competitive harm. "There are IP concerns when talking to pharmaceutical companies . . . [are] they going to lose

145. Telephone Interview with Anonymous Subject #42 (Sept. 2, 2016).

146. 45 C.F.R. § 164.514(e) (2013) (forbidding institutions from disclosing data that contains names, zip codes, treatment dates, and other related information).

147. Telephone Interview with Anonymous Subject #44 (Sept. 11, 2014).

148. *Id.*

149. Telephone Interview with Anonymous Subject #57 (Aug. 29, 2014).

150. Telephone Interview with Anonymous Subject #42 (Sept. 19, 2014).

something if they give away the data?" one subject stated.[151] The former Chief Product Officer of one effort, Project Data Sphere, elaborated on why some pharmaceutical companies may understandably be reluctant to share "core" clinical trial data with a pool composed of competitors:

> If you share the core, central data, that data may remove a little bit of a competitive edge in certain companies, and the competition among those companies is a crucial part of the model of generating innovation.[152]

This interview subject emphasized, however, that the core competitive data is "just a fraction" of the data, and that pooling of "non-core" data—i.e., data that holds no immediate competitive value—can still "advance the field tremendously."[153] The subject indicated that changes in culture and in corporate governance could open the door to greater sharing of this kind of data:

> There are cultural barriers, there is the fear of competitiveness, and also, sharing data is not built into the policies of all the companies, so all the governance reviews and the decision making are awkward because they are completely outside of the regular work that they have. So those are barriers, you see.[154]

These comments are consistent with at least one recent industry report that concluded, "companies are reluctant to share proprietary information—even when anonymity is assured—for fear of losing competitive advantage."[155]

Subjects also reported that academic researchers, a large and important class of cancer data holders, face strong competitive disincentives to disclose their data with pools. As a professor of medicine at a leading U.S. research university involved with oncology data pooling explained:

> There is not a culture in academia to promote [data] sharing because the culture is exactly the opposite: It's protection of information to keep from getting scooped. The only way to protect yourself is to wall-off or circumscribe the information that you

151. Telephone Interview with Anonymous Subject #29 (Aug. 26, 2014).
152. Telephone Interview with Kald Abdullah (August 29, 2016).
153. *Id.*
154. *Id.*
155. James Manyika et al., *Open Data: Unlocking Innovation and Performance With Liquid Information*, MCKINSEY GLOBAL INST. (October 2013), www.mckinsey.com/business-functions/digital-mckinsey/our-insights/open-data-unlocking-innovation-and-performance-with-liquid-information.

> uniquely have access to and even if you don't do anything with it today at least nobody else is doing anything with it either.[156]

According to this subject, academic researchers often feel so reluctant to share data that they will purposefully obscure it even when a federal grant–funding agency requires its disclosure. The subject called this behavior "data–dumping," explaining, "[i]t's very similar to when there is a requirement to share documents [in the course of a litigation] and you essentially overwhelm the other team with too many documents."[157]

In addition to institutions, medical professionals, and academics, patients themselves may fear that contributing to a data pool could lead to reputational, professional, and pecuniary harm. Subjects explained that patients often worry, for instance, that widespread disclosure of their health data will jeopardize their privacy or negatively influence how employers, financial lenders, or insurers will treat them. "There's a lot of fear around the negative externalities of sharing," one subject explained. "Losing health insurance, losing long term care insurance, losing employment opportunities, being embarrassed, being discriminated against, being unable to find dates, that sort of thing."[158] These comments are corroborated by recent press reports on point.[159] When asked to explain how disclosing one's health data could impact employment, for instance, the subject provided an unsettling hypothetical:

> Let's say that we figure out a set of genetic mutations and variations that give you an 85-percent chance of developing Alzheimer's disease early—before age 50. I'm 42 right now. Let's say I have these mutations and in five years, my genome is

156. Telephone Interview with Anonymous Subject #42 (Sept. 19, 2014). These comments are consistent with the observations of Elinor Ostrom and her collaborators. AMY R. POTEETE, MARCO A. JANSSEN & ELINOR OSTROM, WORKING TOGETHER: COLLECTIVE ACTION, THE COMMONS, AND MULTIPLE METHODS IN PRACTICE 263–64 (2010) ("Some scholars have opposed blanket data-sharing policies out of a concern that such policies would either disadvantage research that relies on less standardized forms of qualitative data, or compromise the anonymity of respondents who provide sensitive information.").

157. Telephone Interview with Anonymous Subject #42 (Sept. 19, 2014).

158. *Id.*

159. *See, e.g.*, Stephen F. DeAngelis, *Patient Monitoring, Big Data, and the Future of Healthcare*, WIRED (Aug. 2014), http://www.wired.com/2014/08/patient-monitoring-big-data-future-healthcare/ (reporting a reluctance on the part of patience and opining that "[i]t is more understandable why people would be reluctant to share personal health information in countries like the U.S. where it could be used to justify significant increases in health insurance costs.").

> [available to potential employers] online. Who's going to hire me?[160]

The subject emphasized that *perception* of risk, even if unfounded, may be enough to discourage data sharing.[161] As a result, useful data from healthy individuals may be particularly difficult to obtain. "People who self-identify as healthy have very few incentives to share their health data unless they buy into it philosophically," commented one subject whose field of work focuses on facilitating patient data sharing.[162]

Other subjects expressed the opinion that it is reasonable for patients to have such concerns. A doctor and professor of medicine with prior involvement in an oncology data pool stated, "You know, insurers of course are very interested in having their hands on all of the information because it allows them to know [what] rates to charge."[163] Another subject explained, troublingly, that masking an individual's identity in compliance with HIPAA does not necessarily reduce privacy risks because "[r]e-identifying patients is becoming very easy with the right computer program, even off of genomic data sets."[164] Most subjects explained that, in light of these concerns, patients are likely to either not share their data at all, or to do so with only very highly trusted brokers.

The data pools examined are addressing these widespread perceptions of risk by tailoring their approaches in interesting ways. For instance, Project Data Sphere requires its licensees to agree that they "will not seek, and will not support any third party's effort to seek, patent protection in any jurisdiction for any research procedure or research design that results from . . . use of the Data or User Contributions."[165] The group is also attempting to minimize the competitive risk pharmaceutical firms perceive by collecting only a sub–category of clinical trial data that holds little commercial value for its members but may, once aggregated, yield useful results. "We did it in a way that was relatively low risk by asking for the comparator arms [and] not the experimental [arm]," explained a senior executive with the project.[166] "So you can keep your IP close at hand, but

160. *Id.*

161. *Id.*

162. Telephone Interview with Anonymous Subject #44 (Sept. 11, 2014).

163. Telephone Interview with Anonymous Subject #51 (Aug. 26, 2014).

164. *Id.*; *see generally* Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 1701 (2010) (calling attention to this problem).

165. *Project Data Sphere Online Service User Agreement*, PROJECTDATASPHERE.ORG (on file with the author).

166. Telephone Interview with Anonymous Subject #29 (Aug. 26, 2014).

the opportunities with the data itself are tremendous. I mean, you can work on standards, you can work in how data are collected, you can look at end–point, you can look at progression of end–point selection, you can look at subpopulation . . . you can do all kind of things."[167]

Other interviewees expressed doubts, however, about the likely efficacy of these approaches. Referring to Project Data Sphere's decision to collect only a narrow category of data, one interview subject commented, "I am not trying to say anything negative, but what are they sharing? The data on the control arms for small drugs? How remote can you be?"[168] This subject later amended his comment, however, stating, "But it may have value to patients. In fact, sometimes controls and failed trials can sometimes be more interesting than the trial arms."[169] Not referring to Project Data Sphere, another subject stated that *any* data pool composed of only one type of data risks missing the grand opportunity of Big Data. "If I could [pool] prostate cancer patients' [data] together, I'd understand the story of prostate cancer better," the subject commented, "But that's not the real deal. The real deal is when you find the intersection between many things that otherwise could not be intersected."[170] The subject went on to describe a recent academic study that uncovered a correlation between cancer and certain environmental factors by pooling hospital records with weather data.[171]

Ironically, competitive concerns may also discourage data pools *themselves* from sharing data. Referring to recent efforts to gather cancer treatment data, one subject explained:

> So these organizations [pools] don't necessarily want to share with each other, because they're competing for dollars and oxygen in the community; They're competing to be 'The Group,' right? And so ironically, although it would be better for the people that they purport to serve that they all share data with each other, frequently they don't, right, either for competitive reasons or because it doesn't occur to them.[172]

Other subjects consistently commented that it has become common for burgeoning data aggregators in this field to trumpet the virtues of sharing data to the outside world while maintaining high barriers for researchers who wish to access the data. As one subject remarked, "This is really about externally stating that sharing is important but internally creating hurdles

167. *Id.* ("The majority of the data are phase-three cancer clinical trials, so it is hard end-point data from blood pressure to PSA readings, to basically everything: end points, death, life . . . various things.").

168. Telephone Interview with Anonymous Subject #71 (Aug. 29, 2014).

169. *Id.*

170. Telephone Interview with Anonymous Subject #42 (Sept. 19, 2014).

171. *Id.*

172. Telephone Interview with Anonymous Subject #44 (Sept. 11, 2014).

that make using data nearly impossible. Because now you can, as you set the threshold, gate when people cross over the threshold."[173] Another subject explained that overcoming the high barriers set by data aggregators led him to seek data from individual patients. His organization's goal, he stated, is to "[g]et it out of those silos."[174]

C. COSTS OF OBTAINING AND PREPARING DATA FOR POOLING

Interview subjects reported that accessing, organizing, and preparing data for exchange carries significant upfront costs. Before primary holders of useful health data—hospitals and specialty care centers, doctors, and academic research institutions—can pool data, they must translate it into common formats, confirm its accuracy, correct any errors, and obfuscate details that could be used to identify individual patients. In addition to identifying these technical hurdles, subjects reported a variety of institutional, cultural, and contractual barriers that also discourage the pooling of cancer treatment data.

Patient treatment records constitute one of the most important forms of data sought by the groups interviewed by this study. These are records of care that patients receive from healthcare providers such as hospitals. An interviewee involved with CancerLinQ helpfully divided this type of data into two broad categories: "structured" and "unstructured."[175] Structured patient treatment data, this subject explained, includes objective, machine–recorded facts such as "laboratory test results or the dosages of medicines prescribed, or patient vital signs," whereas unstructured data is generated and recorded more casually and based upon more subjective observations—a clinical physician's handwritten observations, for example.[176]

Healthcare providers typically store structured and unstructured data in patients' electronic health records ("EHRs"). Large hospitals and cancer care centers contract with outside database vendors ("EHR vendors") to store and manage access to patient EHRs.[177] At least some structured data contained in a patient's EHR may also be reflected within a hospital's billing records or in an insurer's customer records: "If the doctor bills for their services, all that sort of information gets converted, if you will, into a

173. Telephone Interview with Anonymous Subject #42 (Sept. 19, 2014).
174. *Id.*
175. Telephone Interview with Richard L. Schilsky, M.D. (Aug. 26, 2014).
176. *Id.*
177. *Id.* This subject explained that EHRs often do not contain all records generated in the course of treating patients: "[I]t's not all completely integrated into the electronic medical records so a lot of it is returned to the doctor in the EHR but some of it is returned to original reports that have to be scanned into the EHR and things of that sort." *Id.*

lot of different kinds of codes that are used to submit the claim to insurance," one subject explained.[178] The overall picture conveyed by subjects is that patient treatment information is usually stored in multiple places and overseen by different stewards, the most important of which are EHR vendors and healthcare providers.

A former president of ASCO involved with the launch of CancerLinQ explained that the business practices of EHR vendors pose a challenge to data–pooling initiatives. "You need to have an EHR vendor or an institution that is willing to share the data," he explained, and "[a] silo mentality is associated with those vendors . . . [T]hey are a notoriously proprietary bunch that are not good at sharing—even within their own system[s]."[179] This problem is aggravated, the subject explained, by the fact that hospitals have few choices when selecting EHR vendors to work with: "A fairly small number of corporations are responsible for the electronic health records in the United States," he said, adding, that these vendors do not store data in common or widely accessible formats.[180] Other subjects made consistent statements about the difficulty of obtaining data from EHR vendors. Some explained that even when data is obtainable, it is not immediately usable.[181] Because EHR vendors store data in proprietary digital formats, some subjects explained, a hospital seeking to share patient data with a data pool must first translate all data it gathers into common, widely–readable digital formats.[182] This involves enlisting engineers with expertise in formatting data to perform the translation.[183]

These findings are consistent with recent press accounts. In a September 2015 industry news report, a doctor who works at a medical practice involved with CancerLinQ commented, "What has happened is that EHR systems don't communicate with each other. Vanguard practices are having to dedicate time and resources, including my entire IT department, to be able to adapt the technology in order to implement CancerLinQ . . . Other EHR vendors are flat-out refusing or making it prohibitively expensive to make the systems communicate with each other."[184]

178. *Id.*

179. Telephone Interview with Anonymous Subject #51 (Aug. 26, 2014).

180. *Id.*

181. Telephone Interview with Anonymous Subject #44 (Sept. 11, 2014).

182. Telephone Interview with Richard L. Schilsky, M.D. (Aug. 26, 2014) (discussing the fact that EHR vendors often store patient data in different formats).

183. *Id.*

184. *See* Frank Irving, *ASCO Calls Out EHR Interoperability as Barrier to CancerLinQ*, XTELLIGENT MEDIA (Sept. 16, 2015), http://healthitinteroperability.com/news/asco-calls-out-ehr-interoperability-as-barrier-to-cancerlinq (discussing

Relatedly, a leading commentator and advocate for health data sharing explained that many EHR vendors contractually forbid the hospitals they serve from sharing data with outside institutions: "they [hospitals] are wrapped up in contracts with their technology vendors—especially their electronic health record vendors—where only the EHR vendors have rights to go outside the hospital with data," he explained.[185] Independent research into the practices of EHR vendors corroborates this statement, revealing that many vendors indeed impose such contractual restrictions.[186] Other subjects echoed this problem, but expressed optimism that such restrictions would lessen over time as a result of certain provisions of the Patient Protection and Affordable Care Act of 2010 that require EHR vendors to make patient data more portable.[187]

The difficulty of accessing data from large EHR vendors has led some nascent data pools to approach smaller practices or academic institutions, where treatment data may be relatively easier to gather. The president of a private oncology practice group explained that CancerLinQ "looked at private practices to start this project rather than academic institutions [because] the data's probably more easily extractable from the private practice EHRs and trying to get discrete information out of a big hospital system can be very tedious."[188] Other subjects consistently reported that data held by smaller private practices is sometimes subject to fewer technical and contractual barriers.

Apart from the barriers presented by EHRs, healthcare providers seeking to contribute patient treatment data to a pool also must carefully remove any information that might identify an individual patient. HIPAA, mentioned earlier, forbids institutions from disclosing data that contains names, zip codes, treatment dates, and other information that could identify an individual patient.[189] Subjects explained that simply stripping these identifiers from a dataset can remove useful information, however, such as the length of time that a patient was treated, or the fact that the same patient received treatments at different institutions.[190] To address this issue, some

---

CancerLinQ's potential and offering recommendations to help eliminate information blocking as a barrier to interoperability).

185. Telephone Interview with Anonymous Subject #44 (Sept. 11, 2014).

186. *Id.*

187. Telephone Interview with Anonymous Subject #51 (Aug. 26, 2014).

188. Telephone Interview with Anonymous Subject #71 (Sept. 8, 2014).

189. 45 C.F.R. § 164.514(e) (2013).

190. *See, e.g.*, INFO. COMM'R'S OFFICE, ANONYMISATION: MANAGING DATA PROTECTION RISK CODE OF PRACTICE 83–84 (Nov. 2012), https://ico.org.uk/media/1061/anonymisation-code.pdf.

health care institutions examined in this study hired engineers to "mask" or obfuscate personally–identifying data with "dummy values" (random numbers) in a manner that preserved underlying information.[191] For instance, engineers could consistently replace a patient's name with unique random numbers so that the same patient or a group of patients could be examined over time.[192] Likewise, a patient's treatment dates might be offset by a consistent period of time—say, 7 weeks.[193] This allows data scientists to know how often a patient was seen and the total period of time the patient was treated. A data scientist involved in this practice explained that these examples are simplified: masking data to preserve privacy requires a nuanced understanding of the content of the data, the ways in which it may be used in the future, and the harm that could result from disclosure of personally identifying patient information.[194] Because the process requires expert judgment, it cannot be automated.

Some subjects reported that manipulating data in this manner can be a costly barrier to sharing data. "I think pretty much everyone in the healthcare community [and the research community] views HIPAA laws as a hindrance rather than a help to most patients," commented one subject.[195] Complying with HIPAA imposes a cost, he explained, "basically because it requires you to get an enormous amount of approval that adds immensely to the expense of these things."[196]

Yet another fundamental upfront cost lies in locating critical data in unstructured sets of information. According to a prominent academic researcher with deep involvement in oncology data pooling, some of the most important oncology data can only be identified by and recorded manually and with great effort. This subject offered a fascinating illustration of this problem by explaining the role that condolence cards play in relaying mortality information:

> It's not just a data-gathering problem. The problem is that, even if you put data together in a soup, there can be huge holes of certain information points that live in doctors' handwritten medical notes. Take condolence cards, for example. As a country, we have the social security death index, the national death index—all of these ways of finding out whether someone has died. The social security death index used to be a mandated activity and now it is not, so

191. *See, e.g.*, Mattioli, *supra* note 13, at 566–68 (discussing data masking in depth).

192. Telephone Interview with Josh Mann, Assistant Director of Oncology Technology Solutions at American Society of Clinical Oncology (Oct. 8, 2013).

193. *Id.*; *see also* Ohm, *supra* note 164, at 1703–05.

194. *See supra* note 163.

195. Telephone Interview with Anonymous Subject #51 (Aug. 26, 2014).

196. *Id.*

> it's a decaying data set, so it only reflects some proportion of people—probably only seventy percent of deaths. But if you want to do research, you need to know with certainty whether someone is alive or not. So, I work with a dataset that is culled from obituaries and funeral homes. That's the main way I get mortality data. It turns out that even with this method, you still only find about ninety percent of the mortality information you need. But if you go into the medical record, and you read the case notes, you see the note from the doctor that says, "I discussed hospice with the family today." Now you know something is going on. You find [a copy of] the condolence card that says, "I am so sorry, Mrs. Jones, to know about what happened to Fred." . . . It actually might be the condolence card that is the best symbol that this patient has died. [Sometimes] we have no other place to figure out that the patient has died. So one of the things that is becoming apparent in the cancer data space is how important this unstructured data morass is. It's there; as clinicians, we know it; but it doesn't get captured in a useful way. [At a point in time about five years ago], we believed that all we needed to do was collect data and we could make sense of it later. That hasn't been borne out. The promise of natural language processing as a scalable solution to turn this kind of unstructured information into meaningful data points is just not bearing out. In oncology, many of the critical data points that you need for research live in places that are completely unstructured.[197]

Upfront costs also appear to hinder the pooling of useful data held by pharmaceutical companies. A scientist involved with Project Data Sphere hypothesized that in the course of his work, he might encounter "five companies collecting information on prostate cancer patients and they are all doing it in a little bit different manner . . . there are little nuances that may be different between my company, and another company, and an academic organization."[198] An expert involved with this effort explained that such differences create an upfront cost, because all data must be translated into a common format.[199] The Chief Product Officer of Project Data Sphere explained that it could take "a programmer and a statistician . . . about 40 to 80 hours" to prepare just one data set for inclusion.[200] Although this amount of work is not tremendous, he explained, "if you . . . scale that up . . . it impacts resources."[201] He later commented that data

197. Telephone Interview with Anonymous Subject #42 (Sept. 9, 2016).
198. Telephone Interview with Anonymous Subject #29 (Aug. 26, 2014).
199. *Id.*
200. Telephone Interview with Kald Abdullah (Aug. 27, 2014).
201. *Id.*

preparation is perhaps "the most significant obstacle" to forming effective data pools in his view.[202]

In summary, the cost of preparing cancer treatment and research data poses a significant barrier to pooling it. These costs stem from hiring and paying engineers and data experts to translate data into useful formats, masking potentially personally–identifying information, searching for useful data points within large sets of unstructured data, and related tasks such as identifying and correcting errors in the data. Care providers may face the additional challenge of obtaining the data from EHR vendors in the first place.

D. UNCERTAIN RETURNS

Subjects interviewed for this study reported that a major challenge in pooling cancer treatment and research data is convincing data holders that their cooperation will yield direct benefits that outweigh the foregoing risks and costs.

The "what's in it for me" question, as one subject termed it, appears to be one of the central conundrums that data pools in the healthcare industry appear to face.[203] As the former Chief Product Officer of Project Data Sphere explained,

> I think there [are] two questions that you need to really work hard to convince them [potential contributors of data]. 'What is the value of sharing this data?'—that's one question. The other question is, 'What is the value *for me* to share my data?[204]

A prominent academic researcher consistently reported that the message some data–pooling projects are communicating to data holders is, in effect, "it's good for *the world* [if you] share your data [because] somebody else can do something with it."[205] But this message, she explained, is less persuasive than "saying you should share your data so that you can collaborate . . . and it's even another . . . move to say 'share your data because it's actually in the sharing that the unique intellectual property comes,'" she continued.[206] "Those degrees of separation—each one requires a step outside the box."[207] A subject involved with Project Data Sphere commented, "I think the biggest problem . . . is getting more data and

202. *Id.*
203. Telephone Interview with Anonymous Subject #29 (Aug. 26, 2014).
204. Telephone Interview with Kald Abdullah (Aug. 27, 2014).
205. Telephone Interview with Anonymous Subject #42 (Sept. 19, 2014).
206. *Id.*
207. *Id.*

convincing people that it is valuable . . . Why am I going to put resources behind this when I am not sure what I am going to get out of it?"[208]

A potential solution that some data–pooling groups are experimenting with is to promise to share useful statistical information with participating data holders. CancerLinQ, for example, plans to report to member hospitals summaries of how their quality of service measures up to that of other members and to national standards.[209] The Chief Medical Officer of ASCO explained,

> We will be able to return to the physician on a regular basis a dashboard report that shows what is the quality of their performance against . . . standard measures . . . . Eventually we will be able to show them how the outcomes of their patients compared to the outcomes of other similar patients in other similar practices. . . . We think that will be a big incentive to them to join.[210]

According to an executive involved with the project, Project Data Sphere is considering offering a similar incentive, in the form of sharing "use cases" that would allow researchers to better design and evaluate their clinical trials.[211]

Some data holders interviewed were optimistic about this approach. "I mean, any quality metrics that you get back to your practice are always helpful to keep you ahead of the curve and to make you practice better," commented the president of a private oncology practice group.[212] Moreover, there may be financial incentives for health care institutions to know how well they measure up to their competitors. Health providers "can use that information to report to insurers what their quality is, how it compares to other physicians," explained the CIO of ASCO.[213] Several subjects explained that the Affordable Care Act's measures for "accountable care," under which participating healthcare providers are reimbursed based upon the relative quality of care they provide, could make the sort of metrics

208. Telephone Interview with Anonymous Subject #29 (Aug. 26, 2014).
209. *Id.*
210. Telephone Interview with Richard L. Schilsky, M.D. (Aug. 26, 2014).
211. Telephone Interview with Anonymous Subject #29 (Aug. 26, 2014) ("We are working on several very interesting projects to publish use cases that can evaluate the validity of clinical-trial data or perhaps predict the potential outcome of phase-three data based on phase-one and phase-two responses.").
212. Telephone Interview with Anonymous Subject #71 (Sept. 8, 2014).
213. Telephone Interview with Richard L. Schilsky, M.D. (Aug. 26, 2014).

CancerLinQ plans to share an effective enticement.[214] In other words, healthcare providers that operate under the accountable care model should wish for more information describing how their quality of service compares to that of their competitors. It is not yet clear whether this incentive will lead to greater data pooling, and in turn, advances in cancer treatment.

Another source of uncertainty stems from a lack of clarity over how profits or valuable patent rights generated by a data pool would be divided among those who contribute to it. "Suppose a new analytical method took a long time to develop," one respondent hypothesized, "and it's significantly predictive and different than what you can buy in the commercial space . . . . Someone is going to want to get paid for that because it took a long time to build . . . . You're going to want to get something back out of that."[215] At the time of this writing, none of the data pools examined have a system in place for dividing such royalties among data contributors.[216]

Lastly, some subjects explained that cultural forces that oppose data disclosure make the potential benefits of pooling all the less certain. Calling back to a theme discussed earlier in this Article, one subject reported that within medical research, cumulative innovation is valued more highly than recombinant innovation: "Science is incredibly reductionist and 'looking down' as opposed to 'looking out and across.' That's one of the big differences. And we celebrate the science that looks down, and we call the science that's collaborative dumb. That's part of the problem."[217] Large hospitals and research centers may similarly fail to see the benefit of data pooling. The CEO of a nonprofit data–sharing group commented:

> We have approached many, many medical institutions, large cancer centers, especially the big ones . . . They are very, very protective of their data. Because they think they are big enough to be able to not need anyone else's data, so they won't share their data . . . It is strongly in the culture.[218]

214. *See generally* Patient Protection and Affordable Care Act, 42 U.S.C. § 18001 (2010); *see also* OLIVER WYMAN, TRACKING THE GROWTH OF ACCOUNTABLE CARE (Sept. 2013), http://www.oliverwyman.com/content/dam/oliver-wyman/global/en/files/archive/2013/OW_HLS_ACO_maps.pdf (providing a helpful summary and map displaying these accountable care organizations).

215. Telephone Interview with Anonymous Subject #17, member, Data Alliance Collaborative (Sept. 10, 2014).

216. For examples of how patent pools divide royalties, see Mattioli, *supra* note 13.

217. Telephone Interview with Anonymous Subject #42 (Sept. 19, 2014).

218. Telephone Interview with Anonymous Subject #57 (Aug. 29, 2014).

IV. ANALYSIS AND RECOMMENDATIONS

The foregoing study reveals several factors discouraging the pooling of cancer treatment and research data. Some of these factors, such as concerns about professional, competitive, and reputational standing, are not widely predicted by theory. This Part explores the policy implications of these findings, and suggests possible avenues for future policy work.

A. IMPLICATIONS FOR INNOVATION POLICY

Should policymakers be concerned by the challenges that private data pools face? To answer this question, it is helpful to first consider the specific problems this study uncovers, and how far they might reach. Why might we think that this study reflects a problem that reaches beyond a few institutions in a single industry?

1. Summarizing the Problem

The problem uncovered by this study can be stated simply: the pooling of cancer treatment and research data—widely thought to be a necessary precondition for certain innovations in cancer care—appears to be hindered by collective action problems playing out between data holders and nascent pools. Some of the most important impediments are not neatly reducible to simple "free–rider" dilemmas predicted by legal scholars and economists.

To consider the possible reach of the data–pooling problem, it is helpful to examine its causes. This problem is spurred in part by data holders' concerns regarding professional, competitive, and reputational standing. Some healthcare providers appear to be reluctant to disclose treatment data to a pool because they fear that doing so will lead to negative publicity and a reduced inflow of patients in the future. Sharing data with a pool could reveal a hospital's relatively poor record of patient outcomes, for example. Similar concerns appear to discourage the pooling of some kinds of clinical trial data held by pharmaceutical companies. A drug firm might hesitate to exchange certain data too widely for fear that a competitor could benefit from information revealed by that data. Academic researchers appear to face similar disincentives. In a profession where research data can pave the way to publications, tenure, and grants, researchers at universities may have little motivation to share the data they collect. To the contrary, according to subjects interviewed, academic scientists tend to guard their data jealously.[219] Although some grant–funding agencies such as the NIH require data to be disclosed, subjects interviewed for this study explained

219. *See supra* Section II.B.

that such requirements are often subverted through strategic "data dumping."[220] (These observations square with recent empirical work on secrecy among clinical biomedical researchers.)[221] Finally, individual patients may believe that sharing their health data too widely could change how insurers, employers, and others view them, possibly opening the door to discriminatory treatment.[222]

Alongside such concerns are the costs of preparing data. For both legal and practical reasons, data holders often must manipulate their data before disclosing it. As several subjects explained, HIPAA creates a risk of liability for health care providers that share patient treatment information.[223] Care providers who wish to minimize that risk without rendering their data useless to research efforts must manipulate it in various ways prior to disclosure (e.g., "masking" personally identifying information, etc.). This may involve hiring highly skilled (and, consequently, highly paid) engineers.[224] Evidence from this study indicates that such costs can discourage pooling.[225]

Electronic health record vendors appear to represent a related, but distinct problem. These companies sometimes intentionally make data access and reuse difficult as a deliberate business strategy—i.e., as an attempt to "lock in" clients. Some may do so by storing data in proprietary formats and by placing contractual restrictions on how their clients (i.e., care providers) may use the data they store. Interview subjects indicated that even when care providers are able to obtain data from electronic health record vendors, they must incur costs translating it into a standard, widely readable format before submitting it to pools.

Data held by pharmaceutical companies also must be prepared prior to sharing—although for slightly different reasons. Unlike healthcare providers, which hire outside vendors to store patient treatment data, private pharmaceutical companies can easily access the data they generate in the course of clinical trials because they store it themselves. This study

220. *Id.*

221. Wei Hong & John P. Walsh, *For Money or Glory?: Commercialization, Competition, and Secrecy in the Entrepreneurial University*, 50 SOC. Q. 145 (2009).

222. There are some notable counter–examples, however, including a cancer survivor who received high–profile press coverage for disseminating his treatment data widely. *But see*, Steve Lohr, *The Healing Power of Your Own Medical Records*, N.Y. TIMES (March. 31, 2015), https://www.nytimes.com/2015/04/01/technology/the-healing-power-of-your-own-medical-data.html.

223. *See supra* Section II.B.; Section III.A.

224. *See supra* Section II.B.; Section III.A.

225. *See, e.g.*, *supra* notes 170–77 and accompanying text. This discussion lays out the character of such costs and includes comments from an interviewee (note 177) who explained their significance.

indicates, however, that different pharmaceutical firms record and store their data differently. As a result, any effort to pool cancer treatment data must include a plan for translating it all into a common format.

2. Assessing the Problem's Possible Reach

If policymakers are to develop solutions to the foregoing problem, they must have a clear sense of the problem's possible reach. As explained earlier, the purpose of this study was not to conclusively prove or disprove the existence of a widespread data–pooling problem. Rather, this study's more modest goal was to investigate whether such a problem, as suggested by theory, is affecting an important field of research. It would be imprudent to assume that the factors that discourage the pooling of cancer treatment and research data will similarly discourage *all* Big Data efforts.[226] How far, then, might the problem reach? A rigorous empirical answer to that question would require similar studies of other settings where data pooling is being attempted. This Article hopes to motivate other scholars to conduct such studies. But until that work is done, it is possible to draw some informed deductions from this study.

At the very least, the problems uncovered by this study appear to present a problem for Big Data efforts in the field of cancer research. This conclusion is based on the observation that the concerns and costs uncovered by this study do not appear to be unique to the specific institutions examined. In accordance with the IAD research framework, interview subjects were deliberately asked to comment on the broader environment in which the pools examined operate. The sources of the problem clearly appear to be more general, and rooted in the perceptions of individuals and dynamics of culture: concerns about patient privacy, the competitive need to draw customers and patrons, reputational interests, the unstandardized formats in which data of some kinds is often stored. Based upon this observation alone, it seems reasonable to expect *other* private efforts to pool cancer treatment data to be susceptible to similar challenges.

The concerns and costs uncovered by this study also do not appear to be unique to cancer data. Stated differently, subjects interviewed did not suggest that there is anything unique about the *content* of cancer treatment data, or the *processes* by which such data is recorded and stored that makes pooling it difficult. Rather, the problem stems from the types of institutions involved, the environment in which they operate, intrinsic qualities of health–related data, and bodies of law that apply to such data.

226. As explained earlier, the decision to focus this study on cancer treatment data was motivated purely by the fact that, at the time of this writing, cancer treatment is a focus of activity and investment among proponents of Big Data.

It seems reasonable, then, to expect that attempts to pool data related to the treatment of other health conditions to be affected by the same concern. Such hypothetical data pools—if they were to form in the future—might include new kinds of data holders, of course. In addition to hospitals, pharmaceutical companies, academic researchers, and individuals, the manufacturers of smartphones and other personal medical devices hold a wealth of useful health–related data. These new companies could introduce new dynamics, new concerns, and new possibilities, of course. But there is no reason to expect they would not be subject to some of the same barriers to data pooling encountered by the firms and institutions studied here.

Could similar problems hinder data–pooling in other industries altogether? Possibly. As explained earlier, an important and somewhat surprising fact uncovered by this study is that cancer treatment data holders are concerned about professional, competitive, and reputational standing. These concerns are motivated by a common idea—namely, that a data holder's data could, at some future time, reflect something unfavorable about them or harmful to them.[227] This anxiety seems particularly timely. A hallmark of Big Data is its power to reveal surprising insights from data generated for no particular purpose. In an age where, as Paul Ohm recently suggested, "everything might reveal everything," it would be unsurprising to learn that data holders of many kinds worry about what their data might reveal about them.[228] This could have important outcomes in developing industries where data pooling could be helpful, but could also reflect poorly on data holders—telemetry and collision data from autonomous vehicles, for instance.

To sum up, the problems uncovered by this study are likely to impact not only cancer data–pooling efforts, but also related efforts to pool health data of other kinds. Moreover, there is at least a basis to expect similar general problems to develop in other industries and research settings where Big Data might soon be embraced.

3. Considering the Problem's Policy Implications

The introduction to this Article explained why data pooling could, in theory, present a problem for policymakers. That explanation focused on two innovation policy goals: first, the federal government's targeted

227. *See supra* Section II.B. As explained earlier, legal scholars such as James Anton and Dennis Yao have highlighted the role that competitive concerns can play in discouraging information exchanges in the context of standard setting. *Id.*; *see also infra* note 231 and accompanying text (referring to concerns over anticompetitive information sharing served as an animating force behind federal legislation).

228. Ohm & Peppet, *supra* note 71.

funding of Big Data research projects; second, the government's broad goal of encouraging innovation through the patent system.

It is clear why the problem uncovered by this study is relevant to the government's investments in Big Data research. As mentioned earlier, federal agencies have committed over $200 million to help develop new methods of gleaning insights from enormous volumes of data. If the private entities that hold some types of data—at least cancer treatment and research data—are unable or unwilling to assemble such volumes of data in the first place, then the full potential of the government's investments in Big Data might go unrealized.

The problem uncovered by this study also appears to have a bearing on federal innovation policy as expressed through intellectual property law. The reasoning behind this conclusion is straightforward: a central purpose of the patent system is to encourage technological progress. Big Data algorithms can enable technological progress of many kinds, and they are also a category of subject matter theoretically eligible for patent protection (even in light of doubts cast by *Alice* and its progeny). If a precondition for the development of such inventions is unmet, then policymakers could view the result as a subversion of the patent system's goals. The foregone innovations could also be described in economic terms, as representing a drop in dynamic efficiency.

Policymakers may wish to correct this problem by seeking to actively encourage data pooling. The U.S. government has a history of encouraging the exchange of technological information for the purpose of fostering innovation. Patent pools have often formed as a result of governmental intervention.[229] On the other hand, the patent pools that the government has encouraged served to untangle knots that the government had arguably created in the first place—i.e., by apportioning patent rights that developed into thickets. By contrast, the data held by private institutions like those examined here is not a form of property created by the government, nor does it enjoy robust intellectual property protections. Opponents of policy intervention might argue, then, that data pooling is not the government's problem to solve because the government played no direct role in creating it.

229. Moreover, patent pools that *have* formed without state intervention have overwhelmingly been geared toward decreasing the cost of producing existing technologies rather than fostering the development of new ones. David W. Van Etten, *Everyone in the Patent Pool:* U.S. Phillips Corp. v. International Trade Commission, 22 BERKELEY TECH. L.J. 241, 254 (2007) (describing express motivations of various actors who formed patent pools without state intervention).

There are other examples of the government stepping in to encourage the sharing of technological information, however. Standard–setting—a process that necessarily involves technological information–sharing—has been encouraged by the government in various ways.[230] The National Cooperative Research and Production Act (NCRPA) is a federal law passed in 1984 that reduces antitrust liability for certain research consortia that make their activities known to the Department of Justice and the Federal Trade Commission.[231] The law is "designed to promote innovation, facilitate trade, and strengthen the competitiveness of the United States in world markets."[232] These examples suggest that a proposal designed to encourage the pooling of data could succeed in being passed into law.

B. POLICY RECOMMENDATIONS

The recommendations and suggestions presented in the paragraphs that follow help to show the value of ethnographic studies like the one presented in Part III of this Article. By understanding how a collective action problem affects an important area of research, policymakers and scholars alike can develop informed solutions.

It is helpful to first touch upon an avenue of policymaking that could present special challenges: exclusive rights. One might conclude that data pooling could be encouraged through new laws that imbue scientific and industrial data with intellectual property–like protections—so–called *sui generis* data protection. This idea is a perennial subject of policy debate and it has an intuitive appeal. Like copyrightable and patentable subject matter, useful scientific and industrial information is costly to create, easy to copy, and subject to free–rider problems. With the power to exclude any and all unwanted users, data holders might be more willing to enter into exchanges.

Some leading commentators have argued persuasively that *sui generis* intellectual property protection for data might actually *reduce* the level of

230. *See generally supra* Part III.

231. Antitrust law's potential chilling effects on information sharing led policymakers in 1984 to pass The National Cooperative Research and Production Act ("NCRPA"). *See* James J. Anton & Dennis A. Yao, *Standard-Setting Consortia, Antitrust, and High-Technology Industries*, 64 ANTITRUST L.J. 247, 247–49 (1995); U.S. DEP'T OF JUSTICE & F.T.C., *Antitrust Guidelines for Collaborations Among Competitors*, *reprinted in* 4 Trade Reg. Rep. (CCH) ¶ 13,161, at 1 (2000), http://www.ftc.gov/sites/default/files/documents/public_events/joint-venture-hearings-antitrust-guidelines-collaboration-among-competitors/ftcdojguidelines-2.pdf (exploring the "likelihood of anticompetitive information sharing").

232. *Id.*

innovation in society, however.[233] Pamela Samuelson and J.H. Reichman, most notably, explained that such laws could give database holders the power to control access to basic scientific research materials.[234] This, they argued, would dampen research and "undermine the competitive ethos on which market economies depend."[235] Current law is consistent with this view. Although Congress has regularly considered data–protection bills since the 1990s, none gained sufficient political goodwill to be enacted into law. But more specifically, this study does not suggest that a lack of property protections is the central reason that data is not being widely shared. There may simply be new and more effective ways of encouraging data pooling.[236]

A second threshold consideration relates to public versus private data. Although this Article is focused primarily on data held in the private sector, it is important to note that a wealth of health data in is funded by government institutions such as the NIH. Arti Rai and Rebecca Eisenberg have explored how such public actors can helpfully influence the pooling of—for instance—federally–funded biomedical data.[237] More recently, Rai has explored the role of risk regulators in the data–pooling context, as well as how some private sector data pooling has been encouraged by threats of regulatory action—her observations, insights, and recommendations are deeply relevant to the problems examined in this Article.[238]

It is also helpful to note that national governments have sometimes worked directly with private companies to create vast databases related to health information. As Peter Lee has documented, one of the most widespread was a joint effort between the Icelandic government and a private company in the 1990s to build a database of "clinical records, DNA,

233. For a valuable related exploration of the merits of property rules versus liability rules with respect to information, see Mark A. Lemley & Phil Weiser, *Should Property or Liability Rules Govern Information?*, 85 TEX. L. REV. 783 (2007).

234. *See, e.g.*, Reichman & Samuelson, *supra* note 119, at 95–113.

235. *Id.*

236. In earlier work, I explored the idea of offering limited exclusivity in data as a means to encourage the disclosure of methods by which data was collected and prepared—i.e., metadata—but that approach addressed a problem that occurs in settings where the use of the data is unknown by the entity collecting and sharing it. The data holders examined in this study, by contrast, are aware of the types of research uses the data pools examined have. As a result, the solution seems unhelpful.

237. Arti K. Rai & Rebecca S. Eisenberg, *Bayh-Dole Reform and the Progress of Biomedicine*, 66 LAW & CONTEMP. PROBS. 289 (2003).

238. Rai, *supra* note 100.

and family histories for the entire country."[239] As Lee explains, this plan was controversial.[240] Because there seems to be a desire within the oncology treatment community to form data pools, it might be most desirable for the government to encourage cooperation through nudges, rather than managing a data pool directly or through such a partnership. More direct involvement may be helpful in other industrial settings, however.

A more hopeful focus of future policy efforts could relate to standards. As discussed earlier, a major obstacle to pooling health data is the cost of conforming patient treatment records into common formats. The federal government could reduce these costs by encouraging the standardization of electronic health records. The best approach is probably not a direct mandate that all healthcare providers and electronic health record vendors adopt a specific set of standards. The federal government mandates standards in only limited settings—the side of the road that we drive on, or permissible uses of radio spectrum frequencies licensed to private users, for instance.[241] As a general policy matter, the government disfavors mandating the adoption of specific technology standards and even specific interoperability requirements, and prefers to instead promulgate standards of *performance*.[242] As a recent publication of the Federal Trade Commission explains, "U.S. Government agencies, such as, for example, the Consumer Product Safety Commission, the Food and Drug Administration, and the Environmental Protection Agency, may set safety, health, and environmental requirements designed to protect the public, but they rely upon voluntary consensus standards, where possible, to meet their regulatory objectives."[243] In short, a gentle approach of encouraging standardization of health data could be more likely to succeed.

One such approach would be for the federal government to lead by example. Federal healthcare institutions could, for instance, adopt certain standards for storing patient treatment records. This approach has been

239. Peter Lee, *Toward a Distributive Commons in Patent Law*, 2009 WIS. L. REV. 917, 990 (2009).

240. *Id.*

241. Joseph Farrell & Paul Klemperer, *Coordination and Lock-in: Competition with Switching Costs and Network Effects*, *in* 3 HANDBOOK OF INDUS. ORG. 2007, 2010 (M. Armstrong & R. Porter eds., 2007).

242. *See* FED. TRADE COMM'N, COMPETITIVE ASPECTS OF COLLABORATIVE STANDARD SETTING, at 5 § 2.2(7) (June 9, 2010), https://www.ftc.gov/sites/default/files/attachments/us-submissions-oecd-and-other-international-competition-fora/usstandardsetting.pdf ("Most government standard setting activities . . . focus on performance standards, without reference to specific technologies or interoperability requirements.").

243. *See id.*

advocated by the Bipartisan Policy Center, a nonprofit policy think tank in Washington. In July 2015, the organization released a report recommending that Congress "require the federal government to adopt standards for health IT."[244] Measures outlined in the report include the federal government's adoption of standards designed to permit patients to be tracked over time and the adoption of common standards by private EHR vendors on contract with the government.[245] This idea could be an effective "nudge" to encourage standardization of private patient treatment records. If all EHR vendors that wish to work with the government must adopt standard ways of organizing patient treatment data, it might be easiest for these same vendors to store private hospital data in the same formats.

The government could offer incentives to institutions that conform to certain data standards. This recommendation was inspired by an expert interviewed for this study, who commented, "I think it may not be that we want the government to set the standards, because sometimes . . . they don't necessarily get it right . . . it may be that we want them to provide the incentives for the standards to be set."[246] One such incentive could be insurance reimbursements offered to health care providers. As explained earlier, the Patient Protection and Affordable Care Act offers monetary rewards to certain healthcare providers that demonstrate a record of high quality care. To reap these rewards, healthcare providers must share patient treatment data with outside institutions, including insurance providers. This law could be amended to offer an *enhanced* reward in the form of higher reimbursements to healthcare providers that share their data in specific, standard formats. In a related article, Rebecca Eisenberg and Nicholson Price explored how insurance companies, which hold claims data could play a role in developing important knowledge about the quality and efficacy of healthcare products—a concept they call "demand-side innovation."[247]

244. BIPARTISAN POLICY CTR., ADVANCING MEDICAL INNOVATION FOR A HEALTHIER AMERICA (July 2015), http://bipartisanpolicy.org/wp-content/uploads/2015/07/BPC-Advaning-Medical-Innovation.pdf.

245. *Id.* at 12 (describing such innovations in Proposal 1.7). A related effort is the Blue Button Initiative, led by the Department of Health and Human Services, which aims to enable patients to securely access personal health data online. *See What is the Blue Button?*, HEALTHIT.GOV (Jan. 15, 2013), https://www.healthit.gov/patients-families/faqs/what-blue-button. Although this program does not explore challenges posed by the lack of standards for data generated by consumer medical devices, the suggestions made here could nevertheless apply to data generated by such devices.

246. Telephone Interview with Richard L. Schilsky, M.D. (Aug. 26, 2014).

247. Rebecca Eisenberg & W. Nicholson Price, *Promoting Healthcare Innovation on the Demand Side* (Univ. of Michigan Law & Econ. Research Paper No. 16-008, 2016), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2766707.

Turning toward industry, this study revealed that data sharing between companies is frustrated by a variety of competitive concerns.[248] In this setting, it may be advisable for the government to take a more hands–on approach by directly mandating data sharing between companies. The FDA, for instance, could require drug manufacturers to make more clinical drug trial data (e.g., data describing safety or effectiveness of drugs) available to certain data-pooling efforts. Similarly, the FDA could require medical device manufacturers and software developers to adopt standards and data-sharing practices as a condition of having their products and services approved. This approach could be viewed as onerous, however. Pharmaceutical companies and manufacturers of devices that collect health data would likely oppose the idea because it places a new burden upon them. Some critics may argue, for instance, that a measure designed to promote innovation should not impose a new set of costs upon innovative companies.

A gentler approach would provide targeted rewards to companies that agree to pool data. There is precedent for this idea. Beginning in 2007, the FDA began offering "Priority Review Vouchers" ("PRV") to drug companies that sought FDA approval for products designed to target tropical diseases.[249] The vouchers, which substantially reduced the time necessary to bring a drug to market, were valued so highly by some corporations that they inspired the formation of at least one patent pool.[250] The FDA could offer a similar expedited review and approval process to companies that submit useful health data (e.g., clinical trial information, data generated by consumer health devices, etc.) to the public or to certain data-pooling consortia. Like PRVs, vouchers for this expedited review process could be transferrable, which would likely enhance their value.[251]

In a similar vein, the NIH and other federal agencies that provide research grants could impose stricter data-sharing requirements on grant recipients, and importantly, greater penalties for failure to adhere to such policies. The practice of "data-dumping," which one interview subject

248. *See generally supra* Part III.

249. *See* Michael Mattioli, *Communities of Innovation*, 106 NW. U. L. REV. 103, 126–27 (2015) (discussing the FDA's provision of such vouchers).

250. *Id.*

251. A similar possibility not directly inspired by this study would be for the United States Patent and Trademark Office (USPTO) to offer a similar fast-track to patent applicants who claim new innovations derived from data pools. This enticement could act as a general incentive. In 2009, the USPTO offered an expedited review process to patent applicants who claimed inventions that would benefit the environment. Pilot Program for Green Technologies Including Greenhouse Gas Reduction, 74 Fed. Reg. 64666 (Dec. 8, 2009).

reported is common, would likely be a helpful area for reform as well. Just as patent applicants are not permitted to hide potentially important information from the Patent and Trademark Office (PTO), federal grant recipients might be forbidden from obfuscating useful data from the public.

Yet another policy intervention could focus on reducing the risk of liability that data holders face for inadvertent disclosure of personally identifying information. The severity of civil and criminal penalties under HIPAA and related bodies of federal and state law designed to protect personal privacy could be reduced, for instance, for organizations that can demonstrate that they disclosed data to a data pool that has identified itself to relevant federal authorities.[252] Such a measure would entail creating a new procedure by which cooperative data pools could notify the FTC (or another federal agency selected) of their cooperative efforts. These measures could include standards to determine if contributed data has been sufficiently de–identified. In exchange, their potential liability for privacy violations under HIPAA and other relevant law could be capped at a percentage of what it would ordinarily be.

This remedy is directly inspired by the National Cooperative Research and Production Act (NCRPA)—a federal law passed in 1984 that reduces antitrust liability for certain research consortia that make their activities known to the Department of Justice and the Federal Trade Commission.[253] Because antitrust liability concerns did not appear to play an important role in discouraging data sharing among the institutions examined here, antitrust is not the subject of any specific recommendation. However, such an approach could be useful in other industry settings. Antitrust authorities have recognized that certain types of information-sharing arrangements between firms can have anticompetitive effects that violate the Sherman and Clayton Acts, however.[254] James Anton and Dennis Yao have posited that risks associated with antitrust liability "may interfere with transmission of information that could improve the joint decision to create a standard."[255]

252. Penalties under HIPAA include fines that can reach as high as $250,000 and up to 10 years in prison. 42 U.S.C. § 1320d(6) (2010).

253. *See supra* note 231.

254. Teece, *supra* note 61, at 474 ("[M]eetings and exchanges of technical information . . . can cause antitrust suspicion."); *see also* Tor & Aviram, *supra* note 101, at 236 ("Due to the potential anti-competitive effect of information sharing, antitrust law frequently analyzes the likelihood that information sharing will facilitate collusion."); HERBERT HOVENKAMP, FEDERAL ANTITRUST POLICY: THE LAW OF COMPETITION AND ITS PRACTICE 171–72 (1994) (discussing the types of information exchanges that may facilitate collusion).

255. Anton & Yao, *supra* note 231, at 264. For instance, an agreement among a group of companies to share pricing information in order to collusively charge consumers supracompetitive rates would be illegal. *Id.*

Tax incentives are another possibility. The IRS could enact a new rule offering charitable tax deductions to individuals who donate their health treatment data to certain qualifying data–pooling efforts, for instance. More nuanced measures could also be possible. Medicare reimbursements could, in the future, be enhanced for patients who agree to share their data more widely. Along the same lines, individuals who donate their data to certain pooling efforts could receive—from private insurers or through Medicaid—reimbursements for medical tests and procedures they would not otherwise be covered for.

Considering the variety of actors that hold data useful to oncology research and the variety of impediments that may discourage each actor from sharing its data, it is helpful to consider how the foregoing suggestions might pair with the various stakeholders discussed earlier in this Article. Table 1 summarizes these relationships visually by comparing various solutions that emerged during the interview process to the various stakeholders discussed earlier in this Article.[256] Cells of the table that contain an "X" indicate a solution primarily designed to address an impediment to data pooling that were consistently reported by a particular actor. This summary shows that (at least among the proposals suggested here), there is no "one–size–fits–all" solution.

256. I wish to credit and express my thanks to Pamela Samuelson for suggesting the use of a table to summarize these connections.

Table 1: Primary Links Between Policy Suggestions and Stakeholders

	Healthcare Providers	Corporations (Drugs and Devices)	Academic Researchers	Patients
Efforts to Standardize Patient Records	X		X	
Insurance Reimbursements	X			
FDA–Mandated Data Sharing		X	X	
Vouchers for Expedited FDA Review		X	X	
Heightened Data–Sharing Requirements for Government Grant Recipients			X	
Reduced Penalties for HIPAA Violations	X	X	X	
Tax Incentives	X	X		X

Each of the foregoing suggestions would present practical challenges, but some of these challenges may be surmountable. An overarching problem would be selective nondisclosure—i.e., the sharing of incorrect or incomplete data. This problem is, in a sense, a cousin of "data dumping," as described by the interview subject from academia. Any policy designed to mandate or encourage the sharing of data could be subject to this form of gaming. In light of this possibility, all such proposals would necessitate some level of monitoring and perhaps penalties for selective nondisclosure—two steps that would introduce complexity and cost. While these challenges should not be downplayed, they are curable. The FDA could levy sanctions, for example, on pharmaceutical corporations that attempt to circumvent a new set of data–sharing requirements. Academic researchers who engage in similar behavior could be denied future grants. In short, this problem is probably solvable.

Some of these measures could also spur litigation. Because data pools would likely need to be bound together by contracts, one would expect to see an increase in contract disputes as the number and size of data pools increase. One would also expect that data pools, if they form in large enough numbers, could raise competition concerns—through tying arrangements, or through pooling substitutive data, for instance. The number of civil disputes and prosecutions under antitrust laws could conceivably increase in the future.[257] This too would represent new costs.[258]

In the area of cancer treatment, it is easy to imagine that the costs of encouraging greater data pooling might well be dwarfed by the social and economic benefits of success. In other industries, the potential gains might be less clear. Through discussion and debate over proposals like those outlined here, policymakers, industry stakeholders, and the public can make informed decisions tailored to specific settings.

V. CONCLUSION

The pooling of data appears to be an increasingly important and unmet precondition for innovation in many settings, and yet it may not occur without government intervention. To gain an empirical view of this problem, this Article presented an ethnographic study of institutions seeking to pool cancer treatment and research data. This study revealed that a variety of costs, risks, and competitive concerns are impeding the useful pooling of data. Some of these findings are not widely predicted by theory: hospitals do not wish to disclose data that reflects poorly on the quality of service they provide, and separately, they voice concerns over potential liability for privacy violations; pharmaceutical companies closely guard data that could reveal their business strategies to competitors; academic researchers have every incentive to hold tight to data that could fuel publications and professional advancement.

Informed by these insights and the important earlier work of other scholars that this study was based upon, this Article proposed a set of policy suggestions. First, lawmakers could encourage the adoption of health data-pooling by requiring federal healthcare institutions and the vendors they contract with to adhere to uniform standards for storing data; second, lawmakers could encourage even greater adoption of health data standards by offering targeted incentives similar to those offered to "Accountable

257. *See* D. Daniel Sokol & Roisin Comerford, *Does Antitrust Have a Role to Play in Regulating Big Data?*, *in* CAMBRIDGE HANDBOOK OF ANTITRUST, INTELLECTUAL PROPERTY & HIGH TECH 271 (Roger D. Blair & D. Daniel Sokol eds., 2017).

258. Even if data pools could avail themselves of the NCRPA, this would only result in reduced penalties.

Care" institutions under the Affordable Care Act; third, the FDA could offer a new expedited review process to pharmaceutical companies that cooperatively pool clinical trial data; fourth, lawmakers could modify HIPAA and other bodies of law designed to preserve patient privacy to encourage the responsible pooling of anonymous treatment data; fifth, income tax incentives could be offered to encourage individual patients to donate their health records to data pools. These suggestions are offered as tailored approaches designed to increase the volume and rate of patient treatment data pooling.

Researchers have glimpsed the future within data pools. It is written in the language of statistics—a language of patterns, signals, and unexpected correlations. Because this new science can spur innovation, the federal government has sought to encourage its development. Missing, however, is a plan for bringing highly dispersed data together, a necessary precondition. This Article has provided theoretical and empirical support for the view that data pools are unlikely to independently form and thrive. A single policy solution seems unlikely to address the many factors that discourage useful cooperation. Policymakers should now seek to understand the collective action problems that stand in the way. From this knowledge, they can assemble new constellations of policies designed to ensure that the cooperative preconditions for innovation are met.