Big Data: Destroyer of Informed Consent

A. Michael Froomkin*

ABSTRACT

The 'Revised Common Rule' took effect on January 21, 2019, marking the first change since 2005 to the federal regulation that governs human subjects research conducted with federal support or in federally supported institutions. The Common Rule had required informed consent before researchers could collect and use identifiable personal health information. While informed consent is far from perfect, it is and was the gold standard for data collection and use policies; the standard in the old Common Rule served an important function as the exemplar for data collection in other contexts.

Unfortunately, true informed consent seems incompatible with modern analytics and 'Big Data'. Modern analytics hold out the promise of finding *unexpected* correlations in data; it follows that neither the researcher nor the subject may know what the data collected will be used to discover. In such cases, traditional informed consent in which the researcher fully and carefully explains study goals to subjects is inherently impossible. In response, the Revised Common Rule introduces a new, and less onerous, form of "broad consent" in which human subjects agree to as varied forms of data use and re-use as researchers' lawyers can squeeze into a consent form. Broad consent paves the way for using identifiable personal health information in modern analytics. But these gains for users of modern analytics come with side-effects, not least a substantial lowering of the aspirational ceiling for other types of information collection, such as in commercial genomic testing.

Continuing improvements in data science also cause a related problem, in that data thought by experimenters to have been de-identified (and thus subject to more relaxed rules about use and re-use) sometimes proves to be re-identifiable after all. The Revised Common Rule fails to take due account of real re-identification risks, especially when DNA is collected. In particular, the Revised Common Rule contemplates storage and re-use of so-called de-identified biospecimens even

* Laurie Silvers & Mitchell Rubenstein Distinguished Professor, University of Miami School of Law.

I am grateful to Jack Balkin, Caroline Bradley, Abbe R. Gluck, Ken Goodman, JoNell Newman, Nicolas Terry, the participants at the Yale Workshop on 'The Law and Policy of AI, Robotics & Telemedicine,' and my student editors Vigjilenca Abazi and James Johnson for their comments on and corrections of earlier drafts, to which I have no doubt added further errors of my own. Thanks to Pam Lucken for research support.

though these contain DNA that might be re-identifiable with current or foreseeable technology.

Defenders of these aspects of the Revised Common Rule argue that 'data saves lives.' But even if that claim is as applicable as its proponents assert, the effects of the Revised Common Rule will not be limited to publicly funded health sciences, and its effects will be harmful elsewhere.

INTRODUCTION

Consent, that is 'notice and choice,' is a fundamental concept in the U.S. approach to data privacy, as it reflects principles of individual autonomy, freedom of choice, and rationality. Informed consent constitutes the most careful and respectful type of notice and consent, the type of consent that, of the options available, best instantiates fundamental ethical principles: Under informed consent we expect the persons collecting data to know what it will be used for, and we expect collectors to make a significant effort to ensure that the people from whom they collect the data have a reasonably accurate understanding of what the data might be used for. Informed consent has problems and limits, but it has been, for some time, the exemplary standard for consent-based data collection, indeed for data-collection in general. A standard-setter and standard-bearer for informed consent has been the informed consent requirement of the so-called Common Rule,[1] the U.S. Federal Government's regulations about federally supported human subjects research.

Big Data makes the U.S. approach to informed consent incoherent and unsupportable, and puts the entire concept of informed consent as currently practiced in the U.S.—and indeed perhaps the entire concept of consent[2]—into question when applied to modern analytics. With modern data mining and analytics, neither party to the original data collection, and to any 'consent' agreement, reasonably can expect to know important facts about how the data might be used. Indeed, the possibility of surprises is part of what makes modern analytics so attractive.

Rather than resist this trend, the U.S. government recently introduced new rules for biomedical research that compromise with it, revising the Common Rule to create a new blanket form of 'broad consent' under which subjects can agree to broad and unforeseen uses of their data and biological samples.[3] Whatever utilitarian justifications exist for this retreat in the context of medicine and public health, the watering down of informed consent as we know it fails to give due consideration to the likely side-effects for consent in other circumstances. Undermining the requirement in the medical context for what was previously the gold standard for consent risks knock-on consequences for many other areas of data collection and use, areas where even meaningful consent, much less informed

1. *See infra* note 33.
2. As explained below, Big Data does not totally kill the concept of consent, but it risks limiting the scope of consent to choosing between signing a blank check and not consenting at all. Since the binary choice remains, consent is not, formally, dead. Giving a blank check, however, is not and cannot be informed consent, especially when neither party to the transaction knows how much the recipient will draw on the account.
3. *See infra* Part II.A.

consent, remains aspirational at best.

Big Data refers to "analytics that can process massive quantities of data in the search for information, including unforeseen information, which can potentially generate unexpected insights. Big Data is characterized by two basic features: first, the possibility of accessing and using large quantities of data (meeting the conditions of the '3 Vs,' namely huge size [volume], created in near real-time [velocity], and diverse [variety]), and, second, the use of data processing techniques that allow for the recognition of previously unidentified patterns."[4] Or, more pithily, "Big Data is less about data that is big than it is about a capacity to search, aggregate, and cross-reference large data sets."[5] A particular data set may or may not contain the subjects' names. But the more data, and the more fine-grained data, available about a person, the easier it becomes to re-identify the person either through reverse engineering the data set,[6] or by matching it with other data sets that do have names included.[7] Thus, in a world of Big Data, even data sets that have been "anonymized" or which appear to lack personally identifiable information carry the risk of re-identification[8] which would undermine consent given to use data on condition that it be de-identified.

Consent is admittedly an imperfect tool. Even before the growth of Big Data, there were good reasons to criticize the regimes in which we required just ordinary, not-especially-informed consent, for most waivers of privacy rights, and did not require even that much before permitting the capture of information streams 'in public.' Sociology-based critiques suggested that most people ignore most notices most of the time.[9] Cognitive critiques suggested that even if people do look at

4. European Parliament, Directorate-General for Internal Policies, Policy Department, Citizen's Rights and Constitutional Affairs, Big Data and Smart Devices and Their Impact on Privacy (2015), http://www.europarl.europa.eu/RegData/etudes/BRIE/2016/589801/EPRS_BRI(2016)589801_EN.pdf.

5. danah boyd & Kate Crawford, *Critical Questions for Big Data: Provocations For a Cultural, Technological, and Scholarly Phenomenon*, 15 INFORMATION, COMMUNICATION & SOCIETY 662, 663 (2011). An estimated 1.7 megabytes of information will be created every second for every human on the planet in the next three years, during which the accumulated digital universe of data will grow from 4.4 zettabytes today to around 44 zettabytes. Laura Bednash, *The Future of Data Storage: From Gigabytes to Petabytes*, RACKTOP (Feb. 12, 2018), https://www.racktopsystems.com/future-data-storage-gigabytes-petabytes/. A terabyte is equal to 1,024 gigabytes. A petabyte is equal to 1,024 terabytes. An exabyte is equal to 1,024 petabytes. A zettabyte is equal to 1,024 exabytes, or one sextillion (10^{21}) bytes. Of course, not all this data is personal information.

6. *See generally* Boris Lubarsky, *Re-Identification of "Anonymized" Data*, 1 GEO. L. TECH. REV. 202 (2017), https://georgetownlawtechreview.org/re-identification-of-anonymized-data/GLTR-04-2017/.

7. *See, e.g.*, Dániel Kondor et al., *Towards matching user mobility traces in large-scale datasets*, IEEE Xplore, DOI: 10.1109/TBDATA.2018.2871693 (Sept. 24, 2018).

8. See *infra* Part II.B.

9. *See* OMRI BEN-SHAHAR & CARL E. SCHNEIDER, MORE THAN YOU WANTED TO KNOW: THE

disclosures, they likely do not understand them,[10] not to mention that as notices and requests for consent proliferate, people become desensitized to them and tune them out.[11] Indeed, it seems likely that even if people suddenly had perfect information about the devices that watch them they would not be able to use that information well. The simplest form of the bounded rationality claim has to do with the amount of time it would take to process all the information and make rational decisions. More far-reaching forms of the claim invoke various cognitive limits constraining our ability to weigh risks and uncertainties,[12] and our tendency to over-optimism.[13]

Yet, despite all its flaws, consent remains one of the best tools we have for limiting excessive collection of personal data and for respecting the autonomy and freedom of the individual. And *informed* consent remains the type of consent best calculated to uphold these values and protect the individual.

Unfortunately, Big Data analytics—the seeking of patterns and correlations in giant datasets—kills the possibility of true informed consent in two ways. At an individual level, Big Data kills consent because by its very nature one purpose of big data analytics is to find *unexpected* patterns in data. Informed consent requires at the very least that the person requesting the consent know what she is asking the subject to consent to. In principle, we hope that before the subject agrees she too comes to understand the scope of the agreement.[14] But with big data analytics, particularly those based on Machine Learning, *neither party to that conversation can know what the data may be used to discover*.[15] Also, given advances in re-identification, *neither party can know how likely it is that any given attempt to de-identify personal data will succeed*. Informed consent, at least as we used to understand it, is simply not possible if medical data is to become part of Big Data, and ever so much more so if researchers intend to link personal health records with data streams drawn from non-medical sources because what we will learn with the

---

FAILURE OF MANDATED DISCLOSURE, 107–18 (2014). *But see* Margret Jane Radin, *Less Than I Wanted To Know: The Submerged Issues in* More Than I Wanted To Know, 11 JERUSALEM L. REV. 51 (2015), https://doi.org/10.1093/jrls/jlu019.

10. BEN-SHAHAR & SCHNEIDER, *supra* note 9, at 101; Daniel J. Solove, *Introduction: Privacy Self-Management and the Consent Dilemma*, 126 HARV. L. REV. 1880 (2013).

11. BEN-SHAHAR & SCHNEIDER, *supra* note 9, at 104–06.

12. *See* DANIEL KAHNEMAN, THINKING, FAST AND SLOW (2011); *Herbert Simon, Theories of Bounded Rationality*, *in* DECISIONS AND ORGANIZATION 161 (C.B. McGuire & Roy Rader eds. 1972); Russell Korobkin, *Bounded Rationality, Standard Form Contracts, and Unconscionability*, 70 U. CHI. L. REV. 1203 (2003); Owen D. Jones, *Time-Shifted Rationality and the Law Of Law's Leverage: Behavioral Economics Meets Behavioral Biology*, 95 NW. U. L. REV. 1141 (2001).

13. *See* KAHNEMAN, *supra* note 12.

14. *See infra* notes 28–29 and accompanying text.

15. *See* Solon Barocas & Helen Nissenbaum, *Big Data's End Run Around Procedural Privacy Protections*, 57 COMM. ACM 31, 32 (Nov. 2014).

information cannot be predicted. Similar—or perhaps worse—problems arise with big data analytics uses outside the context of medical research, especially as, there, informed consent had seemed a plausible solution to the problem of routinized or non-existent consent for non-health-related data acquisition. The more personal data there is, the greater the possible privacy harms, and the greater the re-identification risk becomes.

Big Data potentially also undermines the idea of consent at a group level. Big Data analytics rely on having a mountain of data about many people in order to find the correlations that allow researchers to make relatively accurate predictions about individuals including those whose data did not contribute to building the model. In other words, even people who do not consent to have their data in a database may find that they are just as subject to the models' extrapolations as the people who contributed data to it.[16] Thus, in addition to making true consent impossible in individual cases, when enough people are surveyed to give a sample that is a good approximation of the whole, Big Data threatens to make consent irrelevant for those who resist, as their participation may not be needed. In some very general sense, it was always true that clinical data derived from a small and one hopes representative population ("studies show this vaccination prevents XYZ disease") would extrapolate to a larger one; in such cases human subject consent had important positive externalities.[17] Big Data, however, changes the calculus in that now data derived from a large population can in some circumstances generate particularized predictions about small populations, or even individuals, outside the study group. Some of those will be unwelcome. Currently, these correlations present as risks to personal privacy in the context of much non-medical data;[18] perhaps someday in the medical context they will also present as opportunities for precision medicine, but today that remains more hope or prediction than reality.[19]

Big Data just keeps getting bigger and thus more attractive to researchers in

16. Although most of this essay is about personal rights, the issue of group rights is also significant. *See infra* text at note 108.

17. That said, the externalities caused by some discoveries were not always positive. For example, identifying genetic sources of disease had some adverse externalities: Some people may have preferred not to know; others encountered genetic discrimination in employment or insurance. The risk of genetic discrimination motivated the passage of the Genetic Information Nondiscrimination Act of 2008, Pub. L. No. 110-233, 122 Stat. 881 (codified as amended in scattered sections of 29 & 42 U.S.C.) [hereinafter GINA]. GINA only covers employers and health insurers, but genetic discrimination can manifest in many other scenarios, e.g. housing, schooling, or sports.

18. *See* LINNET TAYLOR, LUCIANO FLORIDI & BART VAN DER SLOOT, GROUP PRIVACY: NEW CHALLENGES OF DATA TECHNOLOGIES 10–17 (2016).

19. "Although some niche applications have been found for precision medicine, and gene therapy is now becoming a reality for a few rare diseases, the effects on public health are minuscule while the costs are astronomical." Michael J. Joyner & Nighel Paneth, *Precision medicine's rosy predictions haven't come true. We need fewer promises and more debate*, STAT NEWS (Feb. 7, 2019), https://www.statnews.com/2019/02/07/precision-medicine-needs-open-debate/.

everything from medicine to marketing. The data arise from multiple sources: some are transactional, including both online and offline commerce; some are communications such as texts, email, video chat; some is collected from personal devices including cell phone, cell phone apps, wearable health monitors, smart watches, so-called smart home technology, and other internet-of-things devices. Another important source of Big Data is self-surveillance, in which people document their activities—and importantly, those of others—via Twitter, Facebook, Instagram, and other platforms. Increasingly, also, personal information is collected via the operation of remote sensors, whether security systems such as ccTV (increasingly paired with facial recognition software), license plate readers, or the myriad data collection programs that form so-called Smart City initiatives. On deck are connected cars, implantable devices, and the dreams of the next startup.

Data collection may occur in situations—such as remote sensing—where consent is not required, or it can arise from situations where consent is a legal prerequisite to data collection—such as in the transactional context. As a general matter, in the U.S. many types of Big Data collection, including many lifestyle and demographic data that might be of interest to public health researchers, do not currently require any consent at all and may be available for sale from data brokers.[20] For other types of data streams, even if consent may be required the consent need not be highly informed.[21] Sometimes a boilerplate clause in a long contract will do. Other times a notice that "these premises are monitored" suffices. But very frequently, nothing is required at all.

It is now a commonplace that the aggregation of all this data creates occasions for modern data analytics to create new information about people. It is not too early to be concerned about the consequences of having governments, marketers, insurers, or political candidates[22] assemble profiles and acquire the ability to make (allegedly) higher-quality predictions about us, even if we do not yet know the full capabilities and consequences of rapidly evolving analytics. It is already clear that there are big, conceivably fatal, problems for the protection of personal privacy in the U.S. Since the U.S. appears unwilling (and may be constitutionally unable[23])

20. *See* Steven Melendez & Alex Pasternack, *Here are the Data Brokers Quietly Buying and Selling Your Personal Information*, FAST COMPANY (Mar. 2, 2019), https://www.fastcompany.com/90310803/here-are-the-data-brokers-quietly-buying-and-selling-your-personal-information (listing 121 data brokers and the information they collect and sell).

21. Examples include the data streams from health monitoring apps, fitness trackers, and even smart cities.

22. Indeed, the early returns here, from the Facebook-Cambridge Analytica scandal, are not encouraging. *See*, *e.g.*, Karl Manheim & Lyric Kaplan, *Artificial Intelligence: Risks to Privacy and Democracy*, 21 YALE J. L. & TECH. 106, 139 (2019) ("Cambridge Analytica is the poster child for inappropriate use of data.").

23. See A. Michael Froomkin, *Regulating Mass Surveillance as Privacy Pollution: Learning*

to adopt European-style data protection law, we must find other means to at least blunt its impact. Short of outlawing analytics, and forgoing the promise of the scientific and medical advances that they promise, we are unlikely to find single-factor solution to the privacy problems they generate.

If the full effects of Big Data analytics have yet to come into focus, the solutions to the problems they will cause are even less certain. Seemingly innocuous data collection may become dangerous when it permits unexpected deductions, or when paired with more private data. Whatever the solution, however, and whether it comes at the point of collection or the time of re-use, it will need to confront questions of consent. Because, as explained further below, the revision of the Common Rule opens the door wider to the commingling of medical data with other data streams, the consent problem needs to be considered in multiple contexts including those where, as noted, some data currently can be collected without a need for any sort of consent. And even when consumers know they are sharing information, when it comes to big data they cannot know—given the nature of modern analytics—what the consequences of sharing may be.[24]

Misuse of data can be addressed with controls on how and when data are collected, or by regulating the use and re-use of data after it is collected, or both. Consent focuses on the collection side, and it is particularly important in the U.S. because the U.S. lacks, and seems unlikely to adopt, systemic controls on the storage and processing of personal data. The EU's GDPR includes significant limitations on data storage and re-use: personal information collected for one purpose cannot in the main be used for another without the data subject's permission.[25] U.S. law contains few such limitations outside the medical and human subjects contexts, and it is unclear to what extent it can or will prevent the sharing of data collected privately[26] and outside certain special relationships that create a duty of confidentiality.

If consent matters, it matters also that it be genuine and knowing consent. But

---

from Environmental Impact Statements, 2015 U. ILL. L. REV. 1713, 1778–81 (describing First Amendment obstacles to U.S. adopting EU GDPR-style rules against data re-use).

24. Thus, big data creates an even more extreme form of what I elsewhere dubbed "privacy myopia." See *Id.* at 1733–37.

25. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ L [2016] 119/1 [hereinafter GDPR]. The GDPR imposes a higher "explicit consent" requirement for disclosures of health data compared to the ordinary "consent" required for disclosures of other data types. See Article 29 Working Party Guidance on Consent Under Regulation 2016/679 (Rev. Apr. 10, 2018), https://ec.europa.eu/newsroom/article29/document.cfm?action=display&doc_id=51030; See also GDPR, *supra*, at Article 4(11), 7.

26. The limits in the Common Rule have traction because they are a condition of various types of federal funding. See *infra* text at notes 33–34.

in the U.S. consent to data gathering can mean many things ranging from the demanding informed consent of medical ethics to seeing a sign saying that cameras are operating on the premises or signing a contract of adhesion. On that spectrum, informed consent as practiced in medicine and human subjects research has been and remains the gold standard. Yet even informed consent has its critics, who argue that it does not achieve much: Fewer than ten percent of patients refuse consent to data collection from their records for research purposes.[27] Furthermore, a significant minority of patients do not understand the informed consent procedure at all; most do not read the form 'carefully,'[28] which may be why it is common to have a researcher explain to the subject what is being consented to.[29] Nevertheless, despite its faults, at least in informed consent the party intending to get and use personal information makes a genuine effort to ensure that the person agreeing understands what they are getting into. That has, or should have value, even if the only thing it does is make the parties aware of what is at stake.[30]

If informed consent is to remain meaningful in the face of Big Data, we will need to find new ways to predict and explain what analytics are doing or are likely to do. Those means do not exist at present, but researchers have only begun to address the issues. While we wait to see how much forward-looking explanation we can engineer into analytics, and how we might create appropriate feedback loops that allow data subjects some say in what their data gets used for—or at least by whom and for what goals their data gets used—we need two things: First, we need more-fully informed consent so that both subjects and investigators understand the stakes. Second, we need better means to meta-tag information so that researchers can go back to subjects to see if they consent to secondary uses. And we also need a way to ensure that the meta-tagging data remains closely held and does become a new channel to re-identify 'anonymized' information shared with researchers.

---

27. R. Baker, et al, *What Proportion of Patients Refuse Consent to Data Collection From Their Records For Research Purposes?*, 50 B.J. OF GEN. PRAC. 655 (2000).

28. Barrie R. Cassileth et al, *Informed Consent—Why Are Its Goals Imperfectly Realized*, 302 NEW ENGLAND J. MED. 896, 896 (1980). *Cf.* Aleecia M. McDonald & Lorrie Faith Cranor, *The Cost of Reading Privacy Policies*, 4 I/S: A JOURNAL OF LAW AND POLICY FOR THE INFORMATION SOCIETY 541 (2008).

29. "[C]onsent is a process, and not just a form that potential study participants must sign." Johns Hopkins Medicine Office of Human Subjects Research—Institutional Review Board, *Guidelines*, https://www.hopkinsmedicine.org/institutional_review_board/guidelines_policies/guidelines/informed_consent_i.html.

30. In this, informed consent may have something in common with Environmental Impact Assessments, and with Privacy Impact Assessments, both of which are argued to make organizations consciously confront the side-effects of their actions. Cf. Kenneth A. Bamberger & Deirdre K. Mulligan, *PIA Requirements and Privacy Decision-Making in U.S. Government Agencies*, in PRIVACY IMPACT ASSESSMENTS 245 (Wright & De Hert eds. 2012).

More importantly, we need to avoid giving up in the face of the challenges of Big Data. We should not reshape our legal and ethical systems to accept less-informed or uninformed consent in order to smooth the way for the greater agglomeration of personal data. Informed consent as originally implemented in the Common Rule not only had value of its own, it had value as an exemplary consent regime; reducing the requirement for informed consent in federally supported human subjects research also reduces the Common Rule's value as a model for other forms of data collection.

The rest of this essay proceeds as follows. Part I begins with a general introduction to the role of informed consent in medicine and human subjects research. It then explains in Part II how the Revised Common Rule weakens the informed consent requirements exacted by its predecessor. One critical change is the introduction of a new alternate form of 'broad consent,' a change Part II.A explains is designed to make it easier for biomedical researchers to use modern data analytics (aka 'Big Data'). As the Revised Common Rule imposes one set of rules on personally identifiable data, but different and less stringent rules on non-identified and particularly so-called de-identified data, Part II.B suggests that re-identification is a much greater risk than the Revised Common Rule seems to anticipate.

Part II.C pivots to the side-effects of the Revised Common Rule on biomedical research outside its purview, and Part II.D considers its likely deleterious effects on privacy policies more generally, which I argue is a serious and unfortunate problem. The essay concludes in Part III with some thoughts about the future interaction of Machine Learning and Big Data and how we might cope with the privacy consequences of that combination.

I. MEDICINE AND HUMAN SUBJECTS RESEARCH MEET BIG DATA

A. Informed Consent as a Legal Default

Medical treatment and federally-funded (but not fully private)[31] human

31. Even if the Common Rule does not apply, medical research on human subjects may be regulated by the Health Insurance Portability and Accountability Act of 1996 (HIPAA), Pub.L. 104–191, 110 Stat. 1936, codified as amended at 42 U.S.C. § 201 (2006) et seq. HIPAA imposes a lesser consent regime than the Common Rule, requiring only "authorization," 45 C.F.R. § 164.508 (2019), not "informed consent" for the use of medical data. HIPAA applies to "covered entities" as defined in 45 C.F.R. § 160.103 (2019), which includes health plans and health care providers that transmit health information in electronic form. *Id.* Firms that are not HIPAA covered entities, such as private pharmaceutical and biotech companies, may still be affected by HIPAA indirectly when they interact with providers, payors, patients, and others that have HIPAA compliance obligations and/or HIPAA-granted rights. *Cf.* Glenn Cohen & Michelle M. Mello, *HIPAA and Protecting Health Information in the 21st Century*, 230 J. AM. MED. ASSOC. 231 (2019).

subjects research are unusual in the U.S. in that for them U.S. law imposes a substantially heightened consent requirement. U.S. law requires physicians in all but the most critical emergency situations to get "informed consent" from patients before they undergo medical treatment. By contrast, the normal, default, rule for commercial transactions, absent some special duty of confidentiality, is that the information is jointly and severally available to both parties.[32] Activities observable in public commonly require no consent to be recorded and analyzed, although certain doctrines such as the right of publicity and may impose limits on commercial publication if not necessarily private use of the data.

The "Common Rule,"[33] the federal regulation that since 1991 (last revised in 2005) has governed federally funded or sponsored human subjects research and, in some cases, all human subjects research at institutions that conduct federally funded research, also requires informed consent from all covered subjects before researchers acquire, much less use, personally identifiable information. The Common Rule required a researcher to get permission from an Institutional Review Board (IRB)[34] before re-using identifiable data for research outside the scope of a

32. A. Michael Froomkin, *The Death of Privacy?*, 52 STAN L. REV. 1461, 1521–22 (2000).

33. Codified (as amended in 2005) at 45 C.F.R. §§ 46.101–.124 (2018).

> The Federal Policy for the Protection of Human Subjects, which is better known as the "Common Rule," was published in 1991 and codified in separate regulations by 15 federal departments and agencies: the Department of Agriculture, Commerce, Defense, Energy, Education, Health and Human Services, Housing and Urban Development, Labor, Transportation, and Veterans Affairs, as well as the Agency for International Development, Consumer Product Safety Commission, Environmental Protection Agency, National Aeronautics and Space Administration, and National Science Foundation. Two other agencies - the Central Intelligence Agency and the Department of Homeland Security - comply with all subparts of 45 C.F.R. Part 46, which are the HHS "Common Rule" regulations. In addition, the Social Security Administration, which was separated from HHS in 1994 and, absent action by the Administrator, must apply all regulations that applied to SSA before the separation.

Mary Bernadette Ott and Gary Yingling, 1 GUIDE TO GOOD CLINICAL PRACTICE ¶ 1250 (2015).

34. IRBs are controversial in ways well beyond the scope of this short paper. For examples of arguments that IRBs systematically under-regulate see Carl H. Coleman, *Rationalizing Risk Assessment in Human Subject Research*, 46 Ariz. L. Rev. 1, 3 (2004); Barbara Evans, *Ethical and Privacy Issues in Pharmacogenomic Research*, in Pharmacogenomics: Applications To Patient Care 328 (Howard L. McLeod et al. eds., 2d ed. 2009). For examples of arguments that IRBs systematically over-regulate see Scott Burris, *Regulatory Innovation in the Governance of Human Subjects Research: A Cautionary Tale and Some Modest Proposals*, 2 REG. & GOVERNANCE 65, 67–68 (2008); Robert Charrow, *Protection of Human Subjects: Is Expansive Regulation Counter-Productive?*, 101 NW. U. L. REV. 707, 708–09 (2007); Todd J. Zywicki, *Institutional Review Boards as Academic Bureaucracies: An Economic and Experiential Analysis*, 101 NW. U. L. REV. 861, 861 (2007).

subject's original, informed, consent. Otherwise, the only way to re-use the data was to ensure it was not identifiable (in which case the Common Rule would not apply) or get fresh informed consent. In the years since 1991, however, we have learned that de-identification is difficult, perhaps impossible, in many ordinary research scenarios; linking patient records with other big data sources makes re-identification all the more likely because it gives would be re-identifiers more fine-grained information to work with.

B. The Revised Common Rule Accommodates Big Data

Changes in the research landscape as well as changes in the modern understanding of informed consent led to a push for relaxation of some of the strictures of the Common Rule and clarification of other parts. As the Preamble to the Revised Common Rule, which took effect January 19, 2019, notes,

> Research with human subjects has grown in scale and become more diverse. Examples of developments include: an expansion in the number and types of clinical trials, as well as observational studies and cohort studies; a diversification of the types of social and behavioral research being used in human subjects research; increased use of sophisticated analytic techniques to study human biospecimens; and the growing use of electronic health data and other digital records to enable very large datasets to be rapidly analyzed and combined in novel ways.[35]

Big Data—the use of "very large datasets . . . rapidly analyzed and combined in novel ways"—loomed large among these changes. The new data sets vary in their nature. Some large consist of purely medical and clinical data, e.g. genomic data or electronic medical records. Others seek to meld medical data with other external information. In so doing they create new dangers of re-identification.

Since the key promise of big data is the discovery of unexpected patterns, medical and public health researchers seek to link patient records with other 'non-patient' data streams. And why not-once one has consent for using the patient data, the temptation to link it when possible to other data, usually acquired without the need for consent, or with mere 'ordinary' rather than informed consent, is sure to be irresistible. Some have argued that existing consent rules, especially as regards information-sharing, should be loosened so that society can "derive the benefits of medical research in the form of improved health and health care."[36] As one critic

35. Dept. of Homeland Security et al., Federal Policy for the Protection of Human Subjects, 82 Fed. Reg. 7149, 7150/1 (Jan. 19, 2017) [hereinafter "Revised Common Rule"].

36. INSTITUTE OF MEDICINE, BEYOND THE HIPAA PRIVACY RULE: ENHANCING PRIVACY,

of the practice summarized the case,

> These data sets are out there, readily available at our fingertips, and so easy to analyze. Why not do this? They are not really research: They have already been collected, or they will be collected anyhow as part of some routine, and nothing different will be done to the patients because of the data collection. At the same time, these data sets and their analyses are praised as extremely important research, capable of yielding tremendous discoveries. Biases such as consent bias[37] are then blocking these fascinating discoveries, science, medical progress, and the public good.[38]

Similar pressures exist abroad. For example, the British government recently announced plans to combine social media records with medical records to create a "predictive prevention" health system.[39] Although the GDPR does not use the term 'broad consent' as such, GDPR Recital 33 contemplates "data processing for scientific research purposes" that it was "not possible to fully identify … at the time of data collection" so long as "[d]ata subjects … have the opportunity to give their consent only to certain areas of research or parts of research projects to the extent allowed by the intended purpose."[40] The fear of competitive pressure from

---

IMPROVING HEALTH THROUGH RESEARCH 35 (2009). *See also* Charles Safran et al, *Toward a National Framework for the Secondary Use of Health Data: An American Medical Informatics Association White Paper*, 14 J. AM. MED. INFORMATICS ASSOC. 1 (2007) ("Secondary use of health data must become a priority for policymakers in the U.S.").

37. "Consent bias" is a type of selection bias said to exist when the group giving researchers access to their data differs from the group denying access. *See* M.A. Rothstein & A.B. Shoben, *Does Consent Bias Research?*, 13 AM. J. BIOETHICS 27 (2013).

38. John P. A. Ioannidis, *Informed Consent, Big Data, and the Oxymoron of Research That Is Not Research*, 13 AM. J. BIOETHICS 40, 40 (2013).

39. Chris Smyth, *NHS Will Use Phone Data to Predict Threats to Your Health*, TIMES (Oct. 6, 2018), https://www.thetimes.co.uk/article/nhs-will-use-phone-data-to-predict-threats-to-your-health-r7085zqfq.

40. GDPR, *supra* note 25, at Recital 33. Opinions seem to differ as to how liberally to actualize this: The Irish Health Research Regulations issued in 2018 do not use the term 'broad consent' but they do permit relatively broad data collection and use when researchers secure "explicit consent" from the research subject for "for the purpose of the specified health research, either in relation to a particular area or more generally in that area or a related area of health research, or part thereof." Data Protection Act 2018 (Section 36(2)) (Health Research) Regulations 2018 (SI 314/2018) (Ir.) § 3(1)(e), http://www.irishstatutebook.ie/eli/2018/si/314/made/en/pdf. In contrast, the Association of German Supervisory Authorities (DSK) recently suggested that under German law broad consent should be used only when "absolutely necessary" and even then with safeguards. *See* Covington & Burling LLP, *Association of German Supervisory Authorities issues paper on broad consent for research*, https://www.insideprivacy.com/data-privacy/association-of-german-supervisory-authorities-issues-paper-on-broad-consent-for-research/ (2019) (summarizing Datenschutzkonferenz paper issued

foreign researchers may also be motivating the U.S. push towards new rules that make Big-Data-based research more feasible and require less paperwork.

II. HOW THE NEW COMMON RULE WEAKENS INFORMED CONSENT

On January 21, 2019, the revised version of the Common Rule came into force.[41] Under the Revised Common Rule, when seeking informed consent human subjects researchers still need to try to explain to subjects what is being asked of them. The Revised Common Rule establishes a 'reasonable person' standard for what investigators must tell prospective participants:[42] Information provided to the subject (or her representative) must "begin with a concise and focused presentation of the key information that is most likely to assist a prospective subject or legally authorized representative in understanding the reasons why one might or might not want to participate in the research."[43] Information provided in an informed consent form "must be presented in sufficient detail relating to the research, and must be organized and presented in a way that does not merely provide lists of isolated facts, but rather facilitates the prospective subject's or legally authorized representative's understanding of the reasons why one might or might not want to participate."[44] Furthermore, to get informed consent a researcher must provide an explanation of the research, foreseeable risks or discomforts, how (if at all) the confidentiality of records will be maintained, and a promise that any biospecimens will either have identifiers removed or that they will not be used or distributed for future research.[45]

So far, so good. Applying the 'reasonable person' standard to data being collected with an eye towards modern analytics likely would lead one to conclude that there is no way for a researcher to tell a subject the full gamut of possible risks of the analytics. Unfortunately, the Revised Common Rule also creates an avenue—'broad consent'—that when invoked will permit substantially less-informed consent. (Readers are cautioned to distinguish between on the one hand the number of matters that require informed consent, i.e. the *quantity* of things subject to informed consent, and on the other hand the amount of information disclosed when informed consent is required, i.e. the *quality* of informed consent.

April 3, 2019 and available in German at https://www.datenschutzkonferenz-online.de /media/dskb/20190405_auslegung_bestimmte_bereiche_wiss_forschung.pdf).

41. The US government delayed the effective date of the Revised Common Rule to January 21, 2019. *See* 83 Fed. Reg. 28497 (June 19, 2018).

42. *See* 80 Fed. Reg. at 7221/1.

43. Revised Common Rule, § __.116(a)(5), 82 Fed. Reg. 7265/3 (codified at 42 C.F.R. § 46.116(a)(5)).

44. *Id.* at 7265/3–7266/1.

45. *Id.* at 7266/1–2.

I will refer to a decrease in the frequency of informed consent as "less-frequent informed consent" and a decrease in the quality of information—which may include less information, or less-complete information, or both—as "less-informed consent").

For biomedical research, the change to requiring both less-frequent informed consent and less-complete advance disclosure is and can be defended on utilitarian grounds: it may permit new valuable life-saving research, speed the path towards precision medicine, and might even open up new realms of public health and epidemiological understanding.[46] Even if one accepts these claim as true to the extent that they concern federally funded medical research, however, one should also consider the side-effects of this change for other types data collection, ranging from biomedical data collection not covered by the Common Rule to privacy in public. The Revised Common Rule, due to its nature as a project focused on biomedical research, does not do this.

A. Enter "Broad Consent"

The Revised Common Rule address the tension between informed consent and Big Data by, in effect, creating a work-around to true informed consent. The Revised Common Rule will permit researchers to get "broad consent"—"prospective consent to unspecified future research"[47]—instead of requiring informed consent, or even ordinary consent, on a case-by-case basis. Researchers armed with sufficiently broad consent for the storage, maintenance, and use of identifiable[48] biospecimens and data may make any secondary research uses of the individual's identifiable biospecimens and data without the need to secure any additional consent. In addition, researchers who de-identify data biospecimens qualify for an exemption to the Revised Common Rule when conducting secondary research on them.

The Revised Common Rule's 'broad consent' contemplates identifiable information and biospecimens being held indefinitely.[49] However, a researcher asking a subject for broad consent must tell the subject that "the subject may

46. *See infra* notes notes 103–106 and accompanying text. *But see infra* notes 107–111 and accompanying text.

47. *Id.* at 7150.

48. A separate option contemplates the removal of identifiers from biospecimens followed by their use for "future research studies" or distribution "to another investigator for future research studies without additional informed consent," see *Id.* at § __.116(b)(9)(i), 82 Fed. Reg. 7266/2 (codified at 45 C.F.R. § 46.116(b)(9)(i)). Since biospecimens might be genotyped, one can reasonably wonder about the long-run stability of de-identification by removal of identifiers attached to the sample. See *infra* note 68 and accompanying text.

49. *Id.* at § __.116(d)(4).

discontinue participation at any time,"[50] and will suffer no negative consequences for withdrawing. So long as consent is not withdrawn, the only limit on the type of research re-use is the imagination of the drafters of the consent form, because the disclosure to the subject must provide a "general description" of the "types of research" that may be conducted with the private information or biospecimens,[51] as well as "the types of institutions or researchers that might conduct research with" them.[52] The disclosures must suffice to put a reasonable person on notice that they might not have consented to some of the specific research studies had they known what they were—a long way from true informed consent. Seemingly cognizant of this, the Preamble to the Final Rule states, in what may be precatory language, that, "It is envisioned that for certain types of research, such as research for which there is reason to believe some subjects will find the research controversial or objectionable, a more robust description of the research will be required in order to meet this 'reasonable person' standard."[53]

However, so long as the researchers make their intentions as whether they will share information clear, there is no obligation to identify the specific studies in which the data and biospecimens are used,[54] nor to share the results (much less the proceeds) with the subjects,[55] so the chances of subjects being concerned after their original broad consent are, to say the least, low. Exactly what researchers will ask subjects to agree to remains uncertain, since the final version of the Revised Common Rule contains no federally mandated, or even default, standard form for getting broad consent. A standard form would have concentrated minds on how little advance information would suffice; it would also presumably have been of value to researchers and institutions who would have a safe harbor to limit any legal liability or impediment to continued federal funding. Institutions appear to be proceeding very cautiously as a result.

In any event, were any subjects to withdraw their consent, that choice would only operate prospectively. Thus, by the time Big Data has served up its surprises, there is nothing the subject can do. Additionally, in some circumstances the Revised Common Rule allows researchers to use information from subjects who

50. *Id.* at § __.116(d)(1) (referencing § __.116(b)(8)).

51. *Id.* at § __.116(d)(2). Could the disclosure be very general ("we will use it to further research in public health")? While that indeed might be considered a request for some sort of consent, I would argue that it is not in any meaningful sense *informed* consent since the distinguishing feature of the disclosure is an absence of any meaningfully detailed information. *See supra* note 2.

52. *Id.* at § __.116(d)(3).

53. 82 Fed. Reg. at 7221/1.

54. Revised Common Rule, § __.116(d)(5). For a very nuanced article which among other things explains why sharing is critical to modern data usage and regulation see Omer Tene & Jules Polonetsky, *Big Data for All: Privacy and User Control in the Age of Analytics*, 11 NW. J. TECH. & INTELL. PROP. 239 (2013).

55. *Id.* at § __.116(d)(6).

did not give their informed or even broad consent, so long as the work is conducted or approved by state or local government officials and studies "public benefit or service programs"—including changes in levels of benefit—and an IRB finds that the research could not "practicably be carried out" otherwise.[56] In other words, sometimes no informed consent may be needed to study the effects of welfare, Medicaid, or Medicare.[57] Peculiarly, the Revised Common Rule instructs IRBs only to allow such waivers if they find that the use of the information "will not adversely affect the rights and welfare of the subjects." And, just in case someone might be inclined to protest that their rights were being violated, the Revised Common Rule does not require the researchers to tell the subjects what their information was used for until after the fact, and then only "whenever appropriate"![58] Neither the previous Common Rule, nor the Revised Common Rule create a private right of action for privacy violations.[59] The creation of new avenues of "broad" consent and consent-by-regulatory-fiat presumably creates some new avenues for data collection and use without fear of litigation or loss of federal funding.

In sum, although the Revised Common Rule modernizes the criteria for informed consent, those requirements do not—and literally cannot—apply to 'broad consent' for use in Big Data research since, again, neither party necessarily can even imagine what the biospecimens and data might be used for in the future. Nor, when data is acquired on the understanding that it will be de-identified, can we know with confidence how the science of re-identification will progress. While getting virtual *carte blanche* from patients and research subjects may solve the formal legal problem from the researcher's point of view, it makes a mock of informed consent and should raise some ethical qualms for human subjects research, especially since "[i]nformed consent has been the cornerstone of conducting ethical research involving humans."[60] Furthermore, given the unforeseeable leaps in technology (who foresaw Big Data 20 years ago?) there are good reasons to think consent should sunset[61] rather than institutionalizing its eternalization.

Of course, much research with large data sets will continue to involve the traditional generation and testing of hypotheses. Researchers may continue to

56. Revised Common Rule, § __.116(e)(i).
57. *Id.* at § __.116(f)(3)(iv).
58. *Id.* at § __.116(f)(3)(v).
59. Ahsin Azim, Note, *Common Sense: Rethinking the New Common Rule's Weak Protections for Human Subjects*, 71 VAND. L. REV. 1703, 1712, 1714 n.80 (2018).
60. John P. A. Ioannidis, *Informed Consent, Big Data, and the Oxymoron of Research That Is Not Research*, 13 AM. J. BIOETHICS 40 (2013).
61. *See, e.g.*, Bart Custers, *Click Here to Consent Forever: Expiry Dates for Informed Consent*, BIG DATA & SOCIETY 1 (Jan-June 2016).

request informed consent, if they choose to. But when it comes to Machine Learning, at least, there is now a new pathway with unpredictable results.

B. Reidentification Risk

The Revised Common Rule excludes all secondary research from its coverage when "[i]nformation, which may include information about biospecimens, is recorded by the investigator in such a manner that the identity of the human subjects cannot readily be ascertained directly or through identifiers linked to the subjects, the investigator does not contact the subjects, and the investigator will not re-identify subjects."[62] This exemption for so-called de-identified data goes beyond its predecessor, in that henceforth the exemption will be available to researchers who collect information or biospecimens and "remove identifiers,"[63] as opposed to the earlier rule that focused on not collecting the identifiers in the first place.[64]

The mere 'removal of identifiers' is not nearly enough to prevent re-identification (and what exactly constitutes de-identification is itself far more complicated than it may seem[65]). New medical research techniques create a great risk of re-identification of both biospecimens and of so-called de-identified personal records.[66] Experience has shown that what we think is de-identified at one time turns out to be identifiable at another, either because of poor technique, or because new techniques, unsuspected at the time of 'de-identification', make later re-identification feasible.[67] The risk of re-identification is particularly great with DNA.[68]

---

62. § __.104(d)(4)(i), 82 Fed. Reg at 7262/2 (codified at 45 C.F.R. § 46.104(d)(4)(i)).

63. *See* 82 Fed. Reg 7194/2 (exemption for when "information is recorded by the investigator in such a manner that the identity of human subjects cannot readily be ascertained directly or through identifiers linked to the subjects, the investigator does not contact the subjects, and the investigator will not re-identify subject.").

64. *Id.* (noting "that prior exemption is being extended to now also cover research with information for which identifiers have been removed when the original collection of information or biospecimens occurs in the future.").

65. *Cf.* Jules Polenetsky, Omer Tene & Kelsey Fintch, *Shades of Gray: Seeing the Full Spectrum of Practical Data De-Identification*, 56 SANTA CLARA L. REV. 593 (2016).

66. "The primary risk raised by research is the unintended revelation of the human subject's identity." Azim, *supra* note 59, at 1725.

67. *See, e.g.*, Linda Carroll, *Anonymous Patient Data May Not Be as Private as Previously Thought*, REUTERS (Dec. 21, 2018), https://www.reuters.com/article/us-health-privacy-idUSKCN1OK24O.

68. *See* Mellisa Gymrek, et al., *Identifying Personal Genomes by Surname Inference*, 339 SCIENCE 321 (Jan 18, 2013), https://science.sciencemag.org/content/339/6117/321.full (demonstrating that "surnames can be recovered from personal genomes by profiling short tandem repeats on the Y chromosome (Y-STRs) and querying recreational genetic genealogy databases" and that "a combination of a surname with other types of metadata, such as age and state, can be used to

The Revised Common Rule addresses this issue in at least three ways. First, it requires that researchers who seek to shelter under its exemption for work with de-identified data promise that they will not attempt re-identification,[69] although it is unclear how this will be enforced.

The Preamble to the Revised Common Rule also argues that to the extent individuals require additional protections, they are provided by HIPAA's much more demanding conditions,[70] although HIPAA in the main does not apply to de-identified data.[71] Whether this will suffice to protect the donors of de-identified data and biospecimens depends in part on the extent to which the data and samples are properly segregated from other data or if they become commingled with data sets that permit re-identification. It will also depend on the extent to which published research products unintentionally permit re-identification. It is possible, for example, that if the result of the research is system or algorithm that returns results based on third-party inputs, then repeated queries by a motivated third party may make it possible to deduce individual data that fed into the creation of the algorithm.[72]

Second, the Revised Common Rule has a special provision defining informed consent as requiring disclosure of research involving biospecimens that "will (if known) or might include whole genome sequencing (i.e., sequencing of a human germline or somatic specimen with the intent to generate the genome or exome sequence of that specimen)."[73] Researchers also must disclose this information when seeking broad consent.[74] The kicker here, however, is the phrase "if known." Later, currently unanticipated, secondary research could involve full sequencing; presumably if suitably capacious broad consent had not been secured in advance, then either additional informed consent would be required or an IRB would have to provide substituted consent, which it can sometimes do. Biological samples, after all, can be stored for decades.

The Revised Common Rule's third response to the danger of technological change is to require covered federal agencies to reexamine the definition of what makes personal information "identifiable" every four years, starting, a year after

---

triangulate the identity of the target.").

69. § __.104(4)(II), 82 Fed. Reg. at 7262/2.

70. Specifically, the Preamble argues that, "Under HIPAA, these protections include, where appropriate, requirements to obtain the individual's authorization for future, secondary research uses of protected health information, or waiver of that authorization by an IRB or HIPAA Privacy Board. . . . We believe that the HIPAA protections are adequate for this type of research." 82 Fed. Reg. at 7194/3.

71. See 45 C.F.R. §§ 164.103, 164.514 (2019).

72. See Nicolas Papernot et al., *Scalable Private Learning with PATE*, ARXIV:1802.08908 [CS, STAT] (2018), http://arxiv.org/abs/1802.08908 for a description of this type of attack.

73. § __.116(5)(c)(9), 82 Fed. Reg. at 7266/2.

74. § __.116(d)(1), 82 Fed. Reg at 7266/3 (referring to § __.116(5)(c)(9)).

the revision takes effect.[75] This matters because the Revised Common Rule constrains research involving 'identifiable private information' as opposed to unidentified (or, de-identified) information. The creation of a mechanism to attempt to keep pace with changes in technology could become a serious attempt to have rules change in light of scientific discoveries; even then, however, in the best case, these changes will always lag those discoveries. On the other hand, one could be forgiven for seeing the creation of this mechanism as a way of punting a controversial problem down the road: despite the issue being raised in earlier drafts,[76] the Revised Common Rule avoids finding that whole genome sequencing produces sufficiently "identifiable" personal information so as to automatically trigger the more onerous consent and data-handling requirements that apply to personally identifiable information.[77] Thus, for example, if researchers collect a biological sample and get consent to fully sequence it, the Revised Common Rule nonetheless permits them to 'de-identify' it and treat it as de-identified information, even though the DNA might well re-identify the donor. At present, "an investigator who possesses information or biospecimens to which such a technology or technique might be applied is not to be considered in possession of identifiable private information or identifiable biospecimens merely as a result of such a circumstance: that would only be true were the investigator to actually apply the technology or technique to generate identifiable private information or identifiable biospecimens."[78]

In the future, additions to the list of what constitutes 'identifiable personal information,' and thus trigger the full human subjects rules, will require not only that a large inter-agency group agree, but that their work product run the full gauntlet of informal rulemaking, namely notice-and-comment plus any subsequent court challenges. As the time it took to write and implement the Revised Common Rule itself illustrates, this is not a swift process, at least not when the rule is controversial[79]—and presuming that genomic data is always at risk of re-identification is likely to retain the controversial nature that kept it out this much-delayed revision of the Common Rule. The Preamble to the Revised Common Rule does state that "[t]he expectation is that whole genome sequencing will be one of

75. § __.102(e)(7), 82 Fed. Reg. at 7260/2.

76. *See* 82 Fed. Reg. at 7163/3–7164/1 (discussing alternative proposals in NPRM), 7166/2–3 (discussing public comments to alternative proposals).

77. That is, requiring that the person has provided their consent which meets the requirements of the Common Rule, or where an IRB has waived the requirement for consent. *Id.* at 7169/2–3.

78. 82 Fed. Reg. 6169/3.

79. *See* Richard J. Pierce, Jr., *Rulemaking Ossification Is Real: A Response to* Testing the Ossification Thesis, 80 GEO. WASH. L. REV. 1493 (2012), (arguing that "'ossification' thesis—that rulemaking takes a very, very long time—is true for the subset of rulemakings that "raise controversial issues where the stakes are high").

the first technologies to be evaluated to determine whether it should be placed on this list,"[80] but given that DNA-based re-identification is already probable for any given sample,[81] it is hard to justify this delay.

The biggest problem, however, remains the broad consent provision. Once obtained, it permits all future research uses of identifiable biospecimens that the researchers were able to include in their consent request. Because we still do not, and conceivably may never, have a good idea of the extent to which Big Data techniques will make re-identification possible, neither party to the agreement can know how and when data granted under broad consent will end up being used, what it will be combined with, and what the results may show about the individual or others. Defenders of broad consent will argue that academic research ethics, and the baleful eye of IRBs will ensure that most if not all academic researchers proceed ethically. Even if this is correct,[82] however, it sets a poor example for the consent to be required by independent commercial research operations both inside and outside the biomedical establishment.

C. Effects on Commercial Human Subjects Research

Writing about an analogous problem—the move of personal health data between the HIPAA-protected zone and the essentially unregulated zone occupied by health data brokers, Nicolas P. Terry noted that data moves remarkably easily between them:

> Some of those data are created within the highly regulated space of health-care practice but legally "exported" (for example, they may have been deidentified). Other big data are created outside the highly regulated health-care domain but are medically inflected, and, once combined with other data points, operate as data proxies for protected HIPAA data. In both scenarios, data triangulation may defeat any de-identification. In the second example, users increasingly generate wellness, fitness, and sickness data on mobile health platforms or by mobile health apps. . . . Some data are created in a highly regulated space but then exported to a mobile device; other data are processed in the opposite direction."[83]

80. 82 Fed. Reg. at 7169/3.
81. *See supra* note 59.
82. *But see* sources cited *supra* note 34.
83. Nicolas P. Terry, *Regulatory Disruption and Arbitrage in Health-Care Data Protection*, 17 YALE J. HEALTH POL'Y, L & ETHICS 143, 147 (2017).

The Revised Common Rule clearly anticipates that researchers will combine public or commercial data with covered data; that is a feature not a bug in the eyes of its drafters. The extent to which data escapes in the other direction remains to be seen, although the Revised Common Rule as drafted appears to contemplate only sharing of covered data among institutions subject to its strictures.[84]

More generally, as a legal matter the Revised Common Rule, like its predecessor, does not apply to private biomedical research undertaken outside the umbrella of a research institution that takes federal funds.[85] Thus, for example, it does not apply to private genetic testing services such as 23andMe. The data acquired by these private services is thus available to be combined with public data, or with data under the Common Rule. And while private services do require consent to testing, indeed they charge for it, their polices as to secondary use vary considerably,[86] and often give the companies very great flexibility to do whatever they want with the consumer's genomic information.[87]

Although the Revised Common Rule does not apply directly to private medical researchers, it will, as it has in the past, stand as the embodiment of what amount to best practices for ethical uses of human subjects data. By lowering the bar in those best practices to permit broad consent,[88] we create a more attainable target for the commercial sector, such as private genetic testing services, but also make it ever more unlikely that the commercial sector will ever do any better.

D. Effects on Privacy More Generally

The problems of use, re-use, and re-identification of human subjects data are not unique to the biomedical research context. Indeed, this is anything but surprising given that a substantial part of the biomedical research issues exposed by the revisions to the Common Rule relate to the merger of biomedical information with non-medical patient databases. It is in fact a much broader problem, one inherent to Big Data more generally.

At present both US law and research regulations tend to treat existing, publicly

84. See § __.114, 82 Fed. Reg. 7265/1, on "cooperative research."

85. See 45 CFR §§ 46.101(a), 46.122. See also *supra* note 31 (citing HIPAA rules that apply to human subjects research projects not covered by the Revised Common Rule).

86. See Valerie Gutmann Koch & Kelly Todd, *Research Revolution Or Status Quo?: The New Common Rule and Research Arising from Direct-To-Consumer Genetic Testing*, 56 HOUS. L. REV. 81, 95–100 (2018).

87. *See, e.g.* Sarah Zhang, *Of Course 23andMe's Plan Has Been to Sell Your Genetic Data All Along*, GIZMODO (Jan 6, 2015), https://gizmodo.com/of-course-23andmes-business-plan-has-been-to-sell-your-1677810999.

88. For a fuller argument that broad consent is not informed consent, see Vilhjalmur Arnason, *Coding And Consent: Moral Challenges Of The Database Project in Iceland*, 18 BIOETHICS 27 (2004).

available, datasets as not creating substantial privacy risks; if the data are public, they are not, a fortiori, private, so where is the privacy issue? But this is, as many have noted, a misconception. "Publicly available data can be put to a wide range of secondary uses, including being combined with other data sets, that can pose serious risks to individuals and communities."[89] European-style data protection deals with these risks by regulating the retention and use of data directly; the lacunae in US regulation stem from our focus on having regulation attached primarily at the moment of collection. The US generally adopts the frame that 'public' data—be it data collected in public, collected data in the public domain, or private data collected outside the strictures that apply to research that falls under the Common Rule—is outside the scope of regulation. There are exceptions, such as the HIPAA constraints on sharing medical data, but most privately collected data about people—marketing info, phone location information, interaction with apps or web pages—is far less regulated, or not regulated at all.

Consistent with its DNA as a rule about the collection and reuse of medical information in government-supported studies and studies by government-supported academic institutions, the Revised Common Rule exempts from coverage all secondary research uses of identifiable private information or identifiable biospecimens if they are publicly available.[90] This limitation has led some to suggest that the reach of the Common Rule should be extended, either as a legal or an ethical norm, to encompass all data science that involves dealing with potentially identifiable data, including supposedly de-identified data, about people.[91] Others have argued that all Big-Data-based research, medical or not, should be subject to a single, universal, set of standards designed to protect personal privacy,[92] or that currently unregulated collections of large data sets should require something akin to an environmental impact statement.[93]

There is no question that very large amounts of data, such as Facebook interactions or ride-sharing data are essentially unregulated in the United States. Even in medicine "not all companies follow—voluntarily or otherwise—the Common Rule's requirement for safe and ethical human subjects research."[94] The story is undoubtedly worse as one moves to non-medical data where ethical data collection and use norms were never as strongly established to begin with. As we

89. Jacob Metcalf & Kate Crawford, *Where are Human Subjects in Big Data research? The Emerging Ethics Divide*, BIG DATA & SOC'Y, June 2016, at 1, 1.

90. § __.104(d)(4)(i), 82 Fed. Reg. at 7262/2 (codified at 45 C.F.R. § 46.104(d)(4)(i)).

91. *See, e.g.*, Metcalf & Crawford, *supra* note 89.

92. *See, e.g.*, Effy Vayena, Urs Gasser, Alexandra Wood, David R. O'Brien & Micah Altman, *Elements of a New Ethical Framework For Big Data Research*, 72 WASH & LEE L. REV. ONLINE 420 (2016).

93. *See* Froomkin, *supra* note 23.

94. Koch & Todd, *supra* note 86, at 113.

roll out Smart Cities, and collect data about the lives of citizens on a 24/7 basis, the amount of available data will only grow. For those of us in the U.S. who believe that the subjects of this data should have more say in how it is used, the core concept has been to try to require more and better consent before that data is collected and before it is shared. Indeed, several scholars, including Jules Polentesky, Omer Tene and Joseph Jermoe, and Ryan Calo have proposed mechanisms by which users of Big Data not subject to the Common Rule could create institutions that would weigh the ethical implications of their proposed research.[95]

Consent is, as noted at the outset, an imperfect tool, but often it is the best tool available. I and others have argued that failing to require even basic consent before allowing unfettered capture of personally identifiable information in 'public' amounts to a form of privacy pollution.[96] Requiring more and better consent—informed consent—seemed one solution to these problems, even if informed consent was far from perfect.[97] (For situations such as cameras in public places where consent is impractical, I have advocated that we require public privacy impact notices from the operators of the surveillance as a way of both tempering and surfacing the data collection.[98])

The Common Rule has been the closest thing to a gold standard of consent, a set of requirements that usually put private industry to shame. By weakening the requirement of genuine informed consent, the Revised Common Rule epitomizes a surrender to Big Data's challenge to informed consent. The surrender is made manifest in the creation of broad consent, which will be as broad as the drafters of a consent form can make it, potentially eternal, and which is certainly not going to be informed as we previously understood the term. Not only does this create an ethical challenge for medical researchers, but it in effect lowers the consent ceiling for all private research whether based on home genetic testing,[99] remote sensing in public places, or Facebook likes.

---

95. See Jules Polentesky, Omer Tene and Joseph Jermoe, *Beyond The Common Rule: Ethical Structures For Data Research In Non-Academic Settings*, 13 COLO. TECH L.J. 333 (2015); Ryan Calo, *Consumer Subject Review Boards: a Thought Experiment*, 66 STAN. L. REV. ONLINE 97 (2013), http://www.stanfordlawreview.org/online/privacy-and-bigdata/consumer-subject-review-boards.

96. Froomkin, *supra* note 23.

97. *See supra* text at note 27. Or perhaps we could generalize Jorge Contreras's suggestion that if the notice and consent model is broken for genetic privacy, we should move from what amounts to a property regime to a liability regime to protect persons from 'abusive research practices.' *See* Jorge L. Contreras, *Genetic Property*, 105 GEO. L.J. 1 (2016).

98. *See* Froomkin, *supra* note 23.

99. See Koch & Todd, *supra* note 86.

III. GOING FORWARD

The problem created by Big Data analytics is a consequence of the unpredictable pattern-recognition that we call Machine Learning. Perhaps someday we can work out better ways to predict what will be found, although currently this seems quite unlikely or even impossible because to predict the outcome would be to do the experiment: an irony of Machine Learning is that just as Big Data seeks correlation from a data set that more or less *is* the population, so too the act of predicting the nature of the outcome of modern analytics may require more or less creating the ML system. Or, perhaps we could develop some way to escrow research results until individual consent of all the members of the data set is secured. That sounds nice, but what it would require is that when some participants withdraw consent, the Machine Learning system would have to be trained a second time on a data set that had been cleaned of the objectors' data, creating not just duplicate work but an increased danger of content bias[100]—a protocol that is unlikely to find favor with researchers. Even less plausibly, we would also have to have some way to unsee, or better yet never see, the research result that upset the data contributors. Indeed, neither of these seems likely either.

More likely, we may develop techniques that at least protect individual privacy from exposure via Big Data research, even if it does not necessarily address any moral objections that persons might have to their information being used for a particular type of research. One way to protect individual privacy, again not addressing the issue of moral objections to participation, would be to create reliable synthetic data.[101] Creating synthetic data takes time and effort, and also introduces new risks if it is not done properly. For example synthetic data might fail to reproduce aspects of the original data that could have led to new discoveries. Or, the synthetic data might unintentionally introduce new artifacts that prove overly-attractive to Machine Learning systems. A different, perhaps more promising strategy to achieve this limited but still important goal of protecting data subjects from identification involves blending differential privacy with Machine Learning, apparently to the benefit of both.[102]

In the absence of a technical solution, we must confront the problem that Big-

100. *See supra* note 37.

101. See Sharon Bassan, *The Ethics in Synthetics: Statistics in the Service of Ethics and Law in Health-Related Research in Big Data from Multiple Sources*, 31 J.L. & HEALTH 87 (2018).

102. *See* Papernot et al., *supra* note 72; Nicolas Papernot et al., *Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data*, ARXIV:1610.05755 [CS, STAT] (2016), http://arxiv.org/abs/1610.05755; *but see* Bargav Jayaraman & David Evans, *When Relaxations Go Bad: "Differentially-Private" Machine Learning*, ArXiv:1902.08874, https://arxiv.org/abs/1902.08874 (warning that differential privacy settings for ML commonly allow too much information leakage, and that correcting problem tends to require an unacceptable limitation on data usage).

Data-based research undermines informed consent as we know it. In that case, we must either make the compromises indicated by the Revised Common Rule, or else individuals seeking to preserve control of the uses that others make of their data and biospecimens will have to choose not to share their potentially personally identifiable information in the first place.[103]

Blocking the re-use of potentially valuable data, the proponents of broad consent and its ilk argue, is a too high a price to pay for privacy.[104] "Data," the argument goes, "saves lives."[105] On what version of ontological ethics, they ask, can the personal interest in data privacy outweigh the collective (and also ultimately personal) interest in public health and in cures? "If society seeks to derive the benefits of medical research in the form of improved health and health care, information should be shared for the greater good, and governing regulations should support the use of such information, with appropriate oversight."[106] The reply is that this is false framing[107] because the interest in data privacy can be understood as a group right as well as a personal one.[108] As Omri Ben-Shahar puts it, "the harms from data misuse are often far greater than the private injuries to the individuals whose information gets released."[109] Framing the costs of sharing as having negative externalities that may greatly exceed the personal costs suggests that at times the cost of sharing may also exceed the benefits.

Perhaps that is right; or perhaps the whole balancing problem is an example of what Julie Cohen argues is a "simplistic vision of the relationship between privacy and innovation."[110] Cohen suggests we should not let an infatuation with hoped-for gains from Big Data analytics blind us from the more likely ways in which research will play out: the gains from a Big Data commons will in the main be privatized, will be subject to the same researcher- and content-biases as other research, and likely will "exacerbate" the problem of using thin models of people that do not reflect their complicated realities (Cohen calls these "constructed subjectivities").[111]

---

103. *Cf.* Froomkin, *supra* note 32.

104. This point was made most forcefully by a number of commentators at the Yale Workshop on The Law and Policy of AI, Robotics & Telemedicine.

105. Mario Romao, *Data Saves Lives*, PRIVACY@INTEL (Dec. 6, 2018), https://blogs.intel.com/policy/2018/12/06/data-saves-lives.

106. INST. MED., BEYOND THE HIPAA PRIVACY RULE: ENHANCING PRIVACY, IMPROVING HEALTH THROUGH RESEARCH 35 (2009).

107. *See* Luciano Floridi, *Open Data, Data Protection, and Group Privacy*, 27 PHIL. & TECH. 1, 1–2 (2014).

108. *See* EDWARD. J. BLOUSTEIN, INDIVIDUAL AND GROUP PRIVACY (2017); Omri Ben-Shahar, *Data Pollution*, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3191231 (Aug, 23, 2018).

109. Ben-Shahar, *supra* note 108, at 3.

110. Julie Cohen, *What Privacy is For*, 126 HARV. L. REV. 1904, 1919 (2019).

111. *Id*. at 1924–25.

Perhaps therefore the cost of making Big Data harder to use will not be as great as its advocates suggest. In the worst case, however, where the social costs of not sharing are indeed as high as Big Data's optimists allege, the alternatives look stark: either people reduce their willingness to share for science (and, less bleakly, in the non-health-related data realm for marketing), or we must learn to view consent in the Big Data era as what it truly has become: the gift that keeps on giving,[112] for which the donor-subject receives neither a profit share nor even a tax deduction.

In the meantime, all these cost/benefit calculations remain very speculative. The one thing we do know is that once information is shared, it is almost impossible to call back. Thus despite the unhelpful example provided by the Revised Common Rule, we should continue to demand more-frequent informed consent, and resist the move to less-informed consent, both in the realm of Big Data and otherwise. We should, for example, demand that investigators go back for more consent if their plans to use information or specimens change. In this context it is notable the final version of the Revised Common Rule rejected a proposal to require investigators to ask if subjects wished to be contacted before using their data in a future study. The drafters argued that "Those who opposed the provision noted that while the intent of the provision was laudable, the ensuing tracking system that would need to be developed by institutions to track who had said ''yes'' or ''no'' to being re-contacted, and in what circumstances, would be difficult to develop and maintain, and would also represent significant costs to institutions without a corresponding tangible increase in the protections afforded to human subjects."[113] While the record-keeping is certainly not trivial, would it really be so difficult and expensive given the availability of open source tools that would do the job and the increasing tendency of people to keep cell phone numbers and email addresses for life? A demonstration project would seem to be in order.

Informed consent is very far from perfect, but at present it remains one of the few tools available to encourage the people to whom we entrust our data to remain proper stewards of it: If they cannot tell us what will become of our information, why should we entrust them with it?

112. *See* Philip J. Nickel, *The Ethics of Uncertainty for Data Subjects* in THE ETHICS OF MEDICAL DATA DONATION 55 (J. Krutzinna & Loridi, eds, 2019), https://link.springer.com/content/pdf/10.1007%2F978-3-030-04363-6.pdf; Kadija Ferryman, *Reframing Data as a Gift* (April 21, 2017), https://ssrn.com/abstract=3000631.

113. 82 Fed. Reg. at 7216/2–3.