A Process-based Approach to Informational Privacy and the Case of Big Medical Data

*Michael Birnhack**

Data protection law has a linear logic, in that it purports to trace the lifecycle of personal data from creation to collection, processing, transfer, and ultimately its demise, and to regulate each step so as to promote the data subject's control thereof. Big data defies this linear logic, in that it decontextualizes data from its original environment and conducts an algorithmic nonlinear mix, match, and mine analysis. Applying data protection law to the processing of big data does not work well, to say the least.

This Article examines the case of big medical data. A survey of emerging research practices indicates that studies either ignore data protection law altogether or assume an ex post position, namely that because they are conducted after the data has already been created in the course of providing medical care, and they use de-identified data, they go under the radar of data protection law. These studies focus on the end-point of the lifecycle of big data: if sufficiently anonymous at publication, the previous steps are overlooked, on the claim that they enjoy immunity. I argue that this answer is too crude. To portray data protection law in its best light, we should view it as a process-based attempt to equip data subjects with some power to control personal data about them, in all phases of data processing.

[*] Associate Dean (Research) and Professor of Law, Tel Aviv University (TAU). Thanks to Nahum Kiryati and participants of the Cegla Conference on The Problem of Theorizing Privacy (TAU, January 2018) and participants of the Privacy Law Scholars Conference (Washington, May 2018) for helpful comments, to Evyatar Ayalon and Omer Levanon for able research assistance, and to the Edmond J. Safra Center for Ethics at TAU and the Australian Friends of TAU for generous support.
Cite as: Michael Birnhack, *A Process-Based Approach to Informational Privacy and the Case of Big Medical Data*, 20 Theoretical Inquiries L. 257 (2019)

Such control reflects the underlying justification of data protection law as an implementation of human dignity. The process-based approach fits current legal practices and is justified by reflecting dignitarian conceptions of informational privacy.

Introduction

Data protection law has a linear logic, in that it purports to trace the lifecycle of personal data systematically from creation, to collection, processing and other uses, transfer, and ultimately its demise, and to regulate each step to facilitate the data subject's control thereof. Big data defies this logic, in that it decontextualizes data from its original environment and conducts an algorithmic nonlinear mix, match, and mine analysis. Thus, applying data protection law to the processing of big data does not work well, to say the least.[1]

This Article examines the case of *big medical data*, focusing on noncommercial medical research.[2] A survey of emerging practices indicates an array of norms among researchers. Researchers either ignore data protection law or assume an *ex post* position, namely that the law permits using anonymous, pseudonyms, or de-identified data that has already been collected in the course of providing medical care. This ethical stance focuses on the end-point of the big data lifecycle, at publication, and claims that such research enjoys a retrospective redemption and immunity. I argue that this position is too crude, as it overlooks the earlier steps of processing, and may compromise human dignity.

We can address this misfit between privacy law and big data by trading off the former's lofty goals for the latter's immense benefits. In most jurisdictions, such a tradeoff would require a legislative action and may be subject to judicial review for its violation of privacy.[3] Privacy might lose, but only after a careful assessment and acknowledgment of the loss. This is the route taken by European data protection law, now embedded in the General Data Protection Regulation (GDPR).[4] The GDPR treats the use of personal data for

[1] Helen Nissenbaum makes a similar argument, framed in her theory of Contextual Integrity. *See* Helen Nissenbaum, *Contextual Integrity Up and Down the Data Food Chain*, 20 Theoretical Inquiries L. 221 (2019).

[2] Post-research commercialization further complicates the picture, as do governmental uses of medical data, which I do not discuss here. *See* Wendy K. Mariner, *Reconsidering Constitutional Protection for Health Information Privacy*, 18 U. Pa. J. Const. L. 975 (2016).

[3] *E.g.*, Convention for the Protection of Human Rights and Fundamental Freedoms (ECHR) art. 8, ¶ 2, Nov. 4, 1950, E.T.S. No. 5.

[4] Council Regulation 2016/679 2016 O.J. (L 119) 1 (EU) [hereinafter GDPR].

scientific research as compatible with the original purpose for which the data was collected, but requires "appropriate safeguards," which are "technical and organizational measures" meant to ensure the principle of data minimization.[5] In other words, the GDPR reflects an explicit preference for the perceived benefits of research over privacy, but acknowledges the latter and hence tries to minimize the harm.

Here, I am searching for philosophical and legal consistency, using the GDPR as well as American law as examples. I argue that in order to portray data protection law in its best light, we should frame it as a process-based attempt to equip data subjects with some control of personal data about them, in *all* steps of data processing. Such control reflects the underlying justification of data protection law as an implementation of human dignity. Subjects' control over their data should also cover the anonymization of the data. We should avoid the narrow and quite negative meaning that *control* has acquired in American privacy studies, which reduces it to the principles of notice and consent, pointing to their numerous shortcomings and practically giving up on control altogether. The process-based approach to data protection law follows Ronald Dworkin's dual requirements—*fit* to current legal practices and *justification*[6]—by reflecting dignitarian conceptions of informational privacy. While I focus on medical data, the discussion is meant to be scalable and apply to other big data cases: we should prefer a step-by-step analysis to the retrospective view.

Part I presents the process-based approach to informational privacy law. *Part II* reviews emerging big medical data studies. *Part III* applies the process-based approach to big medical data.

I. Informational Privacy as a Process

Trying to make sense of informational privacy law, or data protection in European parlance, is complicated. We can turn to underlying theoretical justifications for guidance,[7] or to descriptive taxonomies.[8] We can look to the

[5] *Id.* at art.1(2), 89 & recitals 50, 156.

[6] Ronald Dworkin, Law's Empire 243 (1986). Note that Dworkin suggested these terms in the context of judicial decision-making, where each new opinion should fit the previous ones and be justified in itself. I adapt these criteria to our legal approach more generally. Moreover, fit in itself is not an absolute requirement, as a new government may diverge from previous policies.

[7] For a comprehensive discussion, see Daniel J. Solove, Understanding Privacy (2008).

[8] *Id.* at 101.

social context and decipher its informational norms, as Helen Nissenbaum has suggested.[9] This Article offers some modifications to some of these attempts. I argue that data protection law is best interpreted by following the data throughout its lifecycle, step by step. Following Nissenbaum, I emphasize informational flows; unlike her, I do not insist on context as the organizing structure of the discussion. Following many, I cling to theories that emphasize privacy as a concretization of human dignity, translated into *privacy as control*.[10] Call it a process-based approach to informational privacy law. It meets the fit and justification criteria.

A. Informational Processes

Engineers, system designers, and developers often think about information as a process.[11] They refer to a project's "temporary nature," namely that "each project has a definite beginning and a definite end."[12] Accordingly, when we approach a new socio-technological system, we should follow the data: who creates it, who collects it, and how is it processed? Which measures guarantee security and confidentiality? Is the data transferred? To whom? For which purposes? Does the data expire? Who makes the decisions about each of these steps? This is the process-based approach to data. It insists that before we look at the forest, we must see its trees.

This process-based approach is quite intuitive and guides many interpreters of the law and policymakers, yet not all.[13] To anticipate the next Part, those who design big medical data research projects often opt for the forest view, juxtaposing privacy with the research benefits and prioritizing the latter over the former.

By contrast, a process-based approach instructs us to suspend the urge to balance, and instead be patient and delve into the details. It asks us to break down a particular scenario into its components. Instead of assuming for the

[9] Helen Nissenbaum, Privacy in Context: Technology, Policy, and the Integrity of Social Life (2010).

[10] Alan F. Westin, Privacy and Freedom 7 (1967). Control echoes a property right, but privacy has its independent standing. *See* Paul M. Schwartz, *Property, Privacy and Personal Data*, 117 Harv. L. Rev. 2055 (2004).

[11] *E.g.*, Clive L. Dym et al., *Engineering Design Thinking, Teaching, and Learning*, 94 J. Engineering Educ. 103 (2005).

[12] Paul Sanghera, Fundamentals of Effective Program Management: A Process Approach Based on the Global Standard 5 (2008).

[13] For the discursive gaps between lawyers and engineers regarding privacy, see Michael Birnhack, Eran Toch & Irit Hadar, *Privacy Mindset, Technological Mindset*, 55 Jurimetrics 55 (2014).

sake of the balancing exercise that privacy is harmed, a zoom-in inspection should provide us with a better understanding of what exactly the privacy harms at stake are. A closer look has the potential to direct us to various solutions that might mitigate the tension. It is a fine-tuned approach. Once taken, we can return to balancing and assure ourselves that we chose the least intrusive means. In fact, this is the essence of the principle of proportionality, familiar in some constitutional settings.[14] Once we have a better idea of the process, we can zoom out and examine the system as a whole. Again, this is also a principle of design thinking.[15]

Some of the existing approaches to privacy assume a process-based approach. Solove's famous taxonomy offers four clusters of privacy phases, each containing specific practices: information collection, processing, dissemination, and invasion. The first three reflect the lifecycle of information. His purpose was to provide a "more pluralistic understanding of privacy," and his methodology was what he called "cultural analysis."[16] This led him to follow the data: "The general progression from information collection to processing to dissemination is the data moving further away from the individual's control."[17] I take a normative stance, arguing that we should ensure that the subject has control over her data also once the data moves further away.

Nissenbaum emphasized the importance of *context* as an organizing frame to better figure out complex privacy situations. Each context, she explained, contains norms about information, and within these she focused on transmission norms, namely norms that constrain the flow of information.[18] In examining a context, we should search for "the type of information, the parties who are the subjects of the information as well as those who are sending and receiving it, and the principles under which this information is transmitted."[19] Once a new socio-technological system changes the transmission norms, a red (privacy) flag is raised, indicating that the integrity of the context is challenged.[20] Elsewhere,

[14] *E.g.*, Basic Law: Human Dignity and Liberty, 5752-1992, § 8, SH No. 1454 p. 90 (Isr.); ECHR, *supra* note 3.

[15] "A hallmark of good system designers is that they can anticipate the unintended consequences emerging from interactions among the multiple parts of a system. This kind of foresight is essential for designing engineering systems and managing the design process." Dym et al., *supra* note 11, at 106.

[16] Solove, *supra* note 7, at 101-02.

[17] *Id.* at 103.

[18] Nissenbaum, *supra* note 9, at 145.

[19] *Id.* at 141.

[20] *Id.* at 150.

I have criticized some elements of her theory.[21] Here, I reverse Nissenbaum's order: Whereas she begins with identifying the context and then searches for the transmission norms, I begin with the data and the informational process, and only then turn to the larger picture. She begins with the forest and then reaches the trees; I search for the trees first.

B. Justification: Privacy as Control

Privacy theories abound. Most attempts to justify privacy emphasize the individual: her intellectual needs,[22] psychological needs,[23] the prevention of unwanted access,[24] or rejection of a person's reduction to an arbitrary datum.[25] Other privacy theories emphasize the importance of managing intimate[26] or professional relationships.[27] Moving yet further away from the individual, other theories explain privacy as a crucial social value for the well-functioning of the community[28] or democracy at large, justifying privacy as a public good.[29]

Theories are then translated into specific legal tools, which often converge into mechanisms that equip the data subject with some control over her data, *i.e.*, Fair Information Practices (FIPs). These principles keep evolving.[30] FIPs reflect the notion of *privacy as control*, first articulated by Alan Westin more than fifty years ago.[31] Carried from traditional privacy cases (eavesdropping,

[21] Michael Birnhack, *A Quest for a Theory of Privacy: Context and Control: Review of Helen Nissenbaum's Privacy in Context*, 51 Jurimetrics 447 (2011).

[22] Neil Richards, Intellectual Privacy: Rethinking Civil Liberties in the Digital Age (2015).

[23] Samuel Warren & Louis Brandeis, *The Right to Privacy*, 4 Harv. L. Rev. 193 (1890) (articulating privacy as the right to be let alone); Irwin Altman, *Privacy – A Conceptual Analysis,* 8 Env't & Behav. 7 (1976) (privacy as an interpersonal boundary control process).

[24] Ruth Gavison, *Privacy and the Limits of Law*, 89 Yale L.J. 421 (1980).

[25] Jeffrey Rosen, The Unwanted Gaze: The Destruction of Privacy in America (2001).

[26] Charles Fried, *Privacy*, 77 Yale L.J. 475 (1968).

[27] James Rachels, *Why Privacy is Important*, *in* Philosophical Dimensions of Privacy: An Anthology 290 (Ferdinand D. Schoeman ed., 1984).

[28] *See* Ari Ezra Waldman, Privacy as Trust: Information Privacy for an Information Age (2018).

[29] Priscilla M. Regan, *Privacy as a Common Good in the Digital World*, 5 Info. Comm. & Soc'y 382 (2002).

[30] Robert Gellman, Fair Information Practices: A Basic History (April 10, 2017) (unpublished manuscript), https://bobgellman.com/rg-docs/rg-FIPshistory.pdf.

[31] Westin, *supra* note 10.

revealing someone's secrets, etc.) to informational privacy situations, privacy means that the data subject should have control over her personal data, or as Lisa Austin explains, privacy as control means "individual control over the decision to choose a state of privacy."[32]

Importantly, privacy as control is not an independent theory of privacy. It is a description that captures much of the essence of most of the abovementioned theories and their nuances. Privacy as control does reflect a fundamental principle, though. This is the dignitarian beacon of privacy, echoing Warren and Brandeis' description of privacy as protecting the 'inviolate personality.'[33] To interfere in someone else's life, using personal data without the subject's consent, is to disregard the individual. Unauthorized collection and use of personal data negates the subject's ability to make decisions for herself. Such acts disregard the person as an autonomous agent and disrupt their ability to author their own lives. Accordingly, privacy as control is best understood as a concretization of the overarching idea of dignity, applied to personal issues. Deriving privacy from dignity is now a cornerstone of many legal systems, such as Germany,[34] Israel,[35] and more recently, India.[36]

Breaking down an informational scenario into its components enables us to respect the idea of human dignity, concretized by the notion of privacy as control. The process-based approach requires that for each step of the informational process we should ask what kind of control the subject has over her personal data. This detailed inspection will quite likely yield new suggestions to better enhance control. The process-based approach adheres to the guiding idea of human dignity and enables us to fulfil our right to privacy, or perhaps, as Anita Allen argues, our moral obligation to do so.[37]

Privacy as control has acquired a bad reputation among (mostly American) privacy scholars. The reason is FIPs' focus on notice and consent. With the ever-increasing complexity and overload of information, internet users

[32] Lisa M. Austin, *Re-Reading Westin*, 20 Theoretical Inquiries L. 53 (2019).

[33] Warren & Brandeis, *supra* note 23, at 205. *See also* Edward J. Bloustein, *Privacy as an Aspect of Human Dignity: An Answer to Dean Prosser*, 39 N.Y.U. L. Rev. 962 (1964).

[34] BVerfG, 1 BvR 209/83, Dec. 15, 1983, https://openjur.de/u/268440.html (defining privacy as informational self-determination).

[35] HCJ 8070/98 ACRI v. Ministry of Interior 58(4) PD 842 (2004) (Isr.).

[36] Justice K S Puttaswamy (Retd.) v. Union of India, 2017 10 SCALE 1, https://globalfreedomofexpression.columbia.edu/cases/puttaswamy-v-india/. See discussion in Anita L. Allen, *Synthesis and Satisfaction: How Philosophy Scholarship Matters*, 20 Theoretical Inquiries L. 343 (2019).

[37] Anita L. Allen, *Protecting One's Own Privacy in a Big Data Economy*, 130 Harv. L. Rev. F. 71 (2016).

mistake the heading *Privacy Policy* for a guarantee of privacy,[38] and no one reads privacy notices let alone understands them.[39] Indeed, consent is often meaningless, especially in our dealings with mega-corporations.

Instead of abandoning notice and consent, we should search for ways to strengthen control and (re)empower data subjects.[40] Moreover, the reduction of privacy as control to notice and choice ignores other FIPs. Control should not be limited to the first encounter between data subject and data collector. The subject's control should extend to subsequent phases of the informational process. The law should create additional meeting points between subjects and controllers, between a subject and her data.[41] Thus, control of one's personal data is a continuous effort to make sure that the data subject does not become an object.[42] Adopting a process-based approach to privacy assists in achieving this goal. The Dworkinian criterion of justification is met.

C. Fit: FIPs

A process-based approach fits FIPs. Data protection law is triggered only when threshold conditions are met. In the United States, federal law regulates informational privacy in a so-called sectoral manner: several specific federal laws address data in specific contexts. In the absence of a specific sectoral law, personal data is up for grabs (but perhaps subject to state law). After satisfying the content-based threshold, each law has its own conditions, but most share an identifiability requirement, namely, they apply only to Personal Identifying Information.[43] Identifiability is also the trigger of the GDPR, albeit the only one. The GDPR regulates personal data in a universal manner, regardless of its content.[44]

[38] Joseph Turow, Americans Online Privacy: The System is Broken (2003).

[39] *See* Jonathan A. Obar & Anne Oeldorf-Hirsch, The Biggest Lie on the Internet: Ignoring the Privacy Policies and Terms of Service Policies of Social Networking Services, paper for TPRC 44: 44th Research Conference on Communication, Information and Internet Policy (Aug. 24, 2016), http://sched.co/7jyz.

[40] M. Ryan Calo, *Against Notice Skepticism in Privacy (and Elsewhere)*, 87 Notre Dame L. Rev. 1027 (2013).

[41] Michael Birnhack & Niv Ahituv, Privacy Implications of Emerging and Future Technologies (Dec. 7, 2013) (unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2364396.

[42] Julie E. Cohen, *Examined Lives: Informational Privacy and the Subject as an Object*, 52 Stan. L. Rev. 1373 (2000).

[43] Paul M. Schwartz & Daniel J. Solove, *The PII Problem: Privacy and a New Concept of Personally Identifiable Information*, 86 N.Y.U. L. Rev. 1814 (2011).

[44] GDPR, *supra* note 4, at art. 4(1).

FIPs, developed over time, offer a toolkit for handling personal data. The ultimate landmark was the 1995 European Data Protection Directive.[45] The Directive had, perhaps unexpectedly, an unusual global influence.[46] There are variations among FIPs' many local instantiations, but the core principles are similar. FIPs allow the collection and processing of personal data in order to promote commercial use of data and transborder transfers, and at the same time protect privacy.

Understood in their best light, FIPs trace the timeline of handling personal data. FIPs attempt to empower the data subject by creating several meeting points between subject and processor.[47] FIPs require that the data subject receive a notice prior to data collection. Collection is permitted only if it serves a legitimate purpose, and if based on the subject's consent, which should be given freely, accompanied by a right to withdraw consent. Some kinds of data are considered sensitive and subject to higher demands.[48] The principle of data minimization requires that only the minimum data required to achieve the legitimate purpose be collected. The next step is processing, which should not exceed the stated purpose. Thus, the subject's control extends beyond the first meeting point and continues to limit the processor. The data controller is required to ensure secrecy and data security. These duties guarantee that negligent or malicious parties do not frustrate the subject's initial consent. FIPs provide the subject with some checks: she has a right to access her personal data and, if needed, demand that it be rectified. There should be public oversight, typically in the form of a designated data protection agency (DPA).[49]

[45] Council Directive 95/46/EC, 1995 O.J. (L 281) (On the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data).

[46] *See* Gregory Shaffer, *Globalization and Social Protection: The Impact of EU and International Rules in the Ratcheting up of U.S. Privacy Standards*, 25 Yale J. Int'l L. 1 (2000); Michael D. Birnhack, *The EU Data Protection Directive: An Engine of a Global Regime*, 24 Computer L. & Security Rev. 508 (2008); Christopher Kuner, *An International Legal Framework for Data Protection: Issues and Prospects*, 25 Computer L. & Security Rev. 307 (2009).

[47] Birnhack & Ahituv, *supra* note 41.

[48] Anticipating the discussion in Part II, the GDPR treats "data concerning health" as a special category of personal data, and prohibits its processing, unless the data subject gave explicit consent for a specified purpose or if some other exceptions apply. *See* GDPR, *supra* note 4, at art. 9(1), 9(2). One of the exceptions is that the processing is necessary for scientific research. *See id.* at art. 9(2)(j). Member states may add limitations to such processing. *See id.* at art. 9(4).

[49] In the United States, the FTC has become the *de facto* data protection agency. *See* Chris Jay Hoofnagle, Federal Trade Commission Privacy Law and Policy

Various enforcement options, such as individual suits or class actions, supplement the subject's control with additional checks. Some versions of FIPs include accountability.[50]

FIPs have their shortcomings. We have noted those relating to notice and consent. The other meeting points are far from achieving their purpose. For example, hardly anyone utilizes the right to access data,[51] other than in specific contexts, such as a financial consumer report.[52]

Understanding FIPs as a concretization of *privacy as control* points to the ways to fix it where it is broken. One way is to create additional meeting points between subject and processor. These too are no panacea. Private enforcement is expensive, risky, and, given the often relatively minor and difficult to quantify nonpecuniary damage, it is irrational for one person to sue. Class actions are the obvious procedural solution, but not all legal systems allow it. DPAs are often under-funded.

We are now witnessing the emergence of an invigorated legal toolkit, which looks beyond the law to organizational and technological solutions. The GDPR is a clear manifestation of FIPs' second generation. It fine-tunes FIPs, adding the so called right to be forgotten[53] and tools such as Data Protection Impact Assessment (DPIA); it requires the appointment of a Data Protection Officer (DPO; or in America, Chief Privacy Officer – CPO); and it requires a process of Data Protection by Design (or Privacy by Design – PbD). These means aim to change the data controllers' privacy mindset: an impact assessment draws attention to otherwise unnoticed privacy issues. Engineers bring to the table their process-based understanding of technological systems and the flow of

(2016).

[50] Organization for Economic Cooperation and Development [OECD], *Guidelines on the Protection of Privacy and Transborder Flows of Personal Data*, at art. 14, OECD Doc. C(80)58/FINAL (Jul. 11, 2013); GDPR, *supra* note 4, at art. 5(2) (requiring data controllers to demonstrate compliance with principles of processing personal data); GDPR, *supra* note 4, at art. 13(2)(f) (requiring data controllers to provide subjects with information about profiling).

[51] Cases in which subjects exercised their right to access personal data are so rare that they became news, see, for example, Judith Duportail, *I Asked Tinder for My Data. It Sent Me 800 Pages of my Deepest, Darkest Secrets*, The Guardian (Sept. 26, 2017), https://www.theguardian.com/technology/2017/sep/26/tinder-personal-data-dating-app-messages-hacked-sold.

[52] *See, e.g.*, Fair Credit Reporting Act, 15 U.S.C. §1681b(a)(2).

[53] Case C-131/12, Google Spain SL v. Agencia Española de Protección de Datos (AEPD), 2014 E.C.R.; GDPR, *supra* note 4, at art. 17 (which now anchors this right).

information. Decision makers bring a broad understanding of privacy, which goes beyond data security.

We do not yet know how FIPs' second generation will play out in practice. The point here was to show that a process-based approach to informational privacy is already embedded in data protection law, or to borrow Julie Cohen's metaphor, the process-based approach is the grammar of data protection law.[54] The criterion of fit is met.

D. Big Data

Smart lawmakers attempt to legislate technologically neutral, future-proof legislation.[55] However, experience shows that this goal is frustrated time and again: Even if the legislation does not cite specific technologies, it inevitably reflects the law's (or lawmakers') hidden technological assumptions.[56] This is also the fate of data protection law when it meets big data.

Big Data means accumulating a large volume of data from different sources and formats, rendering it interoperable, and conducting an algorithmic analysis of the data both per-item and in an aggregate manner.[57] Big data analysis provides meta-data, *i.e.*, data about data, which makes it possible to recognize general trends and then predict an individual's behavior.[58] Previously anecdotal data can now be reevaluated on large scales. New correlations can be observed, leading to new research questions. Importantly, such research

[54] Julie E. Cohen, *Turning Privacy Inside Out*, 20 Theoretical Inquiries L. 1 (2019).

[55] Bert-Jaap Koops, *Should ICT Regulation be Technology-Neutral*, *in* Starting Points for ICT Regulation: Deconstructing Prevalent Policy One-Liners 77 (Bert-Jaap Koops et al. eds., 2006).

[56] Paul Ohm, *The Argument against Technology-Neutral Surveillance Laws*, 88 Tex. L. Rev. 1685 (2010); Michael Birnhack, *Reverse Engineering Informational Privacy Law*, 15 Yale J.L. & Tech. 24 (2012); Brad A. Greenberg, *Rethinking Technology Neutrality*, 100 Minn. L. Rev. 1495 (2016).

[57] The conventional definition of big data refers to Volume, Variety, and Velocity, first identified in Douglas Laney, *3D Data Management: Controlling Data Volume, Velocity, and Variety* (Gartner, Working Paper No. 949, 2001), https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf, although he did not use the term Big Data. Popular references now add Veracity and Value. *See, e.g.*, Arvind Sathi, Big Data Analytics 4 (2012).

[58] In the medical field, meta-data often refers to staff's comments on lab results. In the Big Data context, such comments are considered data.

has its own pitfalls and limitations.[59] In the discussion that follows in Part II.B, I use the term big data research to refer to the use of algorithms that are able to observe correlations that a human eye cannot.

The promise of big data lies in its ability to re-contextualize data. However, this is also where its privacy risks lie. For example, consider data collected from millions of people using self-monitoring fitness devices. The data are gathered and recorded as the subjects move around in their daily routine for their own benefit, but the data can reach insurance companies, which might determine that a customer is now risky. The data in this scenario were collected in a personal context but then used in an insurance context.[60] Changing the context means that subjects' prior expectations are frustrated and that legal norms that applied in its original context might be irrelevant in the second context. This change of purpose may happen also in small data cases, but there it is easier for the subject and for the controller to guard against such changes.

I join many scholars who have argued that the first generation of data protection law was unfit to meet the challenges of big data.[61] Notice is vague: all that the data collector can tell the data subject is that he is interested in her data, in order to process it, without further elaborating. Assessing the legitimacy of the purpose is difficult, as it is yet undefined. After all, the processor is interested exactly in finding the unexpected, unanticipated correlations. For the same reason, principles of data minimization and purpose limitation lose relevancy in a big data context. Consent becomes quite empty when one does not know what she is consenting to, other than that her data will be used for data analytics. Accessing one's data after it has been processed is difficult: if the data is properly anonymized, then re-identifying it should not be possible. The meta-data created by using personal data is not necessarily about a particular subject: it is about the relationships among the many subjects in the dataset.

[59] danah boyd & Kate Crawford, *Critical Questions for Big Data*, 15 Info. Comm. & Soc'y 662 (2012) (pointing to the importance of interpretation of big data within its context and to concerns of digital divides); Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 Calif. 671 (2016) (discussing unintended discriminatory effects of algorithmic decision-making).

[60] De-contextualization and re-contextualization pose a challenge to Nissenbaum's framework of Contextual Integrity. *See* Birnhack, *supra* note 21.

[61] Omer Tene & Jules Polonetsky, *Big Data for All: Privacy and User Control in the Age of Analytics*, 11 Nw. J. Tech. Intell. Prop. 239, 243 (2013); Kate Crawford & Jason Schultz, *Big Data and Due Process: Towards A Framework to Redress Predictive Privacy Harms*, 55 B.C. L. Rev. 93, 108 (2014); Viktor Mayer-Schönberger & Kenneth Cukier, Big Data: A Revolution that will Transform How We Live, Work and Think 16 (2013) ("[i]n an age of big data [data protection laws – M.B.] constitute a largely useless Maginot Line.").

These principles and their shortcomings extend to the second generation of data protection law, as exemplified in the GDPR.[62] The new generation of FIPs attempts to fix the shortcomings by strengthening the subject's rights and adding new tools to guarantee them.

FIP's *prima facie* failure to address big data does not obliterate privacy. A data subject's control is compromised when her data is used for purposes she did not anticipate or consent to. Her dignity is violated when she is used as a source of data and not respected as a person. Part III will tackle this issue. Now, let us dive into the medical domain.

II. Patient, Human Subject, Data Subject

Traditional medical research typically discusses specific cases or randomized controlled trials. Meta-analysis aggregates previous studies. Data science now offers a new research paradigm: big data algorithms and the availability of larger quantities of medical data enable researchers to examine much larger populations. Big data enables researchers to observe new correlations and inquire into causations. This is done based on data rather than patients, avoiding physical harm and saving time, money, and inconvenience. The potential is huge.[63]

Big medical data research has its challenges[64] and has drawn some skepticism[65] and criticism.[66] Here I focus on the privacy aspects. I leave aside genetic data, as it has additional complexities, such as that the data about one person indicates

[62] Tal Z. Zarsky, *Incompatible: The GDPR in the Age of Big Data*, 47 Seton Hall L. Rev. 995 (2017).

[63] Harlan M. Krumholz, *Big Data and New Knowledge in Medicine: The Thinking, Training, and Tools Needed for a Learning Health System*, 33 Health Aff. 1163 (2014) ("New big-data methods can turbocharge powers of observation in health care," and "This is a historic moment in medicine.").

[64] John P. A. Ioannidis, *Informed Consent, Big Data, and the Oxymoron of Research That Is Not Research*, 13 Am. J. Bioethics 40 (2013) (mentioning measurement errors, misclassification, confounding by indication, and selection biases); Federico Cabitza, *Breeding Electric Zebras in the Fields of Medicine* (Jan. 15, 2017), https://arxiv.org/ftp/arxiv/papers/1701/1701.04077.pdf.

[65] Ioannidis, *supra* note 64 (arguing that big medical data is "enthusiasm about fools' gold"); Cristian S. Calude & Giuseppe Longo, *The Deluge of Spurious Correlations in Big Data*, 22 Found. Sci. 595 (2017) (insisting on causation). An additional challenge is a distributional one: whose data will be collected and who will benefit thereof?

[66] Neil M. Richards & Jonathan H. King, *Three Paradoxes of Big Data*, 66 Stan. L. Rev. Online 41 (2013).

information about others.[67] I focus on research rather than commercial uses.[68] I begin with a concise overview of the regulatory framework for traditional medical research. I note the partial convergence of rules about human subject research and data protection law. I then describe emerging research practices: at present, there are no uniform practices, but researchers tend to rely on the retrospective nature of the de-identified data; in some jurisdictions this reliance is according to the law, while in others the law is silent or unclear.

A. Traditional Medical Research

Law, institutional norms, and ethical guidelines regulate traditional medical research. Doctors are bound by the Hippocratic Oath to keep their patients' data confidential.[69] There are few exceptions, justified by general policies meant to promote public health.[70]

When research is conducted in the course of clinical trials, data collected for the primary purpose (treatment) are entangled with the secondary purpose (research). When data are collected only for research, the initial meeting point between subject and researcher provides an opportunity to inform the former about the intended uses of her data. The interaction enables the provision of information and inquiries, channeled into the procedure of obtaining the

[67] Ifeoma Ajunwa, *Genetic Testing Meets Big Data: Tort and Contract Law Issues*, 75 Ohio St. L.J. 1225 (2014) (discussing wrongful disclosure of genetic information).

[68] For the commercial side of big medical data, see, for example, Nicolas P. Terry, *Regulatory Disruption and Arbitrage in Health-Care Data Protection*, 17 Yale J. Health Pol'y L. & Ethics 143, 178 (2017); Janine S. Hiller, *Healthy Predictions? Questions for Data Analytics in Health Care*, 53 Am. Bus. L.J. 251, 299-301 (2016).

[69] The oldest version of the Hippocratic Oath reads: "What I may see or hear in the course of the treatment or even outside of the treatment in regard to the life of men, which on no account one must spread abroad, I will keep to myself, holding such things shameful to be spoken about." *See* Peter Tyson, *The Hippocratic Oath Today*, Nova (Mar. 27, 2001), http://www.pbs.org/wgbh/nova/body/hippocratic-oath-today.html.

[70] *See, e.g.*, People's Health Ordinance, 5701-1940, SH No. 2516 § 12 (Isr.), which requires members of the public and doctors to notify the Ministry of Health about a person with a contagious disease. Nissenbaum flags such reporting duties as *prima facie* breaches of contextual integrity, but finds them acceptable, because they "support values of the healthcare context." *See* Nissenbaum, *supra* note 9, at 173.

patient's informed consent.[71] The subject's identity is known to the researcher, but at publication, the subjects' identities are not revealed. They become a "case." The patient agrees to become an anonymous human subject and then a data subject.

Medical research norms reflect the *Ethical Principles for Medical Research Involving Human Subjects*, known as the *Helsinki Declaration*.[72] Research in an academic setting requires the prior approval of an Institutional Review Board (IRB), typically operating under a specific law that implements the Helsinki principles. For example, in the United States, the Department of Health and Human Services (HHS) issued a policy for the protection of human research subjects, known as the Common Rule.[73] Universities have issued Human Subject policies.[74] These policies apply to research that interacts with human subjects in various ways. Surveys, interviews, psychological or educational experiments, monitoring activity, manipulating behavior, and physical procedures require ethical approval.[75]

The Helsinki Declaration states: "Every precaution must be taken to protect the privacy of research subjects and the confidentiality of their personal information."[76] The Common Rule requires the IRB to determine that the risk to the subject is minimized; that it is reasonable in relation to the anticipated benefits; that the selection of subjects is equitable; that informed consent is sought from each subject; and that, "When appropriate, there are adequate provisions to protect the privacy of subjects and to maintain the

[71] Informed consent in the traditional health context has its own shortcomings. *See* Robin Fretwell Wilson, *The Promise of Informed Consent*, *in* Oxford Handbook of U.S. Health Law 213 (I. Glenn Cohen, Allison K. Hoffman & William M. Sage eds., 2017).

[72] *WMA Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects*, World Med. Ass'n dedicated site, https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/ (last visited Mar. 19, 2018).

[73] The Common Rule, 45 C.F.R. § 46, subpart A (1991). These rules apply to federally funded research.

[74] *See, e.g.*, *About Us*, NYU, https://www.nyu.edu/research/resources-and-support-offices/getting-started-withyourresearch/human-subjects-research/about0.html (last visited May 21, 2018).

[75] Scholars have suggested extending the IRB procedures to commercial settings. *See* Ryan Calo, *Consumer Subject Review Boards: A Thought Experiment*, 66 Stan. L. Rev. Online 97, 102 (2013); Omer Tene & Jules Polonetsky, *Beyond IRBs: Ethical Guidelines for Data Research*, 72 Wash. & Lee L. Rev. Online 458 (2016).

[76] *WMA Declaration of Helsinki*, *supra* note 72, at § 24.

confidentiality of data."[77] Importantly, the Common Rule exempts some kinds of research, including "Research involving the collection or study of existing data, documents, records . . . if the information is recorded by the investigator in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects."[78] In other words, research that uses de-identified data to begin with is exempt.

To some extent, the ethical guidelines converge with FIPs discussed in the previous Part. Both derive from and reflect the Kantian notion of human dignity. The Helsinki Declaration and the Common Rule explicitly mention privacy; both the ethical guidelines and privacy mechanisms do not apply when data is non-identifiable; when data is identifiable, both require informed consent.

B. Big Medial Data Research

Every breath we take, every move we make, every step we take, someone is there collecting our personal data.[79] A trail of our medical data is created through direct interactions with various health services (doctors, clinics, hospitals, pharmacies), nonmedical parties (employers,[80] insurance companies), and other individuals (users who share their medical data online, *e.g.*, dating websites such as Grindr enable users to indicate their HIV status).[81] Medical data is also created by the use of various self-monitoring applications and devices,[82] and indirectly, through tracing of our shopping habits and lifestyle in general.[83]

Accumulated and aggregated, individuals' medical data becomes big medical data. The potential for data analytics in the healthcare sector is unprecedented, for example for more efficient management of limited resources.[84] Big medical

[77] 45 C.F.R. § 46.111.

[78] 45 C.F.R. § 46.101(b)(4).

[79] Homage is due to Gordon Sumner, a.k.a. Sting, and The Police (1983).

[80] Ifeoma Ajunwa, Kate Crawford & Joel S. Ford, *Health and Big Data: An Ethical Framework for Health Information Collection by Corporate Wellness Programs*, 44 J.L. Med. & Ethics 474 (2016).

[81] Grindr Privacy Policy, https://www.grindr.com/privacy-policy (last visited Aug. 9, 2017).

[82] Heather Patterson & Helen Nissenbaum, Content-Dependent Expectations of Privacy in Self-Generated Mobile Health Data (May 22, 2013) (unpublished manuscript), https://ssrn.com/abstract=3115768.

[83] Joseph Turow, The Aisles Have Eyes: How Retailers Track Your Shopping, Strip Your Privacy, and Define Your Power (2017).

[84] *See, e.g.*, David W. Bates et al., *Big Data in Health Care: Using Analytics to Identify and Manage High-Risk and High-Cost Patients*, 33 Health Aff. 1123

data analysis enables new insights: A doctor might realize that his diagnosis was correct, but the patient did not buy the medicine. The insurance company might better assess the risks that result from our lifestyle. An employer can better assess the future unavailability of employees due to their health. On an aggregate level, employers, insurance companies, and the government can better identify patterns, understand epidemiologic outbreaks, predict trends, and apply this knowledge to individuals to assess their future path. The beneficiaries are the patients themselves (faster and more accurate diagnosis, better treatment), other interested parties, such as employers and insurance companies, and the public at large.

Increasingly, researchers, health organizations, and governments are realizing this potential.[85] Those who hold the data have won a windfall.[86] There are new projects,[87] collaborations of healthcare providers with the government,[88] academia,[89] and private entities.[90] However, most of the world

(2014) (suggesting the identification of high-cost patients so as to manage their cases better; predicting readmission cases; predicting patients at risk of adverse events such as infection).

[85] *For example*, a background report by the European Commission, Directorate-General for Health and Consumers, *The Use of Big Data in Public Health Policy and Research* (Aug. 29, 2014), https://ec.europa.eu/health/sites/health/files/ehealth/docs/ev_20141118_co07b_en.pdf; Exec. office of the President, Big Data: Seizing Opportunities, Preserving Values 22-24 (2014).

[86] Healthcare payers hold much data and hence are likely to initiate big medical data research. *See* Rebecca S. Eisenberg & W. Nicholson Price II, *Promoting Healthcare Innovation on the Demand Side*, 4 J.L. & Biosciences 3 (2017).

[87] *E.g.*, Ran D. Balicer & Arnon Afek, *Digital Health Nation: Israel's Global Big Data Innovation Hub*, 389 Lancet 2451 (2017).

[88] *E.g.*, in Sweden, Jillian Oderkirk, *Governing Data for Better Health and Healthcare*, OECD Observer (Jan. 2017), http://oecdobserver.org/news/fullstory.php/aid/5780/Governing_data_for_better_health_and_health_care.html.

[89] *E.g.*, *Weizmann Institute of Science and Clalit Health Services Will Collaborate to Advance the Field of Personalized Medicine and Improve Health Care Services*, Weizmann Inst. (Mar. 10, 2014), https://wis-wander.weizmann.ac.il/life-sciences/weizmann-institute-science-and-clalit-health-services-will-collaborate-advance-field.

[90] *E.g.*, the collaboration between the Royal Free London NHS Foundation and DeepMind Technologies Ltd., in which the RFL transferred identifiable patient records for a research on kidney injuries. For a critical assessment, see Julia Powles & Hal Hodson, *Google DeepMind and Healthcare in an Age of Algorithms*, 7 Health & Tech. 351 (2017).

is still behind.[91] There is a growing demand for direction[92] and initial attempts to set and clarify legal standards.[93]

There are a few important characteristics of big medical data that implicate privacy: (1) research is conducted on data, not on human subjects; (2) data are gathered from multiple sources; (3) data are collected primarily for treatment and only secondarily for research (of course, data may be collected directly for research purposes, which is less troublesome in terms of privacy, as long as FIPs are followed and ethical standards are met); and (4) there is no direct meeting point between researcher and subject. I comment briefly on each.

Data, not the body. Big medical data research is conducted *ex post*, after treatment and after the collection of data. Whereas in traditional data the patient becomes a human subject and then a data subject, consenting to the anonymous use, in a big data context the patient becomes a data subject, skipping the status of human subject.

Multiple sources. We have noted that our medical data is scattered all over the place. The data might be in different formats: blood counts, imaging (CT, MRI, X-rays etc.), printed or hand-written information. The data is kept by different kinds of entities: public (*e.g.*, the NHS in the UK), regulated providers (*e.g.*, Israeli Health Maintenance Organizations (HMOs)), or private entities under no specific regulation. This multiplicity poses technological and organizational challenges, such as how to connect the separate bits to one patient. Where a single identifying number is used throughout the medical system, this task is easier, but the risk of re-identification is also higher.

Secondary Purpose. In the course of traditional medical research, human subjects might receive treatment that they would not otherwise have received, such as participating in clinical research. The benefits and risks are medical. By contrast, big medical data research is often ancillary to treatment. Data are created in the course of medical treatment and other activities, as an inevitable part thereof. The risks are not physical but privacy-related.

No direct meeting point. In big medical data studies, volume means that researchers cannot obtain informed consent directly. The absence of a meeting point means that inquiries, explanations, additional information, etc., take a different form than a face-to-face interaction. Creating a meeting point is not

[91] Oderkirk, *supra* note 88 (noting that only half of the OECD countries have policies about big medical data).

[92] *E.g.*, Lisa M. Lee & Lawrence O. Gostin, *Ethical Collection, Storage, and Use of Public Health Data*, 302 JAMA 83 (2009).

[93] *E.g.*, Organisation for Economic Cooperation and Development [OECD], *The Next Generation of Health Reforms* (Jan. 17, 2017), http://www.oecd.org/health/ministerial/ministerial-statement-2017.pdf.

in the hands of the researcher, but rather lies within the power of those who hold the data to begin with, namely HMOs.

The aggregation of these features induces towards the abandonment of the ethical and privacy rules that were developed in the context of clinical trials,[94] and specifically of informed consent.[95] While such an approach will promote big medical data research, it has a cost in terms of privacy.

C. The ColoRectal Cancer Case

Researchers face various challenges.[96] The focus here is on privacy. I begin with one illustrative example, and then turn to an overview of emerging practices in big medical data research.

The research by Goldshtein, Neeman, Chodick, and Shalev is an extraordinary such case.[97] A study that began with Shalev's intuition regarding a patient of hers who suffered from ColoRectal Cancer (CRC) and died, led to an examination of long-term variations in blood hemoglobin levels, which are within the normal range, and hence a physician reviewing blood test results could easily miss them. In a first study, 1074 CRC cases were matched with cancer-free individuals according to age and sex, with 10 controls per cancer case. The results were stunning. The researchers reported: "Our retrospective analysis

[94] *See, e.g.*, Barbara J. Evans, *Big Data and Individual Autonomy in a Crowd*, *in* Big Data, Health Law, and Bioethics 19, 26 (I. Glenn Cohen et al. eds., 2018).

[95] *See, e.g.*, I. Glenn Cohem, *Is There a Duty to Share Healthcare Data?*, *in* Big Data, Health Law, and Bioethics 209 (I. Glenn Cohen et al. eds., 2018) (arguing that consent is unnecessary and advocating a duty to share healthcare data).

[96] *E.g.*, Sebastian Schneeweiss, *Learning from Big Health Care Data*, 370 New Eng. J. Med. 2161 (2014) (lack of uniform data standards); Naren Ramarkishnan et al., *Mining Electronic Health Records*, 43 Computer 77 (2010) (incompleteness of data); Powles & Hodson, *supra* note 90 (lack of transparency, corporate responsibility, the accumulation of market power and more); Sharona Hoffman & Andy Podgurski, *Big Bad Data: Law, Public Health, and Biomedical Databases*, 41 J.L. Med. & Ethics 56 (2013) (Electronic Health Records (EHRs) shortcomings). In the United States, the pace of the adoption of EHRs has improved, following the Health Information Technology for Economic and Clinical Health Act, Title XIII of Division A and Title IV of Division B of the American Recovery and Reinvestment Act of 2009, 42 U.S.C. §§17931-17937. *See* Janine S. Hiller, *Healthy Predictions? Questions for Data Analytics in Health Care*, 53 Am. Bus. L.J. 251 (2016).

[97] Inbal Goldshtein, U. Neeman, Gabriel Chodick & Varda Shalev, *Variations in Hemoglobin before Colorectal Cancer Diagnosis*, 19 Eur. J. Cancer Prevention 342 (2010).

indicates that starting from 4 years prior to cancer diagnosis, a progressive significant (P<0.001) decrement in Hb levels (0.28 g/dl per 6 months) was found among cases but not among controls."[98] In nonmedical language, they found a way to diagnose CRC years before conventional diagnosis.

A follow-up study examined 606,403 Israelis, taken from Maccabi Healthcare Services, the second largest Israeli HMO, with about two million patients, as well as data from the UK Health Improvement Network (THIN).[99] The researchers used the entire population of patients above 40 years old. Within this dataset, they applied computational models. They developed an algorithm that can detect 50% of CRC cases 3-6 months before diagnosis. The findings have tremendous implications: simply put, they can save many lives, as well as time, money and the unpleasant medical checks for those facing less risk.

Needless to say, such a large-scale study was impossible a few years ago. The researchers had access to the Electronic Health Records (EHRs) of all relevant patients. They noted that they anonymized the data and de-identified it prior to analysis, but did not elaborate. The study was approved by Maccabi's Institutional Ethics Committee in a Helsinki review and by THIN. The researchers note that "The Ethics Committees granted waivers of informed consent since this study involved analyses of retrospective data where all patient information was anonymized and de-identified prior to analysis."[100]

The research was feasible because the data was already there; informed consent was avoided by applying anonymization and de-identification, and the internal ethical bodies approved. Privacy was not mentioned in the article, presumably because the researchers believed it was subsumed within the ethical review.

D. Emerging Research Practices

To appreciate the emerging practices, a literature search was conducted in PubMed in January 2018,[101] for terms that indicate big data research.[102] The

[98] *Id.*

[99] Yaron Kinar et al., *Development and Validation of a Predictive Model for Detection of Colorectal Cancer in Primary Care by Analysis of Complete Blood Counts: A Binational Retrospective Study*, 23 J. Am. Med. Informatics Ass'n. 879 (2016).

[100] *Id.* at 888.

[101] PubMed is an open search engine specializing in medical research, operated by the American National Center for Biotechnology Information, and "comprises more than 27 million citations for biomedical literature from MEDLINE, life science journals, and online books." *See* PubMed, https://www.ncbi.nlm.nih.gov/pubmed (last visited May 21, 2018).

[102] Search terms were "Retrospective Studies," "Big Data," "Machine Learning" and "Electronic Health Records."

search included full-text articles in English regarding human subjects that had a medically clinical emphasis. Genetic research, editorials, commentaries, reviews and conference summaries were excluded. The search yielded fifty-three studies that met these criteria.

These fifty-three studies were conducted mostly in developed countries: thirteen in the United States and twelve in South Korea, and the remainder scattered in other countries.[103] We see a fast growth of such studies, indicating that many more are to follow: two studies were published between 2011 and 2014; ten in 2015, twenty-three in 2016 and seventeen in 2017. As for the scope of the studies, twenty-three studies analyzed less than a hundred thousand records (the smallest: 5469); nineteen studies analyzed between a hundred thousand and a million records, and ten analyzed more than one million records. The largest analyzed two billion search queries submitted to a human-guided online service. The studies addressed diverse medical fields, ranging from oncology to psychiatry. Sources varied. Some used governmental datasets, but most used data provided by hospitals; some used special anonymized datasets, or data provided by a device manufacturer.

Ethical Approval: twenty-seven studies explicitly mentioned that they received ethical approval from a hospital or academic IRB, or the IRB of the data controller; nine from governmental agencies, and one from the World Health Organization. Two British studies received the approval of a non-statutory expert advisory board, which reviews requests to access the Clinical Practice Research Datalink (CPRD).[104] Eleven studies did not mention any ethical approval. Of these, eight were conducted in the United States, and one each in Australia, Cyprus, and the Netherlands. Three more studies mentioned that they were exempt from ethical approval, one due to using a dataset that lacked personal identifiers, one because it used public datasets, and another "owing to the retrospective nature" of the study.

Consent: Nineteen articles did not mention consent at all. Two referred to other kinds of consent (to the medical treatment or for drawing a droplet of blood). Thirteen referred to the retrospective nature of the research as a reason for not obtaining consent, or stated that their IRB waived or exempted the

[103] Seven in Australia, six in Taiwan, three each in the UK and New Zealand, two in Japan, and one each in Canada, China, Cyprus, Denmark, Hong Kong, Israel (with the UK), the Netherlands, and Sweden, and a multinational study with data from Guinea, Liberia, and Sierra Leone.

[104] The advisory board is the Independent Scientific Advisory Committee for MHRA database research (ISAC). *See Independent Scientific Advisory Committee for MHRA Database Research*, Gov.UK, https://www.gov.uk/government/groups/independent-scientific-advisory-committee-for-mhra-database-research (last visited May 21, 2018).

consent requirement. In the remainder of the articles, there were multiple kinds of explicit references to consent in concise and sometimes cryptic language. We saw one example, mentioning the retrospective nature of the study and the use of anonymized and de-identified data.[105] A few articles mentioned other kinds of consent: A New Zealand research used a unique identifier to match records (repeat visits to the GP), but added that participants consented to the collection and use of non-identifiable data when they enrolled with their General Practices.[106] In a Canadian study, subjects consented to linking their data from separate datasets into one.[107]

De-identification: Twenty articles reported that they used de-identified data. In five of these, the researchers seem to have conducted the de-identification themselves. For example, in a Korean study, researchers explained that "informed consent was waived because the anonymized data was analyzed retrospectively," Data extracted from hospital EHRs included clinical and demographic data.[108] Two American studies stated that their university IRBs

[105] Kinar et al., *supra* note 99.

[106] Anthony Dowell et al., *Childhood Respiratory Illness Presentation and Service Utilisation in Primary Care: A Six-Year Cohort Study in Wellington, New Zealand, Using Natural Language Processing (NLP) Software*, 7 BMJ Open 1 (2017). Most New Zealanders enroll with a primary HMO when enrolling with their GP. *See Enrollment in a Primary Health Organization*, Ministry of Health, https://www.health.govt.nz/our-work/primary-health-care/about-primary-health-organisations/enrolment-primary-health-organisation (last visited May 7, 2018). The current enrollment form explains that the GP "participates in a national survey about people's health care experience and how their overall care is managed." Patients can opt out. A Fact Sheet explains that health information is collected, *inter alia*, to "carry out authorized research." The Fact Sheet details the patient's rights, including "You have the right to know where your information is kept, who has access rights, and, if the system has audit log capability, who has viewed or updated your information." As for consent, it explains that "Research which may directly or indirectly identify you can only be published if the researcher has previously obtained your consent" and that research that does not identify the person does not require consent. *See Use and Confidentiality of Your Health Information Fact Sheet*, Ministry of Health, https://www.health.govt.nz/system/files/documents/pages/use-of-health-information-statement-november-2016.docx (last visited May 21, 2018).

[107] Dennis T. Ko et al., *High-Density Lipoprotein Cholesterol and Cause-Specific Mortality in Individuals without Previous Cardiovascular Conditions*, 68 J. Am. College of Cardiology 2073 (2016).

[108] Yoon Seob Kim et al., *Extracting Information from Free-Text Electronic Patient Records to Identify Practice Based Evidence of the Performance of Coronary Stents*, 12 PLoS ONE 1 (2017).

considered the research to be non-human subject research. However, the data collected was about patients and included race, age, sex, and additional diagnosis. In thirteen studies, the datasets were handed to the researchers in a de-identified manner to begin with. Nevertheless, the level of de-identification is unclear. For example, an American research team explained that they obtained data from several sources without patient identifiers, one of them from a manufacturer of pacemakers; however, the data included date of implantation, age, sex, patient zip code, and device model numbers.[109]

The studies also indicate various research designs, from the privacy perspective. One option was to use data that was de-identified by the data controller, prior to handing it to the researchers. The two British studies are illustrative. They used CPRD, a governmental research service established in 1987.[110] Patients have a unique NHS identifier, but the CPRD notes that "It is only used by a trusted third party for linkage and is never released to researchers. It is a benefit in ensuring records can be validly linked within the approved governance process."[111] It also explains that "CPRD *never* receives patient identifiable data from GP practices or from NHS Digital."[112]

The absence of references to ethical issues in about a third of the articles and the variety of references in the rest reflect the deeper issues. These studies were conducted in different jurisdictions. The general overview indicates that at present, universal research conventions are yet to emerge. Given the multiple research sites, but that important journals are fewer, publishers are best-located to enforce such norms.

Two frequent explanations were that the study was based on de-identified data and/or that it was conducted retrospectively. Are these convincing explanations? The next Part will apply the process-based approach to this issue.

III. A Process-Based Approach to Big Medical Data

This Part connects the previous two: data protection law and big medical data. It begins with acknowledging the privacy harm in big medical data, and continues with the key issue of de-identification. I return to the process-based

[109] Niraj Varma et al., *The Relationship between Level of Adherence to Automatic Wireless Remote Monitoring and Survival in Pacemaker and Defibrillator Patients*, 65 J. Am. C. Cardiology 2601 (2015).

[110] Clinical Prac. Res. Datalink (CPRD), https://www.cprd.com/home/ (last visited May 21, 2018).

[111] *Id.*

[112] *GP Practice Consent to Datalink*, CPRD, https://www.cprd.com/EmisLinkageConsent/ (last visited May 21, 2018).

approach to data protection law, and show how it can assist us in addressing this issue.

A. Why Is Medical Privacy Important?

In the traditional research mode, answering this question is easy. We trust our doctor and healthcare providers to give us the best treatment. We treat medical data as sensitive and expect the doctor to maintain confidentiality. If we wish to share our medical information with others, it is for us to decide. Should this sensitivity persist when the data becomes part of a large anonymized dataset? What is the privacy harm in a big data context? There are some well-discussed issues of profiling, individual predictions based on patterns identified by big data analysis, and lack of transparency.[113] Here I focus on the unique privacy harms created by big medical data, by examining the critiques that privacy advocates often face. I argue that the harm lies in disregarding the subject in the first step of the informational process.

1. Nothing to Hide

Probably the most popular anti-privacy argument is that if one has not done anything wrong, there is no reason to hide anything. The argument equates privacy with secrecy, to the neglect of its broader understanding, namely subjects' control over their information, which reflects the subjects' human dignity. This argument has met powerful answers.[114] The medical context is a clear situation in which a person has done nothing wrong, and yet wishes to keep the information private.

Taken to the big data context, we should assure subjects that their data is not shared in a way that risks them. One way to do so is anonymization, on the assumption that this is possible, accompanied with preventing leakages and unauthorized access. To enable subjects' real control over their data, we should notify them and ask for permission to use their anonymous data. Emerging research practices are yet to take this route. These practices do not respect one's autonomy to make decisions for herself, unless we place all our cards on anonymization.

[113] Mireille Hildebrandt, *Defining Profiling: A New Type of Knowledge?*, *in* Profiling the European Citizen: Cross-Disciplinary Perspectives 17 (Mireille Hildebrandt & Serge Gutwirth eds., 2008) (profiling); Crawford & Schultz, *supra* note 61 (predictions); Tene & Polonetsky, *supra* note 61; Tal Z. Zarsky, *Transparent Predictions*, 2013 U. Ill. L. Rev. 1503 (2013) (transparency).

[114] Daniel J. Solove, Nothing to Hide: The False Tradeoff between Privacy and Security (2011).

2. *Share to Shatter the Stigma*

Medical conditions often suffer from social stigma: people tend not to share information about mental health, sexually transmitted diseases, etc. Such stigmas are unfortunate. They add misery to the sick and might prompt them to avoid seeking medical care. Privacy, say its critics, reinforces stigma.[115] Such a claim echoes the self-empowerment argument made in the context of outing, namely, that the unilateral exposure of another's sexual orientation is justified as a means to shatter the closet.[116] Can this claim justify using medical data without the person's consent?

Big data carries a promise. Subjects who are concerned about the potential negative social response to their medical condition can still keep it to themselves, but at the same time assist researchers in studying it. Subjects can remain private and contribute to the community, thus expressing solidarity and performing altruism.

However, to decide for a person what to share, with whom, when and how, is to ignore her human dignity. Taking someone's personal data without her knowledge or against her will, or using it for purposes other than those to which she consented, disregards the person. The good intentions, with which the subject might sympathize, should not override the subject's own choice and deprive her of the ability to make decisions for herself. This is the case with outing, and its logic applies to medical data too.

3. *Harm is Speculative*

Another argument against privacy is that the harm is speculative. However, there are enough cases with real harm: when the integrity of data is compromised, whether because of negligence or malicious intent. Medical data leaks.[117] Data

[115] In the case of the HIV status of gay men during the 1980s, see, for example John F. Hernandez, *Outing in the Time of Aids: Legal and Ethical Considerations*, 5 St. Thomas L. Rev. 493 (1993).

[116] Eve Kosofsky Sedgwick, Epistemology of the Closet (1990); Ronald F. Wick, *Out of the Closet and into the Headlines: "Outing" and the Private Facts Tort*, 80 Geo. L.J. 413 (1991).

[117] Most recently, see Taylor Hatmaker, *Healthcare Data Breach in Singapore Affected 1.5M Patients, Targeted the Prime Minister*, TechCrunch (July 20, 2018), https://techcrunch.com/2018/07/20/singapore-hack-health/. According to Gemalto, a data security company that operates a Breach Level Index, about 3% of data breaches are in healthcare, with 2015 leading with 19.3% of all reported breaches. *See* Breach Level Index, https://breachlevelindex.com/ (last visited May 21, 2018). The largest breach was in 2015, with an American insurance company, affecting 78 million records. The HSS lists 408 cases under investigation. *See* Breach Portal: Notice to the Secretary of HHS Breach

in the wrong hands can be used against the person, for example, to expose someone's noncontagious but nevertheless stigmatized disease. Data might be (mis)used to discriminate against a person on bases that are otherwise prohibited by law, such as not hiring a woman based on her having a BRCA1 gene mutation.[118] Medical data can be misused to deny legitimate benefits or services.

In the context of big data, once subjects realize that their data is further transferred and used, some might be less willing to share it to begin with. Without effective notice and consent, big medical data acts as a black box. This is unlikely to produce trust. Some subjects are likely to respond to the new data practices by avoiding medical inquiries, or by turning to undocumented treatments.[119]

4. Implicit Consent

Focusing on data gathered during medical treatment and then used for research, another argument against privacy is that the subject acted in a way that indicates that she did not consider the use of the data harmful. For example, we might say that once the patient gave her data to the clinic, she consented to its processing, and cannot have reasonably expected it to remain private. This is currently the American response, *i.e.*, the third party doctrine.[120]

However, recall that our medical data is scattered in various places. We provided the data ourselves, or allowed others to (sometimes literally) extract the data from our bodies in order to receive medical treatment. If we were to assume that implicit consent in the doctor-patient context extends to the big data context, we would be ignoring the many differences between these contexts, noted above. Put in Nissenbaum's terms, the former context has clear rules about the onwards transmission of information, whereas the latter

of Unsecured Protected Health Information, https://ocrportal.hhs.gov/ocr/breach/breach_report.jsf (last visited Feb. 17, 2018), the largest affecting 36 million patients.

[118] Sharona Hoffman, *Citizen Science: The Law and Ethics of Public Access to Medical Big Data*, 30 Berkeley Tech. L.J. 1741, 1774-79 (2015) (discussing employers' misuse of employees' medical data).

[119] Mark A. Rothstein, *Is Deidentification Sufficient to Protect Health Privacy in Research?*, 10 Am. J. Bioethics 3, 7-8 (2010) (discussing risks associated with losing trust).

[120] Per this doctrine, when a data subject hands personal data to a third party, such as a phone provider, the subject's expectations of privacy vanish, hence they no longer have a protected privacy interest in that data. In Carpenter v. United States, 585 U.S. __ (2018), the Supreme Court narrowly overruled the doctrine in the context of location data obtained from cellphone providers.

still has none. This is why the GDPR explicitly states that the purposes of medical treatment and research are compatible: originally, they are deeply incompatible, and the law has to interfere to reach the opposite result.

Accordingly, we can conclude that privacy does matter also when one's medical data is processed with numerous other data in a de-identified manner. Medical data is sensitive, its collection cannot be done against one's will, the harm is both dignitarian and real, and using data in a different context frustrates reasonable expectations.

B. De-identification

Under emerging big medical data research practices, anonymization bears the heaviest burden. Identifiability is the single criterion that triggers data protection law in the EU, regardless of the content of the data.[121] EU data protection law is universal and applies to sensitive data such as medical data as well as to mundane data about our boring television preferences, routine whereabouts, and practically, any kind of data. In this sense, EU data protection law anticipated the digital condition: the understanding that separate bits of data can be combined together and their aggregation may reveal more than each of the separate bits. Hence, when a person is identified, data protection law applies. Accordingly, for many years, anonymization was key: if the processor could anonymize the data, processing was legitimate, and the law was not triggered at all.

The law covers situations in which a person is not directly identified, but can nevertheless be identified.[122] A processor who wishes to avoid the law's requirements would attempt to anonymize or de-identify the data. There are various techniques for de-identification. Deleting direct identifiers such as names, addresses, and unique (national ID or Social Security) numbers is obvious. Additional measures include using sophisticated mathematical tools, *e.g.*, k-anonymity,[123] or clustering data together. For example, instead of referring to a subject's age (31), we can cluster those in the range of 30-35 together. Adding noise to the data, or replacing some of the data with information that renders it more difficult to re-identify the person, are additional means.

[121] In processing some kinds of data, health-related data included, the law sets higher standards. *See* GDPR, *supra* note 4, at art. 9.

[122] GDPR, *supra* note 4, at art. 4(1).

[123] *Latanya Sweeney*, *K-Anonymity: A Model for Protecting Privacy*, 10 Int'l J. Uncertainty, Fuzziness & Knowledge-Based Sys. 557 (2002).

However, in a digital world, re-identification has become easier, to the level that some have declared the death of anonymization.[124] This does not mean that data protection law is obsolete. On the contrary, it means that the law applies also in situations that have thus far gone under the legal radar. If anonymization is suspicious, the working assumption should be that de-identified data can be re-identified. In Part II.D we saw examples of studies that claimed to use de-identified data, but the kinds of data mentioned are almost an invitation for hackers and other ill-intentioned parties to re-identify the data.

Nevertheless, I submit that anonymization is still a valid legal concept. Indeed, we should be more careful in relying on anonymization and researchers should be diligent in the de-identification measures they apply, but we do not have to throw away the legal threshold. Instead of a binary identifiability dichotomy, we should shift to a spectrum and ask what would be required to re-identify data. This is the approach taken by the GDPR, which applies a *reasonable likelihood* standard for identifiability, taking into account costs, time, and the available technology.[125]

Replacing the anonymous/identified dichotomy with a spectrum means a legal shift from a rule to a standard. The motivated intruder should meet the reasonable data controller. The controller should of course use updated technological means to assure anonymity, and these should be supplemented with suitable organizational measures, such as raising awareness among employees or setting disciplinary procedures. The law should strengthen these measures with background rules, such as reporting to a DPA when breaches occur, imposing civil liability or administrative fines, and in extreme cases, setting criminal sanctions.

Medical research has unique needs. De-identification is inevitable due to the huge quantities of data. But the researchers, as they process the data and before they publish the results, might need to know more. For example, the exact age of subjects might be crucial, because it indicates the vaccinations received at birth. Address might indicate proximity to an environmental hazard. Moreover, aggregating data from different sources is important to have as full and accurate a picture as possible. For example, it is insufficient to know that a patient was prescribed a certain medicine: researchers would like to know

[124] Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. Rev. 1701 (2010). *But see* Felix T. Wu, *Defining Privacy and Utility in Data Sets*, 84 U. Colorado L. Rev. 1117 (2013).

[125] GDPR, *supra* note 4, at recital 26. Similarly, UK law applies a "motivated intruder" standard, see Information Commissioner's Office, Anonymisation: Managing Data Protection Risk, Code of Practice 22 (2012).

whether she actually received the medicine from the pharmacy (this does not yet say she took it, but we are getting closer to the truth).[126] To facilitate such aggregation, and moreover, to track subjects over time, we need to be able to identify them, yet maintain their anonymity. In other words, researchers often need to use a unique identifier. The GDPR treats such data as pseudonymous rather than anonymous, with the implication that the law applies.[127]

C. American Medical Research

In the United States, under the Health Information Insurance Portability and Accountability Act 1996 (HIPAA), the law is triggered based both on the content of the data and its identifiability.[128] HIPAA regulates the processing of "protected health information,"[129] which refers to "individually identifiable health information."[130] This definition includes information that identifies the individual, or such that "there is [a] reasonable basis to believe that the information can be used to identify the individual."[131] The law offers two options for de-identifying health information: (1) when a person with appropriate knowledge and experience in de-identification determines that the risk of re-identifying a person is "very small,"[132] or (2) removing 18 different identifying variables, such as names, geographical subdivision, phone numbers, SSN, medical record numbers, device identifiers, or biometric identifiers.[133]

The HHS' Common Rule allows for an expedited review — not an exemption — "for certain kinds of research involving no more than minimal risk," as determined by the Secretary,[134] for example the collection of nail clippings,[135] or, and relevant to the current discussion, "Research involving materials (data, documents, records, or specimens) that have been collected, or will be collected solely for nonresearch purposes (such as medical treatment

[126] Prof. Ehud Grossman, Remarks at the TAU Big Medical Data Workshop (Dec. 7, 2016).

[127] GDPR, *supra* note 4, at recital 26.

[128] 42 U.S.C. §1320-d(4) (2017) (defining "health information"). Under HIPAA, the HHS promulgated the Privacy Rule. *See* 45 C.F.R. §§164.500-164.534 (2017).

[129] 45 CFR §160.103 (2017). HIPAA does not cover various governmental agencies and private initiatives. *See* Hoffman, *supra* note 118, at 1765, 1769.

[130] 42 U.S.C. §1320-d(6) (2017).

[131] *Id.*; 45 C.F.R. §164.514(a) (2017).

[132] 45 C.F.R. §164.514(b)(1) (2017).

[133] 45 C.F.R. §164.514(b)(2) (2017).

[134] 45 C.F.R. §46.110(a) (2017).

[135] Rules made by the Office for Protection from Research Risks, National Institutes of Health, HHS. 63 Fed. Reg. 60364, 60366 (Nov. 9, 1998).

or diagnosis)."[136] Thus, the law makes an *ex ante* determination that some kinds of research are less harmful; data-based studies are included here. Existing American guidelines are quite friendly to big medical data: they trust anonymization and accept that secondary use is less risky to human subjects than data collected primarily for such research. But they do so at the price of compromising privacy.

D. A Process-Based Approach to Big Medical Data

Let us recap: the emerging practice in big medical data is to assert an exemption from obtaining informed consent. The explanation downplays the importance of privacy, emphasizes the benefits of medical research, and points to the retrospective nature of the research and its use of de-identified data. This section argues that this approach is misguided. It ignores the previous steps in the data-chain, and hence disserves data protection law in word and in spirit. Instead, the process-based approach is better suited to regulating big medical data, as it fits existing law and reflects its underlying understanding of human dignity. This approach requires a change among researchers and their practices, as well as institutional changes to set and enforce norms.

1. Ex Post Tradeoffs or Ex Ante Protection?

One way to view the informational process is to zoom out. This is the forest view. Accordingly, we will take note of the first and last points in time of the medical research. Although in the first step of the information flow, the data subject is identified, by the time we reach the last step, that of the published outcome, subjects are anonymous. The articles that the researchers publish discuss general findings rather than individuals' raw data. Accordingly, if we examine the criterion of identifiability at the last step of the informational process, we can be satisfied that privacy is maintained.

But data protection law is not only about publication. It is about respecting subjects' control of their data. Moreover, medical data is sensitive. When collected during medical treatment, it is collected for the sole purpose of having the patient receive the best treatment. In the absence of informed and free consent, the data should not be used for additional purposes. De-identifying the data after its collection does not obliterate the legal duties that should apply to prior steps in the data lifecycle. Indeed, de-identification might minimize real harms of the data reaching the wrong hands, but it does not address the fact that the person is treated as a means rather than an end. As we saw, the GDPR explicitly exempts research, trading off privacy for research.

[136] *Id.* category 5.

Instead, we should zoom in. We should look at the treatment of the data in each step. At the first step of the informational process, the data had not yet been anonymized. The data was collected in an identified form, used for one purpose, and only then was it anonymized so as to be used for another purpose, that of research. Viewing the process in such a manner, we should insist on applying data protection principles in the first step. Subsequent steps raise fewer privacy concerns. Processing the data is done in the aggregate, referring to (by then) anonymized data, and at the last step, the data is anonymous, and if we are satisfied that it cannot reasonably be re-identified, then there is no privacy issue at that point.

Which view should we take: the forest or the trees? Reverting to the justifications of informational privacy and to the recognition of the privacy harms at stake provides us with a theoretical yardstick. The forest view assumes that the harm to privacy materializes only once data is published: it is then that the data can fall into the wrong hands, be misused or abused against the data subject. Accordingly, *proper* anonymization eliminates this risk. The process-based approach reminds us of the dignitarian harm. Accordingly, the use of the data for a purpose other than that for which the data subject consented to in the first place, constitutes a disregard of the person. The subject should have control over the first step as well: that of anonymization.

2. Consent and Privacy by Design

Consent is the cornerstone of FIPs, as it is the first meeting point between the subject and the processor, and notorious it is for its resounding failure to achieve this goal.[137] But with healthcare there is much at stake; the level of sensitivity is high, the expectation of privacy is strong, and at the same time, so are the potential uses of the data. Can consent fulfill the task? Obtaining consent retrospectively is difficult and expensive, but the typical dataset holders have the means. HMOs maintain ongoing contact with their patients, as do hospitals.[138] Governmental agencies that hold datasets have the means to approach patients. Research-wise, retrospective consent might not be good enough for some medical studies: those who respond positively might not be a representative sample. Nevertheless, as in any smaller data research, researchers should study the dataset carefully and be aware of its limitations.

A better approach is a prospective one, and we are at the point in time where this is still possible. New Zealand's approach is an option: during

[137] Consent appeared in all data protection documents from their beginning. *See, e.g.*, The OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data (1980), art. 7.

[138] *E.g.*, Ko, *supra* note 107.

enrolment for healthcare services, patients are given information about the use of their data for various purposes, including research. Of course, as one article noted, "one cannot effectively communicate the potentially enormous range of testable hypotheses to patients."[139]

The form of consent is important. Obtaining consent for every big data study is costly and practically impossible, as the researchers might not know in advance what exactly they are looking for. After all, this is one of the advantages of big data: by mixing, matching, and mining data from various sources, it yields previously unknown and perhaps surprising correlations. Moreover, if many potential participants refuse, the dataset might be skewed in ways that the researchers will have to carefully pay attention to, as any researcher should in studying any dataset. Alternatives are a blanket ("broad") consent or a tiered consent, enabling the subject to agree to some uses but not to others.[140] Designing the datasets to include metatags about consent is an option worthy of exploration.[141] Such consent should be informed—to the extent possible—and given of free will. This is informed sharing.[142] To ensure the latter, we need to separate the medical treatment from the request to consent in big data studies. Otherwise, patients might be afraid to refuse.

Another such approach is to build designated large datasets that bifurcate the collection of data from its use. Datasets that were created for the purpose of monitoring epidemics, such as Ebola in Western Africa, contain anonymous data. When handed over to researchers for further analysis, the integrity of the informational flows is maintained: the secondary use is compatible with the first; subjects handed over their data either with consent or according to a valid legal requirement. The British CPRD operates in a similar manner. The data is gathered from GPs in an anonymous way, aggregated for research use,

[139] Markus Christen et al., *On the Compatibility of Big Data Driven Research and Informed Consent: The Example of the Human Brain Project*, *in* The Ethics of Biomedical Big Data 199, 209 (Brent Daniel Mittelstadt & Luciano Floridi eds., 2016).

[140] For a critical view of the literature, see Brent Daniel Mittelstadt & Luciano Floridi, *The Ethics of Big Data: Current and Foreseeable issues in Biomedical Contexts*, 22 Sci. & Engineering Ethics 303, 311-16 (2016).

[141] J. Patrick Woolley, *How Data are Transforming the Landscape of Biomedical Ethics: The Need for ELSI Metadata on Consent*, *in* The Ethics of Biomedical Big Data 171 (Brent Daniel Mittelstadt & Luciano Floridi eds., 2016).

[142] Joachim Roski George W. Bo-Linn & Timothy A. Andrews, *Creating Value in Health Care Through Big Data: Opportunities and Policy Implications*, 33 Health Aff. 1115, 1119 (2014).

and each research is subject to an ethical review. The design of the system is privacy-protective, or in other words, this is a case of Privacy by Design.[143]

Conclusion

We increasingly are becoming bits in datasets, beyond our reach. These datasets facilitate big data research, which has huge potential, but also raises privacy concerns. The case of big medical data sharpens this conflict. Big medical data research, if properly done, can save lives, but it can be detrimental to privacy and the dignity of data subjects, who lose control over their data.

A survey of the emerging big medical data scene indicated that researchers often take what I have called a *forest view*: they point to the retrospective nature of their studies and argue that because they were based on anonymous data, traditional data protection principles were irrelevant. The result is that patients become data subjects, retrospectively and without their consent. This is in stark contrast to the traditional mode of medical research, in which patients are asked to provide informed consent to participate in a research, thus becoming human subjects. Big medical data skips this step.

In order to enable data subjects to exercise some control over their personal data, we should suspend the urge to apply the forest view. Instead, we should see the trees first: we should adopt a process-based approach to informational privacy, following data step-by-step. This process-based approach reflects the logic of FIPs, which reflect the understanding of privacy as a matter of one's control over her personal data. Applying the process-based approach to big medical data means that patients should consent to the very first step in using their data: its anonymization for the purpose of research.

At this point in time, as big medical data research is making its first steps, seeing the trees first and the forest later is still achievable. National schemes such as the British CPRD that are carefully designed to enable pre-anonymization control reflect this logic. Technological privacy innovations can also achieve post-anonymization protections.[144] Sharona Hoffman offers a set of technological policies and legal recommendations that can better

[143] See also Gellman's suggestion to design a technological-legal solution for sharing de-identified data. *See* Robert Gellman, *The Deidentification Dilemma: A Legislative and Contractual Proposal*, 21 Fordham Intell. Prop., Media & Ent. L.J. 33 (2010).

[144] *E.g.*, Eric Verheul et al., Polymorphic Encryption and Pseudonymisation for Personalized Healthcare (2016), https://eprint.iacr.org/2016/411.pdf.

improve the quality of research and protect privacy.[145] In the meantime, there is an urgent need to (re)design the ethics of big data research in general, specifically big medical data. Institutionally, given the global diversity of approaches on the matter, academic journals are a convenient bottleneck to promote the adoption and enforcement of such ethical norms.

The discussion offered in this Article focused on medical data, but the process-based approach to informational privacy is applicable to other big data contexts, be it consumer habits offline and online, locational data, or any other personal data that can be gathered from numerous participants and processed in bulk. While each context has its unique features, human dignity remains the same, across the board.

[145] Sharona Hoffman, Electronic Health Records and Medical Big Data: Law and Policy (2016).