Introduction

I. Glenn Cohen, Holly Fernandez Lynch, Effy Vayena, and Urs Gasser

When data from all aspects of our lives can be relevant to our health – from our habits at the grocery store, to our Google searches, to our FitBit data, to our medical records – can we really differentiate between Big Data and *health* Big Data? Will health Big Data be used for good, for example, to improve drug safety, or ill, for example, for insurance discrimination? Will it disrupt healthcare (and the healthcare system) as we know it? Will it be possible to protect our health privacy? What barriers will there be to collecting and using health Big Data? What role will the law play, and what ethical concerns may arise? These questions, and many others, are at the heart of this book.

"Big data" is a term that has been used pervasively by the media and the lay public in the last several years. While many definitions are possible, the common denominator seems to include the "three V's" – volume (vast amounts of data), variety (significant heterogeneity in the type of data available), and velocity (the speed at which a data scientist or user can access and analyze the data). Some would add a fourth "V" of value, the idea that Big Data would allow us to improve healthcare. Defined as such, healthcare has become one of the key emerging-use cases for Big Data. For example, Fitbit and Apple's ResearchKit can provide researchers with access to vast stores of biometric data on users from which to test hypotheses on nutrition, fitness, disease progression, treatment success, and the like. The Centers for Medicare and Medicaid Services (CMS) have vast stores of billing data that can be mined to promote high-value care and prevent fraud; the same is true of private health insurers. And hospitals have attempted to reduce readmission rates by targeting patients who predictive algorithms indicate are at highest risk based on analysis of available data collected from existing patient records. Underlying these and many other potential uses, however, are a series of legal and ethical challenges relating to, among other things, privacy, discrimination, intellectual property, tort, and informed consent, as well as research and clinical ethics.

The contributions in this book examine the promise and perils of Big Data and related technological advancements in the health context from a number of analytically distinct but interrelated perspectives. In so doing, they follow a familiar pattern in assessing novel technologies and their impact on society by addressing three core themes.

The first theme involves development of a rich phenomenological understanding of the new technologies – here Big Data and to some extent the Internet of Things – and their implications for health and society at large: what changes, and what remains the same? What emerges from the nuanced accounts compiled in this volume is a picture of a world in which human lives, from media and communications to education and finance, are increasingly entangled in the data about them. Health and biomedicine are particularly affected by the rise of Big Data, with prominent examples found in clinical care, laboratories, genomic sequencing, and the wider range of genomics research.[1] Most notably, researchers predict a coming genomic data flood.[2] Due to the falling costs of genomic sequencing and an emphasis on genomic data for clinical and research applications, it is estimated that by 2025, between 100 million and 1 billion human genomes will be sequenced,[3] pushing data generation into the exabyte (one billion gigabyte) scale. Advances in bioinformatics and analytics are leveraging personal data to further health and biomedical knowledge and applications. New machine learning techniques, for instance, are now being used to analyze Big Data and help doctors provide diagnosis and treatment to patients.[4]

Health-related data are increasingly being derived from nonbiomedical sources as well. Data from online purchases reveal preferences, opinions, and health statuses, and Facebook "likes" can, with surprising accuracy, predict one's sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender.[5] Retail purchases have been used to predict whether particular customers are pregnant and therefore likely to use coupons for items related to pregnancy.[6] Users of social networks such as Facebook, Twitter, and PatientsLikeMe (a website where patients can "share their health data to track their progress, help others, and change medicine for

[1] Nuffield Council on Bioethics, The Collection, Linking and Use of Data in Biomedical Research and Health Care: Ethical Issues 4–18 (London, 2015).
[2] Zachary D. Stephens et al., Big Data: Astronomical or Genomical?, 13(7) *PLoS Biol.* e1002195 (2015).
[3] Erika Check Hayden, Genome Researchers Raise Alarm over Big Data, *Nature* 312, 312–14 (2015).
[4] See, e.g., Ariana Eunjung Cha, Watson's Next Feat? Taking on Cancer: IBM's Computer Brain Is Training alongside Doctors To Do What They Can't, *Washington Post*, June 27, 2015.
[5] Michal Kosinski et al., Private Traits and Attributes Are Predictable from Digital Records of Human Behavior, 110(15) *Proc. Natl. Acad. Sci. USA* 5802, 5802–5 (2013).
[6] Charles Duhigg, How Companies Know Your Secrets, *New York Times*, February 16, 2012.

good")[7] often reveal health-related information directly,[8] making it possible for their postings to be mined for research purposes.[9] These data can convey sensitive information such as whether an individual is experiencing symptoms associated with a medical condition. Recent pharmacovigilance studies have shown that Twitter posts containing a reference to a medical product can be used to identify adverse events related to a large number of conditions.[10] Other studies have shown that smartphones equipped with sensors allowing the capture of fine-grained geolocation and usage data can be used to detect depression symptoms and potentially with more accuracy than is possible using standard questionnaires. In addition, mobile phone data have been used for contact tracing and other public health surveillance activities, detecting human mobility in affected regions during the Ebola crisis, thus illustrating the value of such data in an infectious disease pandemic or other public health emergency.

At this point, a second core theme of this book emerges, addressing the many normative implications of these new technologies – how to evaluate the various shifts triggered by technological advancements and form new societal consensus around it? In this respect, this book's chapters demonstrate that while the highly penetrative power of Big Data analysis is revealing sought-after patterns in health and biomedicine, it is also challenging traditional approaches, prevailing social norms, and existing regulatory schemes with respect to autonomy, privacy, identity, and other values. Ethical codes and regulations dictate how biomedical research should be conducted when it involves human subjects, their samples, and their data. However, the systems currently in place, as several authors in this volume argue, do not necessarily address many new Big Data activities. For example, data generated on social media platforms are open to a broad range of uses authorized by the terms of service, but it is not clear that users are aware that their postings are often used for research purposes or that they are in agreement with the platform providers and the researchers regarding these uses of their data. There is a fundamental "impossibility of certainty concerning future uses of data" inherent to Big Data because its value largely stems from uses and insights unanticipated at the time of data generation. Hence subjects cannot be "informed" (in the sense contemplated by traditional informed-consent regimes) regarding future and often-unrelated

[7] About Us, PatientsLikeMe, available at www.patientslikeme.com/about [https://perma.cc/8D9L-DJKV] (last visited May 19, 2017).

[8] Susanne Fox, Pew Research Center, The Social Life of Health Information (2011), available at www.pewinternet.org/2011/05/12/the-social-life-of-health-information-2011 [https://perma.cc/VXU7-NM3E] (last visited May 19, 2017).

[9] Kate L. Mandeville et al., Using Social Networking Sites for Communicable Disease Control: Innovative Contact Tracing or Breach of Confidentiality?, 7(1) *Public Health Ethics* 47, 47–50 (2014); Effy Vayena et al., Ethical Challenges of Big Data in Public Health, 11(2) *PLoS Comp. Biol*. e1003904 (2015).

[10] Clark C. Freifeld et al., Digital Drug Safety Surveillance: Monitoring Pharmaceutical Products on Twitter, 37(5) *Drug Safety* 343, 343–50 (2014).

investigations based on shared, aggregated, and reused data from these services. Users often blithely agree to various terms of service or privacy policies without even reading them and cannot predict how the content of their future tweets, for example, will be used and by whom. Information asymmetries between researchers or institutions and subjects are amplified by the ambiguity of legal obligations and ethical practices for researchers using commercial Big Data sources for health-related purposes. Moreover, regulation and ethical guidelines may set requirements that are impossible to meet in the new paradigm of Big Data. An illustrative example discussed in several chapters is that of black-box medicine. These and related examples explored in greater depth in this volume demonstrate the difficult normative questions, including value conflicts, at the intersection of technology, law and ethics, social norms, and market forces that need to be addressed as Big Data is embraced in the health context.

Against this normative backdrop, a third cross-cutting theme emerges: what are the best available approaches and instruments to address the challenges and also embrace the opportunities afforded by new technological capabilities such as Big Data and the Internet of Things? The chapters that follow make clear that a series of approaches and "tools" is available as society responds to and otherwise interacts with the promises and challenges of emerging technologies in health. The set of available instruments ranges from technical to ethical approaches, mirroring the various modes of regulation that have become the standard repertoire when governing emerging technologies in the digitally connected environment.

Solutions to the Big Data challenges just highlighted are explored in detail across this volume's chapters. The solutions can themselves be divided into a few categories. They include market-based approaches, where different ethical and privacy standards compete for consumers in a marketplace. The idea of third-party auditing mechanisms available to users to audit the collection and use of their health data is an example of an approach discussed in this book that combines market-based mechanisms with technical solutions. Similarly, the proposal of implementing a middleware solution that serves as a gateway and point of control between patients and the end points of the data analysis is an illustration of an innovative solution space vis-à-vis Big Data, the Internet of Things, and related technologies. Many chapters in this book identify such novel approaches at both the conceptual and practical levels, not with the goal to replace traditional safeguards aimed at protecting the rights of individuals but to complement and bolster them in light of the phenomenological and normative challenges mentioned earlier. Reading this book as a whole suggests that no single approach or instrument is likely to be a "silver bullet" that solves the myriad challenges or is sufficient to harness the full benefits of the rapidly evolving digital technologies in the health sector. Rather, blended approaches that combine different instruments available in the "toolbox" seem most promising when dealing with both the challenges and opportunities of Big Data and related technologies, as many of the chapters suggest. At least in this respect, the

solution space that emerges from the contributions in this book with its specific focus on health is consistent with an increasingly robust body of knowledge about the promise (and limits) of multimodal and multistakeholder governance of digital technologies more broadly.

Unsurprisingly, legal and regulatory responses to the challenges presented by Big Data and related technologies feature prominently among the chapters in this volume. From a cross-sectional perspective and at a conceptual level, two observations seem particularly noteworthy: (1) the different role law and regulation can (and should) play vis-à-vis the phenomenological changes described earlier in this Introduction and (2) the observation of a familiar set of patterns in how the interaction between law/regulation and technology unfolds in the health and Big Data discourse. With respect to the different functions that law and regulation can play, several chapters highlight the familiar (and often dominant) role of legal norms as a constraint on behavior. For instance, the requirement of informed consent – and what it means under the new technological conditions that set the context of this book – in order to protect the privacy and autonomy of individuals is a much-debated instance in which law clearly has a constraining function. Other chapters indicate a second way in which legal and regulatory approaches might contribute to problem solving by leveling the playing field among actors, for instance, by establishing more comprehensive accountability and oversight schemes that overcome traditional distinctions between regulated and nonregulated actors in a world of blurring lines of what constitutes health data and what does not. Some contributions make visible, or at least allude to, a third role that law and regulation can play: the role of an enabler of technological advancements that benefit the health of individuals and society at large. This enabling function of law finds its clearest expression in the context of the discussion of intellectual property rights and health Big Data and the question of appropriate incentive structures in order to promote the creation and access of beneficial big data frameworks. It also becomes visible where contractual mechanisms are used to enable new types of choice architectures for individuals through the creation of marketplaces for standards or trusted third-party services.

In terms of response modes, finally, the contributions in this volume reveal a pattern that is familiar from other fields where the legal system has to deal with technological developments that have disruptive effects. The default mode is what is known as *subsumption*, that is, the application of old rules to new phenomena. For instance, the discussion of whether new types of actors that are functionally in the health data business should fall under existing regulatory frameworks is such a question of subsumption. The debate about (existing) informed-consent requirements and how they can be applied in a Big Data environment with its new complexities is a good example of an attempt to use existing approaches to deal with a new phenomenon. As many chapters herein demonstrate, such a subsumption approach quickly reaches its limits when confronted with the deeper-layered structural changes in the health technology ecosystem diagnosed in this volume.

In such cases, the legal and regulatory system is itself forced to innovate, either by making gradual adjustments in the sense of evolution or engaging in more radical, paradigm-shifting innovations. An example of gradual adjustment is the updating of existing legislation (including by court interpretation) – such as the False Claims Act – or regulatory processes – such as the Sentinel System used by the Food and Drug Administration (FDA) – to address the unique challenges posed by Big Data and related technologies in the health context. The idea to extend privacy rights to groups that are otherwise inadequately shielded from harms that might result from biomedical Big Data might be seen as an illustration of an innovation within the legal system that goes beyond a gradual adjustment or "update" of existing norms and procedures.

Taken together, this volume provides a sense of the magnitude of the changes at the phenomenological and normative levels and outlines some of the instruments and strategies available when dealing with the rapidly evolving landscape of cutting-edge technologies such as Big Data, the Internet of Things, and artificial intelligence (AI), among others. To make things even more complex, the three challenges addressed in this book and outlined in this Introduction – phenomenological, normative, and designing solutions – cannot be addressed sequentially. In the case of Big Data in health, one needs to simultaneously gain a deeper understanding of the depth of change that has come about in the way we practice medicine and biomedical research, to explore and evaluate the normative pressure points and conflicts, and to develop, test, and revise the necessary legal, technological, incentive-based, and other tools that can help us deal with the challenges. This volume seeks to inform this debate but also demonstrates significant knowledge gaps and reveals remaining degrees of normative uncertainty in terms of the net outcome of the brave new world of health and digital technology. In this sense, it should be read as an open invitation for further research, collaboration, and discourse across a broad range of stakeholders to work toward a use of technology that benefits individual and public health.

This book is divided into seven parts. Part I, introduced by Urs Gasser, describes the ways in which Big Data is shifting existing paradigms of health law and bioethics. The contributions identify and discuss a series of seismic shifts in health-related data collection, aggregation, and use and engage in a thick analysis of their implications, including challenges related to traditional legal definitions and concepts, novel threats to privacy and autonomy, new power asymmetries, and potential ramifications from an epistemological perspective.

Barbara J. Evans in Chapter 1, "Big Data and Individual Autonomy in a Crowd," identifies privacy barriers, such as the risk of reidentification, and related normative challenges that prevent access to large and inclusive data resources, which would be required in order to harness the full benefits of Big Data. Evans' chapter also explores the question of how a new common purpose can emerge in bioethical and regulatory environments that currently emphasize individual autonomy. As a partial

solution, she proposes the concept of consumer-driven commons as a way to empower individuals to protect themselves against research-related risks and to engage, collectively, in civic solidarity.

Privacy also plays a key role in Chapter 2, "Big Data's Epistemology and Its Implications for Precision Medicine and Privacy," in which Jeffrey M. Skopek examines from a broader epistemological perspective the extent to which the nature of scientific knowledge and inquiry is likely to change in the age of Big Data. After a critical examination of the often-discussed shifts from theory to data and from causation to correlation, Skopek explores how a separate shift from explanation to prediction may play out in the field of precision medicines and predicts a series of broad legal and ethical challenges in areas such as intellectual property, torts, and privacy as a result of this shift. Against this backdrop, he focuses on the privacy implications of predictive analytics, arguing that the threat to privacy posed by Big Data is more limited than has been widely thought.

Along similar thematic lines but with partially different results, in Chapter 3, "Correlation versus Causation in Health-Related Big Data Analysis: The Role of Reason and Regulation," Tal Z. Zarsky examines the epistemological implications of the new ways in which data are collected, aggregated, analyzed, and used in today's data-rich contexts that affect health. After analyzing the various forces at play, Zarsky engages in a review of the correlation versus causation debate and cautions, with some important exceptions, against practices in the medical and health context that rely increasingly and at times exclusively on mere correlation and ignore important questions of causation and mechanisms. He also examines the role law and regulation can and, under qualified circumstances, should play in setting requirements as to the appropriate way in which data must be analyzed prior to their use.

Finally, in Chapter 4, "Big Data and Regulatory Arbitrage in Healthcare," Nicolas P. Terry focuses on a series of scenarios concerning the misuse of healthcare data by the Big Data industry. With new commercial actors such as data brokers entering the arena, Terry diagnoses the problem that the current structure of healthcare data protection is insufficient and leaves individuals vulnerable to a series of serious risks, including discriminatory practices and privacy invasions. Terry calls for comprehensive federal legislation to create a level playing field in terms of protection from data-processing actors within and outside the traditional context of healthcare and to fix the current problem of regulatory arbitrage.

Next, in Part II, introduced by I. Glenn Cohen, the focus shifts to overcoming the potential downsides of health Big Data. The contributions in this part lie on a continuum, Cohen argues, between assimilation and disruption. Placement on the continuum depends on how easily the authors see Big Data in healthcare as fitting into existing legal paradigms versus breaking them open and requiring something new in their place.

In Chapter 5, "The Future of Pharmacovigilance: Big Data and the False Claims Act," Efthimios Parasidis looks at the role for Big Data in enforcement of the False

Claims Act, the federal government's main tool for combating healthcare fraud. He argues that the act should apply to fraud relating to Big Data manipulation – structuring data analysis to obscure information that might cause the submission of fewer claims. Among the intriguing questions Parasidis examines are "If Big Data analysis reveals questions about safety or efficacy, does a pharmaceutical company have an obligation to report the information or conduct further research?" and "What is the outcome if a learning algorithm is the sole entity to examine the new information?" Along the way, he also examines a role for Big Data in FDA pharmacovigilance.

In Chapter 6, "Big Data's New Discrimination Threats: Amending the Americans with Disabilities Act to Cover Discrimination Based on Data-Driven Predictions of Future Disease," Sharona Hoffman examines the threat posed by healthcare Big Data to protections from employment (and other forms) of discrimination. She focuses on one of the United States' primary antidiscrimination statutes, the Americans with Disabilities Act (ADA), and shows that it offers no protection to a group imperiled by improved Big Data analysis: "people who are currently healthy but are perceived as being at high risk of becoming sick in the future." Hoffman develops a proposal to expand the ADA to "prohibit discrimination based on predictions of future physical or mental impairment" and "require covered entities to disclose in writing their use of Big Data or other nontraditional means to obtain health-related information."

Chapter 7, "Who's Left Out of Big Data? How Big Data Collection, Analysis, and Use Neglect Populations Most in Need of Medical and Public Health Research and Interventions," by Sarah E. Malanga, Jonathan D. Loe, Christopher T. Robertson, and Kenneth S. Ramos, is a meditation on a major drawback of current healthcare Big Data sets: their failure to include marginalized populations such as racial minorities, people with low socioeconomic status, and immigrants. The authors examine the way these populations are also the ones that face the most acute health disparities and the way in which these gaps imperil moves toward precision medicine. Finally, they consider several ways to remedy these gaps. For example, the FDA, as part of the pharmaceutical and other product approval process, could require data to come from a diverse, inclusive pool of patients; regulation across several agencies could be "implemented, or strengthened, to push users of Big Data to acknowledge and reconcile the possibility for skewed data due to under- and overrepresentation, biases within algorithms, and overreliance on the findings of Big Data." While the authors are candid that these steps will not completely eradicate health disparities, they defend them as important steps for progress.

Finally, in Chapter 8, "Potential Roadblocks in Healthcare Big Data Collection: *Gobeille v. Liberty Mutual*, ERISA, and All-Payer Claims Databases," Carmel Shachar, Aaron S. Kesselheim, Gregory Curfman, and Ameet Sarpatwari focus on the Supreme Court's *Gobeille* decision, which held that the Employee Retirement Income Security Act (ERISA) preempted Vermont's attempt to gather healthcare

data from plans governed by that act, prohibiting states from requiring self-insured employer-sponsored health plans to report data to their All-Payer Claims Database (APCD). The authors critique the decision and lament its effects on attempts to curb healthcare spending and to perform important health services research. At the end of this chapter, the authors explore several potential workarounds that would help to reduce the negative effects of the decision.

Part III, introduced by Nathan Cortez, addresses the Internet of Things and health Big Data. The chapters in this part identify a long list of concerns associated with these emerging technologies and corresponding practices, including interoperability, privacy, safety, transparency, and accountability challenges, among others. Both chapters also propose novel ways in which some of these concerns can be addressed using technological means and market mechanisms that supplement the work (and limited effectiveness) of the familiar regulatory bodies.

After analyzing the broad range of risks associated with mobile health technology such as wearables, smartphones, and other sensing devices that intend to track and assess the health of patients, Dov Greenbaum in Chapter 9, "Avoiding Overregulation in the Medical Internet of Things," proposes the introduction of a third-party clearinghouse, which would act as a transparent and accountable intermediary between Internet of Things devices and end users. The technological role of a third-party clearinghouse would include standardizing health data across multiple platforms and devices, ensuring both interoperability and usability. The regulatory role would be to embed more robust privacy and data security standards into spaces that fall outside the purview of federal privacy and security rules deriving from the Health Insurance Portability and Accountability Act (HIPAA), among others.

Similar in spirit, Marcus Comiter, in Chapter 10, "Data Policy for Internet of Things Healthcare Devices: Aligning Patient, Industry, and Privacy Goals in the Age of Big Data," introduces the idea of nongovernmental *third-party data auditors* (TPDAs) that could be hired or otherwise engaged by individuals to audit the use of their data by corporations, providers, data brokers, and others. Like Greenbaum's clearinghouses, Comiter's TPDAs would serve both a technological and a regulatory (or governance) function, specified in greater detail in the chapter. Comiter sees the primary benefit of independent TPDAs in an increased level of transparency and accountability in health Big Data, which, in turn, would deter misuses and promote consumer trust.

In Part IV, introduced by Effy Vayena, the contributors turn to focus on protecting health privacy in a world of Big Data. The chapters taken together provide a good illustration of the mixed approaches needed to address the privacy challenge. Some involve expanding the scope of existing regulation, such as informed consent; others entail the development of entirely new ethical and legal mechanisms.

In Chapter 11, "Thought Leader Perspectives on Risks in Precision Medicine Research," Laura M. Beskow, Catherine M. Hammack, Kathleen M. Brelsford, and Kevin C. McKenna report on empirical data that revealed shifts in the perception of

harm with uses of biomedical Big Data, specifically an increased acknowledgment of group harm, which is little understood and even less addressed in current protection mechanisms. In the Big Data environment, risks are exacerbated by at least two factors: (1) the never-ending uses of multiple data sets and (2) the difficulty in predicting how data will be used and what they will yield. The uncertainty about what will emerge makes it virtually impossible to anticipate all potential risks a priori.

In his effort to address the issue of group harms, Brent Mittelstadt has formulated a particular right to privacy for algorithmically ad hoc groups: the right to "inviolate personality." In Chapter 12, "From Individual to Group Privacy in Biomedical Big Data," Mittelstadt proposes group privacy as a third interest to balance alongside individual privacy and social, commercial, and epistemic benefits when assessing the ethical acceptability of Big Data analytics. He examines two implementation models for group privacy, wherein group privacy can ground both anticipatory restrictions on processing and reactive redress for groups whose privacy has been violated.

Next, in Chapter 13, "Big Data and Informed Consent: The Case of Estimated Data," Donna M. Gitter addresses regulatory change and the increasing use in biomedical research of estimated data. To protect the privacy and autonomy of these individuals, Gitter argues that they also deserve the protection of the law of informed consent. She also explores the importance of the right not to know of the genetic incidental findings discovered by researchers as another component of the privacy right.

Part V, introduced by Holly Fernandez Lynch, addresses key questions regarding the regulatory and ethical requirements for research use of Big Data, either health research or health data. Contributors to this part agree that consent to such use is often unnecessary as long as appropriate protections of data sources are in place. What those protections ought to look like and how they ought to be implemented and enforced are where the real action is.

In Chapter 14, "Is There a Duty to Share Healthcare Data?," I. Glenn Cohen sketches an argument that there is, rooted in two distinct rationales. His first argument is that healthcare data should not be viewed as a patient's property, given that the patient has not added his or her labor or "sweat equity" to make the data valuable, and therefore, the patient has little claim to profit from or control them. Cohen's second argument, which may be viewed as either buttressing or alternative to the first, is that there is a duty based in reciprocity to contribute healthcare data to systems that generate benefits – health or otherwise – for all of us. Given these rationales, Cohen argues that consent is not required for the collection and use of healthcare data, although public education about these activities would be appropriate.

In Chapter 15, "Societal Lapses in Protecting Individual Privacy, the Common Rule, and Big Data Health Research," Laura Odwazny addresses application of the regulations governing research with human subjects to Big Data health research.