Current regulations will not protect patient privacy in the age of machine learning

Ashwin Narayan^{1,*}

Edited by Jordan Harrod and Anthony Tabet

HIGHLIGHTS

- Tools from machine learning are showing great promise in solving challenging problems in healthcare by leveraging their ability to parse and integrate massive datasets
- Currently, this progress comes at the expense of privacy, since the current state of privacy legislation is not equipped to handle the novel challenges presented by widespread adoption of ML
- Changing the current U.S. “reasonable expectation of privacy” doctrine and ending the restriction of HIPAA to only the healthcare industry are two specific steps local or federal governments can take to protect privacy in the age of ML

Machine learning (ML) has shown great promise in advancing health outcomes by parsing ever more effectively through massive clinical and genomic datasets. These advances are tempered by fears that they come at the cost of privacy. Since data relating to health are particularly sensitive because of immutability and comprehensiveness, these privacy concerns must be seriously addressed. We consider examples (the Golden State Killer, the Personal Genome Project, and the rise of wearable fitness trackers) where the tension between technological progress and lost privacy is already apparent. We discuss, in light of ML capabilities, the current state of privacy regulation in healthcare. We note the Constitutional right to privacy does not yet in general protect voluntary disclosures of data; HIPAA, the current law regulating healthcare data in the US, does not apply to the burgeoning field of healthcare-adjacent companies and organizations collecting health data; and access controls remain subject to re-identification attacks. We then discuss the active research in algorithmic paradigms for privacy, highlighting their promise but also their limitations. In order to encourage technological progress, reframing privacy for the age of ML might involve extending the Constitutional right to privacy, extending the applicability of HIPAA, and/or enforcing transparent privacy policies.

The current state of healthcare is being rapidly disrupted by the advent of sophisticated computational machine learning (ML) tools and massive amounts of data. These tools take advantage of advances both in algorithms and in the availability of data to address health issues in ways that were impossible before. From massive consortium-level genomic studies that crunch through huge datasets to find genetic signals for rare diseases to individual fitness trackers that warn of cardiac arrhythmias or promote better sleep or even just encourage exercise, there is certainly much to find encouraging.

But as with any disruption, there are important fears that need to be addressed, and the current regulatory environment has not caught up to the technology to properly address these concerns. The implications to personal *privacy* of these tools is forefront among the fears. Health data are particularly sensitive, and therefore people — from the users of the fitness trackers to the contributors to big consortia — are strongly invested in its protection. The key privacy-related questions that arise are (i) who has access to the data; (ii) what are those with access allowed to do with the data; (iii) what information can be learned from the data; and (iv) how should they be stored. We contend that current privacy regulations do not offer satisfactory answers to *any* of these questions, as is shown by the following case studies that highlight the tension between the benefits of Big Data healthcare and fears of lost privacy.

Golden State Killer: In a case that reached its climax fittingly on “National DNA Day”, the capture of the suspected Golden State Killer in 2018 brought the tradeoff between the promises and the risks to privacy of genomic data into the societal spotlight.

The Golden State Killer was a notorious serial killer and rapist in the 1970s and 1980s who committed at least thirteen murders, fifty rapes, and hundreds of burglaries [1]. His capture, even after so many years, is certainly a victory for justice and vindicates at least the *efficacy* of using DNA to solve yet unsolved crimes. The method by which he was apprehended, however, raises important questions concerning the privacy of genealogical data. Homicide investigators matched DNA found at crime scenes exhaustively to open source genealogy databases (which record ancestry using DNA data) like GEDMatch, identifying not the killer himself

¹Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA

*Email: ashwin@mit.edu

AN is a PhD candidate advised by Bonnie Berger, whose work in algorithmic privacy is cited in this work.

© 2020 The Author(s)

but rather his *relatives*. Painstaking construction of family trees based on these matches led to the identification of the suspect who was apprehended.

The privacy-motivated questions above are thrown into sharp relief here. Should law enforcement be allowed to search these databases without a warrant? How can one protect their genomic information when their privacy depends not only on their own decisions but also on those of (potentially distant) family members?

Personal Genomes Project: Even for those willing to accept surveillance of health data by law enforcement in the name of security, issues of access to large repositories of these data by private actors cause concern. For example, in order to fully realize the promise of **personalized medicine**, i.e. healthcare tailored to a patient's specific genome, in 2005 the ambitious Personal Genome Project (PGP) was set up, aiming to release to the public the genetic information of 100,000 volunteers. The radical notion underlying the project was "open consent", whereby genetic data and as much additional information as desired by the volunteers would be completely openly released online with the informed consent of the study participants [2]. Participants in the project would thus be aware of and in control of what information would be released. The PGP now contains over two-thousand genomes and has spread to beyond the United States, and the repository of genetic data has been extremely useful to researchers studying personalized medicine.

However, the PGP repository was quickly found to be vulnerable to **re-identification** attacks, i.e. the uncovering of anonymous information in a dataset. Sweeney, Abu, and Winn (2013) [3] were able to recover the names of those on the project with 84 to 97 percent accuracy by using other reported information matched with publicly available records. This calls into question the very notion of open consent: participants who did not consent to revealing their names but *did* provide other demographic information were actually *unwittingly* revealing their names, through relatively straightforward matching with other datasets. The crucial question, "What information can be learned?" thus effectively undermines the open consent paradigm because, as we will see, one of the key features of new algorithms for Big Data is their ability to *integrate* data from many sources to draw unforeseen conclusions.

Other large consortia (e.g. the Thousand Genome Project, the UKBioBank, etc.) have now erred on the side of caution, veering away from openness and instituting access controls on their data. As we will discuss below, even these controls are not sufficient.

Fitness trackers: Many companies are also now amassing their own troves of healthcare data in parallel with and far more opaquely than these research consortia. While this is perhaps unsurprising for those companies embedded in the healthcare industry and which face relatively stringent regulations regarding their use of data (e.g. health insurers), we concern ourselves primarily with those that occupy a gray area: not subject to those stringent regulations but

still in possession of large quantities of sensitive data. For example, consider the increasing ubiquity of "wearable" health devices, like FitBits, that track various health outcomes. These fitness trackers offer benefits to their wearers: they can track sleep, step counts, exercise, and even spot suspicious cardiac patterns. Perhaps even more tangibly, some health insurance companies have started offering discounts to customers for using health trackers to stay healthy [4]. As with the previous examples, these benefits are accompanied by rising concerns with the stewardship of the health data collected. Recent research [4] has found that many popular devices have extremely lax security standards for the data they collect. Other work has found that privacy leakages in the Bluetooth communication between these trackers and a smartphone can track a person's movements (and identify them by gait) [5], or, related to the above discussion of re-identification, can leverage information posted on social media from these trackers with other public information to find home addresses [6]. The relatively lax regulatory environment that governs these trackers does not help mitigate these risks. In fact, the lack of regulation means that the answers to our framing questions are known *only* to the companies themselves.

Contact Tracing: The global COVID-19 pandemic, which is still unchecked as this paper is being written [7], has given new urgency and tangibility to the intersection between health data and privacy. While much of the world economy is stagnating due to the enforcement of lockdowns and "social distancing", policymakers and health experts are desperate to allow public life to resume safely. A powerful tool in the epidemiological arsenal is **contact tracing**, where, when an individual is found to be infected, all of the people they have come in contact with are immediately tested so the infection cannot spread far [8]. Even when contact tracing was restricted to the early stages of an outbreak (because contacts had to be painstakingly traced manually), privacy concerns abounded [9]. But now, Big Data means those infected can be tracked algorithmically. The benefits of these tools being scaled up are clear: if widely taken up, people could go about their lives confident that those infected will learn quickly and can isolate. Several countries and some states, along with Apple and Google, have begun rolling out smartphone apps¹. These tools range from **decentralized**, where device IDs are anonymized and location not collected, to highly **centralized** where location data are available for use by public health officials [10]. Despite a laudable focus on "privacy-by-design" by companies, organizations, and governments involved in developing these applications, many privacy experts are sounding the alarm, worried that safeguards protecting such sensitive data that are lowered during an emergency may stay that way. Collected on apps often made and moderated by private companies, for use by governments to enforce regulations, and potentially a an extremely useful trove for public health researchers, the prospect of algorithmic contact tracing combines all of the questions raised in the above case studies with unprecedented stakes.

¹<https://rb.gy/41mvxy>

These case studies explore the landscape of concerns raised by the interaction between regulation of healthcare data and the technical capabilities of new algorithms. To better understand the gaps in regulation implied by the above examples, we present here a survey of both the uses of machine learning in healthcare and the current state of privacy regulations in this space and close with a discussion of possible changes to the regulatory environment to address the issues discussed.

Machine learning in healthcare

Since the advent of algorithms called “neural networks” that could quickly and accurately identify images, machine learning has entered common parlance and been touted as a panacea for virtually all computational challenges that researchers face [11]. To those not familiar with the literature, it can be difficult to discern what exactly constitutes an application of ML. Here, we consider mainly the subfield known as **supervised learning** and adopt a simple definition. An ML algorithm is a tool that, given a large quantity of *labeled* data, can accurately assign labels to data that have not been labeled yet².

The combination of massive increases in computational power and vast troves of medical data newly available to researchers has encouraged the field to bring the tools from ML to bear on problems in medicine [12]. In fact, the image classification prowess that popularized ML itself has directly translated to medicine, for example in identifying tumors for radiologists [13]. But the framework of supervised learning is easily extended to problems beyond image classification: determining which small molecules (potential drugs) are likely to be effective for binding to proteins [14]; which mutations in genomes are likely to flag potential diseases [15]; or combat unconscious bias in care for patients in chronic pain [16].

A particularly promising application of ML is in the field of **genomics**, i.e. the analysis of DNA. The success of the Human Genome Project in spurring the development of so-called “next-generation” sequencing technologies has provided the massive amounts of raw material that ML algorithms require to draw insights. The PGP and other consortia like the UK Biobank aim to make genomic data available to ML practitioners because of their benefits to healthcare research. But in the case of the Golden State Killer, it was tools from ML that allowed for the generation of open source genealogy databases like GEDMatch and for law-enforcement to feasibly align their DNA fragments with all the contents of those databases.

Private companies, universities, and government organizations have all collected massive amounts of this data. ML algorithms are at their strongest when they are able to *combine* data from these various sources and integrate it

to answer questions that each individual dataset would not have answered alone. For example, while contact tracing alone does not require sophisticated ML, these data can be combined with location data, hospital admission data, social networking data, and so on to learn far richer epidemiological models that go beyond standard assumptions of how people interact [17]. ML tools are able to perform this integration because they do not need to be explicitly told the weights or features of each dataset that are important. Using the given labeled data, ML tools can *themselves* learn the salient aspects of each dataset. But as seen in the PGP example above, this combination of datasets also has a privacy cost in allowing re-identification of intentionally hidden data. Since tools from ML are able to endanger privacy, it falls on regulations to protect that privacy.

Privacy in healthcare

Protecting one’s genetic and healthcare data has particularly strong resonance because the information contained is both immutable and comprehensive: theoretically, virtually all the information about a person is contained in their genome, and this genome is, at least for now, impossible to change [18]. In the age of ML algorithms, the danger is amplified by the propensity of these algorithms to effectively integrate various pieces of disparate information [19], each of which, on its own, might not carry much weight.

Because of the many stakeholders with an interest in the health data ecosystem, each with different incentives, even defining privacy is not straightforward. These data, while obviously valuable to their owner, are also of interest to the state, to researchers, and to companies, not even mentioning various nefarious actors. And while the regulatory framework changes depending on the actor, as we discuss below, a key theme that emerges is that the act of *voluntarily* giving up one’s data precludes it from many forms of protection. These repositories of voluntary health data are proliferating: people give their health data to companies like 23AndMe, participate in consortia like the PGP, and track their own data with wearable technologies, even though the privacy implications of these voluntary disclosures are unclear.

Properly protecting privacy is thus a multifaceted goal. Leaks must be prevented and access must be controlled, but, as importantly, the consequences of *voluntary* releases of data need to be understood. Informed consent is a crucial aspect of collecting health data, but we do not understand the potential of ML algorithms well enough for this to be possible.

Constitutional privacy: Unlike in Europe, where the Right to Privacy is enshrined in the European Convention on Human Rights, there is no *explicit* right to privacy enshrined in the US Constitution. Courts have instead consistently upheld an *implicit* right to privacy using the protections guaranteed primarily by the First and Fourth Amendments. This mode of jurisprudence owes to the seminal Warren and Brandeis article “The Right to Privacy” (1890), which connects a right to privacy to the protection against “unreasonable search and

²On the other hand **unsupervised learning** seeks to find structure and patterns in *unlabeled* data

seizure” without “probable cause” guaranteed by the Fourth Amendment³ [20]. Thus, the privacy afforded to Americans hinges on the interpretation of “unreasonable search” and “probable cause”.

As methods of data collection by the state became more and more sophisticated, the parameters and boundaries of this right to privacy have been repeatedly contested, often in precedent-setting Supreme Court cases. Protections were extended from physical objects owned by a petitioner to *any* data where the owner exhibits an “expectation of privacy” that “society is prepared to accept as reasonable” [21]. This “reasonable expectation of privacy” doctrine has been used by the courts to determine what kind of information the government is allowed to collect about a private citizen.

Crucially for healthcare data, the courts have repeatedly found that data given to an organization *voluntarily* do *not* carry a reasonable expectation of privacy [22]. This is known as the **Third-party Doctrine** and is why law enforcement were able to call upon genealogy databases in the aforementioned Golden State Killer case: all the data in those databases were given voluntarily [1]. In the age of ML, this standard is itself unreasonable: consumers and patients have little idea of the ways that the data they give up voluntarily can be used by these sophisticated ML algorithms. The genomes and biodata collected not just by companies like 23 And Me and now Google and Apple, but also by organizations like the NIH, are all potentially searchable because they were given voluntarily. And again, the potential to combine these data with other publicly available records in the ML crucible raises existential fears about the continued existence of a right to privacy for healthcare data. In the recent *United States v. Jones* case, four of the Supreme Court justices themselves threw up their hands, questioning the appropriateness of the Third-party Doctrine for the digital age [23].

HIPAA and HITECH: While Fourth Amendment protections of medical data from searches by the state remain murky, lawmakers have recognized the sensitivity of healthcare data and have passed legislation that regulates access and storage for *non-state* actors, notably the Health Insurance Portability and Accountability Act (HIPAA) of 1996 [24], later augmented by the Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009 [25]. HIPAA is a cornerstone of healthcare data protection in the United States. It regulates the way that protected health information (PHI) can be shared between major players in the healthcare system: insurance companies, healthcare providers, and healthcare clearinghouses. In addition to regulating access to sensitive data, HIPAA provides some standards for responsible stewardship [26]. Since HIPAA’s adoption, these protections have been crucial in easing the interoperability of different parts of the healthcare system while

preserving privacy.

Unfortunately, the power of ML algorithms undermines the privacy protections afforded by these regulations. Crucially, HIPAA sets no restrictions on *de-identified* data, i.e. data after “the removal of specified identifiers of the individual and of the individual’s relatives, household members, and employers” such that “the covered entity has no actual knowledge that the remaining information could be used to identify the individual” [24]. Especially when combined with other sources, this de-identified data can be re-identified, thereby circumventing the spirit of the HIPAA regulations [3].

HIPAA also does not cover data provided to entities not part of the healthcare system, and these non-covered providers are becoming more significant. Private companies, selling wearable health monitors or gene-sequencing kits, are collecting data that are protected only by their opaque privacy policies [18]. And, as yet, it is unclear the HIPAA-burdens that will be placed on those involved in creating and maintaining contact-tracing apps, which, since users will have to download these apps, will certainly be collecting their data voluntarily. Data are also voluntarily given to research organizations, like the aforementioned Thousand Genomes Project or the PGP. Since there are no federal regulations stipulating the privacy standards these consortia need to uphold, their protocols for privacy are not standardized and cannot easily be evaluated. These major lapses in HIPAA thus still compromise privacy, motivating *access controls* from consortia that hold health data.

Access controls: A tempting solution for addressing the gaps left by HIPAA and HITECH is massively restricting access. However, openness has benefits as well because of the aforementioned promises of ML. To protect patient data and to encourage participation in research projects, consortia often restrict access to the data by establishing some mechanism of trust in those who use them, but this is in tension with the goal of disseminating these data for broad use. For example, the National Institutes of Health (NIH) had instituted a Genome Data Sharing Policy, which stipulated that NIH-funded research needed to be shared; upon recognizing the potential for various re-identification attacks, the NIH was forced to institute access controls under the new database of Genotypes and Phenotypes (dbGaP) [27]. These re-identification methods are effective even when parts of data are hidden or omitted from the dataset since statistical tools can be brought to bear to find correlations between the omitted data and data that are available.

Access controls are an unsatisfying solution for ensuring privacy. Primarily, it is difficult to develop a mechanism for allowing access to all those who would use the data for its intended purpose while rejecting malicious requests. And of course, researchers who gain access can still combine the data with other sources to gain insights into the study participants that might go above and beyond what they consented to. Current regulations do not restrict this integration of data from multiple sources. Thus, these controls

³“The right of the people to be secure in their persons, houses, papers, and effects, against unreasonable searches and seizures, shall not be violated, and no warrants shall issue, but upon probable cause supported by oath or affirmation, and particularly describing the place to be searched, and the persons or things to be seized.”

only serve to stratify researchers into those who are able to meet the thresholds for access and those who cannot, while doing little to protect the privacy of the owners of the data.

Algorithmic privacy: In the absence of more stringent regulations, computer scientists have taken it upon themselves to build privacy directly into their algorithms. Researchers have defined different mathematical notions of privacy in order to develop methods that can *provably* preserve it. These notions build off of Claude Shannon's seminal work in **communication theory** [28], which described information as a mathematical concept, and in cryptography. Thus, from an algorithmist's point of view, a dataset protects privacy if it does not reveal information about a participant in that dataset. Multiple paradigms for algorithmic privacy exist, and while all are promising, all still remain active areas of research.

Differential privacy [29] is one of the most popular paradigms for privacy preservation: a dataset is differentially private if any statistics computed on that dataset would be the same whether any *particular* individual was included or excluded. This helps prevent re-identification attacks when statistics are released from healthcare databases.

While theoretically promising, incorporating differential privacy into healthcare datasets is not a solved problem [27]. In order to guarantee this level of privacy is achieved, noise needs to be added to the results with a higher magnitude than the influence of any particular data point. However, especially for rare genetic variants, adding this amount of noise masks any useful results that the dataset could have provided.

Secret sharing [30] is a privacy protocol that encourages sharing of data currently siloed among different biobanks and consortia. Here, the data are *decentralized*: each data point is chopped up and sent to different, *independent* actors. Crucially, the fragment of the data point sent to an actor gives *no* information about the data point itself. When a researcher wants to perform some analysis on a secret-shared dataset, their query is transformed into operations that each actor can perform on their fragment; the researcher can recombine the transformed fragments from each of the actors, which will be the answer to their query.

In this protocol, neither the researcher nor any of the actors needs to have access to the full dataset; also, recent work [14] has shown that secret sharing can be made practical even for large biological datasets. However, re-identification attacks are still possible, when the results of a query on a secret-shared dataset are combined with other publicly available datasets. In the absence of regulation that forces *all* medically relevant data to uphold some stringent privacy protocol, the availability of public data makes even those datasets with strict privacy protections vulnerable to re-identification.

Homomorphic encryption [31] is an even stronger privacy protocol than secret sharing. Under this paradigm, data are stored *only* in encrypted form. When a researcher queries the database, the query is transformed to a query on the *encrypted* data itself, and the result is returned. Under this

protocol, the data is *never* in plain-text, and so any researcher can query the database and any party can be the steward for the data.

Unfortunately, efficient homomorphic encryption is still far away. While relatively simple operations can be performed in existing homomorphic paradigms, these do not scale well to the complex (and numerous) operations that modern ML queries require.

While research into algorithmic privacy is welcome and promising, until regulations incentivize stewards of data to incorporate these protocols, open-access datasets will remain a weak link for privacy.

Outlook

Disruption has virtually always accompanied technological progress, and the disruption caused by ML in healthcare is no different. Just as it would be a mistake to shun the technological potential completely out of fear for privacy, it would also be undesirable to keep current privacy standards unchanged in the face of this development. Society needs to determine the meaning of privacy in the digital age, and the extent to which this notion of privacy can accommodate the openness necessary for the tools from ML to be widely accessible and successful [32]. We present here some directions that such a process could follow.

A more appropriate expectation of privacy: While the Supreme Court determined in *United States v. Miller* that information given to a third-party voluntarily carried no expectation of privacy, the Court has showed signs of circumscribing this stance in more recent cases. In her concurrence in *United States v. Jones*, Justice Sotomayor famously wrote, "it may be necessary to reconsider the premise that an individual has no reasonable expectation of privacy in information voluntarily disclosed to third parties" [23]. Certainly if, for example, contact tracing apps become widely adopted, the notion that voluntarily given data deserve *no* protection will come under further scrutiny.

The edifice of the Third-party Doctrine crumbled a little bit more in 2018 in *Carpenter v. United States*: the Court ruled that the it was an unreasonable search under the Fourth Amendment for the government to search the petitioner's historical cell-phone records which contained his physical location — despite the fact that these records were technically given by him *voluntarily* to the cell-phone provider [33]. The Court recognized that in the digital world, a cell-phone contract is hardly voluntary in the spirit of the word, while also setting new tests for the reasonableness of trawling historical records.

Extending the applicability of HIPAA: HIPAA currently only applies to healthcare providers, health insurance companies, and clearinghouses. Other private companies that also deal with health data could also be forced to comply. Companies like 23AndMe, open-access repositories of genealogy data like GEDMatch, and makers of fitness trackers like FitBit, for example, do not function as health-providers but certainly hold

sensitive information about their customers. Extending the definition of a “covered provider” in HIPAA to incorporate these groups would force them to treat sensitive data as protected health information, with the restrictions and protections that accompany that designation.

A similar consideration is strengthening the standard of de-identification required for health data to be exempted from the “protected” classification. As discussed earlier, the current standard lets the covered provider off the hook as long as *they* do not think the data can be re-identified. However, the capabilities of ML to combine insights from various disparate sources of data mean that a covered provider's judgment on re-identification risk might not be accurate. Stricter controls, perhaps drawing on the *algorithmic* privacy paradigms discussed earlier could make de-identification more robust.

Transparent privacy policies: Private companies are not subject to Fourth Amendment considerations of privacy; consumer privacy is instead enforced by the Federal Trade Commission (FTC), which ensures that companies abide by their privacy policies [34]. However, the privacy policies of many companies are notoriously dense and difficult to parse for the average consumer. Drabiak (2017) [18] has written an entire paper just parsing the privacy policy of 23AndMe and highlighting the risks to consumers hidden in the dense wording. For consent to be informed and enthusiastic, the consumer must be able to understand exactly what the data they part with can and will be used for. The FTC's guidelines already encourage businesses to make their privacy policies comprehensible⁴, but the standards for comprehensibility and the mechanisms for enforcing this could be strengthened.

There is great scope for harnessing the power of machine learning to further positive health outcomes in a way that protects the privacy of the providers of data. It is imperative that policymakers craft appropriate regulation that will allow ML to thrive while mitigating any risk of harm.

Acknowledgements

The author is grateful to Meet Mehta for providing sources and guidance in understanding jurisprudence and to Bonnie Berger for conversations about algorithmic privacy in computational biology

Citation

Narayan, A. Current regulations will not protect patient privacy in the age of machine learning. *MIT Science Policy Review* 1, 3-9 (2020).

Open Access

This *MIT Science Policy Review* article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to

the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- [1] Dery III, G. M. Can a Distant Relative Allow the Government Access to Your DNA? *Hastings Science and Tech Law Journal* 10, 103–46 (2019).
- [2] Lunshof, J. E., Chadwick, R., Vorhas, D. B. & Church, G. M. From genetic privacy to open consent. *Nature Reviews Genetics* 9, 406–11 (2008). <https://doi.org/10.1038/nrg2360>.
- [3] Sweeney, L., Abu, A. & Winn, J. Identifying Participants in the Personal Genome Project by Name. *arXiv* (2013).
- [4] Fereidooni, H., Frassetto, T., Miettinen, M., Sadeghi, A.-R. & Conti, M. Fitness Trackers: Fit for Health but Unfit for Security and Privacy. In *International Conference on Connected Health: Applications, Systems, and Engineering Technologies*, 19–24 (Institute of Electrical and Electronics Engineers (IEEE), 2017). <https://doi.org/10.1109/chase.2017.20>.
- [5] Das, A. K., Pathak, P. H., Chuah, C.-N. & Mohapatra, P. Uncovering Privacy Leakage in BLE Network Traffic of Wearable Fitness Trackers. In *International Workshop on Mobile Computing Systems and Applications*, 99–104 (ACM, 2016). <https://doi.org/10.1145/2873587.2873594>.
- [6] Aktypi, A., Nurse, J. R. & Goldsmith, M. Unwinding Ariadne's Identity Thread: Privacy Risks with Fitness Trackers and Online Social Networks. In *Multimedia Privacy and Security*, 1–11 (ACM, 2017). <https://doi.org/10.1145/3137616.3137617>.
- [7] Johns Hopkins Coronavirus Resource Center. COVID-19 Dashboard. URL <https://coronavirus.jhu.edu/map.html>.
- [8] Eames, K. T. D. & Keeling, M. J. Contact tracing and disease control. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 270, 2565–2571 (2003). <https://doi.org/10.1098/rspb.2003.2554>.
- [9] Levine, M. L. Contact Tracing for HIV Infection: A Plea for Privacy. *Columbia Human Rights Law Review* 20, 157–202 (1988).
- [10] Cho, H., Ippolito, D. & Yu, Y. W. Contact tracing mobile apps for COVID-19: Privacy considerations and related trade-offs. *arXiv* (2020).
- [11] Lecun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* 521, 436–444 (2015). <https://doi.org/10.1038/nature14539>.
- [12] Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine* 25, 44–56 (2019). <https://doi.org/10.1038/s41591-018-0300-7>.
- [13] Işin, A., Direkoğlu, C. & Şah, M. Review of MRI-based Brain Tumor Image Segmentation Using Deep Learning Methods. In *Procedia Computer Science* (2016). <https://doi.org/10.1016/j.procs.2016.09.407>.
- [14] Hie, B., Cho, H. & Berger, B. Realizing private and practical pharmacological collaboration. *Science* 362, 347 LP – 350 (2018). <https://doi.org/10.1126/science.aat4807>.
- [15] Lee, K. *et al.* Deep learning of mutation-gene-drug relations from the literature. *BMC Bioinformatics* 19 (2018). <https://doi.org/10.1186/s12859-018-2029-1>.

⁴See <https://rb.gy/hwvsmr>

- [16] Obermeyer, Z. & Mullainathan, S. Dissecting Racial Bias in an Algorithm that Guides Health Decisions for 70 Million People. 89–89 (Association for Computing Machinery (ACM), 2019). <https://doi.org/10.1145/3287560.3287593>.
- [17] Martín-Calvo, D., Aleta, A., Pentland, A., Moreno, Y. & Moro, E. Effectiveness of social distancing strategies for protecting a community from a pandemic with a data driven contact network based on census and real-world mobility data. *MIT Connection Science* (2020).
- [18] Drabiak, K. Caveat Emptor: How the Intersection of Big Data and Consumer Genomics Exponentially Increases Informational Privacy Risks. *Health Matrix: Journal of Law-Medicine* **27**, 143 (2017).
- [19] Crawford, K. & Schultz, J. Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms. *Boston University Law Review* **55** (2020).
- [20] Warren, S. D. & Brandeis, L. D. The Right to Privacy. *Harvard Law Review* **4**, 193–220 (1890).
- [21] Katz v. United States, 389 U. S. 347 (1967).
- [22] United States v. Miller, 425 U. S. 435 (1976).
- [23] United States v. Jones, 565 U. S. 400 (2012).
- [24] HEALTH INSURANCE PORTABILITY AND ACCOUNTABILITY ACT 110 Stat. 1936 (1996).
- [25] AMERICAN RECOVERY AND REINVESTMENT ACT 123 Stat. 115 (2009).
- [26] Pasquale, F. Data-informed Duties in AI Development. *Columbia Law Review* **119**, 1917–1940 (2019).
- [27] Shi, X. & Wu, X. An overview of human genetic privacy. *Annals of the New York Academy of Sciences* **1387**, 61–72 (2017). <https://doi.org/10.1111/nyas.13211>.
- [28] Shannon, C. E. A Mathematical Theory of Communication. *Bell System Technical Journal* **27**, 379–423 (1948). <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>.
- [29] Dwork, C. & Roth, A. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends in Theoretical Computer Science* **9**, 211–407 (2014). <https://doi.org/10.1561/04000000042>.
- [30] Berger, B. & Cho, H. Emerging technologies towards enhancing privacy in genomic data sharing. *Genome Biology* **20** (2019). <https://doi.org/10.1186/s13059-019-1741-0>.
- [31] Naehrig, M., Lauter, K. & Vaikuntanathan, V. Can homomorphic encryption be practical? In *Cloud Computing Security Workshop* (ACM, 2011). <https://doi.org/10.1145/2046660.2046682>.
- [32] Vayena, E. & Gasser, U. Between Openness and Privacy in Genomics. *PLoS Medicine* **13** (2016). <https://doi.org/10.1371/journal.pmed.1001937>.
- [33] Carpenter v. United States, 585 U.S. ____ (2018).
- [34] FEDERAL TRADE COMMISSION ACT U.S.C. 41–58 (2006).