Data & Civil Rights: Health Primer

by Alex Rosenblat, Kate Wikelius, danah boyd, Seeta Peña Gangadharan, and Corrine Yu
Produced for Data & Civil Rights Conference / October 30, 2014

Advances in medicine continue to increase people's lifespan while policy-related efforts to make healthcare more affordable have made insurance more accessible to many in recent years. That said, there are widespread and systemic inequities in the health care system. Communities of color, low-income individuals, women, immigrant communities, and other underserved populations experience a number of health and health care disparities caused by factors such as high health care and insurance costs, language challenges, and limited access to providers and benefits—as well as such factors as unconscious biases, intentional discrimination, and the unintended discriminatory effects of many systemic fiscal choices. Inequities in related fields— higher rates of unemployment, jobs without health insurance, and high insurance premiums relative to ability to pay—compound these problems in the field of health.

Data plays a central role in both medicine and insurance, enabling advances and creating new challenges. Although legislative efforts have attempted to protect the privacy of people's health data, many other kinds of data can reveal sensitive health information about an individual. People's medical conditions or health habits can be inferred from many sources, including their purchases, phone call patterns, fitness tracking apps, posts on social media, and browsing histories. Sometimes, medical information that reveals sensitive information about an individual can be linked to the medical state of a relative. However, accuracy of these inferences may be a problem, and inaccurate inference can result in social stigma and harmful reputational effects on the wrongly categorized individual. In addition, the kinds of inferences generated and used by marketers and insurance companies may not be useful when applied to the context of patient care. Not only does misuse of data have consequences for individuals seeking fair access to healthcare, but inappropriate practices also erode productive efforts to use data to empower people, personalize medicine, and develop innovations that can advance healthcare.

Background: Discrimination in Health

Major civil rights concerns and themes in U.S. health contexts:

- *Lack of Insurance*: Minorities constitute more than one-third of the U.S. population, but make up more than half of the 50 million people who are uninsured.[1]
- *National Origin and Language Challenges*: Discrimination on the basis of national origin, which encompasses discrimination on the basis of limited English proficiency

(LEP), creates unequal access to health.[2] The "cumulative effects of race and ethnicity, citizenship status, low education, and poverty" directly interact with LEP, resulting in more barriers to access.[3]

- *Immigration Status*: Mixed-immigration-status families, such as undocumented parents with citizen children, face unique hardships. Deterred by requests for Social Security numbers or other personal information, family members may not apply for coverage or benefits for their eligible children.

- *Sex discrimination*: Sex discrimination takes many forms and occurs at every step in the health care system. This discrimination harms women and threatens their health, causing them to pay more for health care and putting women at risk for improper diagnoses and less effective treatments.[4] Although physical differences may account for some differences in treatment received by men and women, non-biological or non-clinical factors—including overt or unconscious gender bias—also affect clinical decision-making.[5] Discrimination based on gender identity, gender expression, and sexual orientation remains a persistent problem in our health care system. LGBT individuals, particularly transgender and gender nonconforming individuals, face health care discrimination, including verbal abuse, physical abuse, and outright refusals of treatment.[6]

- *Disability discrimination*: Individuals with all types of disabilities report discriminatory physical, programmatic, and attitudinal barriers to accessing health care in hospitals, clinics, diagnostic facilities, and practitioners' offices of all sizes throughout the country.[7]

Open Data and Privacy Concerns

The privacy of health information reflects both a social norm and a legal mandate governed in part through the Health Insurance Portability and Accountability Act (HIPAA) of 1996, which protects patient privacy and health coverage in health, employment, and insurance contexts. HIPPA, as well as the Genetic Information Nondiscrimination Act (GINA) of 2008, which prohibits insurance companies and employers from discriminating against individuals on the basis of their genetic profiles, anticipated how the Big Data phenomenon could render people vulnerable on the basis of their sensitive health information. However, it's become clear that these protections fall short of what is needed to provide protection, especially because health data is often generated outside of the contexts where these acts apply, and because sensitive medical conditions can be inferred from non-medical data, like consumer habits.

While HIPAA is intended to protect people's privacy, deanonymizing techniques make it possible to reidentify individuals in confidential medical records sold or distributed to researchers or other interest parties. According to research by Latanya Sweeney, anonymized medical records can be re-identified by triangulating patient information with other data, like

newspaper events describing patient names and incidents that resulted in a particular medical injury.[8] She also discovered that "87% (216 million of 248 million) of the population in the United States had reported characteristics that likely made them unique based only on {5-digit ZIP, gender, date of birth},"[9] which means that their anonymized medical records can be matched with other databases to reidentify them.[10] Such technical measures prompt serious questions about the efficacy of regulation like HIPAA.

Consumers generate massive amounts of health data outside of protective relationships or institutions, and these data traces allow companies to infer individuals' medical habits and conditions to categorize and target consumers. Target, the retailer, used data analytics to identify pregnant customers based on a combination of items they purchase, like unscented lotion and nutritional supplements.[11] FICO, the credit-scoring bureau, devised a medication adherence score to identify consumers whose data reflect patterns of inconsistent consumption of prescribed medications.[12] Nielsen, a data aggregator company, scraped data from discussion forums at PatientsLikeMe.com where members post about their medical conditions, medicines, and experiences.[13]

Marketing companies compile or purchase data profiles on potential customers. Based on customer registration or displayed interests, marketers target individuals with promotional materials for new drugs or health services.[14] This process can result in reductive, categorical inferences, the "weblining" of consumers into slots of interest, and reputational harms and costs.[15] Moreover, targeted advertisements can create suspicions about individuals who are implicated with diseases, especially those that carry significant social stigmas.[16]

Collection of health data takes place outside of consumer sectors as well. For instance, police can take DNA swabs of the people they arrest in connection with serious crimes in order to identify them, regardless of whether those individuals are charged or convicted.[17] Because African Americans experience disproportionately higher rates of arrest, they risk overrepresentation in these genetic databases. Given a history of unethical use of African Americans in medical research, such as the use of Black men in the Tuskegee syphilis study,[18] many worry that collection of DNA samples at point of arrest will facilitate surveillance and control of marginalized communities.

Companies are seeking new ways to use consumer data to benefit healthcare providers and facilitate timely interventions. For example, Carolinas HealthCare, which operates a large network of hospitals and practices, uses algorithmic risk-assessment to target an asthmatic patient who lives in a neighborhood with a high pollen count and purchases cigarettes.[19] The Food and Drug Administration (FDA) has launched Mini-Sentinel, a pilot program that mines medical records databases to identify and track adverse patient outcomes linked to pharmaceutical drugs in the marketplace, making the agency less reliant on voluntary self-reports of adverse events.[20] In Indiana, the state government recently implemented a data mining

program to sift through different government-owned databases to identify the causes of high infant mortality and pair them with data from other disparate sources, like the Family and Social Services Administration. The program helps the state figure out which programs to fund to reduce infant mortality rates.[21]

Specific Challenges of Data for Public Health Interventions

In the fields of public health surveillance, public health, and epidemiology, privacy considerations are shaping researchers' and practitioners forays into the process of using confidential, sensitive medical information for large-scale analysis and intervention. The stigma attached to many diseases and medical conditions, particularly in the area of mental health or H.I.V. status, raises sharp privacy concerns about the communicability of health records. There are efforts that are now ongoing to sometimes link public health surveillance registries to electronic medical records, which has been a contentious issue regarding particularly sensitive or stigmatized conditions. For example, historian Amy Fairchild has explored the use of New York public health registries that contain H.I.V. data and which remain protected from access – the data goes in, but it doesn't come out again, although there have been some recent reforms to make the data more usable for patient care. The strongest objections to the use of data to co-ordinate care or stage interventions with patients unaware of their H.I.V. status or to encourage their participation in treatment regimes come from public health officials loathe to create identifiable health records with sensitive data.[22]

Real-time, networked data can be used to stage interventions for highly infectious diseases, and to identify individuals who may have been in contact with communicable disease carriers. Mapping the social components of communicable diseases has a surveillance component that comes from non-healthcare settings. For example, the chief medical correspondent of an NBC news crew that had been in contact with a cameraman with Ebola was spotted violating their voluntary quarantine when she was photographed picking up food.[23] While data analytics can be used to better understand how to make healthcare systems more efficient and minimize the spread of disease, the resultant interventions often don't take into consideration the ways in which people pay personal costs that pit them against the public health officials. People who are identified as potential carriers of a disease like Ebola often cannot afford to not go to work, for example. This raises serious questions about who public health data empowers and who is harmed.

More systematic public surveillance could happen outside of healthcare contexts too, Google has access to huge amounts of data, and they tried to use it to identify flu trends, but they significantly overestimated flu trends using the algorithmic analysis they applied to search query terms. However, researchers at Google found that combining their datasets with data from the Centers for Disease Control and Prevention, their results were much more accurate, and could act as a supplementary signaling mechanism about how the flu is trending on a large scale.[24]

Personalized Health and Disaggregated Risk

The application of data analytics tools and techniques for individualized patient care is progressing, though not without challenges. While new forms of data analytics have been a boon for health marketing – where relevant services or products can be targeted at a person whose medical conditions have been inferred from disparate data sources - medical research requires more careful data collection, analysis, and interpretation to create real benefits.[25] The tension between open, accessible data and useful, well-analyzed, and well-interpreted data is divisive in discussions about what increased access to data and data analytics tools offers in tangible benefits to health care. Medicine already offers a great deal of personalized treatments that can be tailored to patients based on factors like genetic variations, ethnicity, a range of blood test results, etc. However, there is a gap between the proliferation of health data on consumers, the hopes it spurs for personalized treatment options, and what is actually useful in a clinical setting.

Self-tracking devices and apps, like Fitbit, that consumers use outside of clinical settings to improve their health are proliferating. Ginger.io, Omada Health, Propeller Health, Iodine, and other new health analytics companies are trying to use apps and sensors to gather data on users to examine how they are affected by their medical conditions and the treatment regimens they participate in at more granular levels.[26] There is some hope that efforts to achieve more personalized medical assessment and treatment regimes will benefit from the surplus of data available on consumers, but currently, fears around what happens to unprotected, identifiable health data that is inferred from consumer habits primarily centers on who, outside of healthcare providers, will have access to it, like insurance companies.

Access to individualized genetic data is available; until the FDA threatened to shut them down for offering a medical device without FDA approval, 23andMe offered direct-to-consumer products genetic health risk profiling.[27] Many physicians objected to it, in part because they cautioned that patients should be counseled through their results.[28] While patients might possess a great deal of data, they are less likely to know which of it is clinically relevant without the help of a physician.[29]

Non-healthcare providers, like insurance companies, are interested in personalizing medical risk estimates, although currently, the policies being created around personalized insurance costs centers on auto insurance and driving-behavior tracking devices.[30] In aggregate, some employers are using employees' Fitbit data (when it's collected, with the employees' consent, generally under the auspices of a Health & Wellness program) to successfully negotiate for lower group insurance rates on the premise that employee participation in fitness programs will reduce the amount employees claim for health costs overall.[31] Rather than relying on current, coarser-grain aggregate groupings of people and their risk factors, insurance companies can theoretically look at more granular data than they normally collect. For example, they could use data provided by health-tracking devices to calculate the costs of individualized health insurance plans. While

personalized risk assessments can potentially involve greater levels of surveillance in exchange for reduced individual costs, the societal implication is that risk becomes personalized instead of socialized; the costs are shared amongst fewer people. Predictive models about health risks and costs may include not just personalized risk factors, like exercise habits, but also the aggregate health scores of others. For example, a healthier person in a low-income neighborhood (poverty being explicitly linked to worse health outcomes) might pay higher costs for the same health coverage as an individual with healthier neighbors. The same might hold for race, where certain high-cost diseases are statistically correlated with race.

Questions for Data, Civil Rights, and Health

1. Given the diversity of data that can be useful in a health and medical context, is the construct of "health data" helpful? If so, for what purposes and how should this concept be constructed?
2. What are the appropriate avenues for addressing potential risks to individuals in non-medical contexts stemming from the availability of their health data? What can be done to safeguard individual privacy while also opening up opportunities?
3. In an ecosystem with data brokers moving health and non-health information around in ways that may shape insurance policies, which populations face the greatest risks when making their health data available?
4. What rights should individuals have over the data that's used to profile them? How can they identify and correct inaccuracies? How can they challenge algorithmic determinations?
5. What protections need to be put in place so that new health-related innovations can emerge without putting people at risk?
6. What is needed for people to be willing to more comfortably share their data for medical or personal health purposes without putting them at risk of being penalized by insurers?
7. When is health surveillance beneficial, and what are its drawbacks? What kinds of oversight or limits should be put on health surveillance?
8. How can personalized medicine efforts be designed to benefit both marginalized and privileged populations? What structures need to be put in place to enable this?

---

[1] Kaiser Family Foundation, "Health Coverage by Race and Ethnicity: The Potential Impact of the Affordable Care Act," (Washington, D.C., 2013). http://kaiserfamilyfoundation.files.wordpress.com/2013/04/8423.pdf.
[2] Lau v. Nichols, 414 U.S. 563 (1974).

[3] Kaiser Family Foundation, "Overview of Health Coverage for Individuals with Limited English Proficiency," (Washington, D.C., 2013), 3. http://kaiserfamilyfoundation.files.wordpress.com/2013/01/8343.pdf.

[4] Leadership Conference Comments in response to the Department of Health and Human Services Office for Civil Rights' Request for Information (RFI) Regarding Nondiscrimination in Certain Health Programs or Activities, September 30, 2013.

[5] Council on Ethical and Judicial Affairs of the American Medical Association, "Gender Disparities in Clinical Decision Making," *The Journal of the American Medical Association* 266, 4 (1991): 559, doi:10.1001/jama.1991.03470040123034.

[6] Jaime M. Grant, Lisa A. Mottet, Justin Tanis, Jack Harrison, Jody L. Herman, and Mara Keisling. *Injustice at Every Turn: A Report of the National Transgender Discrimination Survey* (Washington, D.C.: National Center for Transgender Equality and National Gay and Lesbian Task Force, 2011), 73. http://transequality.org/PDFs/NTDS_Report.pdf.

[7] Disabilities are diverse. As the Surgeon General said in his 2005 Call to Action to Improve the Health and Wellness of Persons with Disabilities stated: "Some disabilities are visible; others are not. Some are physical, some visual or auditory, some developmental or cognitive, and some mental or behavioral. Some persons are born with one or more disabilities; others acquire a disability during the course of a lifetime… No single disabling condition necessarily affects one person in exactly the same way as it does another."

[8] Latanya Sweeney, "Matching Known Patients to Health Records in Washington State Data" (white paper, Harvard University, Data Privacy Lab, 2011). http://dataprivacylab.org/projects/wa/1089-1.pdf.

[9] Latanya Sweeney, "Simple Demographics Often Identify People Uniquely" (working paper, Carnegie Mellon University, 2000). http://dataprivacylab.org/projects/identifiability/paper1.pdf.

[10] Nate Anderson, "'Anonymized' Data Really Isn't—and Here's Why Not," *Ars Technica*, September 8, 2009. http://arstechnica.com/tech-policy/2009/09/your-secrets-live-online-in-databases-of-ruin/.

[11] Charles Duhigg, "How Companies Learn Your Secrets," *The New York Times*, February 16, 2012. http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html.

[12] Tara Parker-Pope, "Keeping Score on How You Take Your Medicine," *The New York Times*, June 20, 2011. http://well.blogs.nytimes.com/2011/06/20/keeping-score-on-how-you-take-your-medicine/.

[13] Lori Andrews, *I Know Who You Are And I Saw What You Did: Social Networks And The Death Of Privacy* (New York, NY: Free Press, 2011), 34.

[14] Frank Pasquale, *The Black Box Society* (Cambridge, MA: Harvard University Press, 2015)*,* 36.

[15] Pasquale, *Black Box*.

[16] Ifeoma Ajunwa, "Genetic Testing Meets Big Data: Tort and Contract Law Issues," *Ohio State Law Journal* 74 (2014), 8. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2460891.

[17] Adam Liptak, "Supreme Court Says Police Can Take DNA Samples," *The New York Times*, June 3, 2013. http://www.nytimes.com/2013/06/04/us/supreme-court-says-police-can-take-dna-samples.html.

[18] Seeta Peña Gangadharan, "Digital inclusion and data profiling," *First Monday* 17, 5 (2012). http://firstmonday.org/ojs/index.php/fm/article/view/3821/3199.

[19] Shannon Pettypiece and Jordan Robertson, "Your Doctor Knows You're Killing Yourself. The Data Brokers Told Her," *Bloomberg*, June 26, 2014. http://www.bloomberg.com/news/2014-06-26/hospitals-soon-see-donuts-to-cigarette-charges-for-health.html.

[20] Nell Greenfieldboyce, "Big Data Peeps At Your Medical Records To Find Drug Problems," *NPR*, July 21, 2014. http://www.npr.org/blogs/health/2014/07/21/332290342/big-data-peeps-at-your-medical-records-to-find-drug-problems.

[21] Mohana Ravindranath, "In Indiana, State Government Tries Using Big Data Project to Reduce Infant Mortality," *The Washington Post*, August 23, 2014. http://www.washingtonpost.com/business/on-it/in-indiana-state-government-tries-using-big-data-project-to-reduce-infant-mortality/2014/08/23/66d57bc0-2973-11e4-8593-da634b334390_story.html.

[22] Ravindranath, "In Indiana, State Government."

[23] Bruce Golding, "NBC's Medical Correspondent 'Violated Ebola Quarantine,'" *New York Post*, October 13, 2014. http://nypost.com/2014/10/13/nbc-crew-ordered-under-mandatory-ebola-quarantine/.

[24] Steve Lohr, "Google Flu Trends: The Limits of Big Data," *The New York Times,* March 28, 2014. http://bits.blogs.nytimes.com/2014/03/28/google-flu-trends-the-limits-of-big-data/.

[25] Catherine Pope, et al., "What's the big fuss about big data?", *Journal of Health Services Research & Policy* 19:67 (2014). http://hsr.sagepub.com/content/19/2/67.full.pdf+html.

[26] Steve Lohr, "To Gather Drug Data, a Health Start-Up Turns to Consumers," *The New York Times*, September 23, 2014. http://www.nytimes.com/2014/09/24/technology/to-gather-drug-information-a-health-start-up-turns-to-consumers.html.

[27] Robert Hof, "Seven Months After FDA Slapdown, 23andMe Returns With New Health Report Submission," *Forbes*, June 06, 2014. http://www.forbes.com/sites/roberthof/2014/06/20/seven-months-after-fda-slapdown-23andme-returns-with-new-health-report-submission/.

[28] Kenneth Offit, "Genomic Profiles for Disease Risk: Predictive or Premature?", *Journal of the American Medical Association* 299, 11 (2008), 1353-1355. http://jama.jamanetwork.com/article.aspx?articleid=181621.

[29] Jae Won Joh, "What Do Doctors Think About 23andme?", *Forbes*, December 11, 2013. http://www.forbes.com/sites/quora/2013/12/11/what-do-doctors-think-about-23andme/.

[30] Peter S. Vogel, "The High Privacy Price of Auto Insurance Monitoring Discounts," *E-Commerce Times*, July 7, 2012. http://www.ecommercetimes.com/story/75600.html.

[31] Nancy Gohring, "This Company Saved $300k on Insurance by Giving Employees Fitbits," *CITEworld*, July 7, 2014. http://www.citeworld.com/article/2450823/internet-of-things/appirio-fitbit-experiment.html.