Corinne McCumber

Barbara Burke and Elena Machkasova

IS 3255H Machine Learning

15 May 2019

Machine Learning in our Genes: Issues with Direct-to-Consumer Genetic Testing

By February of 2019, over 26 million people have had their DNA analyzed with direct-to-consumer genetic tests, and the number has only grown since then (Regalado, "More"). The biggest companies in the direct-to-consumer genetic testing industry, Ancestry□ and 23andMe, spent over $100 million on advertising in 2017 to get consumers to use their products, jump-starting their databases (Regalado, "2017"). In return, these companies promise to help people understand their genealogy and to possibly provide insights into genetic conditions that might affect a person's health, though 23andMe is the only company to offer the latter service without a doctor's approval (*23andMe.com*; *Ancestry.com*). With the exponential growth of this industry, these companies now have at their disposal a massive amount of genetic information that is primarily sorted using algorithms, but the genetic information is not always sorted accurately or accessed in ways that the owner desires. As such, concerns about the use of this data abound, particularly in regard to accuracy and genetic privacy. Current laws allow companies relative freedom regarding how they use genetic data, but to protect individual rights and privacies, these need to be expanded and media coverage of this issue needs to be improved. Specifically, the Genetic Information Nondiscrimination Act of 2008 (GINA), which protects individuals from genetic discrimination by health insurers and employers, needs to be extended to include more types of insurance and thus to more strongly regulate the way companies use genetic data. Before

discussing laws, though, it is helpful to understand what these companies are giving to their customers.

Genomic data, in the most basic sense, is an ordered collection of the nucleotides cytosine (noted as C), guanine (G), adenine (A), and thymine (T), which make up deoxyribonucleic acid (DNA). Different stretches of DNA code for molecules that function within an organism, and segments of DNA are relatively easy to compare. After collecting DNA in the form of spit samples, commercial companies look at single nucleotide polymorphisms (SNPs), which are "around 600,000 positions where people's DNA code commonly differs" (Regalado, "More"). To match individuals as related, genetic genealogy algorithms "search for regions of the genome where two individuals share … segments [that] must contain a minimum number of SNPs (typically ~500) and be over a certain length" (Greytak et al. 2). The more segments match and the longer those segments are, the higher the likelihood that two individuals are closely related. And to gather health information, such as an individual's risk of breast cancer, 23andMe examines a "few breast cancer mutations" on specific genes (Regalado, "More"). All this information is interpreted, and the customer receives a genetic report regarding their family history and possibly their health risks.

As these tests have gained popularity, these companies have greatly benefited, in part because they work with some users' entire genomes after examining specific segments. Using their wealth of data, both 23andMe and Ancestry☐ have teamed up with big names in the health industry. After being paid $300 million, 23andMe partnered with the pharmaceutical giant GlaxoSmithKline to complete drug research, and Ancestry☐ had a partnership (now ended) with Google's spinoff company Calico to research longevity and aging (Brodwin). Research using this much data is also dependent on machine learning. For example, genome-wide association

studies, a "common form of analysis," rely on algorithms that check if a "variant appears alongside a trait or condition significantly more often" than is expected by chance, thus flagging the variant as a possible cause of the trait or condition (Chivers). This can be done with both supervised and unsupervised learning: supervised if specific traits are selected beforehand and unsupervised if the goal is to cluster genomic data into groups that might correlate with any trait or condition. While this can expand research--for example, it has already led to multiple sclerosis being reclassified as an autoimmune disease, among other things--the money these companies are making is still alarming (Chivers). They make a double profit with users' DNA, first selling their analysis to the consumer and then selling the anonymized genetic information of consenting individuals to a third party. This usually isn't a problem, however, because consumers consent to their DNA being used for research.

Yet there are issues with the information that consumers are given in a company's initial analysis. Regarding genealogy, these companies provide more-detailed results for people of European descent. This is because of the training data used with genealogical algorithms. For example, Ancestry☐ bases its ethnicity estimates on reference samples "sourced from the DNA of 16,638 people representing 43 different populations" who are screened for a "long family history in one place or within one group" (Holger). This is the training data given to the algorithm, and it gives the company a broad idea of where a person's ancestors came from. However, because commercial DNA tests were "first available only in the United States and have expanded mostly to European countries or former colonies, the customer base continues to be fairly uniform" (Holger). This means ancestry results are more accurate for people of European descent because there is ever-more training data incorporated that is applicable to these individuals. What's more, challenges with funding, ethics, and infrastructure make it difficult to

gather information on "underrepresented DNA groups like Africans, Asians, and indigenous peoples"--not to mention the cultural differences that might make some adverse to giving up their DNA (Holger). While 23andMe and Ancestry□ are working to expand their reference samples, they still provide more detail for white individuals, which is a major flaw in one of their main marketable services.

For 23andMe specifically, there are also problems with the health information provided to customers. Put simply, direct-to-consumer genetic tests do not tell the complete story of an individual's health. 23andMe claims to give individuals "insights on [their] health predispositions, carrier status, traits, wellness and ancestry," and while their tests "meet FDA requirements for genetic health risks," they may still be inaccurate (*23andMe.com*). This is because the algorithms "analyze only a small selection of all the genes that have been highly studied and are known to relate to disease risk," giving people a fundamentally incomplete picture of their true genetic risk--which is further complicated because genetic predisposition is not the be-all and end-all of health (Brodwin). Other influences, such as "environmental factors ... diet and exercise also contribute," making a simple genetic test only part of the puzzle ("What"). Because of this complexity, people who have used these tests may think they had a medical genetic health test when in actuality they have not (Brodwin). As such, these services might not benefit the consumer if taken to be the whole truth.

Given this possibility for error, the media has been generally skeptical about direct-to-consumer genetic test results, especially regarding health. News article headlines consistently warn that "experts say companies like 23andMe are doing more harm than good" (Brodwin) or that "Consumer DNA testing promises more than it delivers" (Saey). Similarly, in February of 2019, the editorial board of the *New York Times* cautioned readers to "be careful about

23andMe's health test." In surveying these articles, popular media typically frames the issue to suggest that consumers should be wary of using direct-to-consumer tests as anything more than entertainment, generally focusing on the issues with health tests rather than with genealogical information. Regarding health, I believe that the public has been well-warned. However, very few articles I read mentioned issues with racial information, and none discussed the possible impact of direct-to-consumer genetic testing on a person's ability to get insurance--both, in my opinion, flaws in media coverage. To fully understand the ramifications of getting a direct-to-consumer genetic test, then, current laws need to be discussed.

There are some current protections regarding how companies can use the results of direct-to-consumer genetic tests in insurance decisions. The Genetic Information Nondiscrimination Act of 2008 (GINA) made it so "health insurance companies [and employers] cannot use the results of a direct-to-consumer genetic test (or any other genetic test) to deny coverage or require [someone] to pay higher premiums," although GINA does not apply to companies with fewer than 15 employees, individuals in the military, or people using the Veterans Health Administration or Indian Health Service ("Can"). Additionally, this law protects the genetic information of possible insurance recipients' family, including both dependents and "first-degree, second- degree, third-degree, or fourth-degree relative[s]" (United States, Congress 886). This is good because it limits the harmful ways that genetic information compiled by 23andMe or Ancestry☐ could be used, even if someone has never taken one of these tests; however, it does not go far enough. According to the National Institute of Health, GINA does not apply to "other forms of insurance, such as disability insurance, long-term care insurance, or life insurance," which a person may want if they discover they have concerning genetic predispositions ("Can"). Companies that provide these kinds of insurance have "the right to

request medical information, including the results of any genetic testing, when making decisions about coverage and rates" ("Can"). As such, the results of a less-than-perfect genetic test could be used to raise someone's rates or deny them coverage.

To fix these issues, individuals first need to be better protected by the law. Given the limited scope of the health and genealogical information provided by direct-to-consumer genetic tests, I believe that companies should be legally barred from using these test results when making decisions about any type of insurance coverage and rates--thereby expanding a person's genetic privacy. Additionally, while the risks around 23andMe's health information have been covered in detail, the fact that the genealogical information provided is flawed needs to be better publicized by the media and by companies and, if the goal is a better product, fixed by collecting more diverse reference samples. In the end, consumers should be aware, both in the short term and the long term, of what exactly they are getting when they pay to spit in a tube.

Works Cited

23andMe.com. 23andMe, Inc., 2019, www.23andme.com. Accessed 23 April 2019.

Ancestry.com. Ancestry☐, 2019, www.ancestry.com/?. Accessed 23 April 2019.

Brodwin, Erin. "Genetic testing is the future of healthcare, but many experts say companies like

      23andMe are doing more harm than good." *Business Insider,* 19 January 2019,

      www.businessinsider.com/future-healthcare-dna-genetic-testing-23andme-2018-12.

      Accessed 28 April 2019.

"Can the results of direct-to-consumer genetic testing affect my ability to get insurance?"

      Genetics Home Reference, National Institute of Health US National Library of Medicine,

      16 April 2019 ghr.nlm.nih.gov/primer/dtcgenetictesting/dtcinsurancerisk. Accessed 28

      April 2019.

Chivers, Tom. "How Big Data Is Changing Genetic Research." *Digg*, 2 October 2018,

      digg.com/2018/big-data-science-genetics. Accessed 8 April 2019.

Editorial Board. "Why you should be careful about 23andMe's health test." *New York Times*, 1

      February 2019, www.nytimes.com/interactive/2019/02/01/opinion/23andme-cancer-

      dna-test-brca.html. Accessed 13 May 2019.

Greytak, Ellen M. et al. "Genetic genealogy for cold case and active investigations." *Forensic*

      Science International, 2019, doi.org/10.1016/j.forsciint.2019.03.039.

Holger, Dieter. "DNA testing for ancestry is more detailed for white people. Here's why, and

      how it's changing." *PCWorld*, 4 December 2018, www.pcworld.com/article/3323366/

      dna-testing-for-ancestry-white-people.html. Accessed 28 April 2019.

Regalado, Antonio. "2017 was the year consumer DNA testing blew up." *MIT Technology*

      Review, 12 February 2018, www.technologyreview.com/s/610233/2017-was-the-year-

consumer-dna-testing-blew-up/. Accessed 28 April 2019.

---. "More than 26 million people have taken an at-home ancestry test." *MIT Technology Review,*
11 February 2019, www.technologyreview.com/s/612880/more-than-26-million-
people-have-taken-an-at-home-ancestry-test/. Accessed 28 April 2019.

Saey, Tina Hesman. "Consumer DNA testing promises more than it delivers." *Science News*, 22
May 2018, www.sciencenews.org/article/consumer-genetic-testing-dna-genome.
Accessed 13 May 2019.

United States, Congress. *Public Law 110–233*. 21 May 2008, 110th Congress. *United States
Government Publishing Office,* www.govinfo.gov/content/pkg/STATUTE-122/pdf/
STATUTE-122-Pg881.pdf.

"What do the results of direct-to-consumer genetic testing mean?" *Genetics Home Reference*,
National Institute of Health US National Library of Medicine, 16 April 2019,
ghr.nlm.nih.gov/primer/dtcgenetictesting/dtcresults. Accessed 28 April 2019.