The Personal Data is Political

Bastian Greshake Tzovaras¹², Athina Tzovara³

¹: Lawrence Berkeley National Laboratory, Berkeley, CA, USA
²: Open Humans Foundation, USA
³: Helen Wills Neuroscience Institute, University of California Berkeley, CA, USA

Keywords: genetics, personalized medicine, neuroscience, data sharing, diversity, open data

Summary

The success of personalized medicine does not only rely on methodological advances but also on the availability of data to learn from. While the generation and sharing of large data sets is becoming increasingly easier, there is a remarkable lack of diversity within shared datasets, rendering any novel scientific findings directly applicable only to a small portion of the human population. Here, we are investigating two fields that have been majorly impacted by data sharing initiatives, neuroscience and genetics. Exploring the limitations that are a result of a lack of participant diversity, we propose that data sharing in itself is not enough to enable a global personalized medicine.

Personalized or stratified medicine has been one of the hot topics in healthcare, reaching well beyond the launch of the Precision Medicine Initiative in the United States [1]. The promise of personalized medicine is to identify individuals at risk and find optimally tailored healthcare solutions based on their genetic and environmental makeup [2]. Although personalized medicine spans over a variety of medical and biological disciplines, two subfields are particularly promising due to their growing adoption: genetics and neuroscience. Indeed, many current examples of precision medicine come from pharmacogenomics in general, specifically from oncology, where cancer treatments are picked to match the mutations found in tumours [3]–[5].

While this use of genetic data in healthcare is projected to become more central in the next years, its success will depend on multiple factors. As for most things in healthcare, cost plays a huge role. But while the costs for performing a high precision medical examination, like a brain scan, or sequencing a human genome continue to drop [6], their usefulness is bound by both our ability to quickly process these large amounts of data as well as the lack of medically-relevant scientific knowledge we have about individual genetic variants [7], or complex neurobiological processes. As such it is key that science be able to generate genetic knowledge more quickly [8].
Two recent trends in science, big data and artificial intelligence, appear to be promising for not only accelerating our genomic and neurobiological understanding but also for diagnosing in a precision medicine framework [9], [10]. The idea is that artificial intelligence can be used to mine large data sets to find the smallest associations between genetic variants / neuromarkers and disease phenotypes, and to track disease progression or predict optimal treatments. To effectively create such large data collections it thus becomes central to link and share individual data sets [8]. But while the total number of basepairs sequenced per time as well as the total number of participants included in neuroscientific studies have exponentially increased over the last years, sharing practices for such data has not kept up a similar speed [11], despite individual efforts to enable open sharing of genetic [12], [13] or neuroscientific [14] data.

Sharing genomic data

To alleviate these shortcomings individual academic consortia have been founded to pool data sets across institutions and individual researchers. National efforts include the UK10K [15], which aimed to sequence 10,000 participants in the United Kingdom and the similarly structured 100,000 Genomes Project by Genomics England [16]. In the United States the Exome Aggregation Consortium (ExAC) [17] - which has collected over 60,000 exomes - and more recently the All of Us initiative [18] are collecting and aggregating more patient data for research purposes. And it’s not only academic research that is starting to collect large data sets for personalized medicine, commercial companies are starting to explore the field too.

Since deCODE Genetics and 23andMe released the first Direct-To-Consumer genetic tests back in 2007 [19] the market for commercial genetic testing has grown significantly: Not only in terms of companies like MyHeritage, FamilyTreeDNA, AncestryDNA or Veritas that have entered the market, but also in terms of the number of people who have gotten genetic tests through these services. Today AncestryDNA has over 5 million customers and industry veteran 23andMe has genetic data for over 2 million people [20]. These sizable commercial databases are of interest to academic and commercial researchers. 23andMe has collaborated with academic researchers on numerous research papers [21] and has done commercial for-profit collaborations with pharmaceutical companies like Pfizer and Genentech.

Who profits from such large-scale research remains open. As an example, in psychology the need to look into how representative study participants are has been acknowledged. After all, around 80% of
all participants in psychology studies are from WEIRD (Western, Educated, Industrialized, Rich, Democratic) countries and do thus not represent human diversity [22]. As such, only WEIRD participants can fully profit from much of psychological research. To avoid the overrepresentation of WEIRD individuals found in psychology, it is key that our genetic research data resources reflect human diversity across populations. Indeed, this issue of representativeness becomes even more central in the genetic framework of Genome Wide Association Studies (GWAS). These studies are commonly used to inform personalized medicine by identifying genetic risk factors, e.g. for cancer [23]. Unfortunately, most of these identified risk factors are mere correlations, not genes directly causing a disease. As these correlations depend on the ancestry context in which they were found, findings of a GWAS are not necessarily applicable outside the human population in which an association was initially found [24] and cannot be replicated in many cases [25].

Indeed, many data sharing efforts show such a lack of population diversity: More than 50% of the over 60,000 samples in the ExAC consortium come from a European population [17]. Similarly, commercial databases like the ones of 23andMe suffer from ancestry and race biases [26], [27]. Open genomic databases – like the Personal Genome Projects and openSNP – are not fairing much better: 75% of participants in one of Harvard’s Personal Genome Project studies identified as white [12] and amongst a survey of over 500 openSNP participants over 70% come from the US, UK and Canada. Additionally, over 75% of openSNP participants had at least a Bachelor’s degree, hinting at a highly skewed demographic [28].

Sharing neurobiological data

Similar to genetics, neuroscience has gone a long way when it comes to data sharing: While initial attempts to share data mainly focused on post-processed data, like coordinate-based results or statistical maps of magnetic resonance imaging (MRI) [29], more recent initiatives enable sharing of entire functional or structural MRI datasets [30], [31] and magneto- or electro- encephalography (M/EEG) data [32].

As in the case of psychology and genomics, neuroscience research is largely based on data of individuals from WEIRD societies [33], despite a plethora of studies showing that brain development is affected by socioeconomic status, early life stress, or cultural differences [34]–[38]. Indeed, within or across household socio-economic variables during childhood, such as family income, parental education [39], [40] or neighbourhood poverty levels [35], can be traced on trajectories of brain development, and result in differences in brain structure [39] and cognitive functions [41], or gene
expression [42]. Differences in brain networks according to socio-economic status are also evident during adolescence [40] and adulthood [36].

Furthermore, culture has been shown to influence neural functions [38]. Cultural and ethnic differences have an impact on emotion perception and expression, and brain responses to emotional or social cues [43]. Moreover, ethnic differences have been found in physiological responses to fear or novelty [44], [45], which are commonly used to assess anxiety or post-traumatic stress disorders [46]. This situation is aggravated by the fact that ethnicity can influence commonly used laboratory measurements of fear like skin conductance responses [45], potentially leading to the exclusion of ethnicities despite being at higher risk e.g. for post-traumatic stress disorders [47].

How much existing data sharing efforts for neuroscience are affected by these biases is hard to estimate at this point: Although these initiatives generally tend to support standardized data formats for data sharing [48], [49], they only rarely include concrete guidelines for reporting of socio-demographic variables [50].

**Data sharing as a social movement**

All of this paints a bleak picture: The populations we are using to develop personalized medicine are highly WEIRD [22]. Even worse, we might often not be even aware of this, as we are not collecting the needed demographic data to identify our biases. Depending on the field, research studies can furthermore only contain small sample sizes, making it hard to evaluate how ethnicity or social factors influence neurobiological functions and gene expression. Only by sharing diverse datasets, and including rich demographic information will it be possible to make our understanding of disease progression, and neurobiological functions relevant for all individuals, irrespective of their social or ethnic background.

Back in 2005, Thomas Friedman firmly believed that *next great breakthrough in bioscience could come from a 15-year-old who downloads the human genome in Egypt* [51]. Today, we have to acknowledge that there’s a good chance that this 15-year-old would not be able to profit from their own breakthrough. Because of this we are still far away from a truly personalized medicine, making our personal data political. It is up to us, the generators of data and the people sharing data to work on changing this, ensuring that the promise of personalized medicine is equitable. Or to say it with Carol Hanisch’s words: *There are no personal solutions at this time. There is only collective action for a collective solution* [52].


[46] D. R. Bach, A. Tzovara, and J. Vunder, “Blocking human fear memory with the matrix


