The Personal Data is Political

Bastian Greshake Tzovaras [1,2], Athina Tzovara [3]

[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 health care, 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 health care solutions based on their genetic and environmental makeup [2]. Although personal 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 health care 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].

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27079v1 | CC BY 4.0 Open Access | rec: 31 Jul 2018, publ: 31 Jul 2018

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

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27079v1 | CC BY 4.0 Open Access | rec: 31 Jul 2018, publ: 31 Jul 2018

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].

[1] H. Collins, Francis S. ; Varmus, “A New Initiative on Precision Medicine,” *Perspective*, vol. 363, no. 1, pp. 1–3, 2010.

[2] Y.-F. Lu, D. B. Goldstein, M. Angrist, and G. Cavalleri, “Personalized Medicine and Human Genetic Diversity,” *Cold Spring Harb. Perspect. Med.*, vol. 4, no. 9, pp. a008581–a008581, 2014.

[3] S. Kummar *et al.*, “Application of Molecular Profiling in Clinical Trials for Advanced Metastatic Cancers,” *Jnci-Journal Natl. Cancer Inst.*, vol. 107, no. 4, 2015.

[4] R. Smith, “Stratified, personalised, or precision medicine,” *thebmjopinion*, 2012.

[5] C. Tan and X. Du, “KRAS mutation testing in metastatic colorectal cancer,” *World J. Gastroenterol.*, vol. 18, no. 37, pp. 5171–5180, 2012.

[6] L. Wetterstrand, “DNA Sequencing Costs, Data from the NHGRI Genome Sequencing Program,” 2018. [Online]. Available: https://www.genome.gov/sequencingcostsdata/.

[7] F. E. Dewey *et al.*, “Clinical interpretation and implications of whole-genome sequencing,” *JAMA - J. Am. Med. Assoc.*, vol. 311, no. 10, pp. 1035–1044, 2014.

[8] I. S. Kohane, “Ten things we have to do to achieve precision medicine,” *Science (80-.).*, vol. 349, no. 6243, pp. 37–38, 2015.

[9] H. Moon, H. Ahn, R. L. Kodell, S. Baek, C. Lin, and J. J. Chen, “Ensemble methods for classification of patients for personalized medicine with high-dimensional data,” *Moon*, 2007.

[10] S. E. Dilsizian and E. L. Siegel, “Artificial intelligence in medicine and cardiac imaging: Harnessing big data and advanced computing to provide personalized medical diagnosis and treatment,” *Curr. Cardiol. Rep.*, vol. 16, no. 1, 2014.

[11] N. V. Kovalevskaya *et al.*, “DNAdigest and Repositive: Connecting the World of Genomic Data,” *PLoS Biol.*, vol. 14, no. 3, 2016.

[12] Q. Mao *et al.*, “The whole genome sequences and experimentally phased haplotypes of over 100 personal genomes,” *Gigascience*, vol. 5, no. 1, 2016.

[13] B. Greshake, P. E. Bayer, H. Rausch, and J. Reda, “openSNP--a crowdsourced web resource for personal genomics.,” *PLoS One*, vol. 9, no. 3, p. e89204, Jan. 2014.

[14] J.-B. Poline *et al.*, “Data sharing in neuroimaging research,” *Front. Neuroinform.*, vol. 6, no. 9, 2012.

[15] “UK10K.” [Online]. Available: https://www.uk10k.org/.

[16] “Genomics England.” [Online]. Available: https://www.genomicsengland.co.uk/.

[17] "ExAC." [Online]. Available: http://exac.broadinstitute.org/faq.

[18] "All of Us." [Online]. Available: https://allofus.nih.gov.

[19] D. Vorhaus, "The Past, Present and Future of DTC Genetic Testing Regulation," *Genomics Law Rep.*, 2010.

[20] B. F. McAllister, "Exponential Growth of the AncestryDNA Database," 2017. [Online]. Available: https://wiki.uiowa.edu/display/2360159/2017/09/15/Exponential+Growth+of+the+AncestryDNA+Database.

[21] "23andMe Research." [Online]. Available: https://research.23andme.com/publications/.

[22] J. Henrich, S. J. Heine, and A. Norenzayan, "The weirdest people in the world?," *Behav. Brain Sci.*, vol. 33, no. 2–3, pp. 61–83, 2010.

[23] A. Agyeman and R. Ofori-Asenso, "Perspective: Does personalized medicine hold the future for medicine?," *J. Pharm. Bioallied Sci.*, vol. 7, no. 3, p. 239, 2015.

[24] W. S. Bush, J. H. Moore, J. Li, S. McDonnell, and K. Rabe, "Chapter 11: Genome-Wide Association Studies," *PLoS Comput. Biol.*, vol. 8, no. 12, p. e1002822, 2012.

[25] U. M. Marigorta *et al.*, "High Trans-ethnic Replicability of GWAS Results Implies Common Causal Variants," *PLoS Genet.*, vol. 9, no. 6, p. e1003566, 2013.

[26] "Problems with 23andMe Ancestry Composition." [Online]. Available: http://koreanhistoricaldramas.com/23andme-ancestry-composition/.

[27] Euny Hong, "23andMe has a problem when it comes to ancestry reports for people of color." [Online]. Available: https://qz.com/765879/23andme-has-a-race-problem-when-it-comes-to-ancestry-reports-for-non-whites/.

[28] T. Haeusermann, B. Greshake, A. Blasimme, D. Irdam, M. Richards, and E. Vayena, "Open sharing of genomic data: Who does it and why?," *PLoS One*, vol. 12, no. 5, pp. 1–15, 2017.

[29] P. T. Fox and J. L. Lancaster, "Mapping context and content: The BrainMap model," *Nat. Rev. Neurosci.*, vol. 3, no. 4, pp. 319–321, 2002.

[30] K. J. Gorgolewski *et al.*, "NeuroVault.org: a web-based repository for collecting and sharing unthresholded statistical maps of the human brain," *Front. Neuroinform.*, vol. 9, 2015.

[31] R. A. Poldrack *et al.*, "Toward open sharing of task-based fMRI data: the OpenfMRI project," *Front. Neuroinform.*, vol. 7, 2013.

[32] G. Niso *et al.*, "OMEGA: The Open MEG Archive," *Neuroimage*, vol. 124, pp. 1182–1187, 2016.

[33] E. Falk, H. Luke, C. Mitchel, J. Faul, and E. Al, “What is a representative brain? Neuroscience meets population science,” *PNAS*, vol. 110, no. 44, 2013.

[34] D. A. Hackman, M. J. Farah, and M. J. Meaney, “Socioeconomic status and the brain: Mechanistic insights from human and animal research,” *Nat. Rev.*, vol. 11, pp. 651–659, 2010.

[35] N. A. Marshall, H. A. Marusak, K. J. Sala-Hamrick, L. M. Crespo, C. A. Rabinak, and M. E. Thomason, “Socioeconomic disadvantage and altered corticostriatal circuitry in urban youth,” *Hum. Brain Mapp.*, vol. 39, no. 5, pp. 1982–1994, 2018.

[36] M. Y. Chan, J. Na, P. F. Agres, N. K. Savalia, D. C. Park, and G. S. Wig, “Socioeconomic status moderates age-related differences in the brain’s functional network organization and anatomy across the adult lifespan.,” *Proc. Natl. Acad. Sci. U. S. A.*, p. 201714021, 2018.

[37] E. R. Duval *et al.*, “Childhood poverty is associated with altered hippocampal function and visuospatial memory in adulthood,” *Dev. Cogn. Neurosci.*, vol. 23, pp. 39–44, 2017.

[38] B. J. Liddell and L. Jobson, “The impact of cultural differences in self-representation on the neural substrates of posttraumatic stress disorder,” vol. 1, pp. 1–13, 2016.

[39] M. E. Ellwood-Lowe, K. L. Humphreys, S. J. Ordaz, M. C. Camacho, M. D. Sacchet, and I. H. Gotlib, “Time-varying effects of income on hippocampal volume trajectories in adolescent girls,” *Dev. Cogn. Neurosci.*, vol. 30, pp. 41–50, 2018.

[40] D. G. Weissman, R. D. Conger, R. W. Robins, P. D. Hastings, and A. E. Guyer, “Income change alters default mode network connectivity for adolescents in poverty,” *Dev. Cogn. Neurosci.*, vol. 30, pp. 93–99, 2018.

[41] D. Hackman and M. Farah, “Socioeconomic Status and the Developing Brain.” [Online]. Available: https://bb.wustl.edu/bbcswebdav/pid-2052428-dt-content-rid-4878238_1/courses/FL2015.L12.Educ.102.01/Hackman_Farrah_SES and the Developing Brain.pdf.

[42] N. Parker *et al.*, “Income inequality, gene expression, and brain maturation during adolescence,” *Sci. Rep.*, vol. 7, no. 1, p. 7397, 2017.

[43] B. Derntl *et al.*, “Culture but not gender modulates amygdala activation during explicit emotion recognition,” *BMC Neurosci.*, vol. 13, no. 1, 2012.

[44] K. G. Martínez, J. A. Franco-Chaves, M. R. Milad, and G. J. Quirk, “Ethnic differences in physiological responses to fear conditioned stimuli,” *PLoS One*, vol. 9, no. 12, 2014.

[45] M. Alexandra Kredlow *et al.*, “Assessment of skin conductance in African American and Non–African American participants in studies of conditioned fear,” *Psychophysiology*, vol. 54, no. 11, pp. 1741–1754, 2017.

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

metalloproteinase inhibitor doxycycline," *Mol. Psychiatry*, 2017.

[47] A. L. Roberts, S. E. Gilman, J. Breslau, N. Breslau, and K. C. Koenen, "Race/ethnic differences in exposure to traumatic events, development of post-traumatic stress disorder, and treatment-seeking for post-traumatic stress disorder in the United States.," *Psychol Med.*, vol. 41, no. 1, pp. 71–83, 2011.

[48] G. Niso *et al.*, "MEG-BIDS, the brain imaging data structure extended to magnetoencephalography," *Sci. Data*, vol. 5, p. 180110, 2018.

[49] K. J. Gorgolewski *et al.*, "The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments," *Sci. Data*, vol. 3, 2016.

[50] C. R. Madan, "Advances in Studying Brain Morphology: The Benefits of Open-Access Data," *Front. Hum. Neurosci.*, vol. 11, 2017.

[51] D. H. Pink, "Why the World Is Flat," *WIRED*. [Online]. Available: https://www.wired.com/2005/05/friedman-2/.

[52] C. Hanisch, "The Personal Is Political," 1969. [Online]. Available: http://www.carolhanisch.org/CHwritings/PIP.html.