Approved on February 14. 2024

REWS Diversity in Datasets Policy Draft v2.4
Diversity in Data Sets

Chair: **Anna Lewis (**Brigham and Women's Hospital)
Coordinator: **Beatrice Kaiser** (Centre of Genomics & Policy, McGill University)

Task force: Anna Lewis, Brigham and Women's Hospital

Co-Authors

Maili C. Raven-Adams, Nuffield Council on Bioethics
Tina Hernandez-Boussard, Stanford Medicine
Yann Joly, Centre of Genomics & Policy, McGill University
Bartha Maria Knoppers, Centre of Genomics & Policy, McGill University
Subhashini Chandrasekharan, National Institutes of Health
Adrian Thorogood, The Terry Fox Research Institute
Judit Kumuthini, Centre for Proteomic and Genomic Research
Calvin Ho, Centre for Medical Ethics, University of Hong Kong
Ariana Gonzlez, Dasa Genómica LATAM
Sarah C. Nelson, Genetic Analysis Center, University of Washington, Department of Biostatistics
Yvonne Bombard, Institute of Health Policy, Management & Evaluation, University of Toronto
Donrich Thaldar, University of KwaZulu-Natal
Hanshi Liu, Centre of Genomics & Policy, McGill University
Saumya Shekhar Jamuar, KK Women's & Children's Hospital
Alessia Costa, Wellcome Genome Campus
Vijaytha Muralidharan, Stanford Medicine
Sasha Henriques, Wellcome Connecting Science
Jamal Nasir, University of Northampton
Aimé Lumaka, African Rare Disease Initiative

Considerations for how Genetic and Genomic Researchers Should Approach Thinking About Diversity in Data

Calls to increase the diversity in datasets used for research are prominent within the genomics community. By increasing diversity in genomic data, the hope is that the benefits of advances in genetics and genomics can be applied and distributed more equitably and reliably. However, many of these calls do not specify what exactly is meant by diversity in datasets, and when this

Approved on February 14. 2024

*is specified, usually the language of ancestral diversity is used. However, diversity of genetic ancestry is only part of the need. Moreover, much care is needed in how human diversity is framed: the concepts and terms we use matter. This policy makes two contributions. First it proposes that researchers identify **what** types of diversity are important in data, via considering the outcome that this diversity is designed to deliver. This will often be linked to achieving equity. A consequence of this approach is that it demonstrates that diversity is not necessarily about representativeness. Second, this policy proposes a systematic way for **how** researchers can enable achieving the benefits that the identified diversity is designed to enable. Researchers must reckon with the complex reasons for the lack of diversity present in current data. We emphasize that data gathering is only one stage within the data lifecycle, and consideration is needed across all stages. We also emphasize that attention to lawful, contextually appropriate benefit sharing is needed across each of these stages. We illustrate this proposal using two examples.*

Table of Contents

Overview 2
The need for diversity in datasets 3
What exactly is meant by “diverse data”? 5
- A lack of specificity in the calls for diversity in data 5
- The problems of framing in terms of “ancestral” diversity 6
- The limitation of thinking about diversity just in terms of genetic diversity 7
- What types of diversity should matter in genomics research? 8
  - Question 1a: What are the goals of the project, and how can these goals be framed in a way that maximally aligns with ethical norms? 9
  - Question 1b: Given these goals, what type(s) of diversity in data are the most important? 10

How can diversity in data be achieved? 10
- Complex reasons for a lack of diversity 10
- How should these types of diversity in data be pursued? 12
  - Question 2a: What are the implications for choices made throughout the data lifecycle? 12
  - Question 2b: How can lawful, contextually appropriate benefit sharing be supported at all the relevant stages? 14

Two worked examples 17
- Example 1: Minimizing reporting false positive pathogenic variants 17
- Example 2: The clinical impact of differential predictive performance of polygenic scores 20

Overview

In this policy we propose a process for how genetic and genomic researchers should approach the achievement of diversity in data. The process is designed to help enable the outcomes that calls for diversity aim to achieve. We acknowledge that researchers are very differently situated from one another, and that aspects of this proposed process will be differentially salient. These

Approved on February 14. 2024

differences include, for example, whether the research is based in a Low and Middle Income Country (LMIC) or a high income country, whether the research is embedded in a healthcare system, the accessibility of related resources, geography, and cultural practices.

The first intention of this policy is to help researchers consider the multiple dimensions of diversity that are important for their research goals and transparently show how diversity factors into their topic. We begin this discussion by highlighting that calls for more diverse data often do not specify what type/types of diversity is/are important (for example, racial diversity, genetic ancestry diversity, geographical diversity). Furthermore, we acknowledge issues both with calling for “ancestral” diversity and with focusing just on genetic diversity. We argue that there is no single definition of “diversity in datasets”, and that diversity is not necessarily about representativeness. Rather, we argue that researchers should view diversity in data as a means to achieve ethical ends, including public benefit, equity, justice, and fairness. This alignment with the ends of the research clarifies *what* types of diversity in data are important. We suggest researchers ask themselves two questions to address what types of diversity are important for a given project: 1) What are the goals of the project, and how can these goals be framed in a way that maximally aligns with ethical norms? 2) Given these goals, what type(s) of diversity in data are the most important?

The second intention is to help researchers plan *how* they can go about achieving the types of diversity they have identified as important for a given project. To enable the outcomes that diversity in data is a means to achieve, it is necessary to consider stages of the data lifecycle beyond data collection. We begin this discussion by acknowledging the complex reasons for the lack of diversity in data, including an often well-founded lack of trust in genomics research specifically and health research more broadly.Helping researchers achieve the benefits of more diversity involves prioritizing transparency, accountability, culturally appropriate behavior and language, as well as recognition and respect for the role of marginalized communities asked to participate in research. This may involve ensuring benefits are shared with communities, local researchers, and possibly research participants. For example, in the case of researchers from high-income countries working in LMICs, this involves focusing on equitable collaborations with local researchers, likely including capacity building and infrastructure investments. We suggest researchers plan how to achieve the type(s) of diversity identified as most important for their project by asking themselves two further questions: What are the implications for choices made throughout the data lifecycle? How can lawful, contextually appropriate benefit sharing be supported at all the relevant stages?

The four questions we suggest are summarized in Figure 1. Of note, consideration of these different points might enable additional realizations about the types of diversity in data necessary to achieve the project goals (and maybe even prompt a refinement of these project goals). Iteration between these questions may hence be necessary and desirable. Explicitly answering

Approved on February 14. 2024

these questions can help researchers transparently communicate their thought processes, and should also help ensure that research is aligned with ethical ends.

In the final section of this document, we illustrate how this process could be used by giving two examples specific to genomics related research: minimizing reporting false positive pathogenic variants, and the clinical impact of differential predictive performance of polygenic scores.

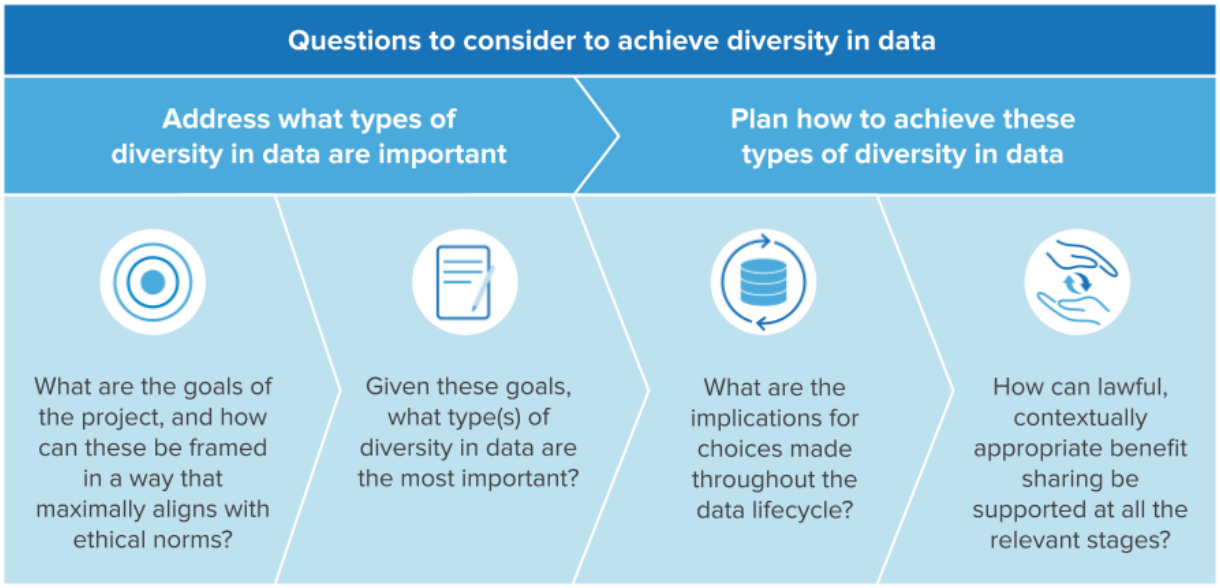


Figure 1: Questions genetic and genomic researchers should answer when thinking about diversity in data. *We note that these steps can be iterated on.*

We acknowledge that the project-specific approach we take in this policy runs the risk of obscuring the shape of research being pursued by the overall biomedical enterprise. There will also be important considerations for those overseeing and shaping many research projects, such as funders, policy makers, ethics committees, and publishers. These actors have a key role here to ensure that projects address the highest order goals, such as justice and fairness. This is particularly true for ensuring a balance within basic/translational research, and the types of conditions that are the focus of research projects. We stress that there will always be a mandate to understand and integrate diverse perspectives as research agendas are shaped. There will also be a role for certain communities that have been marginalized in genomics research to lay forward terms of engagement necessary to achieve the outcomes they desire.

The need for diversity in datasets

On an international scale, calls for ‘diversity in datasets’ are identifiable across a range of actors aiming to ensure the benefits of genetics and genomics can be realized for all. Although human

Approved on February 14. 2024

genomes are 99.9% identical, differences in the 0.1% can have significant impacts on human health.[1] These differences are not distributed equally throughout human kind; studying the breadth of human genetic diversity is thus important.[2] A frequently cited statistic is the percentage of genomics studies conducted in individuals of 'European descent', which was 86% in 2021, a 5% increase from 2016.[3] The need to better understand the relevance of the human genome to human health and wellbeing has led to pleas for more 'diverse' data, and for changed practices, from a range of players in the global genomics field: academic papers,[3] funders (e.g. Wellcome[4] and National Human Genome Research Institute (NHGRI),[5] professional societies (e.g. US National Academies),[6] genomics initiatives (All of Us Research Program,[6] SG10k Health,[7] CanPath,[8] H3Africa,[9] Our Future Health[10] and Genomics England[11]) and publishers (e.g. Nature).[12]

The benefits of increased diversity in datasets are twofold. First, increased power of scientific advances e.g. locus discovery, fine mapping, and causal inference.[13] Such advances have the potential to benefit all.[3,14,15] Second, health equity in the context of genomic medicine, where this has been defined as the global and unbiased applicability of genomic knowledge e.g. diagnosis, pharmacogenomics, genetic counseling, clinical care, and drug development. For example, the increased rate of false positive diagnosis of rare genetic diseases, the lower predictive performance of polygenic scores (PGS), and higher rates of adverse drug reactions in certain groups are linked to a lack of diversity in datasets, and all have been framed as contributing to health disparities.[16–18] Improving diversity in datasets is therefore essential to ensure that health disparities are not increased for people who are already underserved by research and healthcare.[14]

It is not enough to simply call for "diversity" within datasets: there needs to be ways to understand what types of diversity matter (and why), and how to act on this understanding throughout the research process.

What exactly is meant by "diverse data"?

A lack of specificity in the calls for diversity in data

As we have outlined, there is widespread acknowledgement of the need for 'diversity in datasets'. However, the calls for diversity in datasets often lack specification and there is little clarity on exactly what is meant by diversity.[19]

Most of the calls for diversity in datasets and subsequent work addressing these calls are framed in terms of "ancestral" diversity, with a focus on improving the representation and analysis of datasets from beyond those of European ancestry.[19]

Approved on February 14. 2024

The term “representation” — and the allied “underrepresentation” — is often used in tandem with ‘diversity.’[5,14,20,21] However, there is often a lack of acknowledgement around what sort of representation is needed, for what purpose(s), in what contexts, and how this should be achieved.[22]

The problems of framing in terms of “ancestral” diversity

How diversity in datasets is framed, e.g. using concepts of race or genetic ancestry, also has ramifications for how we think about human groups. There are dangers of reinforcing the false notion that humans come in distinct biological “types”.[23,24] When geneticists conceptualize human differences they are not merely structuring an academic enquiry, but shaping how ideas of human differences will be utilized outside the academy.[25]

The birth of genetics was intertwined with that of eugenics, and concepts that justified racism in the late 19th and early 20th century.[26] We need to be aware of this history to guard against current and future arguments. Unfortunately, ideas about human diversity from genetics continue to be appropriated by those seeking to justify racial prejudice and hatred.[27]

The framing of diversity in terms of ancestry suffers from ambiguity, because the term “ancestry” is itself highly ambiguous. Empirical work shows that researchers offer varied and divergent definitions of the term, spanning from explicitly a social phenomenon to explicitly a biological one determined by genetic analysis.[28] This allows for the conflation of social and genetic categories, particularly given that the dominant labels used are continental ones (e.g. African ancestry, European ancestry), which are conflated with racial categories.[24]

We note that the current focus on ancestry can lead to obfuscation of the role of geographical location. For example, researchers can (and do) claim to have adequately represented Latino individuals (framed as an ancestry category), even though most of these individuals were recruited in the US.[26] This risks the results of genotype-phenotype and pharmacogenetic studies being substantially less predictive in individuals living in Latin America and the Caribbean, where environmental conditions could be very different to those experienced in the US.[29]
Framing diversity in terms of genetic ancestry (i.e. more specifically than “ancestry”) is a step in the right direction, but we should now be thinking of genetic ancestry in a continuous rather than categorical way. Genetic ancestry can be defined as the way in which segments of DNA have been passed down through the human family tree.[30] For an individual, their genetic ancestry is then defined as the subset of paths through their family tree through which they have inherited DNA from specific ancestors. It is possible to produce coarse summaries of the information in this human family tree in various ways, for example by allocating each individual to one or a percentage of more than one group. These abstractions have the disadvantage that they risk

Approved on February 14. 2024

essentializing genetically defined groups, when we know that human genetic variation varies in a continuous fashion.[24,31] [1f]

The limitation of thinking about diversity just in terms of genetic diversity

Moving from “ancestral diversity”, thought of mostly in terms of continental level categories, to considering diversity of genetic ancestry, thought of in a continuous manner, is a necessary but insufficient reconceptualization of diversity in data for genetic studies. Diversity in the environments that an individual experiences is also important — after all, we sample genomes and environments concurrently. “Environment” includes everything non-genetic, including an individual's social environment. For many health conditions the influence of the social determinants of health outweighs other factors, including genetic predisposition.[32] Capturing diversity in the amount/type of environmental exposure will be important.[33]

The problems of a lack of diversity along these other dimensions of difference - e.g. sex, socio-economic status - is well recognized for clinical trials,[34] and for healthcare studies using artificial intelligence.[34] A complication to adequately conceptualizing diversity in datasets is that environmental factors and genotypes often co-vary.[35] Genotypes are often taken as a proxy for unmeasured non-genetic variables that could confound genetic association tests. However, there is no guarantee that this will work well — environmental variables should be directly and robustly measured and incorporated into genetics studies.

What types of diversity should matter in genomics research?

We have pointed out that what types of diversity in data are important often goes unspecified, and that there are problems with the most common conceptualizations. Now we turn to our proposal for how researchers should identify what types of diversity in data are important. This starts with the acknowledgement that diversity in data is a means to achieve certain outcomes. On the clinical application side, progress towards equitable benefit for all from genomic medicine is the end to which diverse data is a means.[15] On the basic research side, diverse data can improve our understanding of the manifold contributions to complex traits and outcomes, through increasing power to detect effects, and making it more likely that effects detected are robust. Such advances can (though need not) be motivated by a particular type of impact, such as drug discovery, or policy development.

[1f] We also note that even when limited to considering diversity within genetic data, there is a distinction between diversity of observed genotypes, and diversity in how genomes are recombined over generations. The rate of mutations of individual variants is 100-times lower than the rate of recombination, and it is the latter which is most important for causal allele mapping.

Approved on February 14. 2024

Because the goals of each project are different, the types of diversity that are important will vary with each project.

We propose a way to identify what types of differences are relevant, and why: start with the goals of the project, ensuring these are in pursuit of ethical norms. Once the goals of a given project are appropriately delineated, the type of diversity in data needed to make progress towards those goals should become clearer. We give more specifics below in the "Proposal" section, but here note some corollaries of thinking of diversity in datasets as a means to achieve certain goals:

- **There is no one definition of "diversity" in datasets.** Rather, this will be project specific.
- **Equity will often be the most important concept.** Diversity in datasets is often desired in order to achieve, or at least make progress towards, equitable health outcomes. "Equity" is also not a concept that can be given a clear, context-free definition in this case. Researchers who appeal to the need for equity should attempt to spell out what they mean by that.
- **Diversity is not necessarily about representativeness.** There are likely to be projects where a sampling approach that leaves some groups of individuals under-represented increases progress towards the intended goals. For example, to maximize some genomic discovery projects, over-sampling individuals with recent genetic ancestries from Africa compared to the rest of the world may be needed. For a wide range of projects, a project that aimed for data that was representative would not achieve equity.
- **There continues to be a mandate for inclusive practices.** While what diversity means will vary by project, inclusion — i.e. making it as easy as possible for differently situated people to participate should they wish to do so — remains a constant mandate for every project.

We hence propose two questions that research teams should ask themselves:

Question 1a: What are the goals of the project, and how can these goals be framed in a way that maximally aligns with ethical norms?

Team members for every project — be it basic science, translational, or clinical research — should take the time to clarify the goal, or goals, of the project. As they do so, they should consider how these goals align, or fail to align, with generally recognized ethical norms, fundamental human rights and the improvement of healthcare. The goals should then be reframed in order to maximally align with these norms. Perhaps, for example, the goals as currently framed have implications that detract from one of the norms.

Which ethical norms matter? Most research in genetics and genomics is motivated by the prospect of the ethical principles of public benefit, that the research ultimately benefits patients. The focus on diversity in datasets highlights another consideration, that this benefit should be as

fairly distributed as possible, which would be in accord with the ethical principle of justice and fairness. There are a set of core principles that should be considered for all genetics and genomics projects: public beneficence, justice and fairness, and respect for individuals and communities. Research teams should take the time to consider if there are additional principles that they consider relevant in their particular case. It is particularly relevant to consider the norms that matter most to the communities studied.

Relevant human rights include rights to health, privacy, autonomy, and the right to science. The right to science is the ability of everyone to enjoy the benefits of scientific progress and its applications, and it is one of the principal ways to ensure realization of the right to health.[36] While the scope and content of the right has steadily evolved, its underlying principle provides a firm starting point for how researchers may better account for the impact their innovations have. It encompasses the need for informed consent, and suggests that where such consent is not possible, research should be carried out only for the person's health benefit or the health benefit of others. The particular needs of disadvantaged communities should further guide each stage of the research process, to ensure that the benefits of science, including access to medical innovations, are universally accessible.

Researchers should call on bioethicists and lawyers to help think through the norms that matter, and what the norms imply. They should also consult with the communities that will be most impacted by their research. For example, to ensure equal access to the benefits of science likely requires engagement with disadvantaged groups.[37] Clarity on the goals of the project increases the chances that the hoped for benefit will result. Conversely, a failure to consider how research could detract from ethical norms can (and has) led to harmful practices and mistrust in science. This delineation of goals that align with ethical norms defines the outcome(s) for which diversity in data is the means to achieve.

Question 1b: Given these goals, what type(s) of diversity in data are the most important?

The goals of the project should give an indication of the type(s) of diversity in data that are the most important. Team members should ask themselves, why is diversity in data important to achieve these goals? This should include consideration of who should be represented in the data, and what data points are most important to know about them. We note that sometimes it may not be clear what types of diversity are important for a given research question. If so, this needs to be transparently acknowledged. Even in these cases, starting with what is known in the literature may give indications of what dimensions of diversity are likely to matter.

Approved on February 14. 2024

How can diversity in data be achieved?

In the preceding section we discussed *what* data is needed to enable the benefits that calls for diversity in data are designed to achieve. We now turn to *how* this should be achieved. This starts with a discussion of why the field of genomics suffers from a lack of diverse data. These reasons are not the same across settings — in particular there may be differences between LMICs and high income countries — though there are also many commonalities.

Complex reasons for a lack of diversity

All of biomedical research suffers from a lack of diversity in research participation. This is connected to embedded inequities in the scientific enterprise. The reasons for a lack of diversity include ongoing sociocultural and economic barriers to research participation, such as childcare responsibilities, transport availability, cost, and availability of resources for scientific research amongst countries.[38] They also include language barriers, especially for informed consent documents and a lack of knowledge/understanding of local culture and religion on the part of researchers.[39] Additionally, continued inequities in health care access impact who gets diagnosed, and hence who is eligible for many studies.[40] Beyond these reasons, there have unfortunately been several instances where individuals — and particularly the most vulnerable members of society — have been used as a mere means to the end of scientific research. Famous examples include the U.S. Public Health Service's studies of Syphilis in Tuskegee[41] and of sexually transmitted disease in Guatemala.[42] Such instances inevitably erode trust in the scientific enterprise. The reasons for mistrust can be context-specific, and must therefore be addressed in a context-specific way.

Genomic researchers have not succeeded in fostering an environment of trust with equal success for all members of society, negatively impacting engagement from individuals who may otherwise have participated in research. During the 1990s, the most prominent project explicitly aimed at capturing “diversity”, the Human Genome Diversity Project, came to be dubbed as the “Vampire Project” by the World Congress of Indigenous Peoples.[43] This surprised the researchers involved, who did not foresee that their efforts to target indigenous populations (which they viewed as representing “pure” populations) would be viewed in this way.[44] To give just one other example, DNA was taken without consent for studies on schizophrenia and migration from the Havasupai Tribe, resulting in a lawsuit and the Navajo Nation placing a moratorium on genetic research studies.[2] In both cases, the groups sampled were viewed as harboring information valuable to others, and the projects were designed to extract this information, without consideration of benefiting the communities sampled.[42] Such examples give reason for individuals to be fearful not just to harms they may experience themselves directly, but to indirect harms to the groups they identify with.

Approved on February 14. 2024

Diverse data efforts have to be contextualized by such examples and within the wider historical setting of structural racism and colonialism.[19] As noted above, genetics has a long entanglement with racist belief systems. Without such contextualization, we will not be able to acknowledge and understand current problems with how we conceptualize, design, and carry out genetics and genomics research.[46]

Additionally, in some countries, genetic discrimination is a concern.[47] For example, in the US, any genetic results that are returned as part of research must then be disclosed to insurers, and life insurers are allowed to use genetic test results to decline coverage or set higher premiums.[45] While this is a concern that affects everyone, it can disproportionately affect some communities.

Legal barriers and uncertainty about the law can also inhibit diverse data efforts. For example, a growing number of African countries have enacted data privacy statutes that create legal barriers to genomic data sharing – especially cross-border. Given the relative novelty of these statutes, there is also uncertainty regarding exactly how their legal rules should apply. There are academic efforts underway to create more legal certainty in this regard in Africa.[49]

The lack of diversity in the genomics workforce has likely also affected recruitment of participants into studies. It has been demonstrated that research staff with similar race and ethnicity to that of prospective study participants helps with enrolment,[50] yet racial and ethnic minorities are underrepresented in research. The lack of diversity in research leadership is particularly important.[51]

Finally, in the early 2000s and even 2010s it was felt to be best practice to include only genetically homogeneous groups in research studies, as including mixed ancestry groups would cause population stratification or reduce power if analyzed separately. For analysis techniques based on linkage disequilibrium, such as GWAS, genomes of non-African ancestry were much easier to represent and analyze, which resulted in a self-perpetuating cycle of commercially available GWAS microarrays representing non-Europeans much less well and control datasets being mainly European ancestry.[52]

How should these types of diversity in data be pursued?

Once it is clear what types of diversity ought to be present in the data for a given project, attention to the above reasons is needed to support data gathering. In this section, we emphasize that it is not only important who is in datasets, nor what data is gathered about each of them. Rather, beyond data gathering, to achieve the ends for which diversity in data is a means will require strategies to be consciously adopted throughout the data life cycle. For example, there may be relevant considerations around data sharing, or around how data is processed. The choices we make with regard to the language and concepts we employ are also particularly important. We also emphasize that consistent attention is needed on lawful, contextually

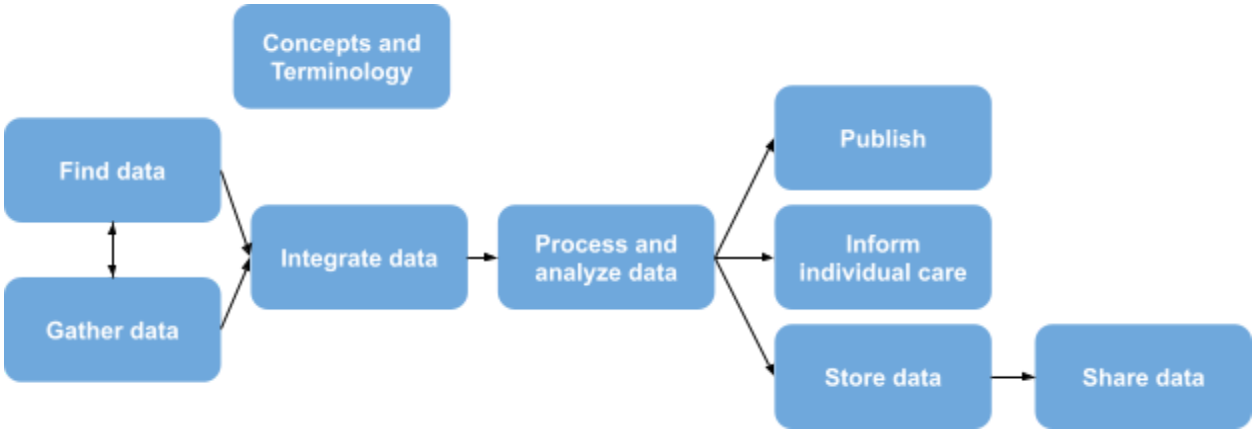
Approved on February 14. 2024

appropriate benefit sharing throughout these stages. We hence propose a question that researchers ask themselves two further questions.

Question 2a: What are the implications for choices made throughout the data lifecycle?

Armed with some degree of clarity about the types of diversity in data needed to achieve the project goals (via answering Questions 1a and 1b), teams should then consider what options they need to pursue at decision points throughout the data lifecycle, see Figure 2. We emphasize that not all stages will be relevant to all projects, for example "gather data" is not relevant for secondary users of data. There are some considerations that will hold across many (but not all) projects, see Table 1**.** There are also considerations that will be unique to the given project, which we take up in the Examples section.

Figure 2: Stages of the data life cycle. *Note that not all stages may be relevant to all projects. The arrows capture when decisions made at one stage influence other stages; there may be additional arrows for some projects.*



We draw attention in particular to the importance of data sharing. Across every branch of genomic medicine and for many types of genetic research, results are optimized by integrating data from as many individuals as possible, requiring data sharing across institutions and jurisdictional borders. A lack of sharing of already generated data contributes in a major way to the lack of diversity in available data.[53] However, such sharing needs to be done responsibly and in accordance with human rights: setting the foundation of the Global Alliance for Genomics and Health.[54] There are both legal restrictions and ethical guidelines that limit what data can be shared. These limitations necessitate careful consideration and awareness, especially since they

may be increasing in some countries.[55,56] [2f] It is important to understand what data is most important to prioritize for sharing, to maximize the benefits often called for when the need for diversity in datasets is highlighted. For example, where it is not possible to share all, or even most, de-identified individual level data, there may be particular aspects of the data that could be prioritized as shareable, or sought under the consent framework, to maximize the benefits of aggregating diverse data. Data access can also be federated and limited in its query scope (single rare variant variant queries to obtain an associated phenotype) to minimize risk yet allow some benefit.

We also draw attention to principles and practices that indigenous communities have highlighted and pioneered, in particular the sovereign-based governance models created by these communities. This includes notions of data stewardship and control by the communities from whom samples are taken.[45] By providing such communities with the tools to engage in research development, and the questions posed with their data, sovereignty enables disadvantaged communities to have influence in the research cycle. Tools such as the Biocultural Labels project are designed to help enable such control.[57] Diversity in data, however, is also about inclusion and must go beyond the research lifecycle to include the broader context in which policies have been created. Disadvantaged communities, including members of indigenous communities, sexual minorities etc. are often left out of the data governance process. To this end, it is critical that the needs of particular communities guide data governance frameworks.

Throughout all the stages of the data lifecycle, researchers need to consistently pay attention to lawful, contextually appropriate benefit sharing, prompting the final question we propose researchers ask themselves.

Question 2b: How can lawful, contextually appropriate benefit sharing be supported at all the relevant stages?

Although there is no authoritative definition of benefit sharing — whether it is with participants, with local communities, and with local researchers — it seems to have wide support in the bioethics literature.[58,59] Many have argued that benefit sharing should span across different domains. Bedeker et al identify these as including: Financial Health and Wellbeing, Infrastructure, Equipment, Skills capacity, Services capacity, Career development and attribution/recognition, Knowledge.[60]

[2f] For example, new data protection legislation has been passed since 2018 in at least 7 African countries: Botswana (Data Protection Act 32 of 2018), Kenya (Data Protection Act 24, 2019 and The Data Protection (General) Regulations, 2021), Nigeria (Nigerian Data Protection Regulations (NDPR) 2019; Nigerian Data Protection Regulations Implementation Framework 2020), Rwanda (Law N° 058/2021 of 13/10/2021 relating to the Protection of Personal Data and Privacy), Tanzania (The Personal Data Protection Act No. 11 of 2022 was passed in November 2022), Uganda (The Data Protection and Privacy Act 9 of 2019), and Zimbabwe (Data Protection Act [Chapter 11:22] No. 5 of 2021)

In particular, it has been recognised that equity cannot be achieved without building research capacity and equitable partnerships globally.[15] This includes avoiding 'helicopter research,' understood as when researchers from wealthy countries undertake research in LMICs without working with the local researchers as equal partners, or without sharing results and any benefits of the outcomes with the individuals and communities being studied.[60] There will inevitably be differences in the responsibilities and priorities of those in developed countries and LMICs.

However, the long-term social consequences of benefit sharing within recipient communities have not been studied. It is important to recognize that some well-meaning benefit-sharing efforts can be harmful in a number of ways. For example, providing local students financial opportunities to pursue PhDs in genetics may lead to them emigrating if there are no in-country opportunities for their onward careers. Another example is if an international research institution does research in a specific LMIC, and provides high-value benefits to a given community. If a local university wants to conduct research in the same community, the community representatives have high expectations for benefit sharing that the local university has no budget for, and hence may not be able to conduct research in its own country.

Additionally, benefit sharing with research participants may be unlawful. This is already the case in a number of African countries, including three countries that are prominent in international genomics projects, namely South Africa, Rwanda, and Nigeria. For example, in South Africa, providing financial or other rewards to research participants for donating biospecimens is legal only if it is for reimbursement of reasonable costs incurred.[61] Accordingly, when it comes to research participants, a focus on return of value to research participants may in some countries be unlawful.[62]

It is vital that researchers know how the law of a jurisdiction in which they intend to operate will apply to their intended research. To enable this, researchers should work with a local research partner responsible for compliance, and consider processes for sharing and discussing applicable norms, including resources to support local norms assessments if needed. In assessing local laws, particular attention should be given to: health law (with relation to gathering samples, ethics clearance, what kinds of benefit sharing are allowed, if any, etc.), privacy law (with relation to the rights of research participants qua data subjects, criteria for de-identification/anonymization, etc.), and property law (with relation to ownership of samples and data, etc.). If local laws do not legally permit benefit sharing with research participants, an appropriate course of action for the local research community may be to lobby to change the law.[61] Conversely, if local laws permit benefit sharing with research participants, the local research community may have good reasons to lobby for more restrictive laws, depending on local context and culture. International partners should respect local political processes.

Table 1: Considerations for both Questions 2a and 2b broadly applicable to all projects. *We note that not all projects will have the ability to influence all the decisions listed here.*

Approved on February 14. 2024

Stage of data life cycle (see Figure 2)	Implications for choices at this stage, including lawful, contextually appropriate benefit sharing
Concepts and terminology	Avoid reference just to “ancestry”, and instead refer to “genetic ancestry” if this is in fact what is meant. Avoid use of “population” terminology. Most terms used to describe a ‘population’ or ‘population category’ such as race or ancestry have a wide variety of meanings under different contexts. If the appropriate term is “sample”, this should be used. If not, the term “group” is preferred, because this invites the question “why were these individuals grouped together?” Engage with the communities contributing samples to understand the appropriate and most informative method of referring to samples or groups and to inform the determination of how to label samples.
Find data	Attempt to identify datasets that adequately capture the different dimensions of diversity identified as important. Attempt to use all data from existing selected sources, rather than excluding individuals. This can often help address the goals of a project. Attention should be paid to legal and regulatory frameworks the data is subject to. Concepts such as data ownership, data stewardship, data sovereignty and control, are ways that govern data use once gathered.
Gather data	Identify culturally and contextually appropriate ways to recognize the value provided by participants, to the extent that this is allowed by the relevant local legal framework Avoid helicopter science, whereby developed country researchers extract data from LMIC research teams and communities they sample from without providing anything of lasting value in return. Identify the barriers to participation, especially for those who are already underrepresented, and take steps to reduce these (e.g. enable ‘at home’ donation, reimburse travel costs, provide childcare).

Approved on February 14. 2024

	The goal is to ensure non-participation should be a free and informed choice. Aim to have researchers/nurses/technicians represent the communities being recruited. This helps ensure that local understanding of the diversity within these communities is not misrepresented. Ensure adequate skills capacity training for these individuals
Integrate data	Integrating data across multiple cohorts can give a natural way to increase diversity across many dimensions, and tools and practices that enable this should be encouraged Data harmonization may be required when integrating datasets if relevant factors are measured differently (or not at all) in different data sources. Efforts should be made to retain the fidelity of the source information, and to not inappropriately collapse or equate different concepts.
Process and Analyze data	Wherever possible, genetic ancestry or genetic similarity should be operationalized in a continuous way, i.e. without imposing categories.
Store data	Carefully consider where data should be stored, i.e. in host country, home country, both or something else? Considerations of data sovereignty may be applicable. Data should always be stored alongside appropriate metadata, as this can include key information about relevant dimensions of diversity. See for example the Biocultural labels project.[57]
Share data	Focus on enabling sharing of the data needed to achieve the project's goals (which will likely not be “all data”). Consider the most appropriate ways to share data, e.g. allowing federated access, allowing query of limited elements of the data for discrete purposes, sharing knowledge extracted from data (ClinVar submissions of classified variants, aggregate allele frequencies in gnomAD, etc). Obtain a professional legal opinion on whether the intended data sharing will be legal in the jurisdiction of the data subjects.

Approved on February 14. 2024

	Community engagement is required to understand prospective data donors' attitudes to data sharing, including cultural sensitivity, level of education/knowledge, histories, etc.
Publish	Carefully state how results are expected to generalize or fail to generalize Research that centers diversity in data can bring too much attention to the small amount of genetic variation that divides us. It can be used to tell over-simplified stories about human groups, and their histories.[63] And it can contribute to the stigmatization of already marginalized groups.[64] Researchers need to think through future potential nefarious uses of their work (this applies to all stages of the research lifecycle, but is particularly relevant at the publication stage). For all these reasons, great care is needed in how results are reported.
Inform individual care	Be aware that patients and health care providers will conflate continental ancestry categories and racial categories

Two worked examples

Example 1: Minimizing reporting false positive pathogenic variants

Around 80% of all rare diseases are genetic, and most of these diseases are monogenic/Mendelian.[65,66] Variant interpretation is a key step in returning genetic diagnoses to individuals suffering from rare diseases, and the results of variant interpretation directly impact clinical decision-making, and patient outcomes. A key piece of information in interpreting variants is knowledge of whether a given variant is found in unaffected individuals. Originally, it was proposed that a variant frequency of 5% was sufficient evidence that a variant was benign; now it has been proposed that a much lower frequency cut-off is reasonable,[67] and this cut-off is now integrated into the most referenced genomic database.[68] Evidence that a variant is at or above this frequency cut-off in ANY appropriately defined group is taken as evidence towards the variant being benign no matter who harbors it. In the absence of reference frequency data suitable for a given individual, a benign variant can look very rare, and any spurious evidence that it may be associated with disease can lead to a misdiagnosis, for example high rates of false positives have been reported in those “for ancestry groups that have historically been less well

studied" leading to inappropriate medical interventions and erroneous identification of at risk family members.[16,69] [3f]

Question 1a. What are the goals of the project, and how can these be framed in a way that maximally aligns with ethical norms? To help in the achievement of equity, understood in terms of equal likelihood for all humans that variants that are not causative of disease are not falsely identified as such.

This project's goals are clearly aligned with the principles of both public beneficence and justice and fairness. The former because reporting false positive variants can cause harm, and the latter because of the transparent connection to equity. Respect for communities is also upheld via not simply grouping people by continent, but instead focusing on the genetic diversity that makes a meaningful difference to the particular task at hand.

Question 1b. Given these goals, what type(s) of diversity in data are the most important? As outlined above, it is important to identify if a variant occurs at a frequency inconsistent with being implicated in rare disease. Diverse data can help in enabling the definition and calculation of a large set of frequencies suitable for this type of filtering. What matters here is the structure of the overall human family tree — due to human demographic history, the processes of genetic drift[5] and to a much lesser extent natural selection, alleles are not spread uniformly across our human species. This motivates defining the groups in genetic terms, rather than in terms of socially created population labels that may accompany data. The question is, how can these frequencies be defined in a way that allows for maximal identification of evidence they are benign? Defining frequencies based on the smallest number of individuals possible makes it easier to identify which variants are benign. Because for much reference data we have no guarantees that the individuals are healthy, it is also important to avoid misclassifying pathogenic variants as benign that have ended up at high frequency because of a bottle-neck that correlates to a high rate of disease. The challenge is hence to create frequency data defined within *relatively genetically-homogeneous groups, absent of bottle-necks***.** Importantly, we need to ensure that we have appropriate individuals in the data such that we can define such groups suitably for the whole human family tree. In particular, groups with higher rates of uncommon and rare variants, and variants not well represented in current databases.

Questions 2a and b. What are the implications for choices made throughout the data lifecycle? How can lawful, contextually appropriate benefit sharing be supported at all the relevant stages? See Table 2 for considerations, and note that responses in Table 1 are also relevant.

[3f] We have discussed above the way in which frequency evidence can help establish that a variant is not associated with a rare disease (i.e. is benign) through being too common. Frequency data can also be used as contributory evidence that a variant is associated with a rare disease (i.e. is pathogenic) through being very rare. We note that use of such evidence no longer counts so much towards this end, based on the observation that we all harbor many very rare variants..

Approved on February 14. 2024

Table 2: Considerations throughout the data lifecycle for Example 1

Stage of data life cycle	**Implications for choices at this stage, including contextually appropriate, lawful benefit sharing**
Concepts and terminology	Define, and appropriately label, genomically defined groups (see "Analyze data" below)
Generate data	In terms of prioritizing who to gather data from: ● Maximize the probability of assaying genomic variation which is currently absent or rare in known data sources. ● Operationalizing this requires understanding who is in current databases and how to identify the next 'best' person to sequence for use in a datasource (allele frequency and/or reference genome). ● This requires understanding the history of peoples and regions to understand how likely a given individual or group of people (whether defined by geography or ethno-linguistic affiliation) is to be related to those currently in such datasources Be upfront with potential research populations about the reasons they were selected, the importance of data sharing (both to their own community and to patients more generally)
Integrate data	Need to support the role of centralized portals, chiefly GnomAD
Process and Analyze data	Encourage efforts to work out how to suitably define groups using genomic information Carefully decide how to label the groups so defined. This could be done just by listing the inclusion criteria for those sampled. Another option, offered as a possible strategy by the National Academies population descriptors committee, would be to compare to a reference population, in the following manner: "1KG-YRI-like individuals", see their report for details.
Share data	Streamline data sharing of the minimal necessary information to enable producing allele frequencies [Could we provide some standard text to projects just starting?]
Publish	Clearly explain how the groups on which the frequencies were based are defined, why these were chosen, and why it is important not to over interpret them.

Approved on February 14. 2024

Example 2: The clinical impact of differential predictive performance of polygenic scores

Polygenic scores (PGS), sometimes called Polygenic Risk Scores, aggregate association signals between genotypes and a given trait or condition. PGS do not capture all the genetic variation that increases the presence of a trait or condition, but they do capture some amount of predictive signal. PGS for common, complex conditions — such as diabetes, heart disease, and common cancers — have thus been proposed as a useful tool for preventative medicine. The amount of predictive signal captured by a PGS for a given trait or condition, as assessed by the amount of variance of the trait or condition in a group that can be explained by the PGS, depends on how similar the group is compared to the groups used to construct the PGS (often termed the transferability or "portability problem").[17] Largely because most existing data is from individuals of predominantly European genetics ancestries, PGS currently perform best for these individuals. Recent work shows that the performance of PGS falls off in a continuous fashion with genetic distance from the training data.[70] Predictive performance of PGS can vary by more than just genetic ancestry/genetic distance. For example, this performance can vary by sex, age, and socioeconomic status.[71] It is known that major genomic data sets are skewed in terms of sociodemographic variables.[72] A major reason for this differential performance is because the effect sizes of the associations between genetic variants and traits that form the input to PGS are sensitive to non-genetic factors. (This stands in contrast to whether or not a given variant can be robustly associated with a trait; this tends to be constant across samples.) If a diversity of environments is not represented in the training data, it is not clear how well the PGS will perform for those who experience substantially different environments. Complicating the picture, environmental variables can covary with genetic ancestry/genetic distance. We consider a project designed to make progress on this issue of differential predictive performance as a second illustrative example.[73] [4f]

Question 1a. What are the goals of the project, and how can these be framed in a way that maximally aligns with ethical norms? To make progress towards achieving comparable predictive accuracy of PGS across suitably defined groups, understood as comparable proportion of total variance of the trait/outcome explained. Exactly similar predictive ability should not be expected, because sometimes genetics will be less predictive because there may be a different overall fraction of environmental variation that contributes to the variance of the trait/outcome.

This goal is clearly aligned with the ethical principle of justice and fairness. It is, however, important to stress that this is a very limited definition of equity targeted at the question of the datasets that would be needed. Equity in terms of potential benefit from the clinical use of PGS

[4f] Of note, the gold standard for sharing PGS for a given condition in a clinical setting would be to contextualize it, ideally in an integrated score, alongside other information that is relevant to that individual's risk of developing the condition (for example, age, sex, smoking status, blood pressure). These other factors are often more important than genetic effects in determining overall risk.

will depend on many other factors, for example, access to PGS testing, infrastructure and training needed to conduct PGS testing, and access to recommended follow up actions. These factors are likely to have a larger impact on health and healthcare disparities than the differences in predictive performance.

Question 1b. Given these goals, what type(s) of diversity in data are the most important? PGS performance should be assessed across multiple dimensions of difference. One key dimension of difference is genetic ancestry (which should be assessed in a continuous fashion),[70] but other dimensions — which could include e.g. sex, socioeconomic status, geographical location — are also important to assess. Identifying which dimensions of difference are important for a given trait or condition would be a first task of researchers. This should start with a review of which factors are known to influence who develops the condition. Ensuring diversity of these factors — or at a minimum, well-evidenced and justifiable proxies for these — would be an obvious starting point.[5f] Because where someone lives and has lived are good proxies for many such environmental variables, they are examples of dimensions of difference that should often be considered.

Questions 2a and b: What are the implications for choices made throughout the data lifecycle? How can lawful, contextually appropriate benefit sharing be supported at all the relevant stages? See Table 3 for considerations, and note that responses in Table 1 are also relevant.

Table 3: Considerations throughout the data lifecycle for Example 2

Stage of data life cycle	**Implications for choices at this stage, including contextually appropriate, lawful benefit sharing**
Concepts and terminology	Identify the dimensions of difference that are important for the condition or trait
Find data	GWAS data: describe, as far as possible, how the data captures the different dimensions of difference PGS validation data: identify data that allows for assessing the performance of the PGS along as many of the dimensions of difference as possible
Generate data	GWAS data: capturing data representing the dimensions of difference is ideal, though this should be balanced against the ability to generate large enough sample sizes to detect any effects. Because diversity in genetic

[5f] In addition to helping with assessing and achieving equitable predictive performance, a wider diversity of types of data (non-genetic) could help solve the interpretation problem of PGS, whereby we lack a good understanding of the nature of the predictive signal captured by the scores.

	ancestry is known to have a very large impact on PGS performance, diversity along this dimension should be prioritized.[17] PGS Validation data: capture variability along as many of the identified dimensions of difference (or good proxies for them), as possible. This will likely include geographic location.
Integrate data	Ensure preservation of sample level data, to enable consideration of as many of the relevant dimensions of difference as possible
Process and Analyze data	Ensure patterns of genetic similarity are controlled for in adjusting “raw” PRS[74] Provide performance metrics across each dimension of difference. Address limitations if any of the data that permit this are missing. Methodological development is needed to: ● Compute dimensions of difference across a range of factors deemed important for a study ● Provide access to easy-to-use software that enables an assessment of how genetic ancestry/distance effects PGS performance (and hence avoiding using genetically defined categories) ● Improve assessment of GWAS effect sizes in groups with recent genetic admixture[17]
Store data	Accuracy of PGS relies on closer integration of genetic and non-genetic data, increasing risks of identification. This risk needs to be assessed, mediated, and accurately communicated to individuals
Share data	
Publish	Researchers should be transparent about what data representation - in terms of the dimensions of difference - was included in the GWAS data and in the PGS validation data Researchers should contextualize predictive performance (i.e. amount of variance explained) against other known variables that predict the trait or condition. This should include all the variables along the identified dimensions of difference.
Inform individual care	If a published PGS is used in a clinical report, many further considerations apply for those developing a clinical report. Consider

	community/stakeholder engagement to establish appropriate normative practices.[75]

Acknowledgements

This Policy was developed by the Diversity in Datasets Subgroup of the GA4GH Regulatory and Ethics Work Stream, and is the result of the collaborative work, comments, and input of many individual and organizational contributors.

Bibliography

1. Cox, D. The search for the world's 'missing' genomes. https://www.bbc.com/future/article/20230227-the-search-for-the-worlds-missing-genomes.
2. de Villiers, C. B. Bringing diversity to the reference genome. *PHG Foundation* https://www.phgfoundation.org/blog/bringing-diversity-to-the-reference-genome.
3. Fatumo, S. *et al.* A roadmap to increase diversity in genomic studies. *Nat Med* **28**, 243–250 (2022).
4. CMS & Wellcome Trust. *Data Diversity in Human Genomics Global Landscaping*. 1–8 https://cms.wellcome.org/sites/default/files/2023-01/RFP-Data-Diversity-Genomics-Global-Landscaping.pdf (2023).
5. National Human Genome Research Institute Home, NHGRI. Population Architecture Using Genomics and Epidemiology (PAGE) Consortium. https://www.genome.gov/.
6. National Institutes of Health. All of Us Research Program | The Future of Health Begins with You. *All of Us Research Program | NIH* https://allofus.nih.gov/future-health-begins-all-us (2020).
7. Singapore National Precision Medicine Program. SG10K_Health. *SG10K_Health* https://www.npm.sg/collaborate/partners/sg10k/.

Approved on February 14. 2024

8. CanPath. Canadian Partnership for Tomorrow's Health: A Study. *Canadian Partnership for Tomorrow's Health* https://canpath.ca/about/.
9. H3Africa. Human Heredity & Health in Africa. https://h3africa.org/.
10. Our Future Health. Why diversity is essential to our mission. *Why Diversity is Essential to our Mission* https://ourfuturehealth.org.uk/our-research-mission/why-diversity-is-essential-to-mission/ (2023).
11. Genomics England. Diverse Data. *Genomics England* https://www.genomicsengland.co.uk/initiatives/diverse-data (2023).
12. Nature Editorial. Genetics for all. *Nat Genet* **51**, 579–579 (2019).
13. Wojcik, G. L. *et al.* Genetic analyses of diverse populations improves discovery for complex traits. *Nature* **570**, 514–518 (2019).
14. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The Missing Diversity in Human Genetic Studies. *Cell* **177**, 26–31 (2019).
15. Jooma, S., Hahn, M. J., Hindorff, L. A. & Bonham, V. L. Defining and Achieving Health Equity in Genomic Medicine. *Ethnicity & Disease* **29**, 173 (2019).
16. Manrai, A. K. *et al.* Genetic Misdiagnoses and the Potential for Health Disparities. *The New England Journal of Medicine* (2016) doi:10.1056/NEJMsa1507092.
17. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet* **51**, 584–591 (2019).
18. Gross, A. S., Harry, A. C., Clifton, C. S. & Della Pasqua, O. Clinical trial diversity: An opportunity for improved insight into the determinants of variability in drug response. *Br J Clin Pharmacol* **88**, 2700–2717 (2022).
19. Hardcastle, F. *et al.* Ethical, Legal and Social Issues in Diversifying Genomic Data: Literature

Approved on February 14. 2024

Review and Synthesis. Preprint at https://doi.org/10.31235/osf.io/myjtd (2022).

20. PRIMED Consortium. About PRIMED. https://primedconsortium.org/about/primed.

21. Bentley, A. R., Callier, S. & Rotimi, C. N. Diversity and inclusion in genomic research: why the uneven progress? *J Community Genet* **8**, 255–266 (2017).

22. Gaziano, J. M., Concato, J., Brophy, M. & Fiore, L. Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *Journal of Clinical Epidemiology* **70**, 214–223 (2016).

23. Epstein, S. *Inclusion: The Politics of Difference in Medical Research*. (University of Chicago Press, 2009).

24. Lewis, A. C. F. *et al.* Getting genetic ancestry right for science and society. *Science* (2022) doi:10.1126/science.abm7530.

25. Lewis, A. C. F. *et al.* An Ethical Framework for Research Using Genetic Ancestry. *pbm* **66**, 225–248 (2023).

26. Saini, A. *Superior: The Return of Race Science*. (Beacon Press, 2019).

27. Molteni, M. Buffalo shooting ignites a debate over the role of genetics researchers in white supremacist ideology. *STAT* https://www.statnews.com/2022/05/23/buffalo-shooting-ignites-debate-genetics-researchers-in-white-supremacist-ideology/ (2022).

28. Dauda, B. *et al.* Ancestry: How researchers use it and what they mean by it. *Front. Genet.* **14**, (2023).

29. Prieto-Garcia, J. M., Graham, L., Alkhabbaz, O. & Mazzari, A. L. D. A. Potential Pharmacokinetic Interactions of Common Cardiovascular Drugs and Selected European and Latin American Herbal Medicines: A Scoping Review. *Plants* **12**, 623 (2023).

30. Mathieson, I. & Scally, A. What is ancestry? *PLOS Genetics* **16**, e1008624 (2020).

Approved on February 14. 2024

31. Bliss, C. Conceptualizing Race in the Genomic Age. *Hastings Center Report* **50**, S15–S22 (2020).

32. Commission on the Social Determinants of Health. *Closing the gap in a generation: health equity through action on the social determinants of health*. https://www.who.int/publications-detail-redirect/WHO-IER-CSDH-08.1 (2008).

33. Marmot, M. Status syndrome. *Significance* **1**, 150–154 (2004).

34. Edinburgh Friends. The heart attack gender gap. *The University of Edinburgh* https://www.ed.ac.uk/edinburgh-friends/supplements/the-heart-attack-gender-gap (2019).

35. Machlitt-Northen, S., Keers, R., Munroe, P. B., Howard, D. M. & Pluess, M. Polygenic risk scores for schizophrenia and major depression are associated with socio-economic indicators of adversity in two British community samples. *Transl Psychiatry* **12**, 1–9 (2022).

36. United Nations General Assembly. *International Covenant on Economic, Social and Cultural Rights*. Art.15 https://www.refworld.org/docid/3ae6b36c0.html (1966).

37. Committee on Economic Social and Cultural Rights. *General comment No. 25 (2020) on science and economic, social and cultural rights (article 15 (1) (b), (2), (3) and (4) of the International Covenant on Economic, Social and Cultural Rights)*. https://docstore.ohchr.org/SelfServices/FilesHandler.ashx?enc=4slQ6QSmlBEDzFEovLCuW1a0Szab0oXTdImnsJZZVQdxONLLLJiuI8wRmVtR5Kxx73i0Uz0k13FeZiqChAWHKFuBqp%2B4RaxfUzqSAfyZYAR%2Fq7sqC7AHRa48PPRRALHB (2020).

38. Mapes, B. M. *et al.* Diversity and inclusion for the All of Us research program: A scoping review. *PLoS ONE* **15**, e0234962 (2020).

39. Council For International Organizations Of Medical Sciences. *Clinical research in resource-limited settings*. (CIOMS, 2021).

40. Centers for Disease Control and Prevention. Toward Achieving Health Equity for People

Approved on February 14. 2024

with Genetic Disorders. https://blogs.cdc.gov/genomics/2021/06/28/toward-achieving-health/ (2021).

41. Center for Disease Control and Prevention. The Syphilis Study at Tuskegee Timeline. https://www.cdc.gov/tuskegee/timeline.htm (2022).

42. Spector-Bagdady, K. & Lombardo, P. A. U.S. Public Health Service STD Experiments in Guatemala (1946-1948) and Their Aftermath. *Ethics Hum Res* **41**, 29–34 (2019).

43. Harry, D. Indigenous Peoples and Gene Disputes. *Chicago-Kent Law Review* **84**, 47 (2009).

44. Reardon, J. *Race to the Finish*. (Princeton University Press, 2004).

45. Garrison, N. A. *et al.* Genomic Research Through an Indigenous Lens: Understanding the Expectations. *Annual Review of Genomics and Human Genetics* **20**, 495–517 (2019).

46. Popejoy, A. B. *et al.* The clinical imperative for inclusivity: Race, ethnicity, and ancestry (REA) in genomics. *Human mutation* **39**, (2018).

47. Joly, Y. & Dalpe, G. Genetic discrimination still casts a large shadow in 2022. *Eur J Hum Genet* **30**, 1320–1322 (2022).

48. Chapman, C. R., Mehta, K. S., Parent, B. & Caplan, A. L. Genetic discrimination: emerging ethical challenges in the context of advancing technology. *J Law Biosci* **7**, lsz016 (2019).

49. DS-I Africa. Providing legal guidance to health researchers in Africa. *Providing Legal Guidance to Health Researchers in Africa* https://www.datalaw.africa/ (2023).

50. George, S., Duran, N. & Norris, K. A systematic review of barriers and facilitators to minority research participation among African Americans, Latinos, Asian Americans, and Pacific Islanders. *Am J Public Health* **104**, e16-31 (2014).

51. Forrester, N. Diversity in science: next steps for research group leaders. *Nature* **585**, S65–S67 (2020).

Approved on February 14. 2024

52. Hindorff, L. A. *et al.* Prioritizing diversity in human genomics research. *Nat Rev Genet* **19**, 175–185 (2018).

53. Conley, J. M., Cook-Deegan, R. & Lázaro-Muñoz, G. MYRIAD AFTER MYRIAD: THE PROPRIETARY DATA DILEMMA. *North Carolina journal of law & technology* **15**, 597 (2014).

54. GA4GH. Genetic Discrimination: Implications for Data Sharing Projects (GeDI) (Ver. 1.0). (2022).

55. Republic of Kenya. *Data Protection Act No. 24 of 2019*. (2019).

56. Communications Regulatory Authority of Botswana. *Data Protection Act*. (2018).

57. Anderson, J. & Hudson, M. The Biocultural Labels Initiative: Supporting Indigenous rights in data derived from genetic resources. *Biodiversity Information Science and Standards* **4**, e59230 (2020).

58. Schroeder, D. Benefit sharing: it's time for a definition. *Journal of Medical Ethics* **33**, 205 (2007).

59. Schroeder, D. & Gefenas, E. Realizing benefit sharing – the case of post-study obligations. https://philpapers.org/rec/SCHRBS (2012).

60. Bedeker, A. *et al.* A framework for the promotion of ethical benefit sharing in health research. *BMJ Global Health* **7**, e008096 (2022).

61. Thaldar, D. & Shozi, B. Is benefit sharing with research participants lawful in South Africa? An unexplored question in the governance of genomics research. *Journal of Law and the Biosciences* **10**, lsad018 (2023).

62. Wilkins, C. H. *et al.* Understanding What Information Is Valued By Research Participants, And Why. *Health Aff (Millwood)* **38**, 399–407 (2019).

63. Terrell, J. E. 'Plug and Play' Genetics, Racial Migrations and Human History. *Scientific American Blog Network*

Approved on February 14. 2024

https://blogs.scientificamerican.com/observations/plug-and-play-genetics-racial-migrations-and-human-history/.

64. Nature Editorial. Science must respect the dignity and rights of all humans. *Nat Hum Behav* **6**, 1029–1031 (2022).

65. Wright, C. F., FitzPatrick, D. R. & Firth, H. V. Paediatric genomics: diagnosing rare disease in children. *Nat Rev Genet* **19**, 253–268 (2018).

66. Kmth, R. & M, T.-G. Genetic Modifiers and Rare Mendelian Disease. *Genes* **11**, (2020).

67. Whiffin, N. *et al.* Using high-resolution variant frequencies to empower clinical genome interpretation. *Genet Med* **19**, 1151–1158 (2017).

68. Broad Institute. Filtering allele frequency | gnomAD. https://gnomad.broadinstitute.org/help/faf.

69. ClinGen SVI Working Group. ClinGen Sequence Variant Interpretation Recommendation for PM2. (2020).

70. Yi Ding *et al.* Polygenic scoring accuracy varies across the genetic ancestry continuum in all human populations. *bioRxiv* 2022.09.28.509988 (2022) doi:10.1101/2022.09.28.509988.

71. Mostafavi, H. *et al.* Variable prediction accuracy of polygenic scores within an ancestry group. *Elife* **9**, e48376 (2020).

72. Fry, A. *et al.* Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am J Epidemiol* **186**, 1026–1034 (2017).

73. Abu-El-Haija, A. *et al.* The clinical application of polygenic risk scores: A points to consider statement of the American College of Medical Genetics and Genomics (ACMG). *Genetics in Medicine* **25**, 100803 (2023).

74. Casalino, S. *et al.* Genome screening, reporting, and genetic counseling for healthy

Approved on February 14. 2024

populations. *Hum Genet* **142**, 181–192 (2023).

75. Wand, H. *et al.* Re-envisioning community genetics: community empowerment in preventive genomics. *J Community Genet* (2023) doi:10.1007/s12687-023-00638-y.