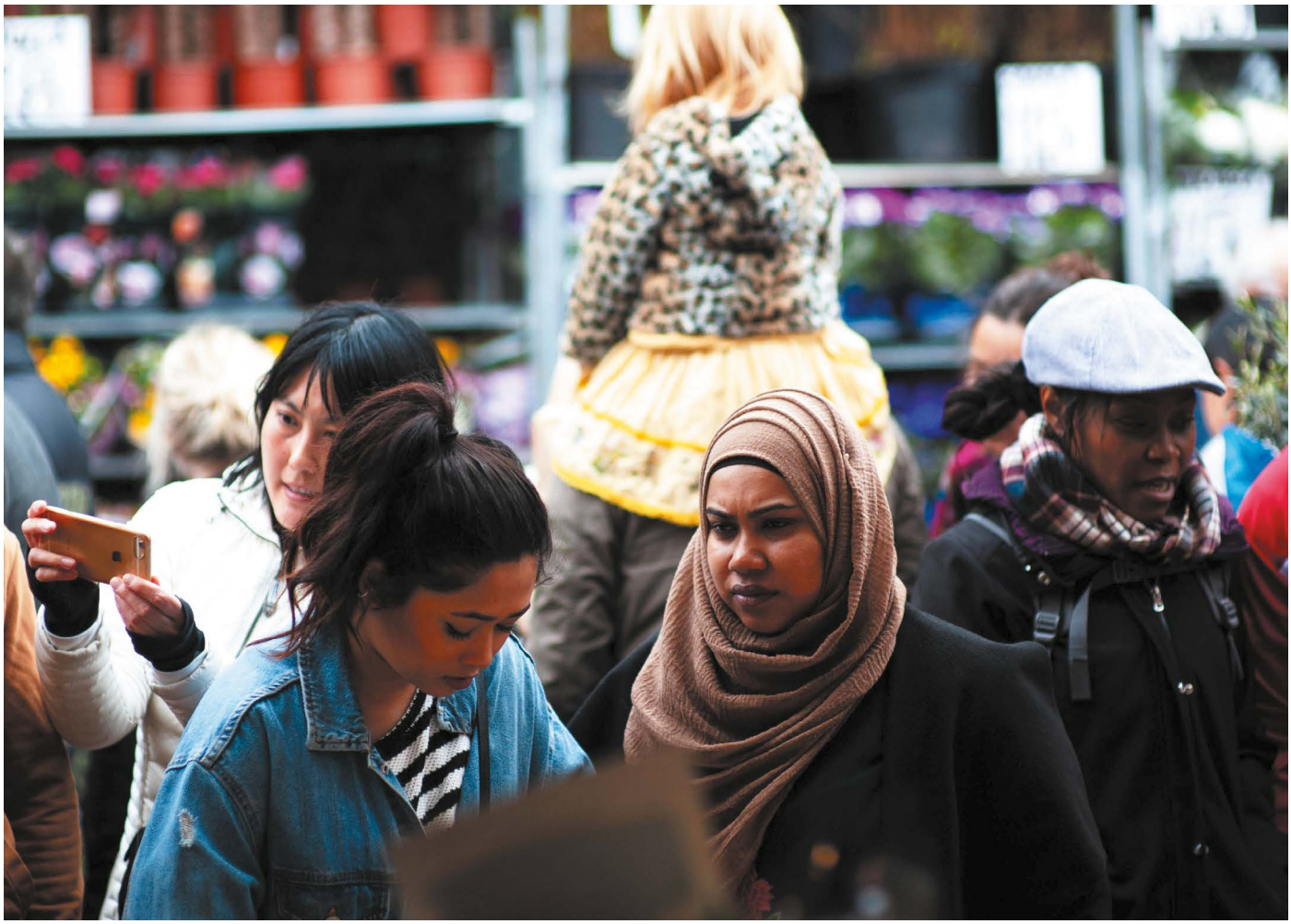
# Comment



ALAMY

Customers at Columbia Road Flower Market, UK.

# Don't ignore genetic data from minority populations

**Chief Ben-Eghan, Rosie Sun, Jose Sergio Hleap, Alex Diaz-Papkovich, Hans Markus Munter, Audrey V. Grant, Charles Dupras & Simon Gravel**

Efforts to build representative studies are defeated when scientists discard data from certain groups. Instead, researchers should work to balance statistical needs with fairness.

Geneticists have known for more than a decade that their focus on people with European ancestry exacerbates health disparities[1]. A 2018 analysis of studies looking for genetic variants associated with disease found that under-representation persists: 78% of study participants were of European ancestry, compared to 10% of Asian ancestry and 2% of African ancestry. Other ancestries each represented less than 1% of the total[2]. Several projects, such as H3Africa[3], are starting to increase participation of under-represented groups, both among participants and among researchers. Large biobanks assembled in Europe and North America, combining biological samples with health-related data, also set sampling targets to increase diversity[4,5,6].

But even when data from minority groups are available, many researchers discard them[7].

Although there can be valid reasons to restrict analyses to a particular population, discarding such data by default is ethically problematic: it worsens under-representation and negates participants' efforts to contribute to research.

Funding agencies have taken steps to improve the diversity of participants who are recruited for studies — notably, this has led to better representation of women in clinical trials since the 1990s. But agencies have less control over researchers' decisions of what to analyse. Scientists are pulled towards statistical convenience and publishing incentives, which can both conflict with the collective goal of greater equity.

Here we suggest that an approach used in health care can help researchers to make analysis decisions that are ethically as well as scientifically sound.

## Ruled out

To estimate how often minority data are excluded, we examined publications that used data from either the UK Biobank (UKB; which contains material from 502,655 individuals) or the US Health and Retirement Study (HRS; 12,454 individuals). Both biobanks support genome-wide association studies (GWAS). These scan data from thousands of participants to find genetic variants associated with disease.

To compare the criteria researchers used to include or exclude data types across studies, we distinguished between participants from majority (MAJ) and minority (MIN) groups in the United States and the United Kingdom. We used MAJ regardless of whether a study focused on self-declared ethnicity, such as 'white', or on the location of an individual's ancestors, such as 'European ancestry'. We used MIN to refer to all other individuals, including those of mixed ancestry or ethnicity. This coarse labelling helps to describe how data were used in statistical analyses, and does not imply that either group is uniform. We counted MIN data as 'included' if any analysis reported linking traits or diseases to genotypes in the relevant samples.

First, we reviewed 21 articles from the GWAS catalogue (www.ebi.ac.uk/gwas) that contained the keywords 'UK biobank' (see Supplementary information). Twenty restricted their analysis to only MAJ individuals in the UKB database (two of these also analysed data from a broader range of ancestries in other databases). We also queried online repositories and randomly sampled another 20 GWAS that used UKB data. Only one used MIN data. Finally, we reviewed 17 GWAS listed on the HRS online publications list. Here, only six studies limited analysis to MAJ populations, perhaps because the proportion of MIN participants in the US biobank (24%) was higher than in the UK one (5%).

Overall, 45 of 58 studies in our sample excluded MIN data. If we weight representation by the number of times data from an individual were actually analysed, MIN representation in the UKB falls to 0.06% (see 'Left out'; details are in Supplementary information). This problematic situation will surprise few genetics researchers[6,7].

> **"By omitting data, scientists squander an opportunity to build useful knowledge about minority populations."**

Both the UKB and the HRS made efforts to represent their national populations. However, including individuals from minority groups in data cohorts but not in analyses can be seen as de facto tokenism. Unused data do not help under-represented groups.

## Why exclude?

Of the 45 studies that excluded data, 31 gave no reason. The remaining 14 studies provided 15 explanations for exclusion.

The most common explanation was fear of confounding (11/15). If a genetic variant happens to be more common in an ancestry group, and that group happens to have a higher rate of a particular trait, there will be a correlation between having the variant and having the trait. An example is childhood asthma, which is influenced by both genetic and environmental factors. Researchers might confuse the correlation as evidence that this variant causes childhood asthma. Although statistical methods to avoid confounding exist, they are not foolproof, and confounding is a legitimate concern[7].

It is not necessary to exclude data to reduce the risk of confounding. Data from different groups can simply be analysed separately. However, because samples from minority populations are so much smaller, they have less statistical power and are therefore less likely to reveal new genetic associations.

This lack of power was the second-most-cited reason for exclusion (3/15). An underpowered study can be seen as a waste of time because it might not yield statistically significant results. Because finding a genetic association can be enough to garner a publication, adding analysis of other populations comes at a cost. It takes time, makes the manuscript more complicated, gives reviewers one more thing to criticize, and so could delay publication.

## Grounds for inclusion

**There is value in data from minority populations.**

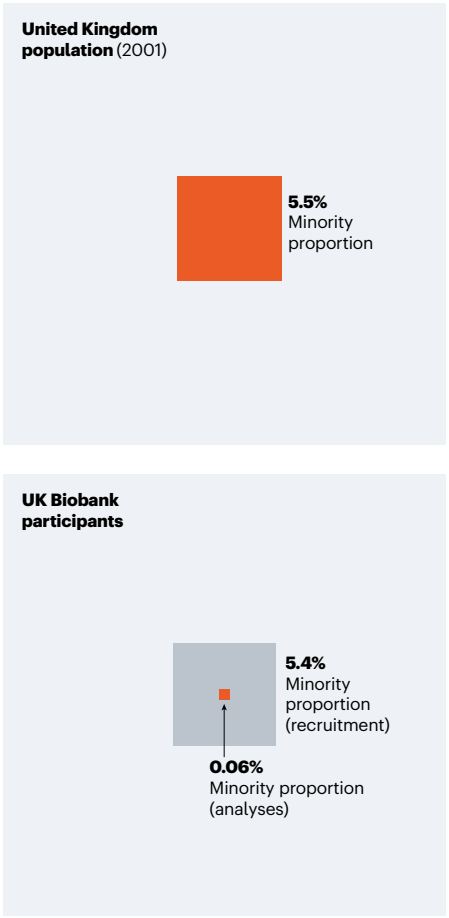
As part of a study on asthma, we performed a genome-wide association study for eosinophil cell counts. (Eosinophils are a subset of white blood cells and are often elevated in individuals with asthma.) We did three separate analyses. One was of the majority (MAJ) population; two were of the minority (MIN) populations defined using the UK Biobank self-reported ethnicity categories (participants who identified as Black or Black British, and those who identified as Asian, Asian British or Chinese).

The MAJ analyses identified 432 genetic loci (1,510 independent genetic variants). The two MIN analyses independently identified 3 loci (at genome-wide significance, $P \leq 5 \times 10^{-8}$), all of which were identified in the MAJ analysis. The MIN analysis enabled validation of more than one-quarter of the identified variants in the MAJ population at nominal significance ($P = 0.05$). It also showed overall consistent results across ethnicities, except for one variant that showed nominal significance, but opposite effects in Asian, Asian British and Chinese populations, relative to the MAJ analysis. Without further evidence, this variant should probably not be used to predict genetic risk outside Europe. (See Supplementary information for details.)

These analyses took 10 hours of computing time as well as some forethought. This is insignificant compared with the cost of accessing the data. Evidence of association for the millions of variants we tested can now be compared across populations and can be made available for meta-analyses. Such data are particularly important for studying minority populations, when samples in individual cohorts might lack statistical power.

**LEFT OUT**

UK biobank recruitment reflected diversity (in 2001; ref. 11). Analyses do not.




SOURCE: ANALYSIS BY C. BEN-EGHAN *ET AL.*

Just one study explicitly mentioned following methods from past publications as grounds for exclusion (1/15), but we suspect that this is common. There are good reasons to follow precedent: using standard analytical pipelines reduces development cost and the need for extensive validation and explanation.

Together, these three reasons drive researchers to discard data from MIN populations.

## Lost opportunity

By omitting data, scientists squander an opportunity to build useful knowledge about minority populations. If researchers perform GWAS on populations of European ancestry, they can often use previously published results in the form of summary statistics to strengthen their findings. Because summary statistics present little privacy risk to participants, they can usually be downloaded freely in just a few minutes. Doing the same comparison with MIN population data that have not been previously reported requires accessing individual-level information. This involves obtaining institutional ethics approval, requesting data access from the cohort, plus cleaning and processing data — all before finally performing GWAS. This can take months. If MIN data are not analysed alongside MAJ data, they might never be used.

When done as part of the primary study, by contrast, MIN analyses add little cost and can be informative (see 'Grounds for inclusion').

## Four criteria

Analysing MIN data is important for equity and discovery. But how should we weigh that against the immediate, individual burden of statistical analysis and delayed publication? General rules that apply to all studies are hard to define, but there is an approach that should help.

Over the past two decades, governments and ethicists have leant on a framework called accountability for reasonableness (A4R) to help allocate scarce resources in health care, such as new or expensive treatments. A4R recognizes that individuals in a pluralistic, democratic society give different weight to different considerations, and so might never agree on broad principles. Instead, A4R focuses on the decision-making process itself, and sets out criteria that encourage fairness and legitimacy[8]. In short, reasons for decisions should be transparent and relevant. Adherence to these criteria should be enforced and measured in a way that adapts to new information.

The A4R criteria suggest small changes in analysis and publication conventions that would improve fairness and accountability.

**Transparency.** In their publications, researchers should state reasons for excluding participant data. More generally, they should explain design and analysis choices that have the potential to worsen inequalities.

**Relevance.** The stated reasons for exclusion should explain how the decision sought to best serve society, given the real-world constraints of research. Reasons such as fear of confounding, limited power and precedent might not meet this requirement if they can be circumvented by a particular analysis method (using stratified or meta-analysis, for instance). Barring more compelling reasons, we recommend that researchers compute association statistics for MIN populations and report them as part of the primary study.

**Enforcement.** We propose that journals mandate that submitted manuscripts justify any exclusion of participant data in analyses. Forms should ask reviewers whether relevant reasons were provided.

The goal is not to turn reviewers into moral arbiters. Rather, they should simply assess whether the reasons provided are relevant to the analyses under review. This modest requirement would encourage analyses to be more inclusive, foster broader discussion about legitimate grounds for exclusion and clarify expectations for authors.

Importantly, reviewers should not require results of analyses of MIN and MAJ populations to be consistent. Discrepancies should be discussed, but forcing researchers to explain all observations would prevent useful results from being shared.

**Revisions.** How researchers assess transparency and relevance should change with society and methodology. Our recommendations that data from MIN populations be analysed by default might become moot if sufficient data become available in cohorts that focus on under-represented groups[9]. The field might also move to a model in which specialized teams analyse MIN data across multiple phenotypes (see, for example, https://pan.ukbb.broadinstitute.org). This would change both the costs and benefits of performing subsequent analyses of MIN data. It could reduce the impetus for analysis by individual studies while providing tools that reduce the analysis burden and risk of confounding for subsequent researchers.

Statistical analyses that are more inclusive cannot overcome fundamental inequities in representation among study participants, let alone solve the broader issues of equity and data sovereignty[10]. But they are a step in the right direction. By acknowledging the tension between ethical and practical considerations, researchers in genetics and other fields can hold themselves accountable for making scientific advances more efficient and more fair.

## The authors

**Chief Ben-Eghan**, **Rosie Sun**, **Hans Markus Munter**, **Alex Diaz-Papkovich**, **Audrey V. Grant** and **Simon Gravel** are genomics researchers and **Charles Dupras** is an ethicist at McGill University in Montreal, Quebec, Canada. **Jose Sergio Hleap** is a high-performance-computing technical consultant in bioinformatics at SHARCNET in Guelph, Ontario, Canada.
e-mail: simon.gravel@mcgill.ca

1. Bustamante, C. D., De La Vega, F. M. & Burchard, E. G. *Nature* **475**, 163–165 (2011).
2. Sirugo, G., Williams, S. M. & Tishkoff, S. A. *Cell* **177**, 26–31 (2019).
3. Mulder, N. *et al. Pharmacogenomics Pers. Med.* **11**, 59–66 (2018).
4. Ollier, W., Sprosen, T. & Peakman, T. *Pharmacogenomics* **6**, 639–646 (2005).
5. Sonnega, A. *et al. Int. J. Epidemiol.* **43**, 576–585 (2014).
6. Tutton, R. *Race/Ethnicity: Multidiscip. Glob. Contexts* **3**, 75–95 (2009).
7. Peterson, R. E. *et al. Cell* **179**, 589–603 (2019).
8. Daniels, N. & Sabin, J. E. *Br. Med. J.* **337**, a1850 (2008).
9. Wojcik, G. L. *et al. Nature* **570**, 514–518 (2019).
10. Fox, K. *N. Engl. J. Med.* **383**, 411–413 (2020).
11. Fry, A. *et al. Am. J. Epidemiol.* **186**, 1026–1034 (2017).

Supplementary information accompanies this article (see go.nature.com/2dwmgka).