HHS Public Access

Author manuscript

Lancet. Author manuscript; available in PMC 2021 January 23.

Published in final edited form as:

Lancet. 2019 March 30; 393(10178): 1297. doi:10.1016/S0140-6736(18)33067-8.

UK Biobank, big data, and the consequences of non-representativeness

Katherine M Keyes,

Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, NY 10032, USA

Center for Research on Society and Health, Universidad Mayor, Santiago, Chile

Daniel Westreich

Department of Epidemiology, UNC Gillings School of Global Public Health, Chapel Hill, NC, USA

UK Biobank is an unparalleled resource of extensive health information from 500 000 individuals and with more than 400 peer-reviewed publications to date. The sampling population is volunteer-based and is not representative of the UK population.[1] Investigators state that although the estimates of prevalence and incidence should be interpreted with caution, valid measures of association and estimates of causal effect can be more readily interpreted as they do "not require participants to be representative of the population at large".[2]

This statement is a puzzling claim: sample selection can indeed influence measures of association. Specifically, whether or not an association observed in a study is similar in some other target population (ie, has external validity) depends on a number of factors, including the distribution of effect measure modifiers of the exposure–outcome relationship in the study sample and target population.[3] Critically, a study can have restricted external validity even when it has internal validity, which might occur in a randomised trial.[4] Thus, researchers should not be quick to set aside issues of representativeness in interpreting UK Biobank results.

The situation can be illustrated with a numerical example. The appendix shows the unobservable truth with regard to an exposure–disease relationship in a target population of 66 million individuals (approximately the current population of the UK). Those exposed to X (a gene, or an environmental factor) have twice the risk of the outcome Y compared with those that are unexposed. Suppose further that X is only causal among people who are also exposed to a third factor A (eg, lack of exercise, present in 50% of the population). Among those with A, the exposed population have 3 times the risk compared with those who are unexposed; among those without A, there is no association between X and Y. We hypothetically recruit 500 000 participants (approximately the size of the UK Biobank

kmk2104@columbia.edu.

See Online for appendix

DW received personal fees for consulting from Sanofi-Pasteur within the past 24 months, outside of this work. KMK declares no competing interests.

sample). But, during recruitment, those unexposed to A (healthier volunteers) are more likely to join,[1] than those exposed to A are (by about 2:1, appendix). In the study sample, the risk ratio is 1·67 (appendix). The association between those exposed to X and the disease differs between the study sample and the target population because the prevalence of A differs; indeed, the association would differ in any population with a different prevalence of A. The magnitude of an exposure's association with an outcome depends on the prevalence of other factors that interact with the exposure; in our case, A interacts positively with X, and because the prevalence of A decreases, the magnitude of the association between X and Y also decreases.

When the target population is well defined and the relevant modifiers are measured, weighting methods[5] can be used to map the results in the study sample to match the target population. But increases in sample size alone cannot overcome selection issues. Indeed, larger sample size in a skewed sample only leads to confidence in answers that might not apply to the target population. Thus, it is paramount that external validity be taken more seriously in the UK Biobank and other large data resources.

We suggest that investigators in the UK Biobank and elsewhere consider these assumptions when making inference, and reconsider the idea that associations are generalisable to all possible target populations, or relevant to public health and clinical medicine, simply because the sample size is large.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by National Institute on Alcohol Abuse and Alcoholism K01AA021511 (KMK) and National Instititues of Health OD/NICHD DP2-HD-08-4070 (DW).

References

1. Fry A, Littlejohns TJ, Sudlow C, et al. Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. Am J Epidemiol 2017; 186: 1026–34. [PubMed: 28641372]
2. Biobank UK. Researchers. http://www.ukbiobank.ac.uk/scientists-3/ (accessed Dec 6, 2018).
3. Lesko CR, Buchanan AL, Westreich D, Edwards JK, Hudgens MG, Cole SR. Generalizing study results: a potential outcomes perspective. Epidemiology 2017; 28: 553–61. [PubMed: 28346267]
4. Westreich D, Edwards JK, Lesko CR, Cole SR, Stuart EA. Target validity and the hierarchy of study designs. Am J Epidemiol 2019; 188(2): 438–43. [PubMed: 30299451]
5. Westreich D, Edwards JK, Lesko CR, Stuart E, Cole SR. Transportability of trial results using inverse odds of sampling weights. Am J Epidemiol 2017; 186: 1010–14. [PubMed: 28535275]