THE LANCET

Supplementary appendix

This appendix formed part of the original submission. We post it as supplied by the authors.

Supplement to: Keyes KM, Westreich D. UK Biobank, big data, and the consequences of non-representativeness. *Lancet* 2019; **393:** 1297.

	Diseased (Y+)	Diseased (Y-)	Total	Risk ratio
Association between exposure (X) and outcome (Y) in the population				
Exposed (X+)	16 500 000	16 500 000	33 000 000	$\frac{\left[\frac{16\ 500\ 000}{33\ 000\ 000}\right]}{\left[\frac{8\ 250\ 000}{33\ 000\ 000}\right]} = 2 \cdot 0$
Unexposed (X-)	8 250 000	24 750 000	33 000 000	
Total	24 750 000	41 250 000	66 000 000	
Observed association in the volunteer sample, with a 1% sample from those unexposed to A, and 0·5% sample drawn from those exposed to A				
Exposed (X+)	103 125	144 375	247 500	$\frac{\left[\frac{103\ 125}{247\ 500}\right]}{\left[\frac{61\ 875}{247\ 500}\right]} = 1 \cdot 67$
Unexposed (X-)	61 875	185 625	247 500	
Total	165 000	330 000	495 000	

Table: **Association between exposure (X) and outcome (Y) in a hypothetical population of 66 million residents, and a study sample of approximately 500 000, with differential recruitment based on healthiness**