Science Translational Medicine
AAAS

Supplementary Materials for

Prospective study design and data analysis in U.K. Biobank

Naomi E. Allen *et al.*

Corresponding author:Naomi E. Allen, naomi.allen@ndph.ox.ac.uk

Sci. Transl. Med. **16**, eadf4428 (2024)
DOI: 10.1126/scitranslmed.adf4428

This PDF file includes:

Table S1. UK Biobank's design approach and potential analytical approaches to reduce error in exposure-disease associations

Design characteristic	Rationale and error to be addressed	UK Biobank's design approach	Potential analytical approaches
Large-scale prospective design	Exposure data collected prior to disease (i.e. to reduce recall bias and reverse causation bias)	Exposures measured at recruitment and participants' health followed up over time via linkage to longitudinal healthcare records	Perform longitudinal analyses to determine exposure-disease associations
	Large numbers of participants are needed to provide sufficient statistical power for reliable assessment of risk factors with health outcomes (i.e. to reduce random error)	Recruitment of 500,000 participants	Pool data and /or results with other studies to increase sample size
Participation rate and comparing cohort characteristics with that of the wider population	Comparison of results from UK Biobank with studies in other populations needed to determine generalizability of research findings	Postal invites sent to all 9.2M individuals age 40-69 years and living within travelling distance of an assessment centre	Control for factors associated with study participation and retention
	The study population should be sufficiently heterogeneous to include a wide range of risk factors under investigation to enable	Recruited participants with widely varying risk factor levels[1]	Use Directed Acyclic Graphs (DAGs) to investigate the underlying assumptions and to identify potential sources of bias of the association(s)

	generalizable assessments of exposure-disease associations High participant engagement needed to obtain high response rates for continued data collection activities (i.e. to reduce systematic error due to responder bias)		Perform sensitivity analyses to assess impact of missing data resulting from non-random participation Compare results with studies from different populations, including meta-analytical approaches that assess the impact of UK Biobank data on the overall findings
Reliable assessment of a wide range of exposures:			
Depth and breadth of exposure measurement	Comprehensive characterisation of participants' behavioural, environment and germline genome are needed to identify independent risk factors for disease (i.e. to reduce confounding) Data on exposures need to be complete and accurate to improve precision (i.e. reduce random and systematic error).	Comprehensive (i.e. cohort-wide) assessment of exposures (incl. genomics and other biomarkers) to reduce missing values of variables Standardised data collection protocol used to ensure data were collected accurately and in a consistent manner	Use of DAGs to clarify the presence and direction of potential confounders and mediators Adjust for multiple relevant factors Use of genetic causal inference models

		Sample assays performed in the full cohort at the same time facilitate quality control Supplemented crude measures with detailed objective assessments (e.g. accelerometer to assess physical activity)	Use different analytical approaches to triangulate evidence Use of simulations (e.g. probabilistic bias analysis) to assess likely impact of measurement error of the exposure and confounder(s) on the risk estimate Calibrate variables for the full cohort based on more precise measures performed in a subset
Repeated exposure measures	Repeated exposure measures are needed to enable accurate assessment of long-term average exposure (i.e. to reduce regression dilution bias).	Repeat assessments performed in subsets of the cohorts	Correct for regression dilution bias using repeated measures
Reliable assessment of a wide range of health outcomes:			
Comprehensive ascertainment of health outcomes	Passive cohort-wide collection of health outcomes is needed to minimise ascertainment	Cohort-wide linkage to routine administrative health records	Control for factors associated with differential ascertainment of health

	bias and reduce loss-to-follow-up (or attrition) bias.		outcomes that are based on participant characteristics
Specificity of health outcomes	Accurate ascertainment of outcomes (and their subtypes) needed to increase their specificity and positive predictive value (i.e. reduce random error associated with false-positives)	Detailed ascertainment using diagnostic codes and other biochemical, imaging data etc.	Perform subgroup analyses by disease sub-type, where these data are available Develop research agenda to implement novel approaches for accurate disease classification
Long duration of follow-up	Long-term follow-up of participants' health needed to enable assessment of temporality of associations (i.e. reduce reverse causation bias) and to accrue large numbers of incident disease (i.e. reduce random error)	Linkage to routine administrative health records since recruitment: ~15 years complete follow-up	Consider impact of exclusion/inclusion of prevalent disease cases on analysis Perform sensitivity analyses that exclude initial follow-up periods

[1] UK Biobank does not cover the full range of racial/ethnic diversity (given the study is based in the UK) or age (the study recruited individuals aged 40-69 years).