Understanding Society
THE UK HOUSEHOLD LONGITUDINAL STUDY

Understanding Society Biomarker and health data

Meena Kumari, Professor of Biological and Social Epidemiology
mkumari@essex.ac.uk
Health Users Conference
1st July 2024

University of Essex
ISER
UKRI Economic and Social Research Council

An initiative by the Economic and Social Research Council, with scientific leadership by the Institute for Social and Economic Research, University of Essex, and survey delivery by NatCen Social Research and Kantar Public

General overview

- Overview of *Understanding Society*
  - Currently available Health and biomarker data
  - Future data collection

- Data access processes

Special features of Understanding Society: health and biomarker

- Longitudinal health data
- All adults in household
- All UK countries
- All ages
- Ethnic minority boost
- Multi-topic

Topics areas

- Biomarker, Genetic and Epigenetics
- COVID-19
- Education
- Employment
- Ethnicity and Immigration
- Family and Household
- Money and Finances
- Politics and Social attitudes
- Health and Wellbeing
- Young people
- Transport and Environment

Health & wellbeing

Health data from two sources:

1. Questionnaire i.e. individual reports it, e.g.:

- Current prescriptions
- Currently has chronic condition
- GHQ measures mental health

2. Objective data measured by a nurse (in Waves 2/3, IP12):

- Height/weight
- Lung function, grip strength
- Blood analytes – for cholesterol, raised glucose, kidney & liver function, anaemia
- Genetic data

Adult and children mental well-being

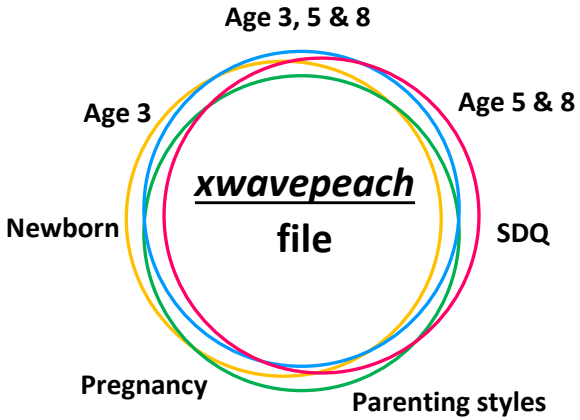
- Adults: Annual
  - General Health Questionnaire – 12 (continuous, categorical or caseness scales; individual items
  - Satisfaction
    - Domains: Overall life, job, leisure time, income, health
- Waves 1, 4, 7, 10, 13
  - Edinburgh-Warwick Positive well being (short form)
- Children: Annual
  - Happiness with different aspects of life (school work, appearance, family, friends, school, life as a whole - can be combined to create an overall happiness score)
- Biennially (start at wave 1)
  - Strengths and Difficulties Questionnaire (SDQ) - can create a total difficulties score or 5 subscales

Pregnancy and Early Childhood file (PEACH)

- Our primary goal is to facilitate researchers interested in child development topics to engage with Understanding Society data seamlessly.
- The file brings together most of the already available information on pregnancy and early childhood through mainstage, into a single location, enhancing its quality and value.
- It has been constructed by leveraging data from all children reported in the child file based on their eligibility for the child development questions.
- The information provided is at the child level, using the child's identifier (pidp) to ensure each row is uniquely identifiable.
- All variables within each category are time-specific and remain constant across all waves. *For instance, information such as the child's birth weight, the frequency of fussing or crying during infancy, and the frequency of reading to the child at ages 3, 5, and 8 are age-specific and remain constant over time.*
- A total of 303 variables are included, created from 164 questions.

Pregnancy and Early Childhood file

- Its structure is designed to facilitate easy tracking of various aspects of a child's development, including whether they were born prematurely or past their due date, their ability to speak in full sentences at age 3, and their health condition at ages 3, 5, and 8, amongst other variables.
- In addition, the file includes the cross-wave identifier of the person who provided the information for the child and the wave when it was collected.
- We have classified the information available about children into 7 categories, based on content and when these were asked.

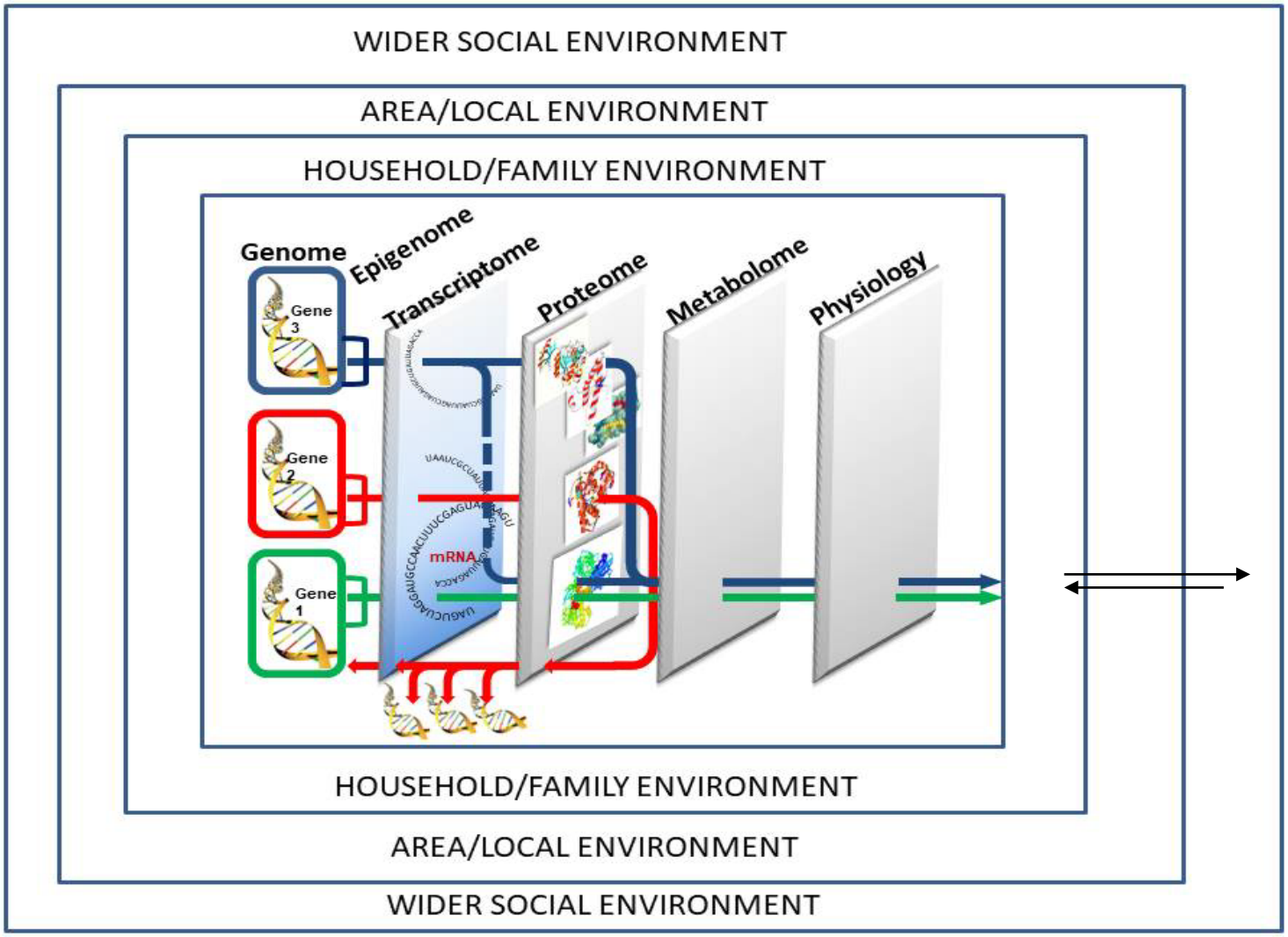


- Plans are to release an updated version of the file every year, after a new wave is released, as PEACH is not part of mainstage releases.

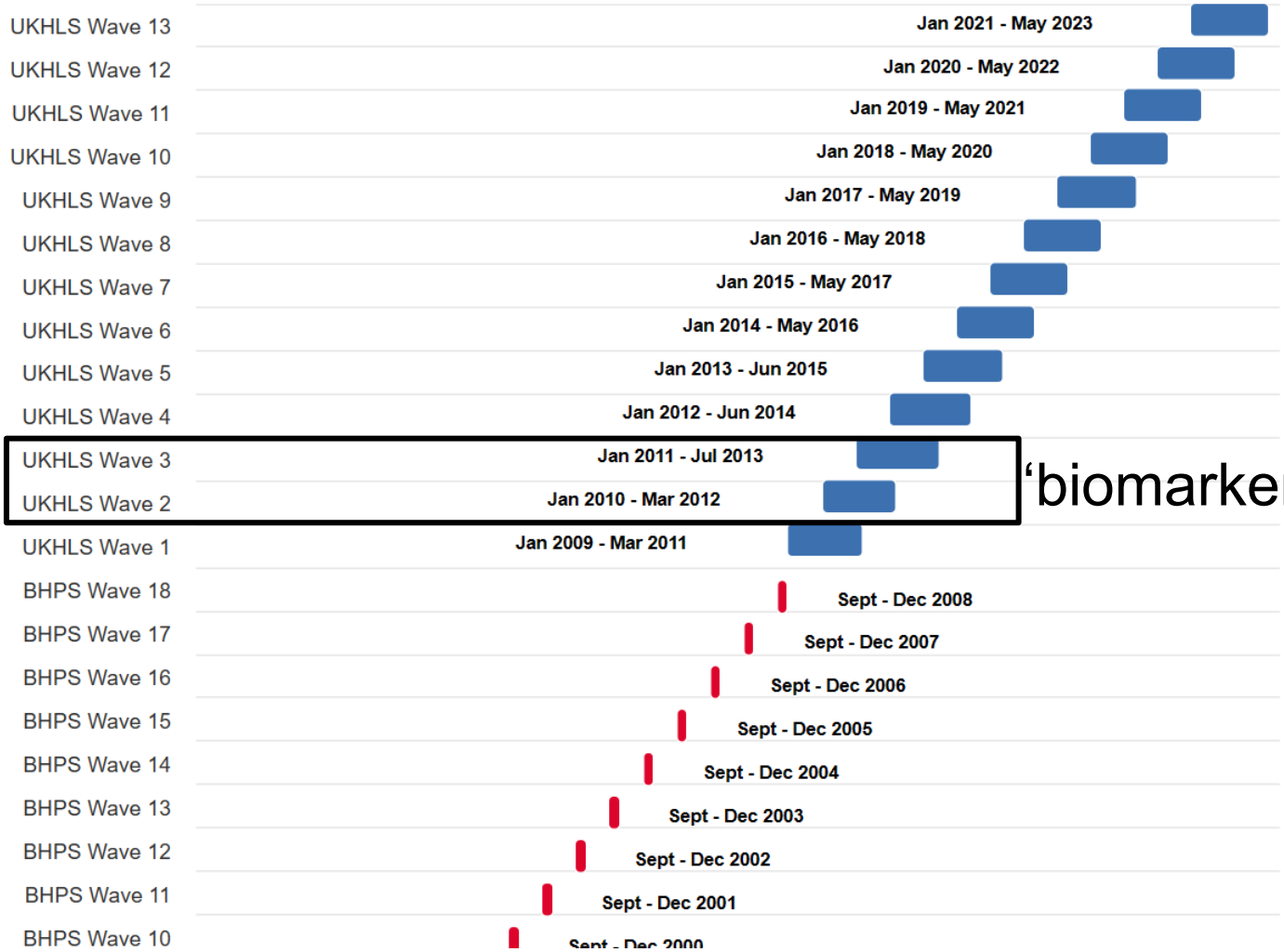
Adult health and well being

- SF-12
- GHQ-12
- Health behaviours
  - Smoking status
  - CAGE – alcohol consumption
  - IPAQ - Physical activity
  - ModifiedPSQI – sleep behaviours

Biomarker data: analytes



Understanding Society: Fieldwork Schedule



'biomarker wave'

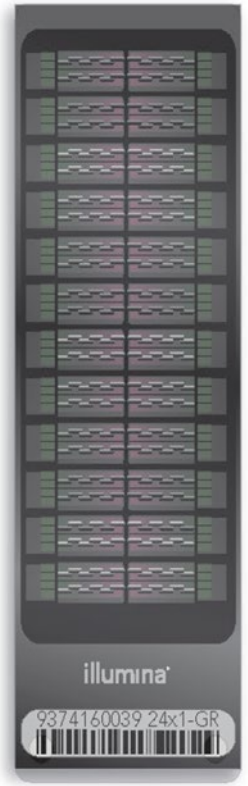
Understanding Society

Genetic Data

- **9,944** Individuals who also have biomarker data.
- **Population:** England, Scotland and Wales.

- Genotyped at the **Wellcome Trust Sanger Institute** (Genome Research)
- **Illumina** Human Core Exome BeadChip

- **>500,000 SNPs**
- **>8,000,000 Imputed SNPs**



56% 44%

wellcome trust

Eleftheria Zeggini
Karoline Kuckenbäcker
Bram Prins

www.illumina.com

-OMIC Data
DNA methylation

- Selected from those who had genetic Data
- Measures in two batches:
  - Batch 1: BHPS Blood samples processed within 1-2 days
  - Batch 2: remaining BHPS samples and random sample from GPS
- **Data Normalised and cleaned by Tyler Gorrie-Stone, Professor Leo Schalkwyk**

58% 42%

- **3105 samples**
- **Infinium MethylationEPIC BeadChip**
- **>850,000 methylation sites across the genome**
- **857,071 sites after QC**

UNIVERSITY OF EXETER

Professor Jonathon Mill
Dr Eilis Hannon
Dr Joe Burrage

-OMIC data: biomarkers of age

- DNA Methylation ‘clocks’:
- built with supervised machine learning methods, such as penalized regression (e.g., lasso or elastic net) trained against chronological age or other phenotypes of age to identify an informative and sparse predictive set of methylation sites

- Understanding Society has deposited 5 such biomarkers of age

DNA methylation age

First generation: *generated against age*

- Horvath*: used a large collection (n > 8000) of publically available Illumina HumanMethylation array data on multiple tissue types to train and test a model for age prediction from 353 CpG loci. (Horvath 2013)
- Hannum*: Blood based predictor from 666 White or Hispanic American adults (Hannum et al., 2013).

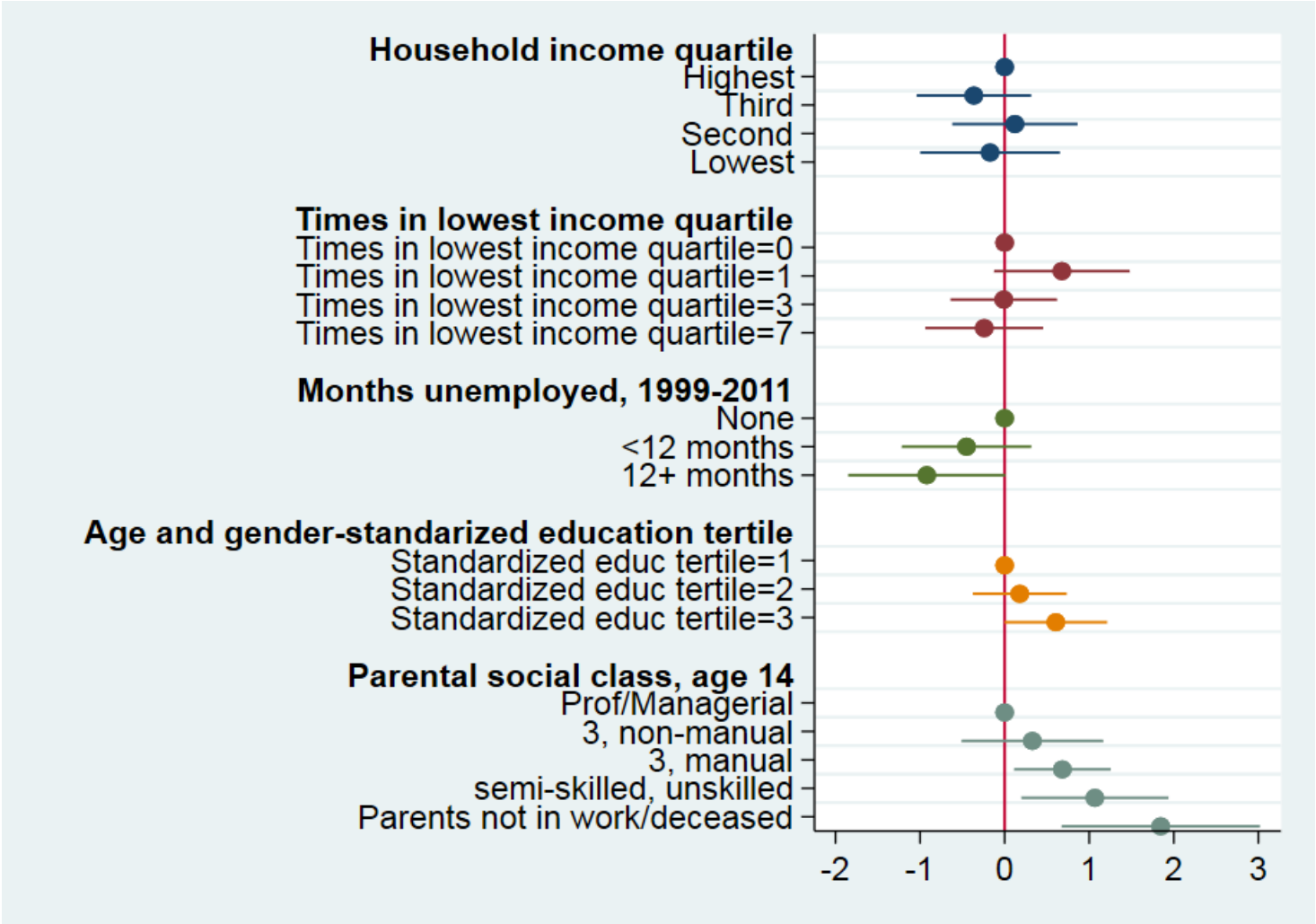
Second generation: *generated against mortality*

- Levine et al., 2019 (pheno*): Blood based algorithm estimated with chronological age and a 'phenotypic aging measure' which consists of nine clinical biomarkers (albumin, creatinine, glucose and C-reactive protein levels, lymphocyte percentage, mean cell volume, red blood cell distribution width, alkaline phosphatase and white blood cell count)

Second generation: generated against change in biomarkers – speed or Pace of age

- DunedinPoAm (2020)*; change-over-time in 18 biomarkers tracking organ-system integrity across 12 years
- DunedinPACE (2022); within-individual decline in 19 indicators of organ-system integrity across four time points spanning two decades
- (others: Weidener 2014; Bockland et al., 2011; Koch et al., 2011)

Education and father's social class and 'accelerated age'



Analyses adjusted for age, age2, batch, cell type, smoking status, BMI

From: Hughes et al., AJE 2018

-OMIC Data

Proteomics

Cardiometabolic

Complementing and expanding on our previous portfolio of disease area-focused biomarker panels, our exploratory panels offers scientists the chance to cast a wider net in the quest to identify new biomarkers and relevant protein signatures, offering many new assays focused on important biological processes with wide-ranging clinical relevance. assays in this panel include proteins involved in key biological processes such as *cellular metabolic process, cell adhesion, immune response* and *complement activation.*

https://www.olink.com

Neurology

The panel offers a mix of established markers related to neurobiological processes and neurological diseases (e.g. neural development, axon guidance, synaptic function, or specific conditions such as Alzheimer's disease), as well as some more exploratory proteins with broader roles in processes such as cellular regulation, immunology, development and metabolism. This balanced selection provides an ideal basis for protein biomarker discovery in the neurology area.

- 184 biomolecules that are NOT established measures of health

- Includes proteins related to
  - cell metabolism
  - cell adhesion
  - immune response
  - complement (an immune pathway)
  - neuron formation and function
  - neurodegenerative diseases

-OMIC Data
Proteomics

- 184 biomolecules that are NOT established measures of health

- Includes proteins related to
  - cell metabolism
  - cell adhesion
  - immune response
  - complement (an immune pathway)
  - neuron formation and function
  - neurodegenerative diseases

Exemplar analyses

Social differences in proteins: association with educational attainment

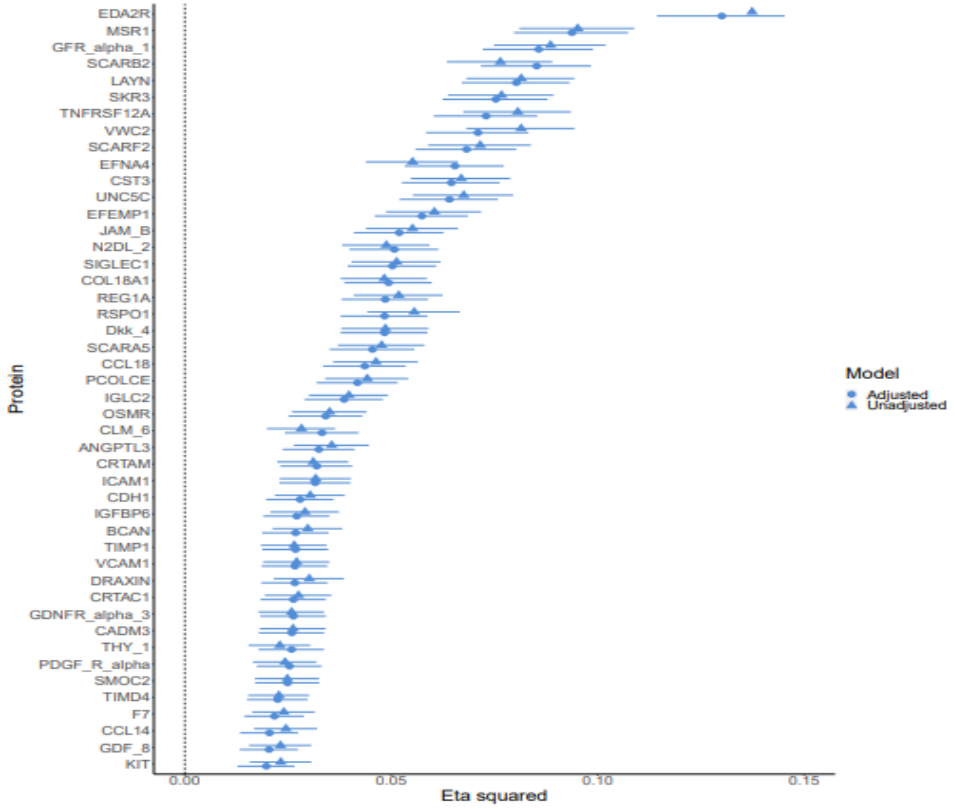


Figure 4: Associations between proteins and educational attainment

Understanding Society Data deposit and access procedure

- Managed Access
  - Genetic Data deposited in EGA
  - Access to genetic data only managed by EGA
  - Access to linked genotype and phenotype data managed by committee hosted by ESRC
    - www.understandingsociety.ac.uk

mkumari@essex.ac.uk

Websites

- www.data-archive.ac.uk
- Understandingsociety.ac.uk

- genetics@understandingsociety.ac.uk
- mkumari@essex.ac.uk