The CLS Genetics Data Resource

David Bann, Gemma Shireby and Tim Morris

CENTRE FOR LONGITUDINAL STUDIES

Economic and Social Research Council

CENTRE FOR LONGITUDINAL STUDIES

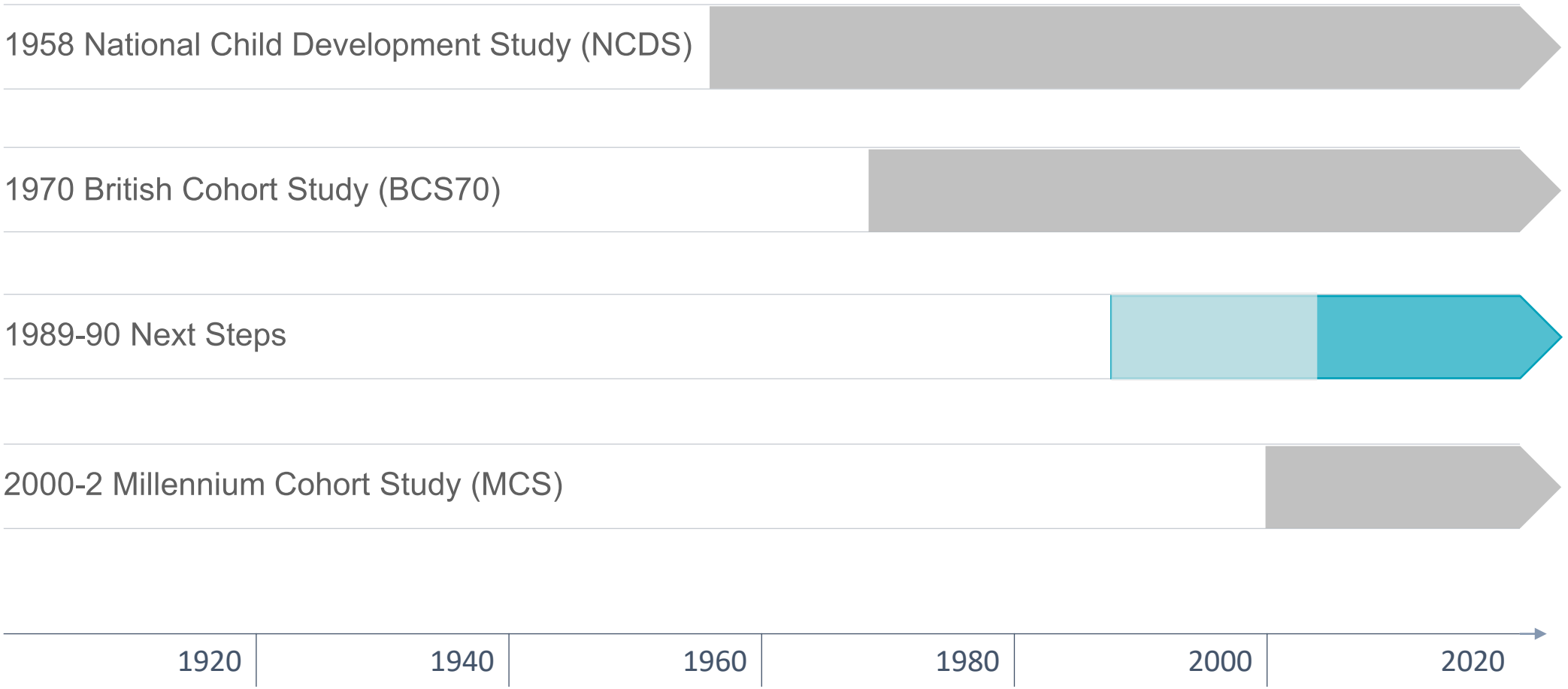
Housekeeping

- We are recording this session so it will be available online at a later date
- If you have a question, please use the chat function, and please note your question will be visible to all attendees
- Technical issues – please email us: ioe.clsevents@ucl.ac.uk
- We would be grateful for your feedback. Please follow the link in the chat at the end of the event for the short survey – we have also emailed this to you

Thank you for joining us today

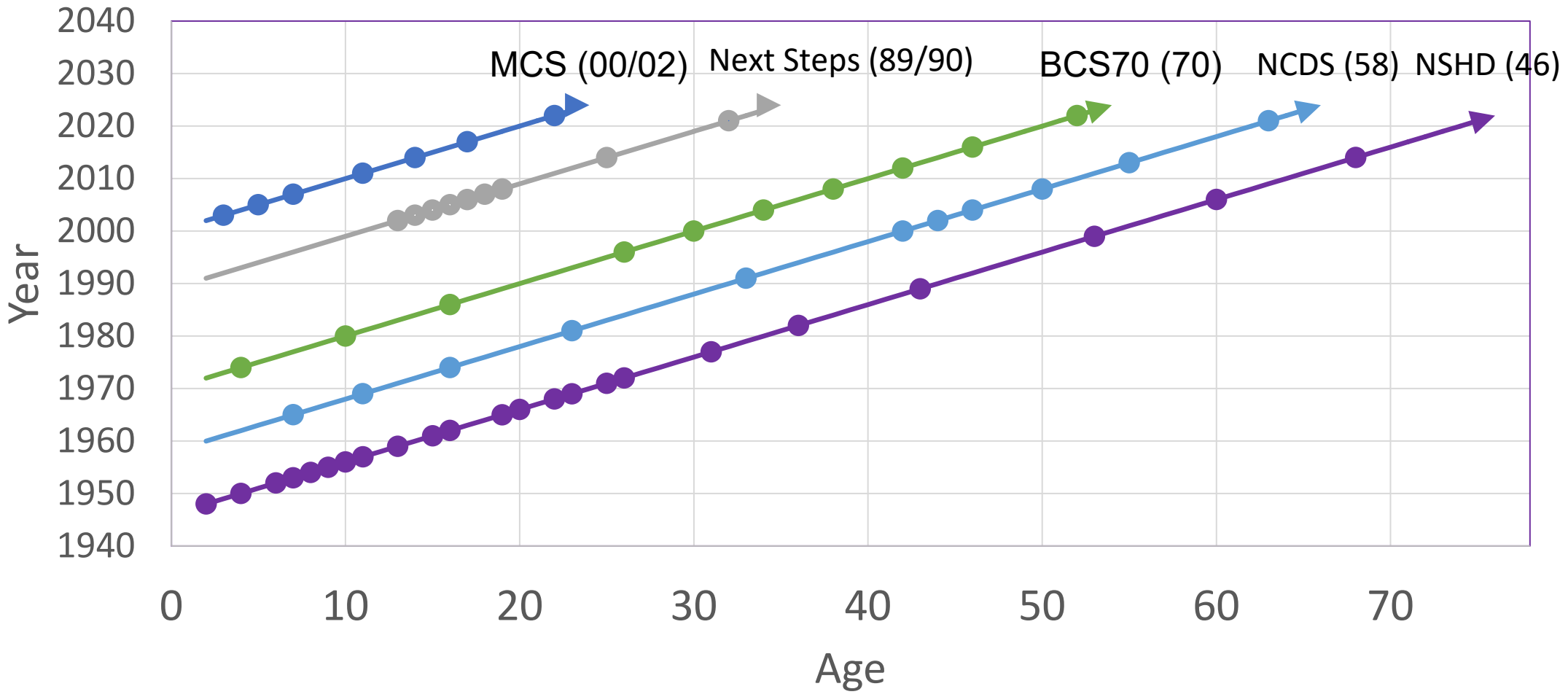
Centre for Longitudinal Studies (CLS) current core studies

CENTRE FOR LONGITUDINAL STUDIES



Study timelines and future 2020-2030

CENTRE FOR LONGITUDINAL STUDIES



Example: NCDS

A study of everyone born in one week in 1958 (GB)

CENTRE FOR LONGITUDINAL STUDIES

	1958	1965	1969	1974	1981	1991	2000	2003	2004	2008	2013
	Birth	7	11	16	23	33	42	44	46	50	55
main respondent	mother	parents	parents	cohort member / parents	cohort member	cohort member	cohort member	cohort member	cohort member	cohort member	cohort member
others		school	school	school		children (1 in 3)					
medical	medical exam	medical exam Ht/Wt	medical exam Ht/Wt	medical exam Ht/Wt	Ht/Wt	Ht/Wt		Ht/Wt blood - DNA biomedical		Ht/Wt	Ht/Wt
survey instruments		cognitive mental h.	cognitive mental h.	cognitive mental h.	mental h.	mental h.	mental h.			cognitive mental h.	
linked data				area of residence (census)	area of residence (census)					consent for health and economic records	
response rate	17,415	15,425	15,337	14,654	12,537	11,469	11,419	9,377	9,534	9,790	9,137

CENTRE FOR LONGITUDINAL STUDIES

Typical information covered

Birth	School years	Adult
Household composition	Household composition	Household composition
Parental socio-economic situation	Parental social class & education	Employment
Obstetric history	Parental employment	Social class
Smoking in pregnancy	Financial circumstances	Income and wealth
Pregnancy (problems, antenatal care)	Housing	Housing
Labour (length, pain relief, problems)	Family relationships	Family and partnership history
Birthweight, length	Health	Health (including biomarkers)
	Cognitive tests	Well-being and mental health
	Emotions and behaviour	Health-related behaviour
	School	Training and qualifications
	Views and expectations	Basic skills
	Attainment	Views and expectations

COVID-19 and serology surveys

Available via the UKDS (EUL)

https://cls.ucl.ac.uk/covid-19-survey/

COVID-19 surveys response					
	NCDS	BCS70	Next Steps	MCS CMs	MCS parents
Wave 1	5,178	4,223	1,907	2,645	2,831
Wave 2	6,282	5,320	3,664	3,274	5,707
Wave 3	6,809	5,758	4,239	4,474	5,251

Serology survey response					
	NCDS	BCS70	Next Steps	MCS CMs	MCS parents
Invited	6,939	6,594	4,826	5,266	7,143
Consented	4,156	3,741	2,090	1,397	3,214
Blood sample returned	3,222	2,547	1,267	1,140	2,266

Serology Survey:

- Participants who took part in one of three COVID-19 Surveys were invited to provide a finger-prick blood sample
- Two antibody tests conducted - N-assay and S-assay
  - N-assay more likely to identify naturally occurring antibodies through exposure to virus
  - S-assay more likely to identify antibodies occurring following vaccination
- Same antibody tests conducted in multiple longitudinal studies including ALSPAC, USoc, ELSA, TwinsUK and NSHD (1946 cohort), funded by National Core Studies.

https://cls.ucl.ac.uk/covid-19-survey/covid-19-antibody-testing/

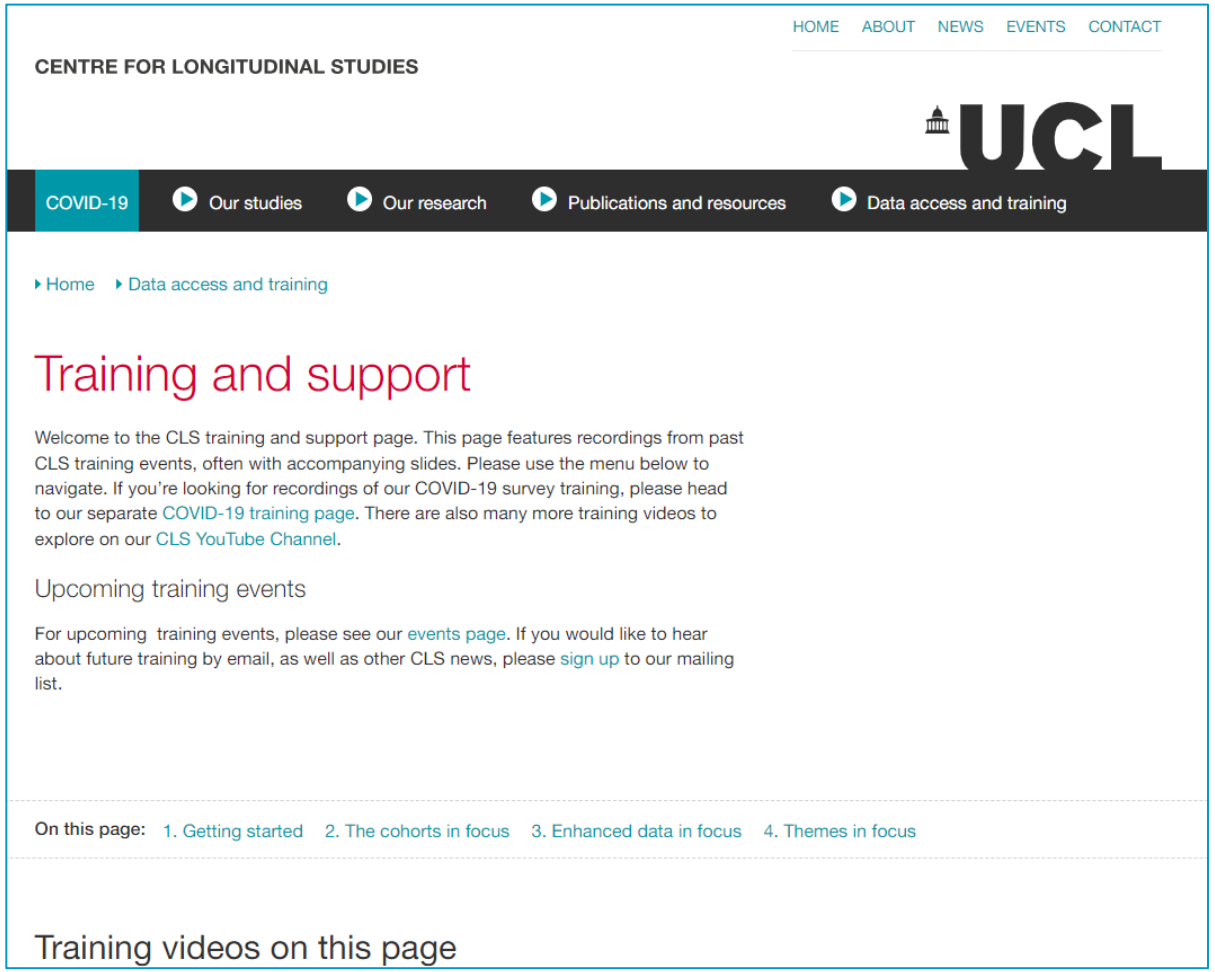
Linked administrative data in the cohorts https://cls.ucl.ac.uk/data-access-training/linked-data/

	Country	Study	Data set	Access
Health	England	**NCDS, BCS70**, Next Steps, MCS	**Hospital Episodes Statistics (HES)** • Admitted Patient Care (APC) • Critical Care (CC) – linked to APC • Accident & Emergency (A&E) • Outpatient Care (OP)	Available at UKDS (e.g. link) via Secure Lab
	Scotland		**Scottish Medical Records (SMR)**	Available at UKDS (e.g. link) via Secure Lab
		NCDS, **BCS70** MCS	• Inpatient, Outpatient, Prescribing information	
		NCDS, BCS70 only	• Maternity inpatient	
		MCS only	• Immunisation (SRS), Child Health Review , Birth and neonatal records	
	Wales	MCS	• Health data assets from SAIL Databank (e.g. emergency department, outpatient) up to age 14 and for CM's parents	Available at Secure Anonymised Information Linkage (SAIL)
			• Hospitalisations & no. of diagnoses from ICD-10 < age 11	Available at UKDS via Secure Lab
Education	England	Next steps, MCS Next Steps	• KS1 to KS4 • KS5, Individual Learner records (ILR), Student Loan Company (SLC)	Available at UKDS (e.g. link) via Secure Lab
	Scotland	MCS	• NPD KS1	Available at UKDS via Secure Lab
	Wales	MCS	• Welsh NPD KS1 To KS4, Post 16 education	Available at SAIL

Coming soon:

- HES data refresh in Next Steps, BCS70 and NCDS (beyond years 2017)
- Refresh of Welsh health dataset linked to MCS (up to age 14 and parents) UKDS and SAIL post age 14
- Mental health data in MCS, Next Steps, NCDS, BCS70 (Early 2025)

CLS training and support



HOME ABOUT NEWS EVENTS CONTACT

CENTRE FOR LONGITUDINAL STUDIES

UCL

COVID-19 | Our studies | Our research | Publications and resources | Data access and training

Home > Data access and training

Training and support

Welcome to the CLS training and support page. This page features recordings from past CLS training events, often with accompanying slides. Please use the menu below to navigate. If you're looking for recordings of our COVID-19 survey training, please head to our separate COVID-19 training page. There are also many more training videos to explore on our CLS YouTube Channel.

Upcoming training events

For upcoming training events, please see our events page. If you would like to hear about future training by email, as well as other CLS news, please sign up to our mailing list.

On this page: 1. Getting started 2. The cohorts in focus 3. Enhanced data in focus 4. Themes in focus

Training videos on this page

Upcoming training events	
Methods: Cross-cohort analyses	May 2024
Handling missing data in the BCS70	June 2024

https://cls.ucl.ac.uk/events/

Coming later in the year
Webinar: Polygenic scores in the cohorts

CENTRE FOR LONGITUDINAL STUDIES

https://cls.ucl.ac.uk/data-access-training/training-and-support-2/

The CLS Genetics Data Resource

Gemma Shireby
Genetics Data Manager/ Bioinformatician,
Centre for Longitudinal Studies, UCL Social Research Institute

CENTRE FOR
LONGITUDINAL
STUDIES

Economic
and Social
Research Council

Centre for longitudinal study (CLS) genomic datasets

NCDS

1958 National Child Development Study

Following the lives of 17,000 people born in a single week in 1958 in Great Britain.

BCS70

1970 British Cohort Study

Following the lives of 17,000 people born in a single week in 1970 in Great Britain.

Next Steps

Following the lives of 16,000 people in England born in 1989-90.

MCS

Millennium Cohort Study

The most recent of Britain's cohort studies, following 19,000 young people born in the UK at the start of the new century.

Data Availability

MCS:

- Genotype + imputed
- Whole exome

NCDS

- Genotype + imputed
- Exome
- Epigenetic

BCS70

- Genotype + imputed
- Epigenetic

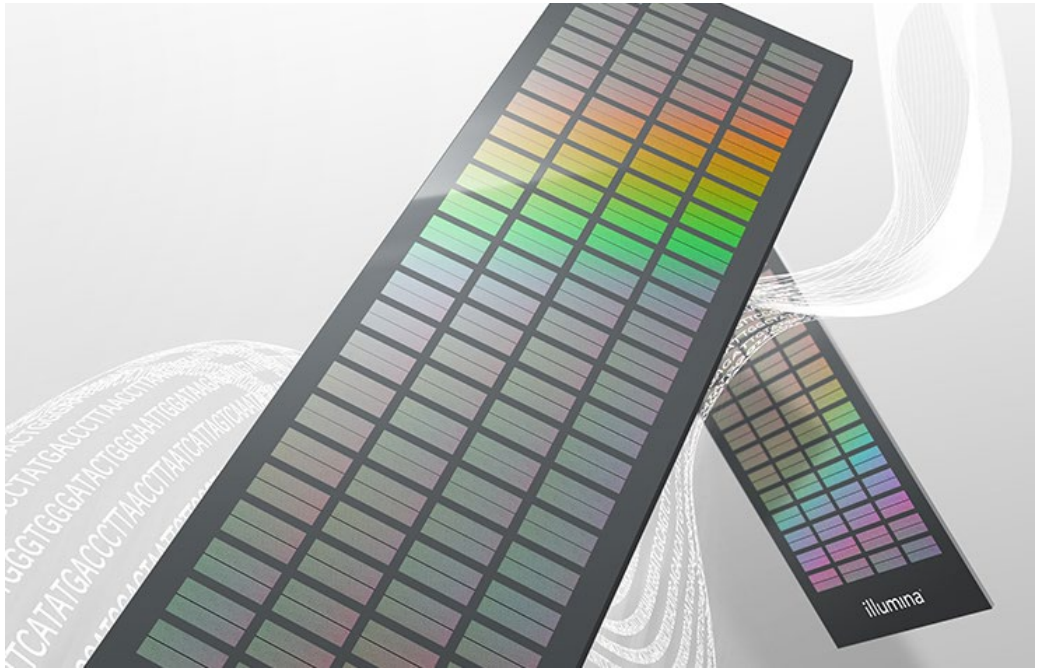
Next Steps

- In prep

Polygenic risk scores

Epigenetic clocks

Array-based sequencing

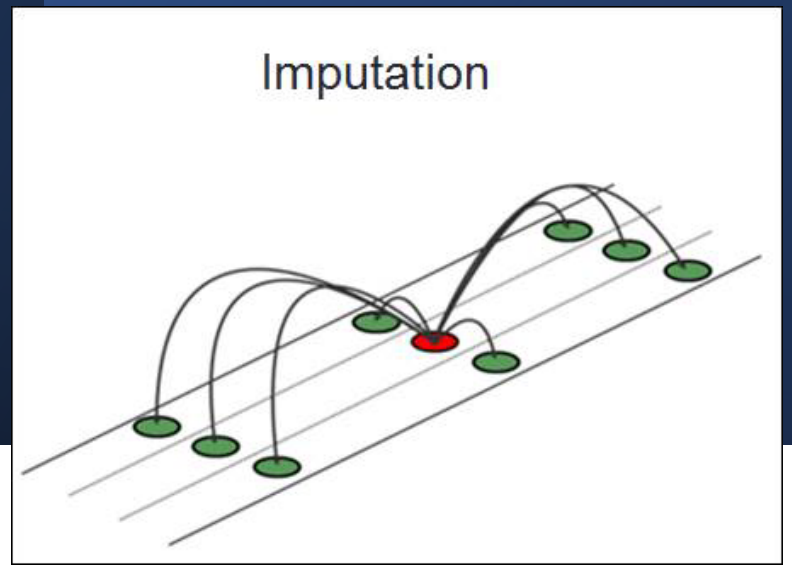


- Genotyping has been performed using microarray technology

MCS Genetics data

Genotype data	21,169 samples (21,064 individuals) 618,540 genetic variants
Imputed, QC'd data (TOPMed)	20,257 individuals 8,720,874 genetic variants

MCS Genetics data



Genotype data	21,169 samples (21,064 individuals) 618,540 genetic variants
Imputed, QC'd data (TOPMed)	20,257 individuals 8,720,874 genetic variants

MCS Genetics data

Imputed, QC'd data (TOPMed)

Category	Count
Mother [M]	7,781
Father [F]	4,635
Child [C] (% female)	7,841 (50%)
Trios	3,119

MCS Whole Exome data

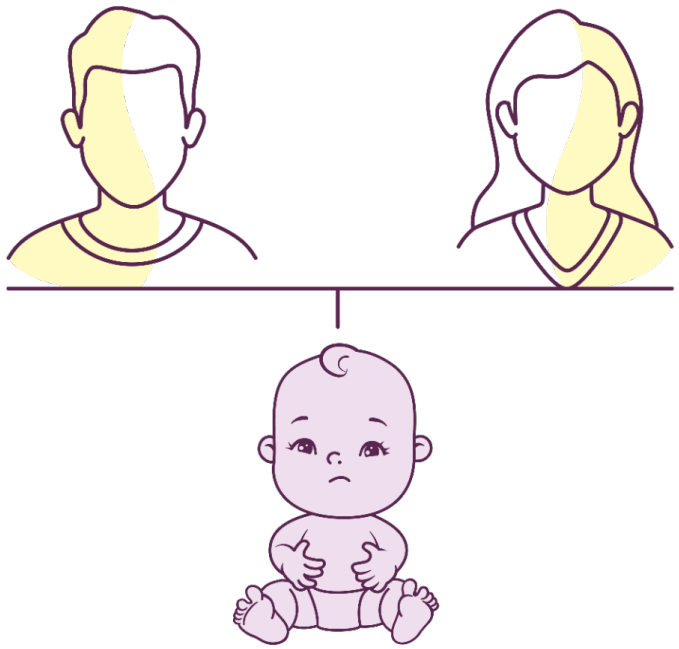
- Wellcome Sanger sequenced 15,240 samples
- 15,055 were processed after sequencing
- 14,753 samples passed QC and 1,916,636 sites

Category	Count
Child [C]	7,620
Mother [M]	3,423
Father [F]	3,482

MCS Genetics data: advantages of the trio design

Trio design + longitudinal data:

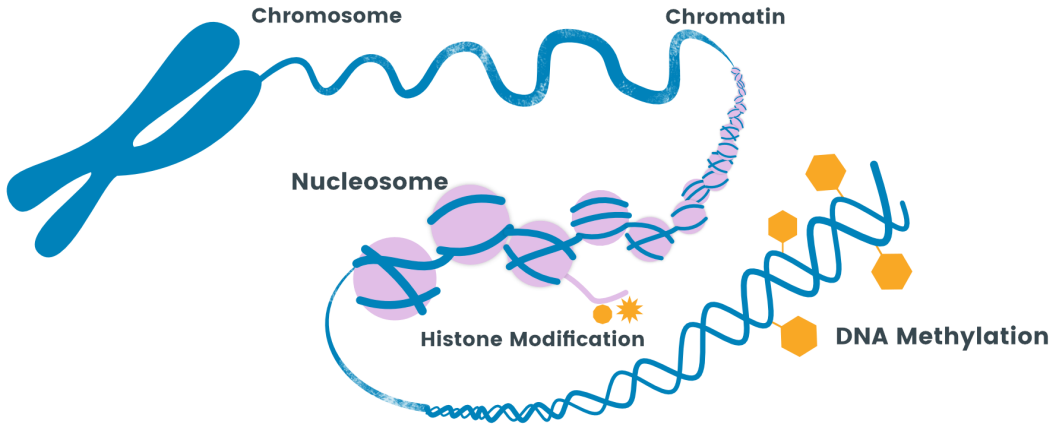
- Parent-of-origin effects
- Improves identification of de novo and novel disease-associated variants
- Gene/ environment interactions

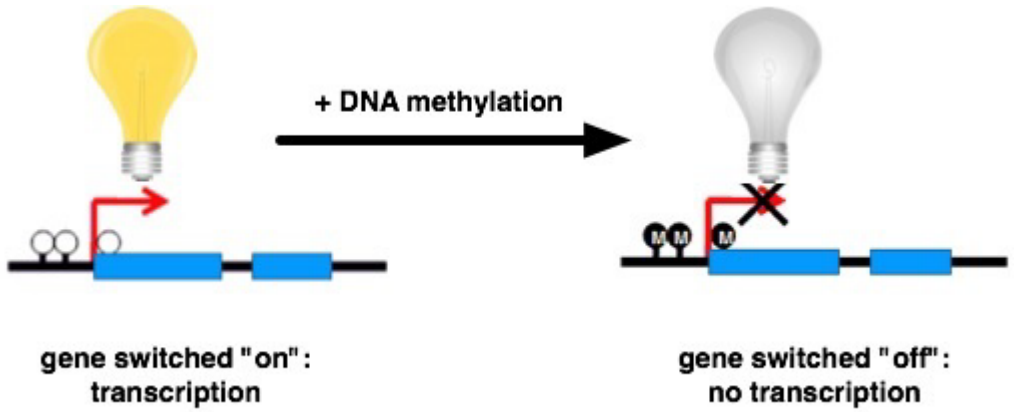


NCDS Genomics data

Genotype data	13,738 samples (6,431 individuals), across seven arrays Illumina 1.2M; Illumina 15k Custom Chip; Illumina Human 660-Quad; Infinium HumanHap 550K v1.1; Infinium HumanHap 550K v3; Affymetrix 500k; Affymetrix v6
Imputed, QC'd combined data (TOPMed)	6,382 individuals (50% female) 7,496,556 genetic variants
Exome data	1000 individuals on Illumina HiSeq 2500
Epigenetic data	Batch1 541 samples; Batch2 1,377 samples (1,169 individuals), 2 time points (ages 45 and 62) ~800,000 DNA methylation sites

NCDS Genomics data





Epigenetic data

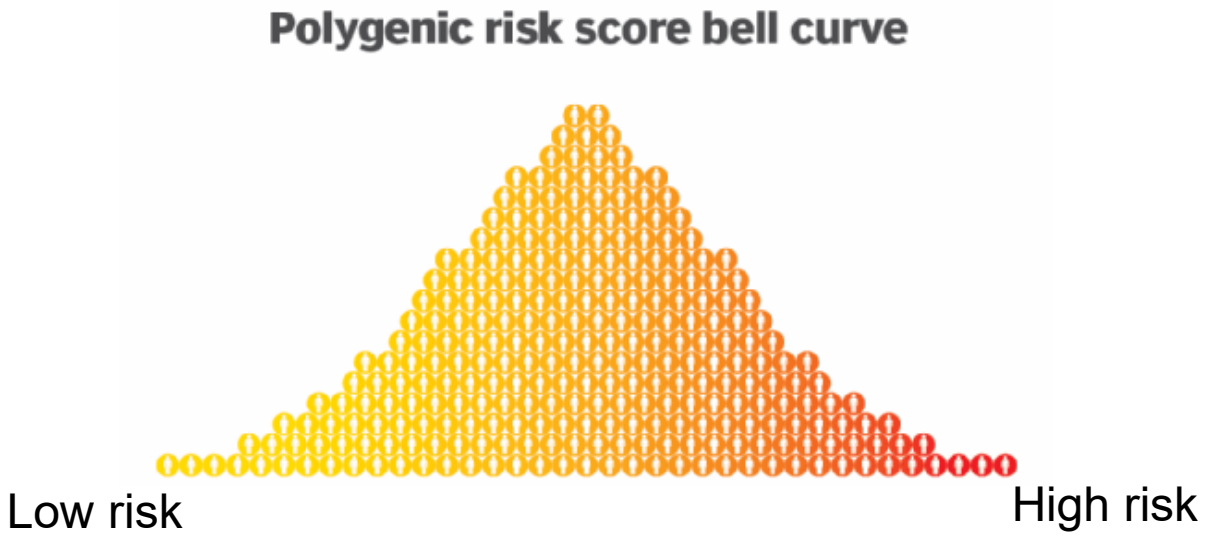
Batch1 541 samples
Batch2 1,377 samples (1,169 individuals), 2 time points (ages 45 and 62)
~800,000 DNA methylation sites

BCS70 Genomics data

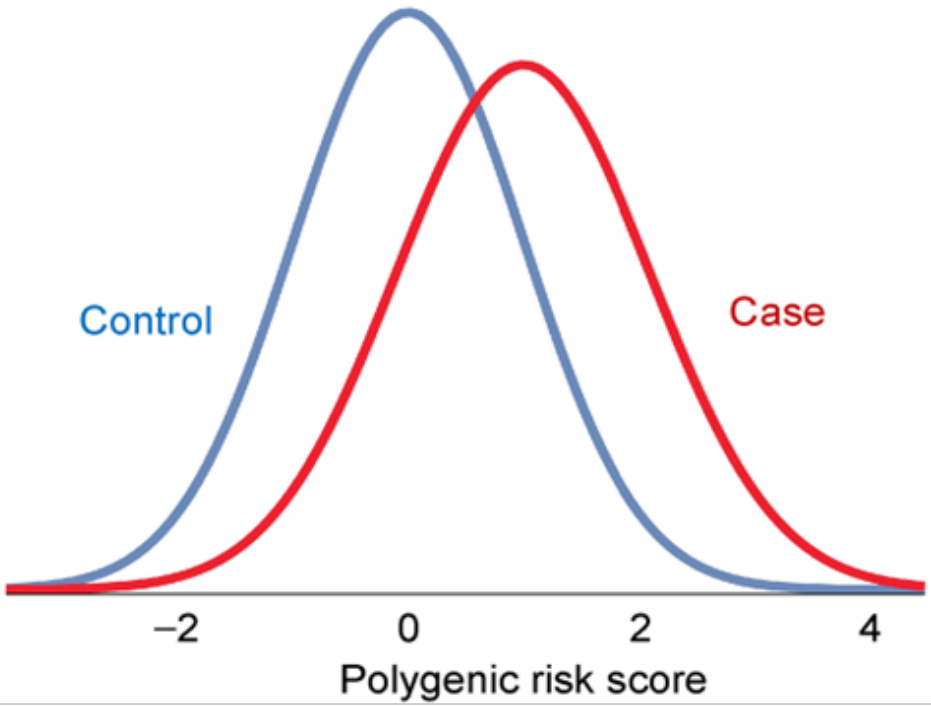
Genotype data	5,830 samples (5807 individuals) 654,027 genetic variants
Imputed, QC'd data (TOPMed)	5,598 individuals (51% female) 8,604,230 Genetic variants
Epigenetic data	255 samples, ~800,000 DNA methylation sites

Polygenic risk scores (PRS)

PRS summarise into a single score, the number of genetic variants an individual has linked to a particular trait



Polygenic risk scores (PRS)

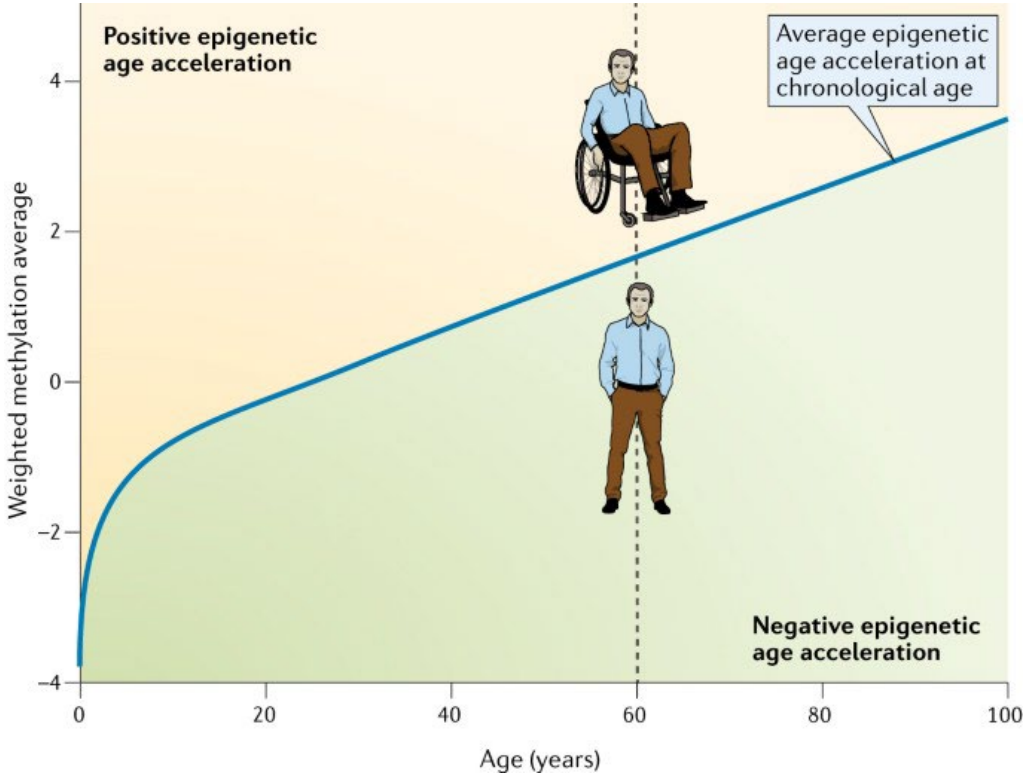


Studies have shown there are differences between case/ control groups but the distributions overlap

Generating PRS in house and using The Polygenic Index Repository for a range of health and social traits

Epigenetic clocks

DNA methylation data have been used to develop biomarkers of ageing, referred to as ‘epigenetic clocks’



Horvath & Raj (2018)

Epigenetic clocks

First Generation Epigenetic Clocks:

- Primarily rely on DNA methylation patterns, which correlate with chronological age
- Include the Horvath and Hannum clocks

Second Generation Epigenetic Clocks:

- Designed to predict lifespan and health span
- Incorporate additional biomarkers of aging, such as indicators of immune system function
- Better reflect biological age and the risk of age-related diseases
- Include the PhenoAge, GrimAge and DunedinPACE clocks

Epigenetic clocks

- We will generate epigenetic clock measures for the DNA methylation data we have (NCDS and BCS70)
- Epigenetic clocks we will use: Horvath, Hannum, GrimAge, PhenoAge and DunedinPACE

Data application/ further information

- To apply for the data:https://cls.ucl.ac.uk/data-access-training/data-access/accessing-data-directly-from-cls/
- There is a monthly data access committee meeting where applications are discussed
- Aim to get data out within 3 months of application
- For further information on the quality control and availability of CLS genomic data please see our github pages site: https://cls-genetics.github.io/docs/intro.html
- Another webinar will be run on the polygenic risk scores once the repository has been finalised

Data application/ further information

- Can apply for all cohorts in one application

13.3 Phenotypic sample requested for genetics applications

Phenotypic variable data are provided by default only for the genotyped cases. If you wish to request data from the whole cohort sample, please indicate this below.

NCDS	**BCS70**	**MCS**
☐ Genotyped sample	☐ Genotyped sample	☐ Genotyped sample
☐ Whole cohort	☐ Whole cohort	☐ Whole cohort

Data application/ further information

- PRS and Epigenetic clocks to be made available via special licence on the UK data service (UKDS) https://ukdataservice.ac.uk/
- PRS should be available on UKDS by Q4 2024
- Epigenetic clocks available on UKDS by 2025
- DAC applications to be used before then

Data application/ further information

CLS Genomics Data

Search CLS Genomics Data

- Introduction
- MCS
- NCDS
- BCS70
- Next Steps
- Polygenic risk scores
- Epigenetic Clocks
- Glossary
- Statement on Ancestry
- Contact us

MCS

The Millennium Cohort Study (MCS), known as 'Child of the New Century' to cohort members and their families, is following the lives of around 19,000 young people born across England, Scotland, Wales and Northern Ireland in 2000-02. The study began with an original sample of 18,818 cohort members. Cohort members were genotyped at age 14.

Data availability

Data type	Array / Imputation panel	Number samples	Coverage
Genetic (non QCd)	GSA Array v1	21,169	618,540 genetic variants
Imputed (QCd)	TOPMED	20,257	8,720,874 genetic variants
Whole exome sequencing	TWIST	14,753	1,916,636 sites

https://cls-genetics.github.io/docs/intro.html

Potential research using the CLS cohort studies: MCS trios analysis

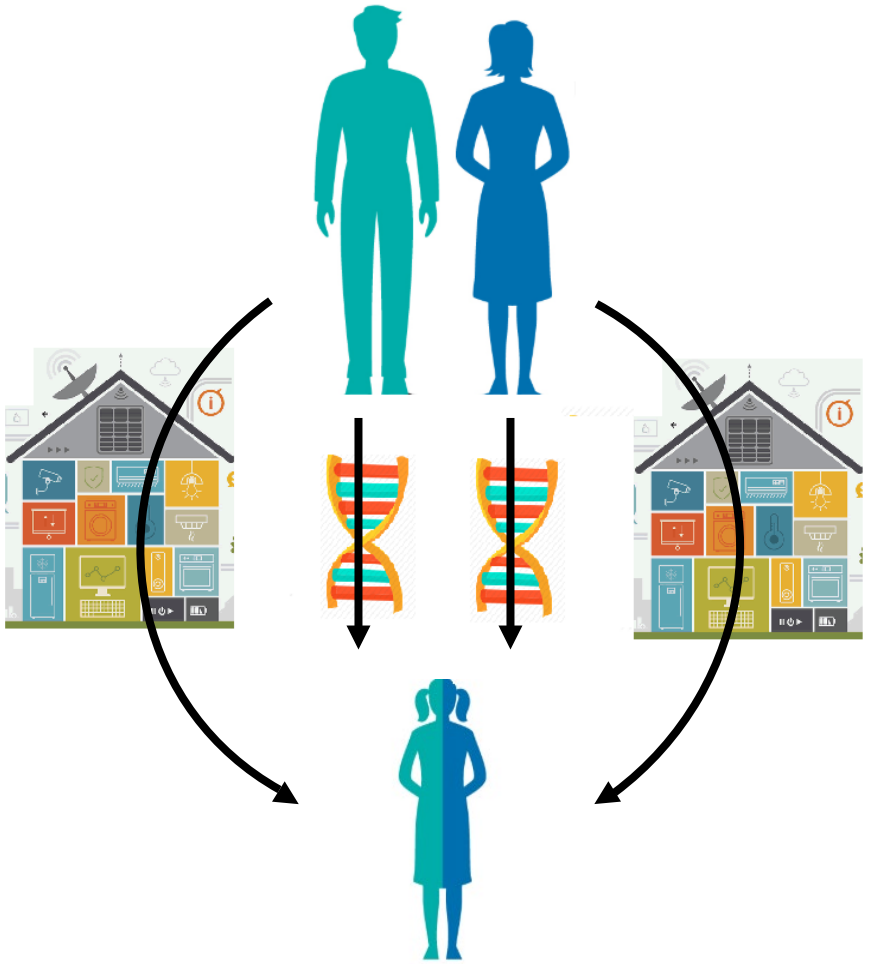
Tim Morris
Centre for Longitudinal Studies, UCL Social Research Institute

CENTRE FOR LONGITUDINAL STUDIES

Economic and Social Research Council

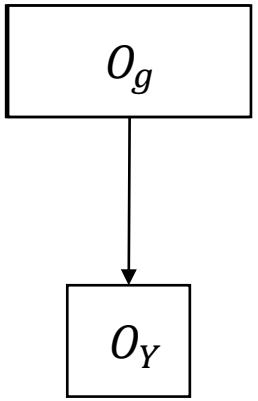
Environmentally mediated genetic effects

- Children inherit both genes and environment
- Genetic and environmental effects may differ in complex ways (gene environment interaction)
- These effects may also be ‘contaminated’ by each other; social effects appear genetic / genetic effects appear social
- Need genotyped family data to study



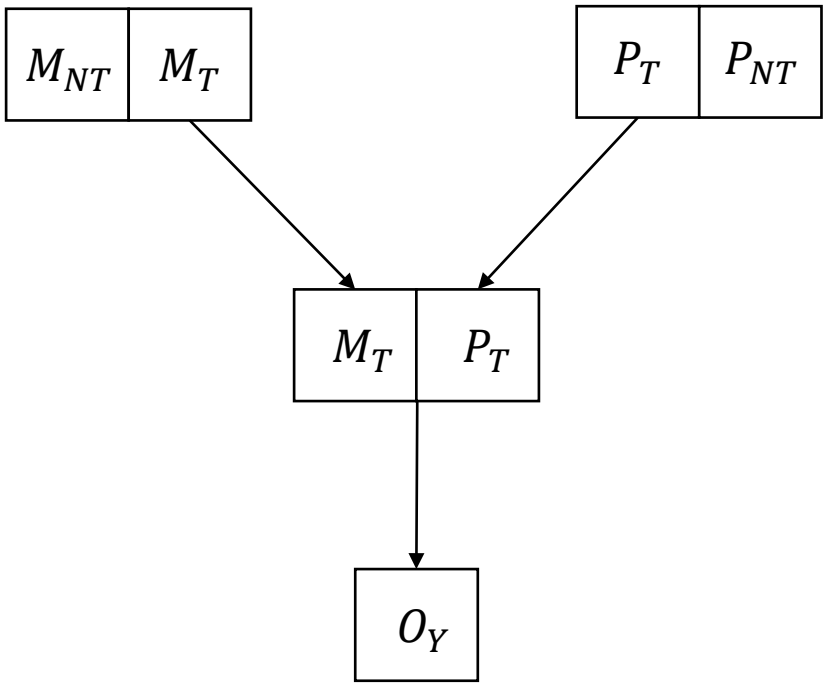
Non/transmitted genotype

- RQ: How does genotype affect BMI
- RQ: How does genotype affect education



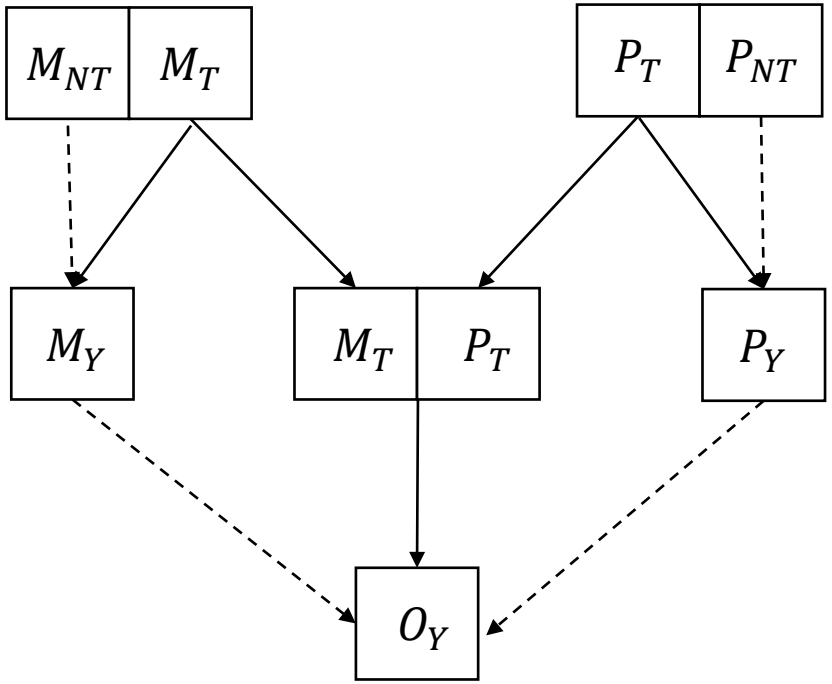
Non/transmitted genotype

- RQ: How does genotype affect BMI
- RQ: How does genotype affect education



Non/transmitted genotype

- Pathways may be more complex
- BMI: childhood food environment; intrauterine effects
- Education: household learning environment; parental investments
- Non-transmitted genotypes from each parent can be combined to create pseudo-offspring
- Can explore pathways of mediation



Trios in MCS

- Educational records from National Pupil Database
- Family information from MCS self-reports; age, sex, parenting behaviours; family socioeconomic background
- Polygenic score of educational attainment

Category	Count
Mother [M]	7,781
Father [F]	4,635
Child [C] (% female)	7,841 (50%)
Trios	**3,119**

CHILD OF THE NEW CENTURY

Department for Education

Secure Lab at the UK Data Service

UK Data Service

Thank you for listening!

Please follow the link in the chat for the feedback survey – thank you!

CENTRE FOR LONGITUDINAL STUDIES

Economic and Social Research Council