Analysis plan

for GWAS on selected phenotypes at different stages in the lifecourse

Last updated: February 19, 2025

Authors: Grace M. Power, Gibran Hemani, Eleanor Sanderson

Background

Acute, chronic, and recurring, adverse health conditions that emerge in later life are often shaped by processes experienced throughout life. Gaining a better understanding of how exposures at different stages in the lifecourse influence health outcomes is critical to developing more effective disease prevention and treatment strategies. This is of key public health importance.

Lifecourse stratified effects are of primary interest for identifying critical periods in Mendelian randomisation (MR). Inherited genetic variants may have different effects on some exposures at different time periods across the lifecourse (within a population). As a result, we are seeking to establish a consortium dedicated to generating and integrating data using an age-stratified genome-wide association studies (GWAS) approach.

To robustly run MR, valid instrumental variables must be employed which require large-scale datasets comprising phenotype and genotype data. By aggregating data from a wide range of cohorts, we will be able to access larger sample sizes without requiring repeated measures. This approach additionally allows us to remain agnostic about the shape of the Gene-by-Age (GxAge) interaction during the analysis stage and enables us to model it in greater detail at the post-meta-analysis stage.

This consortium will allow us to develop a more comprehensive set of instruments for future MR analyses to be better able to estimate the effects of a range of phenotypes at multiple time periods across the lifecourse on later life outcomes.

Aim

To explore how selected phenotypes at different stages in the lifecourse modify risk, we will:

1. Generate GWAS summary data of age x phenotype across the lifecourse for 24 phenotypes
2. Combine results by meta-analysing at the age x phenotype level across studies
3. Model changes in genetic effects over time at the post-meta-analysis stage
4. Evaluate the influence of biasing mechanisms on GxAge interaction estimates
5. Conduct post GxAge analyses e.g. time-varying MR, and provide a usable tool for researchers interested in using the data generated

Document scope

This Analysis plan will guide you through the Lifecourse GWAS consortium analysis. Here, you will find instructions for data preparation and running code to generate GWASs on time-varying phenotypes. We are collecting data on a comprehensive list of phenotypes (see Supplementary 1 below) every year up until 19 years of age and every five years thereafter.

We have prepared the pipeline to minimise time and energy required by analysts to contribute data to the overall effort, ensure harmonisation across cohorts, and minimise errors. The use of standardised procedures across all samples is critical in order to increase the effectiveness of the subsequent meta-analyses that we be run internally upon receipt of these GWAS. Because there is always a chance of error, we may ask some analyses to be re-run, although we will attempt to keep such requests to a minimum. We encourage analysts to adhere to the data and file organisation structures proposed for the pipeline to facilitate debugging and ease of any subsequent analyses that might be required.

Inclusion criteria

To be included, a cohort will need to provide at least one phenotype for at least one age range. The list of phenotypes is somewhat aspirational and if bandwidth limits the number of phenotypes that can be contributed, that will not be a barrier to contributing where possible. Please discuss which phenotypes to contribute with the core Lifecourse GWAS consortium group if you are uncertain (lifecourse-gwas-group@bristol.ac.uk). More details on the definitions of each phenotype are given in the Phenotype definition section (Supplementary 1).

Overview of analysis required

Below we outline the methodological steps required for generating the GWAS summary statistics for each cohort, summarised in Figure 1. We will provide code for each step in the pipeline. Further details on these steps are given below.

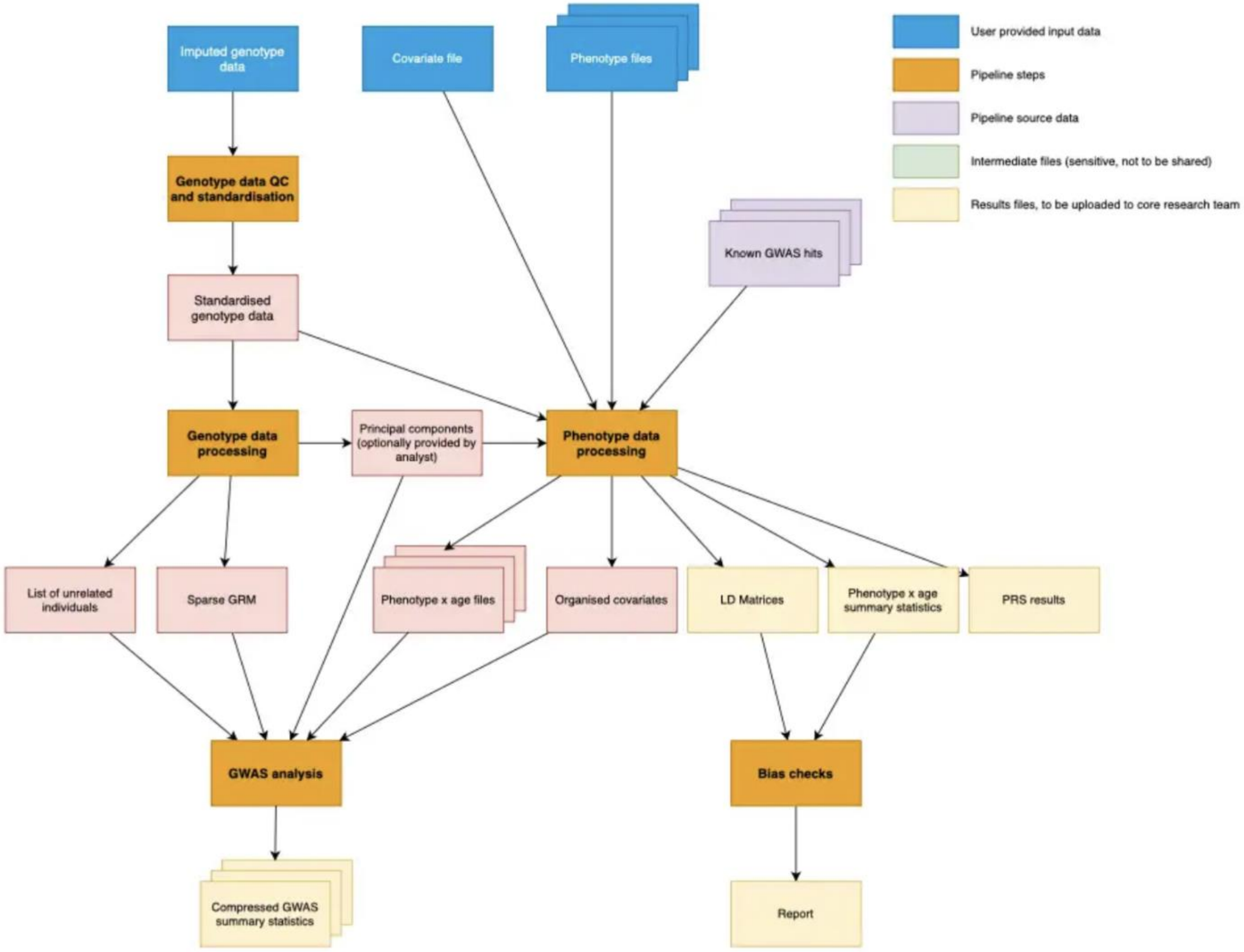
Imputed genotype data
Covariate file
Phenotype files
User provided input data
Pipeline steps
Pipeline source data
Intermediate files (sensitive, not to be shared)
Results files, to be uploaded to core research team
Genotype data QC and standardisation
Known GWAS hits
Standardised genotype data
Genotype data processing
Principal components (optionally provided by analyst)
Phenotype data processing
List of unrelated individuals
Sparse GRM
Phenotype x age files
Organised covariates
LD Matrices
Phenotype x age summary statistics
PRS results
GWAS analysis
Bias checks
Compressed GWAS summary statistics
Report

Figure 1. Diagram depicting the steps involved in the pipeline for generating the GWAS summary statistics.

For more detailed guidance on any of the following steps, please visit:
https://github.com/MRCIEU/Lifecourse-GWAS/wiki/

Data Preparation

Example data formatting templates are available here:
https://github.com/MRCIEU/Lifecourse-GWAS/tree/main/test/example-data.

Covariate Data

There are three types of covariate data used in the pipeline:

1. Genetic Principal Components (PCs): Generated by the pipeline or optionally provided by the analyst.
2. Static Covariates: Provided by the analyst in a file called static_covariates.txt.
3. Time-Varying Covariates: Provided by the analyst for relevant phenotypes within long-format phenotype files.

Static covariates

There should be a single covariate file for the 'static' covariates in the analysis. It must be named `static_covariates.txt` and stored in the `phenotype_input_dir` folder specified in config.env, with the following minimum format (note you may need to provide more covariate columns as explained further below):

```
FID IID sex yob my_covariate1 my_covariate2
ID1 ID1 1 1970 1 1
ID2 ID2 1 NA 0 1
ID3 ID3 2 1978 0 0
```

Note if year of birth is not available then you can omit this. Don't include the variable in the `static_covariates.txt` and make sure it's removed from the `static_covariates` field in `config.env`.

If there are individuals that have missing values, indicated by NA, these will be omitted.

Phenotype Data

Requirements:

A file is required for each phenotype (e.g., bmi.txt for body mass index) and must be stored in the phenotype_input_dir folder

File format example:

```
    FID IID value age
    ID1 ID1 23.000 15.25
```

```
ID2 ID2 27.000 23.33
ID3 ID3 18.000 5.00
```

- FID and IID: Family and individual IDs.
- value: Phenotypic value in the units specified in phenotype_list.csv.
- age: Age of measurement in years (non-integer values supported, e.g., age in months should be converted to years using age = months / 12).

Age-specific covariates for some phenotypes

In the 'Phenotype defintions' section, each phenotype lists the age-specific covariates required for the analysis. For those phenotypes, please make sure that the phenotype file also has the additional covariate column(s) required. The pipeline will be expecting those age-specific covariates in the phenotype file for that phenotype.

```
FID IID value age cholesterol_med
ID1 ID1 2.6 15.33 0
ID2 ID2 2.9 23.50 0
ID3 ID3 2.5 65.83 1
```

Missing values in the time-varying (medication related) covariates should be coded as 0, i.e. the same as if the individual is not taking any medication. Please remove any individuals where data for the value and/or age column **are missing.**

Additional notes:

- FID and IID may repeat if individuals are measured at multiple time points, but duplicate rows with the same FID IID age will be removed.
- Extra columns can be included but will be ignored.

Imputed Genotype Data

Requirements:

- Impute to recent reference panels (e.g., TOPMED, HRC v1.1).
- If using ancestry-specific panels, contact the pipeline team.
- Data must be in .bgen format, with one file per chromosome.

File Naming:

- No spaces in file names.
- Include all 22 autosomes; chromosome X (optional) must be in .pgen format.

Creating the Genotype Input List:

- Create a file (e.g., geno_files.txt) listing .bgen and .sample file paths:

```
/path/to/data_01.bgen /path/to/data.sample
/path/to/data_02.bgen /path/to/data.sample
```

- Update config.env to include the list location, e.g.:

```
genotype_input_list="/path/to/geno_files.txt
```

bgen-1.2 and Indexing:

- Use .bgen version 1.2 for faster analysis and ensure .bgen.bgi index files are present.
- Use the update_bgen.sh script to migrate files to version 1.2 and create indices:

```
./utils/update_bgen.sh /path/to/new/bgen1.2/directory
```

Step 00. Organise genotype data

This step checks the input genotype data, extracts an LD pruned subset for subsequent analysis, and identifies a subset of high quality variants to use for subsequent analysis.

Step 01. Ancestry

This step will generate principal components (PCs) if required. The pipeline is designed for a single ancestral group per run. For cohorts with multiple ancestral groups, run the pipeline separately for each group. Admixed cohorts can generally be treated as a single group unless you have specific expertise suggesting otherwise.

Step 02. Phenotype organisation

This step will organise the phenotype and time-varying covariate data ready for the GWAS analysis. Phenotype data is provided by the analyst in long format with each individual's phenotype measure annotated with the age at measurement. In order to ensure comparability of associations across studies we will create summaries and plots for each phenotype so that we can evaluate the distributional overlaps across cohorts. Efforts have been made to ensure that no individual-level will be shared, and no disclosive data will be included in the summary data that is shared (e.g. jitters are applied to plots etc).

Step 03. PRS

This step focuses on quality control of the GxAge interactions. It will evaluate the associations of externally weighted polygenic scores for the traits as positive controls. It will also generate LD matrices at each phenotype x age time point for the known variants for that phenotype. This will help to evaluate if age-differential selection bias could be driving GxAge interactions.

Step 04. GWAS

This step will perform the GWAS analysis using fastGWA. It will attempt to perform a linear mixed model using the sparse kinships and PCs generated in Step 01. However if that fails to converge it will revert to running a linear model on unrelated samples. Because there are potentially a large number of GWASs that will be run the pipeline will store the results in a .gzfile that contains a standardised variant ID, beta, SE, N, p-value and effect allele frequency for each variant. It will also create a summary file .summary.rds with lambda QC values etc.

Uploading results

Each pipeline step will package the logs and shareable results ready for upload. A central server will be provided for analysts to upload the results from each stage. We will encourage analysts to upload the packaged logs and results for inspection by the core research team, so that if there are any issues they can be identified before analysts potentially spend time on subsequent steps.

Contact details

Questions about this analysis plan, or any aspects of the project, can be directed to: lifecourse-gwas-group@bristol.ac.uk

The working group for this consortium consists of:

Grace M. Power grace.power@bristol.ac.uk

Eleanor Sanderson eleanor.sanderson@bristol.ac.uk

Gibran Hemani gibran.hemani@bristol.ac.uk

Genevieve Leyden genevieve.leyden@bristol.ac.uk

David Carslake david.carslake@bristol.ac.uk

Supplementary 1.
Phenotype definitions

All traits will use the following static covariates

- sex (1=Male, 2=Female)
- yob (Can be approximate to distinguish generations if exact values are not available)

Any phenotype-specific covariates required are listed below. Note that Phenotype-specific covariates can be time-varying (e.g. medications may start and stop over the lifecourse if there is longitudinal data for an individual).

To see an example of the relevant data files and their columns / formats required for each phenotype please see here: https://github.com/MRCIEU/Lifecourse-GWAS/tree/main/test/example-data

To ensure consistency and comparability across cohorts, we request that measures collected during pregnancy be excluded, as physiological changes during this period may introduce systematic differences.

Additionally, measures should be sourced from regularly scheduled cohort visits rather than opportunistic data capture, such as emergency room visits or other unplanned medical encounters. This approach enhances data reliability and ensures greater standardisation across cohorts.

Anthropometric

Body Mass Index (BMI)

- Units: kilograms/metres^2 (kg/m^2)

  Note: These data will also be used to calculate BMI-for-age z-scores, applying the WHO growth reference standards for individuals aged 0 to 19 years. This standardised approach ensures accurate age- and sex-specific assessments based on internationally recognised benchmarks.

Height

- Definition: Standing height
- Units: centimeters (cm)

Waist circumference

- Units: centimeters (cm)

Waist to hip ratio (WHR)

- Definition: Divide the waist measurement (cm) by hip measurement (cm) (cm/cm)
- Units: Absolute ratio

Bone mineral density

- Definition: Total body bone mineral density (TB-BMD) measured using Dual-Energy X-ray Absorptiometry (DXA) scanning equipment.
- Units: grams/centimeters^2 (g/cm^2)

Lung function

Forced expiratory volume in 1 second (FEV1)

- Units: Litres (l)

FEV1/FVC ratio

- Definition: Ratio of Forced expiratory volume in 1 second (litres) to Forced Vital Capacity (litres)
- Units: Please state whether absolute ratio value or Z-scores are being provided.

Cardiovascular

Systolic blood pressure (SBP)

- Definition: Automated reading
- Units: millimetres of mercury (mmHg)
- Covariates
  - `bp_med`: blood pressure medication (1=Yes, 0=No). We will adjust for blood pressure medication by applying a constant increase of 15mmHg in individuals taking medication.

Diastolic blood pressure (DBP)

- Definition: Automated reading
- Units: millimetres of mercury (mmHg)
- Covariates
  - `bp_med`: blood pressure medication (1=Yes, 0=No). We will adjust for blood pressure medication by applying a constant increase of 10mmHg in individuals taking medication.

Blood measures

Low density lipoprotein (LDL) cholesterol

- Definition: blood measure
- Units: millimoles per liter (mmol/L)
- Covariates:
  - `cholesterol_med`: statin medication use (1 = Yes, 0 = No). We will adjust for statin/cholesterol lowering medication by applying a relative reduction to the measured value of 40%.

High density lipoprotein (HDL) cholesterol

- Definition: blood measure
- Units: millimoles per liter (mmol/L)
- Covariates:
  - `cholesterol_med`: statin medication use (1 = Yes, 0 = No). We will adjust for statin/cholesterol lowering medication by applying a relative reduction to the measured value of 40%.

Triglycerides

- Definition: blood measure
- Units: millimoles per liter (mmol/L)
- Covariates:
  - `cholesterol_med`: statin medication use (1 = Yes, 0 = No). We will adjust for statin/cholesterol lowering medication by applying a relative reduction to the measured value of 40%.
  - `insulin_med`: Insulin medication use (1 = Yes, 0 = No).

Glycated haemoglobin (HbA1c)

- Definition: blood measure
- Units: millimoles per liter (mmol/L)
- Covariates:
  - `insulin_med`: Insulin medication use (1 = Yes, 0 = No).

Leptin

- Definition: blood measure
- Units: nanogram per milliliter (ng/ml)

Insulin

- Definition: blood measure. Conversion factor 1 µIU/mL = 6.00 pmol/L.
- Units: pmol/L
- Covariates:
  - `insulin_med`: Insulin medication use (1 = Yes, 0 = No).

Adiponectin

- Definition: blood measure
- Units: micrograms per milliliter (µg/ml)

Calcium

- Definition: blood measure
- Units: millimoles per liter (mmol/L)

Vitamin D

- Definition: blood measure
- Units: nanomoles per liter (nmol/L)

Immune markers

C-reactive protein (CRP)

- Definition: blood measure
- Units: milligrams per liter (mg/L)

Interleukin-6 (Il6)

- Definition: blood measure
- Units: picograms per millilitre (pg/ml)

Sex hormones

Estradiol

- Definition: blood measure
- Units: picomoles per liter (pmol/L)
- Covariates:
    - `hormone_med`: hormonal contraception or hormone replacement therapy (1 = Yes, 0 = No). If data is missing or not available then set this value to 0. We will conduct a sensitivity analysis at the meta-analysis stage, to assess whether potential heterogeneity in GxAge interaction effects observed are attributed to unadjusted drug use.

Total testosterone

- Definition: blood measure
- Units: nanomoles per liter (nmol/L)
- Covariates:

    - `hormone_med`: hormonal contraception or hormone replacement therapy (1 = Yes, 0 = No). If data is missing or not available then set this value to 0. We will conduct a sensitivity analysis at the meta-analysis stage, to assess whether potential heterogeneity in GxAge interaction effects observed are attributed to unadjusted drug use.

Bioavailable testosterone

- Definition: blood measure
- Units: nanomoles per liter (nmol/L)
- Covariates:
    - `hormone_med`: hormonal contraception or hormone replacement therapy (1 = Yes, 0 = No). If data is missing or not available then set this value to 0. We will conduct a sensitivity analysis at the meta-analysis stage, to assess whether potential heterogeneity in GxAge interaction effects observed are attributed to unadjusted drug use.

Sex hormone binding globulin

- Definition: blood measure
- Units: nanomoles per liter (nmol/L)
- Covariates:
    - `hormone_med`: hormonal contraception or hormone replacement therapy (1 = Yes, 0 = No). If data is missing or not available then set this value to 0. We will conduct a sensitivity analysis at the meta-analysis stage, to assess whether potential heterogeneity in GxAge interaction effects observed are attributed to unadjusted drug use.