Goals of this Document:

This document details how I prepared the accelerometer cohort, processed the genetic data and created the polygenic risk score, and created the PAEE exposure

Last updated: 2/22/2023

This document gives an overview of all data processing, sanity checks, and analyses performed for paper 3. I created an accelerometer cohort, a phenotype cohort, and a polygenic risk score as the three and merged them as the primary outputs of the document. The accelerometer cohort and phenotype cohort are distinct because I needed to keep the phenotype cohort intact to allow direct comparisons with the PRS score’s performance in Tamlander et al., 2022 before merging them. Because I am using coronary artery disease here, I had to redefine the outcome of interest.

**Creating the Accelerometer Cohort**

Associated Files:

“Paper 3 Inclusion Criteria Flowchart.docx”

“UPDATED ACCELEROMETER Spark JL Code for Getting CSV of Data in UKB.txt”

“Paper 3 Analysis.R” (thru Line 162)

Output:

“PACOHORTCAD.csv”

“dathes.csv”

“datdeath.csv”

“datdeathcause.csv”

“PACOHORTprocessedPAVars.csv”

External Resources:

<https://github.com/OxWearables/rap_wearables>

<https://github.com/OxWearables/rap_wearables/blob/main/1_Extract_Data.ipynb>

Summary of section:

The first step in this process was to create the accelerometer cohort. I followed the inclusion criteria outlined by the Oxford Wearables Group, who are leading the charge in terms of analyzing the UK Biobank accelerometer data. Their quality control inclusion criteria do not appear to differ substantially from those elsewhere in the literature. Because the variables required to perform this screening were largely already available in the UK Biobank, I used the Cohort Browser to create the PA Cohort to restrict to only individuals who wore accelerometers and met the inclusion criteria detailed in the flow chart. Through the code inspired by the Oxford group in “UPDATED ACCELEROMETER Spark JL Code for Getting CSV of Data in UKB.txt,” I ran a Spark JupyterLab notebook that created separate datasets for the covariates (PACOHORTCAD.csv), hospital inpatient data (dathes.csv), and death date and cause of death (datdeath.csv and datdeathcause.csv). This separation of datasets allowed for a simpler process to incorporate surgery, hospital inpatient, and death CAD cases.

Lastly, ran Paper 3 Analysis.R but BECAUSE I USE PAEE (which is processed later), only relevant part here was converting overall activity and hours worn. CAD cases are also formally defined later. There were 96,660 individuals in this cohort before further screening by kinship done later.

NOTE: ARRIVED AT CAD CASES SEPARATELY - THIS IS HANDLED FURTHER DOWN IN WORKFLOW W DIFF SPARK JL

**Genomic processing and creating the Tamlander et al., polygenic risk score**

Associated Files:

“ukb\_mfi\_chr1\_v3.txt” (1 to 22)

“multiallelicsnps.csv”

“Final Aligning UKB and Scoring File Variants.R”

“FINAL CODE FOR BED CONVERSION TO PRS.txt”

“Code to Allow Final Plink Merge.R”

“Merging Final Set of PRS Files.txt”

“plinking PGS code.txt” (starting at FINAL - RUNNING THIS FOR TAMLANDER 22 CODE)

“PGS001780GRCH37.txt”

Output:

“FINALRSIDsChromXXNEWSCORERSIDS.txt”

NEWCHRXXBED.bed

BEDXXMAFREST.bed

BEDXXMAFRESTNoDup.bed

FlippedSNPsChrXX.txt

BEDXXMAFRESTNoDupFlip.bed

NODUPFLIPCHROMXXMISMATCH.txt

BEDXXFINAL.bed

“BEDFINAL\_1\_22\_merged.bed/.bid/.fam”

“FINALPGS.log”

“FINALPGS.sscore”

“FINALPGS.sscore.vars”

External Resources:

<https://www.youtube.com/watch?v=762PVlyZJ-U&t=2027s>

https://www.pgscatalog.org/score/PGS001780/

<https://github.com/ninamars/INTERVENE-PRS-transferability/blob/main/prs_calculation.sh>

“Integration of questionnaire-based risk factors improves polygenic risk scores for human coronary heart disease and type 2 diabetes”

Associated Link: https://www.nature.com/articles/s42003-021-02996-0

Summary of section:

This section involved the creation of a polygenic risk score to match the one used in Tamlander et al., 2022 (that was the most recent and powerful to date, as well as tested on but not trained on UKB data, which makes it an ideal choice to avoid winner’s curse bias). I first downloaded the 22 ukb\_mfi\_chrX\_v3.txt files from UK Biobank data showcase. These files contain minor allele frequencies and INFO (imputation quality) scores but for our current purposes the most relevant aspect of the files is that they contain *ALL OF THE VARIANTS IN EACH CHROMOSOME*. I also downloaded the Scoring File from the PGS Catalog (ID: 001780), which contains all of the variants and the weights required to implement the polygenic score (and applied to the correct genomic build of GRCH37 - same as UKB). I then used “Final Aligning UKB and Scoring File Variants.R” through line 1735 to find variants that are contained in both the UK Biobank and polygenic score (by matching RSIDs). There were 1,089,716 matched variants (out of 1,090,048 in the polygenic score), so only 332 unmatched.

Starting at line 1739 until line 4403 of “Final Aligning UKB and Scoring File Variants.R,” I take these existing matched datasets and screen out multi-allelic SNPs (genetic variants with more than 2 alleles) because these cannot be handled by the plink software and this is a small fraction of the variants. I also restrict to only variants with an INFO score > 0.6 (per recommendation of Nina Mars) to exclude variants that were imputed with a low level of certainty. This process reduces the number of variants only marginally to 1,089,653.

Imputed genetic variant files are in bgen format in the UK Biobank, which is difficult to use with most of the standard genomic analysis software. Thus, I next converted the chromosome files from bgen to bed format (a standard plink binary file format that produces 3 files - bed/bim/fam) using “FINAL CODE FOR BED CONVERSION TO PRS.txt.” The unfortunate part of bed files that is not true of pgen files (the newer binary file used in plink) is that it does not keep track of dosage information and instead hard codes (that is, instead of storing probabilities it makes these imputations look certain with some threshold chosen). This drawback did not seem like a major issue since many of the recent PRS codes I’ve read (including Mars et al.) make use of bed files and (as we’ll see later) this ultimately does not appear to substantially affect PRS performance. This produces bed files NEWCHRXXBED.bed (from 1 to 22).

We next have to merge all of these bed files to apply the PRS weights. I kept the initial merge code at the top of “Merging Final Set of PRS Files.txt” as a reminder of the error. In brief, there were 2.5k “multi-allelic” SNPs (not actually since we screened these out earlier). My two theories here are either that part of the population experienced strand flipping (for more on strand flipping: <https://www.snpedia.com/index.php/Ambiguous_flip>) or that different ancestry groups had different allele combinations at these SNPs. In case these were only extremely rare occurrences, I restricted to variants that had a minor allele frequency > 0.01, which is a fairly typical standard and screens out any variants with extremely low levels of variation. I experimented with this from lines 1 to 196 in the file “Code to Allow Final Plink Merge.R.” After restricting to MAF > 0.01 (creating bed files BEDXXMAFREST.bed), there were only a few variants that still exhibited issues for the merge. I would then remove one of these duplicates. This is imperfect because force-first arbitrarily chose the first instance to keep. HOWEVER, this would affect such few variants that it likely has no effect on the performance of the PRS at all.

I then created code in “Merged Final Set of PRS Files.txt” to perform this MAF restriction, remove ‘duplicates’ (creating BEDXXMAFRESTNoDup bed files), and flip variants. This process involved constant interactions between the R file and plink. I first restricted to MAF > 0.01, then I removed all ‘duplicates’. I then downloaded the bim file and in R compared allele alignment between the Scoring File and the variants in the chromosome. I saved the vectors of mismatched alleles (FlippedSNPsChrXX.txt) then back in the txt file used plink to flip the strand of any mismatched alleles in the UK Biobank (BEDXXMAFRESTNoDupFlip.bed). I then checked whether alleles matched once more and, if mismatches still existed, I removed these variants (“BEDXXFINAL.bed”). Lastly, I merged all of the chromosome files into a single binary file (bed/bid/fam) using plink in (BEDFINAL\_1\_22\_merged.bed/.bid/.fam). There are now 1,087,647 variants (2,069 variants fewer than in last checkpoint mostly owing to MAF screening). NOTE: IN TAMLANDER THEY ACTUALLY ONLY USED 1,087,714 FOR CHD IN THE UK BIOBANK!!!! This means that we only actually removed 67 variants…

I finally fit the polygenic score using plink with code in “plinking PGS code.txt,” which produced the key output FINALPGS.sscore. This polygenic score performed nearly as well as the one in Tamlander et al., with an OR of 1.67 (95% CI: 1.65 to 1.70) compared to 1.72 (95% CI: 1.70 to 1.75) in the original paper. We used a similar but smaller population in the UK Biobank (337,005 vs 343,652) and used a slightly smaller set of variants, which together likely explain the small differences in performance (comparison is done in “Creating Final PA Dataset.R” which is discussed further below).

**Creating other phenotype cohort data (CAD cases) & merging phenotype datasets**

Associated Files (R and resulting output files):

“Created **Processed Cohort ALL ANCESTRIES** Cohort w/ genotype QC”

“CAD Spark JL Code for Getting CSV of Data in UKB.txt”

“CAD Definition in Tamlander et al.docx”

“PROCESSING FOR MARS PGS TEST.R”

“Spark JL to bring PCs to CAD Analysis.txt”

Output:

“FULLCADCOHORT.csv”

“dathes.csv”

“datdeath.csv”

“datdeathcause.csv”

“FULLPHENODATAALLANCESTRIES.csv”

“PCsforCAD.csv”

External Resources:

<https://github.com/OxWearables/rap_wearables/blob/main/2_Further_Prep_in_R.ipynb>

Summary of section:

In this section, I created a cohort containing people of all ancestries and merged the genetic and phenotypic data. I also formally create the coronary artery disease variable here. The goal of creating a cohort ***DISTINCT FROM THE ACCELEROMETER COHORT*** was to allow me to evaluate the performance of the PRS in a population similar to that used in Tamlander et al.

I started by creating this new cohort “Processed Cohort ALL ANCESTRIES” in the Cohort Browser. I applied the typical genotype quality control procedures WITH THE EXCEPTION OF KINSHIP, which I handled later. Briefly, I excluded individuals who were not genotyped, withdrew consent, exhibited sexual aneuploidy, or were outliers for missingness or heterozygosity.

I then use Spark Jupyter Lab to create a CSV of the phenotype cohort that contains ALL possible codes for CAD and some of the covariates (more added later) with “CAD Spark JL Code for Getting CSV of Data in UKB.txt” and saved as “FULLCADCOHORT.csv.” This creates a few datasets: phenotype dataset, hospital episode, and cause and time of death datasets. Separating the data in this way like Oxford Wearables did allows for much simpler case definition.

I next load all these datasets into “PROCESSING FOR MARS PGS TEST.R” and use this code to create the CAD case definition, death, region, status, and time variables and apply censoring by region. I apply the CAD case definitions in “CAD Definition in Tamlander et al,” which makes use of ICD-10 and -9 codes for deaths, hospital inpatient episodes, and OPCS-3 and OPCS-4 codes for surgery. This R file ultimately yields “FULLPHENODATAALLANCESTRIES.csv.”

I had to add in the genetic principal components (forgot earlier), which follows a similar process. I used “Spark JL to bring PCs to CAD Analysis.txt” to add the principal components to the original cohort and arrive at PCsforCAD.csv. Next, I had to get the PA variables correctly defined and replicate the existing studies, which I do in the next section.

**Creating the Physical Activity Exposure Variables (PAEE, %MVPA, %Vigorous) and Replicating Results**

Associated Files:

“Paper 3 Prelim Analyses and Final Processing.R”

Output:

“PACOHORTprocessedPAVarsPAEE.csv”

External Resources:

“https://academic.oup.com/eurheartj/article/43/46/4789/6770665#supplementary-data”

Summary of section:

This section is where I formally created the physical activity energy expenditure (PAEE) variable used in recent studies that have utilized this data that has been validated in previous studies. The advantage of this as an exposure is that it has been validated to closely align with thigh accelerometer and heart rate monitor in a doubly labeled water study. I replicate the results in the recently published article “Physical activity volume, intensity, and incident cardiovascular disease” in the European Heart Journal to ensure that I correctly processed these variables. Starting at line 298 in “Paper 3 Prelim Analyses and Final Processing”, I created % MVPA and % vigorous as PAEE from mgs > 125 and PAEE from mgs > 400 following the recent literature, which virtually exactly matched the results in Dempsey et al., 2022. The overall physical activity measure after a conversion (1440/1000) matches very closely to their measure Euclidean Norm Minus One (ENMO). These measures are generally the best available in epidemiologic studies of physical activity. This process yielded “PACOHORTprocessedPAVarsPAEE.csv.”

**Evaluating PRS and Merging Accelerometer Cohort/PRS/Phenotype**

Associated Files:

“Creating Final PA Dataset.R”

Output:

“FinalPADataset.csv”

“FinalPhenoDataset.csv”

CAN SIMPLY USE p22020, which limits to individuals that passed kinship screenings of UKB

External Resource:

<https://www.biorxiv.org/content/biorxiv/suppl/2017/07/20/166298.DC1/166298-1.pdf>

Summary of section:

Finally, I had to merge all of the datasets together, evaluate the PRS, and restrict kinship (all in “Creating Final PA Dataset.R”). I started by restricting kinship to only individuals not related third degree or closer (following standards of UK Biobank). While this step is strangely not done in many GxE studies, the logic is similar to that of Mendelian randomization analyses: “If the data contains participants who are closely related then their genomes would be more similar than those of unrelated individuals, which can lead to biased estimations in population-level analyses." I then merged the principal component dataset with the phenotype dataset and then ultimately merged these datasets with the accelerometer wearing cohort. I tested the PRS in the original larger phenotype dataset, which showed that the performance was extremely similar to that found in Tamlander, et al., which suggests that it worked well. I finally save the datasets as FinalPhenoDataset.csv and (MOST IMPORTANTLY) the **FinalPADataset**.csv, which is used for subsequent analyses.

**Adding Further Covariates that are often controlled for in these types of analyses and MERGING w/ overall dataset**

**Associated Files:**

Spark JL to bring OTHER COVARIATES to CAD Analysis.txt”

“CovarsforCAD.csv”

“FinalPADataset.csv”

“Merging Other Covariates w PA Dataset.R”

FINALIZEDPADATASET.csv

Summary of section:

I kept this section extremely simple to make adding further covariates down the line easy. I first ran Spark JL to add all of the potential covariates of interest not included in the original dataset, which results in a dataset CovarsforCAD.csv. I then merged this dataset with the FinalPADataset.csv via the “Merging Other Covariates w PA Dataset.R” file to produce a final csv FINALIZEDPADATASET.csv. I perform data processing and prepping the dataset for analyses in the next section.

PROCESSING COVARIATES AND TIME AND INCIDENT CAD AND MAKING FINAL ANALYSIS DATASET (all of this code exists in one form or another…)

**Associated Files:**

FINALIZEDPADATASET.csv

Final Analysis Dataset Processing Code.R

Dempsey Replication and Sanity Checking.R

Time Varying Dempsey Results.R

Output:

FINALANALYSISDATAPAPER3.csv

DempseyDataset.csv

Summary of section:

This section involves a few diverging paths depending on the analyses we intend to run. First, I ran “Final Analysis Dataset Processing Code.R” with FINALIZEDPADATASET.csv. This R code was reshaped on 3/5/2023 to account for the ‘1’ factors and to make sure covariates are interpretable on a useful scale. In this code, I standardized the polygenic score to zero mean and unit variance, created age at baseline, created the time since accelerometer wear date, accounted for region-specific censoring, made sure that the right instance was chosen for covariates (most recent non-missing visit BEFORE wear time started), created the pooled veggie and fruit variables, created the medicine use variables, created the parental heart disease history variables, created season of wear covariate, and recoded several diet variables to be consumption over a week (year for alcohol). This yielded FINALANALYSISDATAPAPER3.csv.

I then use “Formal Analysis Code for Paper 3.R” but ONLY through line 154. This code limits PAEE to be at zero at the lowest end and constrains fruit and vegetable consumption to at least less than 50 (STILL pretty high but hard to arrive at good choice here without being arbitrary). TRIED and have not yet implemented imputation successfully…

I then went to “Formal Analysis Code for Paper 3” originally to run the restricted quadratic spline but this was before Patrick’s solution and so this file also contained several tests. INSTEAD, the process is now that I implemented the restricted quadratic spline at the beginning of “Dempsey Replication and Sanity Checking.R” and compared results. I THEN starting at line 513 got the covariates to match the way they were coded up in Dempsey for maximal comparability and saved these results as DempseyDataset.csv on line 1554. From lines 1620 to 1688 I assessed the proportional hazards assumption and found that it did not hold for PAEE or several confounders. Then without adjusting for this point (since they do not note doing so), I ran a restricted cubic spline and compared performance.

I then used “Results n Figure from Dempsey Data.R” to get my quintile comparisons and create these initial figures that took into account genetics and overall PA.

Lastly, in “Time Varying Dempsey Results.R,” I compared results for individuals at age 65 vs at age 75 (where a discontinuity in risk appears to exist) and found much stronger associations at 75 than 65.

ACTUAL RESULTS

“Getting Cubic Spline Right.R”

This is THE primary analysis R code. I use age as the time scale CORRECTLY here and find that I no longer have significant proportional hazards violations (outside biological sex). I compared restricted cubic spline, restricted quadratic spline, and linear model for PAEE and found that the linear spline performs best based on AIC/BIC. This remains true even if knots are defined based on the distribution among those who experienced a CAD event. The results were similar to those in Dempsey but even stronger for PA volume. I then tested for interaction and whether a spline would make sense for % MVPA and concluded that there are no significant interaction effects and % MVPA works best linearly too. I then compared the results of a model with %MVPA and PA together to the interaction results in Dempsey and got similar results, which reinforces in my view the fact that any interactive effect that might exist is extremely small. We have now confirmed that we have a sensible model, the results cohere with what we’d expect relative to Dempsey, and NOW we can perform finishing touches and recreate figures and tables with accurate numbers.

BRINGING IN REMAINING VARIABLES, GETTING IMPUTATION RIGHT, AND FINAL RESULTS

Variables:

SUBJECTIVE PA FROM ALL 4 WAVES (IPAQ Activity Group - High/Med/Low - 22032; 22039 - MET minutes of vigorous; 22038 - MET minutes of MVPA; 22040 - Summed MET minutes per week for all activity)

ETHNICITY???

OCCUPATION - Manual labor or not - HEAVY MANUAL OR PHYSICAL WORK (816) and walk or stand (806)

IMPUTATION:

Spark JL to bring OTHER COVARIATES to CAD Analysis.txt”

“CovarsforCAD.csv”

“FinalPADataset.csv”

“Merging Other Covariates w PA Dataset.R”