# INTERVENE

# Genetic validation of disease endpoints

Deliverable 2.5

Dissemination level: public

Part of:

WP 2:

**Data harmonisation, integration, and evaluation of genetic scores**

## Project summary

| | |
|---|---|
| **Project acronym:** | INTERVENE |
| **Project full title:** | International consortium for integrative genomics prediction |
| **Project Coordinator** | Institute for Molecular Medicine Finland FIMM, University of Helsinki; Prof. Samuli Ripatti and Dr. Andrea Ganna |
| **Project start date:** | 1.1.2021 |
| **Project end date:** | 31.12.2025 |
| **Project duration:** | 60 months |
| **Action type:** | RIA |
| **Call identifier:** | H2020-SC1-FA-DTS-2018-2020 (Trusted digital solutions and Cybersecurity in Health and Care) |
| **Grant number** | 101016775 |

**Document descriptors**

| | |
|---|---|
| **Deliverable No.** | 2.5 |
| **Work package** | WP2 |
| **Deliverable lead** | UTARTU |
| **Contributors** | UH-FIMM, QMUL, NTNU, HUS, MGH, UNISI |
| **Dissemination level** | Public |
| **Expected delivery date** | 30/06/2022 |
| **Submission date** | 30/06/2022 |

**Change history log**

| Version | Changes made | Prepared by | Reviewed by | Date |
|---|---|---|---|---|
| 0.1 | First draft shared with all partners | UTARTU | UH-FIMM, QMUL, NTNU, HUS, MGH, UNISI | 20.06.22 |
| 0.2 | Major revision based on feedback | UTARTU | Reedik Mägi, TARTU;<br>Samuli Ripatti, UH-FIMM | 21-28.06.22 |
| 1.0 | Minor revision - Formatting and references | Kristi Läll, UTARTU | Julius Anckar, UH-FIMM | 29.6.2022 |

# Contents

## Summary of findings

1. Genetic correlations estimated based on GWAS results from three cohorts are very high on average for EstBB-FinnGen and FinnGen-UKBB pairs and slightly lower for EstBB-UKBB pair.
2. For the majority of endpoints, heterogeneity was not detected. Out of 831 pruned genome-wide significant SNPs from 16 meta-analyses, 39 SNPs showed significant heterogeneity.

## Introduction

The purpose of work package two so far has been mapping of types and availability of existing data among participating cohorts and providing input to partners to define flagship endpoints in partners' datasets. Even though the definitions of endpoints are harmonized, the origin of the information (i.e hospital-based records, national registries, self-reported diagnoses, or combinations of many sources) on disease codes is different among partners, the accuracy of ICD codes varies across biobanks or some information might be missing for some endpoints. Therefore, even using harmonized definitions for endpoints does not guarantee that endpoints represent exactly the same phenotypes across cohorts.

Genome-wide association studies (GWAS) are meant to test genotype-phenotype associations, using genetic variants across the genome. Usually millions of genetic variants are included in each study. These types of analyses provide insights into genetic architecture of diseases/traits, allowing researchers to detect novel disease-genetic variant associations[1]. This type of study allows hypothesis-free exploration of the genetic background of the disease and facilitates many post-GWAS analyses for further investigation. Post-GWAS analyses can be used to pinpoint causal variants and genes, map implicated biological pathways, cell types and tissues, and explore genetic architecture shared by traits both on single variant level as well as on a genome-wide level.

Biobanks hold many different types of data about individuals, which allows researchers to construct a definition of a disease in several ways. Most of the biobanks have disease related information stored in a standardized way using either the ICD-10 or 9 classification. Drug prescription data and clinical measurements are also sometimes available. There may be other types of coding systems than ICD-10 or 9 sometimes, including self-reporting of a disease. However, how one defines a phenotype - which combination of disease codes/clinical measurements/prescribed drugs should be included for a case and which codes excluded for a control - can be set up in a variety of ways.

The focus of this delivery is to study the consistency of the GWAS results for multiple phenotypes across cohorts participating within the INTERVENE project. If there are large inconsistencies, then it is possible that some ICD codes are used differently between different cohorts, different coding methods create slightly different phenotype or disease definitions (e.g. ICD10 code vs. ICD9, self-reported phenotypes etc.), the precision of coding is different (digits given in ICD10 code) etc. Furthermore, other sources of heterogeneity (like environmental factors) contribute to a phenotype or for example, some information necessary for defining an endpoint is missing in the cohort database. If large inconsistencies are found, it might be necessary to modify the definition of the flagship endpoint to achieve a better concordance of GWAS results among partners. To address this possibility, alternative definitions of endpoints are considered using FinnGen data and the correlation between flagship endpoints genome-wide association study results and alternative endpoints genome-wide association study results are systematically explored.

# Methodology

This task includes three large parts:

1. Running GWAS on flagship endpoints defined in deliverable 2.3.
2. Heterogeneity between studies will be studied for available endpoints using genome-wide significant SNPs.
3. Genetic correlations between individual biobanks' GWAS results will be estimated using the LDSC (LD score regression) methodology.

We will describe each of them separately in the following chapters.

## Genome wide association study plan:

We proposed that all partners run genome-wide association analyses using Regenie or Saige[2–5]. These two were selected for the following reasons:

- They allow inclusion of related individuals which maximises power to detect association
- They are computationally efficient
- They have been developed specifically for biobank-scale data

Both methods use saddle-point approximation to account for case-control imbalance, which can arise in the biobank setting if the binary trait in question is very rare. It is important to take this potential imbalance into account, since it can affect the effect estimates, especially for rare variants.

Our GWAS study plan is based on the Global Biobank Meta-analysis initiative (https://www.globalbiobankmeta.org/; GBMI) GWAS plan. This was chosen because the GBMI includes 23 biobanks around the world and they have had similar efforts to run GWASs across multiple cohorts in a unified way while including relatives.

Both SAIGE and Regenie perform single-variant association tests for binary traits and quantitative traits. For binary traits, it is possible to use the saddlepoint approximation (SPA) to account for case-control imbalance.

According to the analysis plan, the following standard association model was used:

*Phenotype ~ variant + age + sex + PCs + biobank_specific_covariates*

Here, biobank-specific covariates stand for covariates used to correct technical artifacts (such as genotyping batch) and not risk factors or other comorbidities.

Cohort-level analyses were filtered based on imputation INFO score and/or MAF values, specifics in Results.

Due to the heavy computational burden of each GWAS analysis (for example, a logistic mixed model with SAIGE for a binary phenotype in UKBB requires 517 CPU hours and 10.3G of memory[6], and although Regenie is somewhat better in this respect, whenever possible, already calculated summary statistics were used to reduce computational costs and associated CO2 footprint[4].

## GWAS meta-analysis

After collecting cohort-level GWAS summary statistics, the files were converted to b37, if necessary, using the binary liftover tool (https://liftover.broadinstitute.org/), and formatted to contain the following necessary columns:

1. Marker name in the format of chromosome:position:A1:A2, whereby A1 and A2 stand for the two alleles of this marker in alphabetical order. This format was chosen to make sure the marker names in each cohort file have a standard and uniform format. Alleles were ordered alphabetically, because the effect allele might differ in each cohort.
2. Chromosome
3. Position
4. Effect allele
5. Non-effect allele
6. Rs-number
7. Beta (effect size)
8. SE (standard error of the effect size)
9. EAF (effect allele frequency in the cohort)
10. N (sample size for the cohort)

We then proceeded to conduct a GWAS meta-analysis. Current data freeze includes only individuals with genetic European ancestry.

For meta-analysis, we chose the GWAMA software[7], since the method is open-source, and provides a wide range of meta-analysis summary statistics, including two different measures for heterogeneity. Evaluating heterogeneity between different cohorts gives a proxy measure for estimating consistency in endpoints‘ definitions across cohorts. The Cochran's statistic provides a test of heterogeneity of allelic effects for each genetic variant. An alternative statistic, $I^2$, quantifies the extent of heterogeneity in allelic effects across studies, over that expected by chance, and is more robust to variability in the number of studies included in the meta-analysis.

After the meta-analysis for each endpoint, we defined genome-wide significant signals (variants with an association p-value < $5x\ 10^{-8}$). This threshold was chosen because it is the gold standard for statistically reliable associations in GWAS. Then we evaluated the heterogeneity statistics for these variants and created forest plots for variants with significant heterogeneity to evaluate effect sizes across cohorts and detect sources of heterogeneity.

## LD Score regression for genetic correlations

LD Score regression (LDSC)[8] is a method for estimating heritability and genetic correlation[9] ($r_g$) from GWAS summary statistics. This method was chosen because it is open source, takes summary statistics as an input, and there is no need for individual level data. The method is also computationally efficient and thus suitable for the purposes of this delivery. It is widely used in the genetic epidemiology field, as illustrated by the current number of citations (1654), and therefore thoroughly tested. In short, the method aims to quantify the separate contributions of polygenic effects and various confounding factors, such as population stratification, based on summary statistics from GWAS. The LDSC is suitable for ancestry-matched data (for example, all GWAS summary statistics are from European ancestry samples). If one should have multi-continental GWAS data, the right approach is to estimate genetic correlation for each continent separately then average the results[10]. Current data freeze includes only individuals with genetic European ancestry.

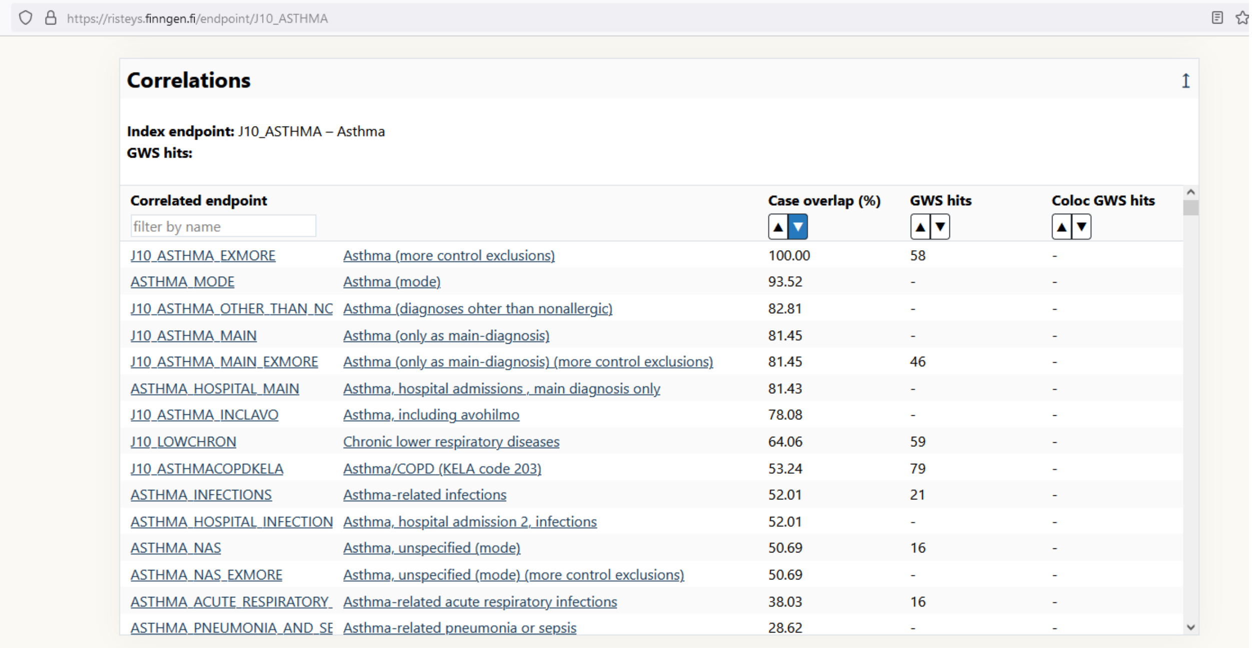
LDSC provides both the SNP-heritability estimate ($h^2_{SNP}$) based on each GWAS summary file together with the standard error (SE) estimate and genetic correlation ($r_g$) estimate between GWAS summary results. We also calculate confidence intervals (1.96* *LDSC* standard errors on either side of the point estimate) for $h^2_{SNP}$ and $r_g$.

By calculating the genetic correlations between endpoint GWAS summary statistics, we can estimate the degree to which genetic variants have a consistent effect across biobanks. There are a couple of notes about the results and behaviour of LDSC. First, LDSC is not a bounded estimator with +-1, so it can produce estimates outside of these values due to sampling variation[11]. Therefore, unlike Pearson correlation estimates, the ld score regression based estimates of genetic correlations may have values below -1 and over 1, but the interpretation of these values is similar to the Pearson correlation coefficient being close to -1 or 1, in case they do not deviate too much from abs(1). Second, standard errors of genetic correlations are roughly a function of the sample size of GWAS and heritability estimates. Therefore, when sample size is small and/or heritability estimates are low, SE will be large and it is likely not possible to get stable $r_g$ estimates[11]. It is generally said that if heritability results are not significantly different from zero for either trait (can happen when cohort-level GWAS results are under-powered or when loci of large effect size exist within the GWAS), then genetic correlation results are not likely to be reliable. To address the possibility of heritability being statistically indifferent from zero, we define a subset of endpoints in that way that $h^2_{SNP}$ estimate is not zero (ie 95% CI do not include 0) for either GWAS included in LDSC, calling them “reliable subsets”. We also set a threshold for pairwise $r_g$, defining cohort-wise correlation for an endpoint high if it was above 0.8 (lending the threshold for “high” from imputation related articles, where $r^2$ is squared correlation between known genotypes and imputed allele dosages, high often defined as 0.7 or 0.8[12–14]) and would suggest considering alternative definition for an endpoint if pairwise correlation for an endpoint between any two cohorts were lower and meta-analysis results would support heterogeneity among genome-wide significant SNPs.

LDSC analysis was carried out centrally for biobank-wise genetic correlations. All of the GWAS summary statistics were reformatted before running LDSC using the *munge_sumstat* tool included in the LDSC github repository[15]. Because imputation quality is a confounder for LDSC and as the GWAS summary statistics did not include information about imputation quality, the SNPs were filtered using the HapMap3 SNPs (https://data.broadinstitute.org/alkesgroup/LDSCORE/w_hm3.snplist.bz2) as suggested, because these are usually well imputed[16]. The default parameters were used for *munge_sumstat* as well as for LDScore regression analysis. Pre-computed LD Scores (https://data.broadinstitute.org/alkesgroup/LDSCORE/eur_w_ld_chr.tar.bz2) from 1000 Genomes European data were used for estimating genetic correlations as advised in the tutorial for LDSC.

## Workflow description for calculating genetic correlations between index flagship endpoints and alternative definitions

For each flagship endpoint, we queried two alternative definitions using the FinnGen’s database developed to explore results per phenotype (https://risteys.finngen.fi/). Alternative endpoints were selected that way that the overlap with the flagship endpoint cases were maximized as we are most interested in the endpoints as they are defined already and were looking for alternatives which would slightly alternate the definitions, preferably definitions of controls. Overlap was queried from the table “Correlations”, while searching for flagship endpoint FinnGen’s name (as given in Table 1) and sorting for “case overlap” (see Figure 1, an example for Asthma). Queries were done during 12.05-24.05.2022 and on 17.06.2022 (for I9_STR and I9_VTE, for which the database gave an “internal server error” earlier). For eight alternative definition endpoints the GWAS summary statistics were not available, each of these endpoints was substituted using the next endpoint in the correlation table for which the summary statistics were available.



https://risteys.finngen.fi/endpoint/J10_ASTHMA

## Correlations

**Index endpoint:** J10_ASTHMA – Asthma
**GWS hits:**

| Correlated endpoint<br>filter by name | | Case overlap (%)<br>▲▼ | GWS hits<br>▲▼ | Coloc GWS hits<br>▲▼ |
|---|---|---|---|---|
| J10_ASTHMA_EXMORE | Asthma (more control exclusions) | 100.00 | 58 | - |
| ASTHMA_MODE | Asthma (mode) | 93.52 | - | - |
| J10_ASTHMA_OTHER_THAN_NC | Asthma (diagnoses ohter than nonallergic) | 82.81 | - | - |
| J10_ASTHMA_MAIN | Asthma (only as main-diagnosis) | 81.45 | - | - |
| J10_ASTHMA_MAIN_EXMORE | Asthma (only as main-diagnosis) (more control exclusions) | 81.45 | 46 | - |
| ASTHMA_HOSPITAL_MAIN | Asthma, hospital admissions , main diagnosis only | 81.43 | - | - |
| J10_ASTHMA_INCLAVO | Asthma, including avohilmo | 78.08 | - | - |
| J10_LOWCHRON | Chronic lower respiratory diseases | 64.06 | 59 | - |
| J10_ASTHMACOPDKELA | Asthma/COPD (KELA code 203) | 53.24 | 79 | - |
| ASTHMA_INFECTIONS | Asthma-related infections | 52.01 | 21 | - |
| ASTHMA_HOSPITAL_INFECTION | Asthma, hospital admission 2. infections | 52.01 | - | - |
| ASTHMA_NAS | Asthma, unspecified (mode) | 50.69 | 16 | - |
| ASTHMA_NAS_EXMORE | Asthma, unspecified (mode) (more control exclusions) | 50.69 | - | - |
| ASTHMA_ACUTE_RESPIRATORY_ | Asthma-related acute respiratory infections | 38.03 | 16 | - |
| ASTHMA_PNEUMONIA_AND_SE | Asthma-related pneumonia or sepsis | 28.62 | - | - |

**Figure 1.** An example of querying alternative endpoint definitions for Asthma from FinnGen's database, using case overlap for sorting possibilities.

After collecting the GWAS summary statistics for possible alternative definitions, the files were converted to b37, reformatted using the *munge_sumstats* tool and the pairwise genetic correlations were calculated with the LDSC method between each flagship endpoint and its alternative definition endpoints.

# Results

## Overview of cohorts participating in GWAS meta-analysis

By the beginning of June 2022, data freeze was done to perform the first round of analyses. By that time, UKBB, FinnGen (which includes the Biobank of Helsinki data, therefore not separately analysed) and EstBB had delivered the GWAS results. Total number of cases for endpoints delivered by all participating biobanks separately and together is presented in Table 1. All endpoints, including ones only present in a subset of GWAS are in Supplementary Table 1.

*Table 1. Sample sizes for evaluated endpoints present in all participating biobanks*

| Endpoint name in FinnGen manifest | Partners (n of cases/controls) | | | Total | |
|---|---|---|---|---|---|
| | FinnGen | EstBB | UKBB | cases | controls |
| C3_COLORECTAL | 4401/256004 | 1725/ 198143 | 7269/ 400837 | 13,395 | 854,984 |
| C3_BREAST | 11573/135488 | 2690/ 128295 | 13947/ 206669 | 28,210 | 470,452 |
| T2D | 37031/214308 | 12425/187443 | 28668/386926 | 78,124 | 788,677 |
| C3_PROSTATE | 8709/104635 | 2266/ 66617 | 10825/176665 | 21,800 | 347,917 |
| I9_CHD | 25707/234698 | 17594/ 182274 | 26972/347119 | 70,273 | 764,091 |
| I9_STR | 14351/238854 | 4575/ 195293 | 10394/ 360479 | 29,320 | 794,626 |
| G6_AD_WIDE | 7329/252879 | 558/ 199310 | 2160/ 405946 | 10,047 | 858,135 |
| F5_DEPRESSIO | 28098/228817 | 51926/ 147942 | 36838/382918 | 116,862 | 759,677 |
| RHEUMA_SEROPOS_OTH | 5793/254548 | 2389/ 197479 | 985/ 373007 | 9,167 | 825,034 |
| I9_VTE | 11288/249117 | 13202/ 186666 | 13273/ 360818 | 37,763 | 796,601 |
| M13_OSTEOPOROSIS | 3960/249139 | 9714/ 190154 | 16327/ 355286 | 30,001 | 794,579 |
| AUD_SWEDISH | 14864/245541 | 10119/ 189749 | 8941/ 365150 | 33,924 | 800,440 |
| E4_HYTHYNAS | 32871/221975 | 14238/ 185630 | 24458/ 345172 | 71,567 | 752,777 |
| G6_EPLEPSY | 7224/208845 | 4997/194871 | 5402/ 402704 | 17,623 | 806,420 |
| GE_STRICT | 1751/253181 | 276/199592 | 617/ 414698 | 2,644 | 867,471 |
| FE_STRICT | 728/253181 | 1875/197993 | 553/ 407553 | 3,156 | 858,727 |

Summary Data for FinnGen is publicly available (https://www.finngen.fi/en/access_results, release R6[17]). Total sample size of the FinnGen cohort used in GWAS was ~260,000 and in total, ~17 million variants were included in the analyses. Analyses were performed with SAIGE v0.39.1. All models were adjusted for sex, age, 10 PCs and genotyping batch (see https://finngen.gitbook.io/documentation/methods/phewas).

Only variants with minimum allele count of 5 (SAIGE optionminMAC = 5) and INFO>0.6 are included (more details about QC here: https://finngen.gitbook.io/documentation/methods/phewas/quality-checks)

The Estonian Biobank cohort had a total sample size of ~200,000 individuals and approximately 30M variants in the GWAS analyses. Analyses were performed with Regenie v2.2.4 and all models were adjusted for sex, age and 10 PCs. Regenie default settings filter out variants with a MAC <5, and before meta-analysis, all variants with an imputation INFO score < 0.4 were excluded.

For UKBB (with sample size of ~400,000), we had 2 phases of GWAS were performed: first set consisted of 10 diseases (AUD_SWEDISH, E4_HYTHYNAS, F5_DEPRESSIO, GE_STRICT, I9_CHD, I9_STR, I9_VTE, M13_OSTEOPOROSIS, RHEUMA_SEROPOS_OTH, T2D). Those analyses were done with SAIGE and a filter comparing the MAF reported in the gnomAD EUR population and UKBB population was applied (removing SNPs with difference greater than 0.2), leaving sumstats with ~25 million SNPs. Second phase was performed for 6 diseases with REGENIE v3.1.1 to analyze the data with about 61M SNPs included. SNPs with INFO>0.3 were included in the shared sumstats. Models were adjusted for sex, age 10PCs, genotyping batch and assessment centre. Two different programs were used as GWAS were performed in two stages by two different people. Some endpoints did not have summary statistics available yet by the time of the first data freeze but will be added in the near future.

EstBB and UKBB data used only ICD codes to define the phenotypes whereas FinnGen additionally used KELA prescription codes.

## Meta-analysis and heterogeneity in SNP effect estimates

We conducted GWAS meta-analyses for the phenotypes described above using GWAMA as described in the Methods section. Only endpoints with all three cohorts contributing were analyzed. To evaluate the heterogeneity between different cohorts (and thus different endpoint definitions), we first looked at the distribution of the $I^2$ statistic from the meta-analysis. In brief, $I^2$ quantifies the extent of heterogeneity in allelic effects among cohorts, estimating the probability (from 0 to 1) that there is heterogeneity in this locus. For this part of the analysis, we extracted variants with a meta-analysis p-value less than $1 \times 10^{-5}$. This threshold was chosen since it is often used in GWAS studies as the so-called threshold of suggestive significance. We only considered sentinel variants from each hit locus by pruning those significant loci and removing variants in 100Kb distance from the top variant.

We identified a total of 4567 regions with a sentinel variant p-value less than 1 x $10^{-5}$ in the meta-analysis. The distribution of $I^2$ values for these signals can be seen on Figure 2. While for the majority of phenotypes, the median $I^2$ value for these variants was 0, RHEUMA_SEROPOS_OTH, I9_VTE and I9_CHD exhibited higher $I^2$ values (0.23, 0.46, 0.24, respectively).

We then focused on only genome-wide significant variants ($p<5x10^{-8}$). In this analysis we only considered sentinel variants from each hit locus by pruning the genome-wide significant loci and removing variants in 500Kb distance from the top variant. We identified 831 genome-wide significant variants. Of these, 39 show statistically significant heterogeneity (Cochran's p-value < 0.05/897=5.6 x $10^{-5}$) and 90 loci have an $I^2$ value larger than 0.8. The forest plots showing effect estimates in individual cohorts for the 39 statistically significantly heterogeneous loci can be found https://docs.google.com/document/d/10cJmHaNvXtbpQGHiF3SHIvMFtYgTOz0_iZY_z9IRwvI/edit?usp=sharing).

While for some loci, the heterogeneity might be caused by allele frequency differences in participating cohorts, for others the allele frequencies do not differ substantially and cohort- or endpoint-specific effects are likely. For example, C3_PROSTATE, where the effect allele frequencies differ 2 to 5 times between cohorts for both rare (minor allele frequency <1%) and common (minor allele frequency >1%) variants and largest effect estimates are seen in FinnGen, where the heterogeneous loci also have the highest minor allele frequency. For phenotypes with an autoimmune etiology (E4_HYTHYNAS, J10_ASTHMA, RHEUMA_SEROPOS_OTH), the detected heterogeneous loci are in the *HLA* region on chromosome 6, which is very diverse across populations. Although for these loci we see no differences in allele frequencies across the evaluated cohorts, it is still possible we are capturing some population-specific effects with our analysis. As to cohort-specific effects, for certain later-onset phenotypes (such as cardiovascular traits or autoimmune diseases), the age-distribution in the studied cohorts might also affect the observed heterogeneity.

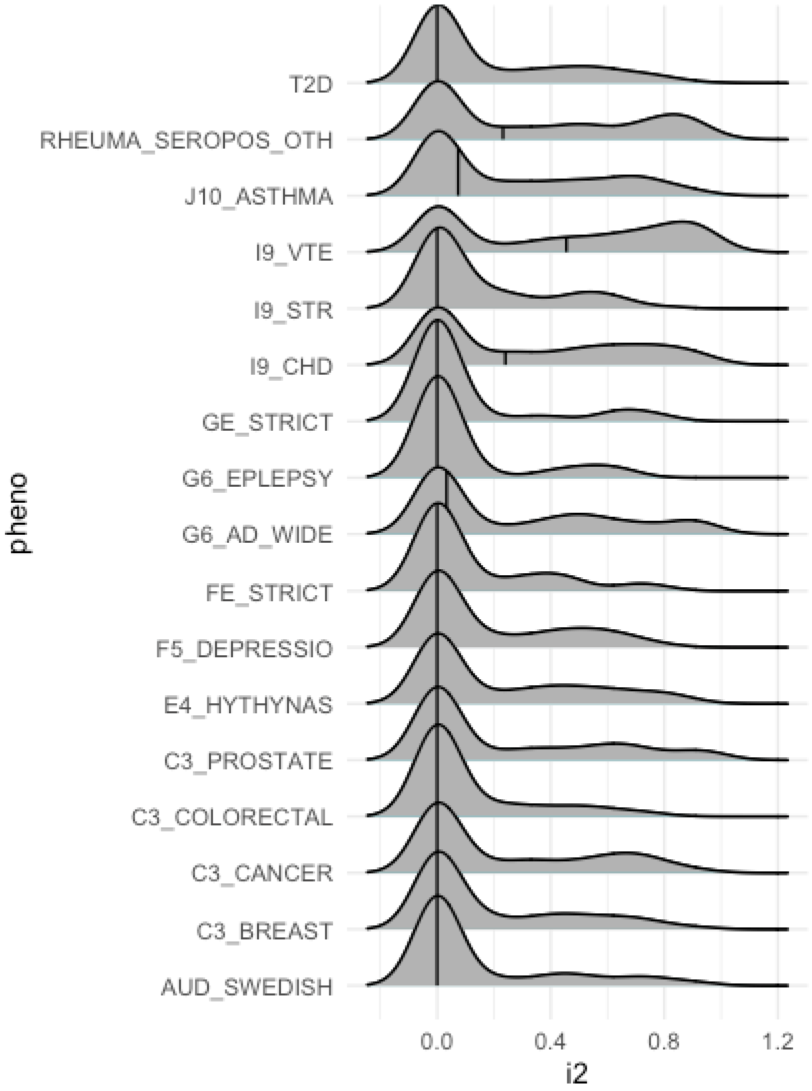


**Figure 2.** Density plots of $I^2$ values for sentinel variants with a p-value less than 1 x $10^{-5}$ in the meta-analysis. Median values are shown with a vertical line.

## Genetic correlations

Some of the $r_g$ estimates are out of bound due to too small sample size of GWAS, statistically insignificant $h^2_{SNP}$ estimate (or two small $h^2_{SNP}$ values) or due to too large SE-s of $h^2_{SNP}$. This was expected, as some of the flagship endpoints are with low prevalence in individual biobanks. We did not filter out any endpoints based on effective sample size or ratio of expected/observed variability.

For each analysis, some warnings about unstable $r_g$ or $h^2_{SNP}$ were reported (see Tables here: https://docs.google.com/spreadsheets/d/1EVTEbMO2fJwr6AzbAb_gGWio8SKgICU1/edit?usp=sharing&o

uid=105250623214509639753&rtpof=true&sd=true, column "warnings"). For EstBB-FinnGen analysis, I9_SAH, C3_MELANOMA_SKIN, C3_BRONCHUS_LUNG and FE_STRICT had $r_g < 0.8$ and for SAH and MELANOMA, the correlation was even negative (see Table 2). However, for all of them, the $h^2_{SNP}$ estimates were not different from 0, se of $r_g$ are large and therefore $r_g$ estimates unstable. For EstBB-UKBB analysis, C3_PROSTATE and I9_VTE had $r_g$ ~ 0.75 and $h^2_{SNP}$ estimates were statistically different from zero.

**Table 2.** Endpoints with $r_g$ estimates below 0.8 in any cohort-pair analysis

| FinnGen_name | rg | rg_se | cohort_pair |
|---|---|---|---|
| I9_SAH | -1.1129 | 1.3908 | EstBB_FinnGen |
| C3_MELANOMA_SKIN | -0.0736 | 1.2720 | EstBB_FinnGen |
| C3_BRONCHUS_LUNG | 0.3423 | 0.6734 | EstBB_FinnGen |
| FE_STRICT | 0.5559 | 1.9132 | EstBB_FinnGen |
| C3_PROSTATE | 0.7476 | 0.1871 | EstBB_UKBB |
| I9_VTE | 0.7478 | 0.1413 | EstBB_UKBB |

We further investigated $r_g$ estimates in "reliable" subsets as defined in the methods section. For EstBB-Finngen, 19/34 endpoints had $h^2_{SNP}$ for both biobanks significantly different from zero and their $r_g$ estimates varied between 0.91-1.24 (median 1.043), indicating significant overlap in genetic effects across biobanks. While estimating genetic correlations for EstBB-UKBB GWAS, 9/17 endpoints had $h^2_{SNP}$ estimates based on both biobank GWAS statistically different from zero and for those 9, $r_g$ estimates varied 0.748-0.968 (median 0.878). For FinnGen-UKBB, in total 11/17 endpoints belong to the reliable subset. For those endpoints, $r_g$ estimates varied between 0.875 and 1.263 (median 1.14). Genetic correlation estimates with 95% CI are presented in Figure 3 for endpoints, which $h^2_{SNP}$ estimate in each cohort was not zero.

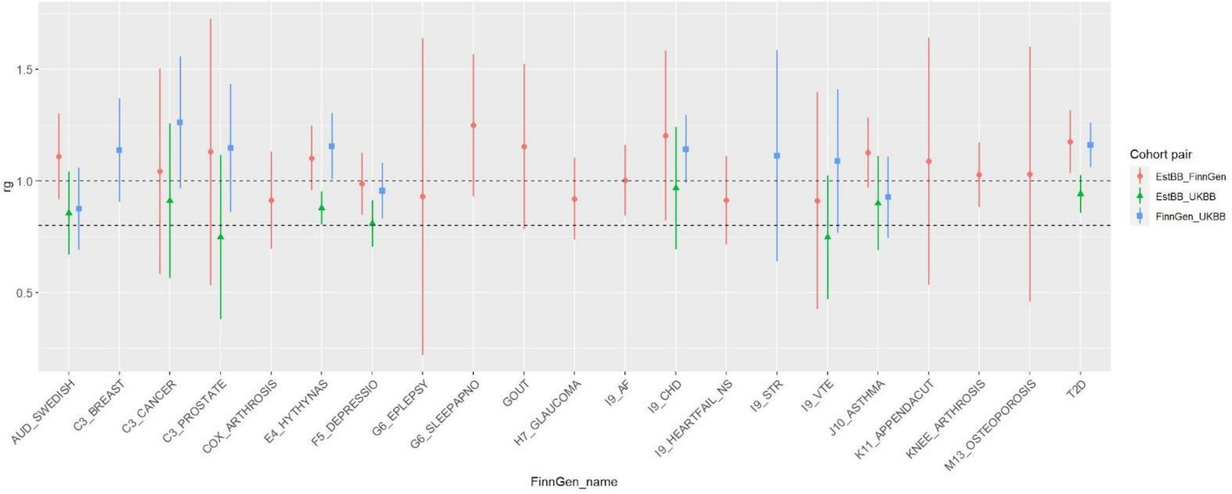


**Figure 3.** Pairwise genetic correlation estimates for endpoints, which had $h^2_{SNP}$ estimate statistically different from zero based on both input GWAS. Large variability for some endpoints with very low $h^2_{SNP}$ estimate in at least one cohort can be observed. Dashed lines are r=1 and r=0.8.

Overall, it seems that FinnGen-EstBB and FinnGen-UKBB GWAS datasets are very strongly correlated, whereas EstBB-UKBB datasets show slightly lower $r_g$ estimates on average. Majority of genetic correlations are high (median 1.03, interquartile range (IQR) 0.913-1.146), lending support to the hypothesis that flagship endpoints' definitions tend to measure the endpoints similarly in different biobanks. However, one must note that a. only about half of the endpoints belonged to the "reliable set" based on cohort-specific $h^2_{SNP}$ variability and b. some of the standard errors of $r_g$ are still quite large (Figure 3) in "reliable set", often for those endpoints, which have at least one $h^2_{SNP}$ estimate close to zero. Even when $r_g$ itself is >0.8, 95% CI lower half for several endpoints are below 0.8. All pairwise $r_g$ estimates are plotted in Supplementary Figure 1.

We also hypothesized that $r_g$ may depend on the age of onset for diseases (Figure 4), as this may reflect different disease etiology (more severe cases may have earlier onset). For endpoints like VTE, ASTHMA, G6_AD_WIDE and RHEUMA_SEROPOS_OTH exhibiting more heterogeneity in effect sizes (Figure 2), age at first available diagnosis also seems to vary between cohorts. For some endpoints (such as GE_STRICT and T1D), which generally are early-onset diseases, we see significantly later age at first diagnosis in the UKBB. However, it is likely that this is an artefact that stems from the fact that ICD-based electronic NHS hospital inpatient records are available starting from 1996 (England from 1996, Scotland from 1997 and Wales from 1998) in the UKBB, and given the age distribution for this cohort (aged 40-69 at enrolment[18]), the initial diagnoses are simply missing from the data.

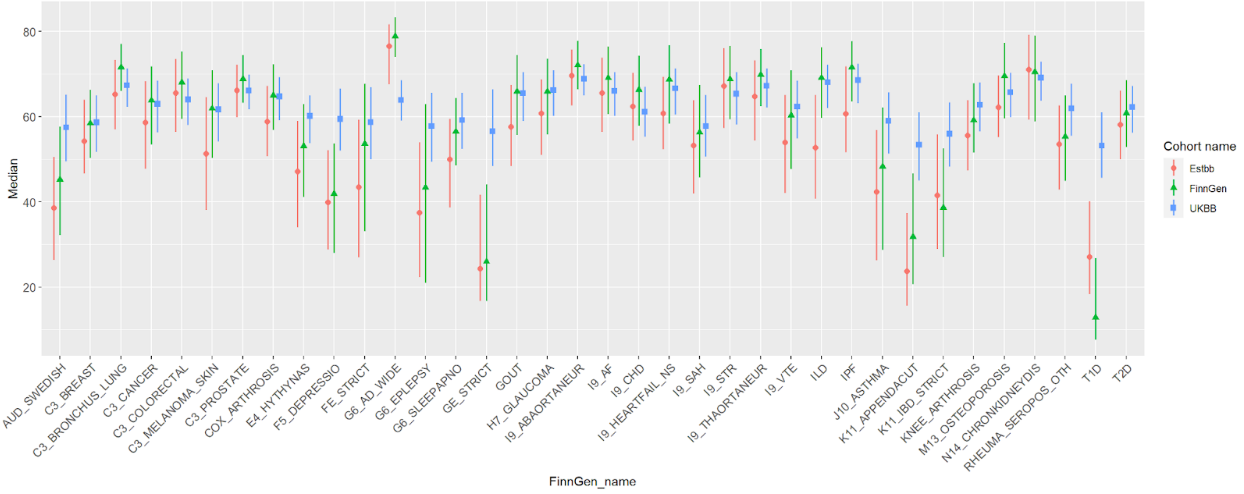


**Figure 4.** Age at first diagnosis for cases for each evaluated endpoint in all three cohorts.

## Genetic correlations between flagship endpoints and alternative considered definitions using FinnGen data

We had proxies for 32 endpoints - for two endpoints (I9_ABAORTANEUR, I9_THAORTANEUR), GWAS results from the FinnGen R6 database were not found, therefore no alternatives could be looked for. As mentioned before, we prioritized selecting alternative endpoint definitions (two per each index endpoint) in a way that preferably cases remained as similar as possible, and controls were modified.

The case overlap for first alternative endpoints varied between 27%-100% (IQR 63.4%-100%), second alternative endpoints had case overlap between 2.8%-100% (IQR 51.8%-89.7%). The median genetic correlation between first alternative and index endpoints as well as second alternative and index endpoints was 1, IQR varied roughly between 0.99-1.06 in both scenarios. For some endpoints, estimating $h^2_{SNP}$ or $r_g$ was not possible due to technical reasons. All genetic correlations between index phenotype and alternatives for each phenotype are seen in Figure 5.

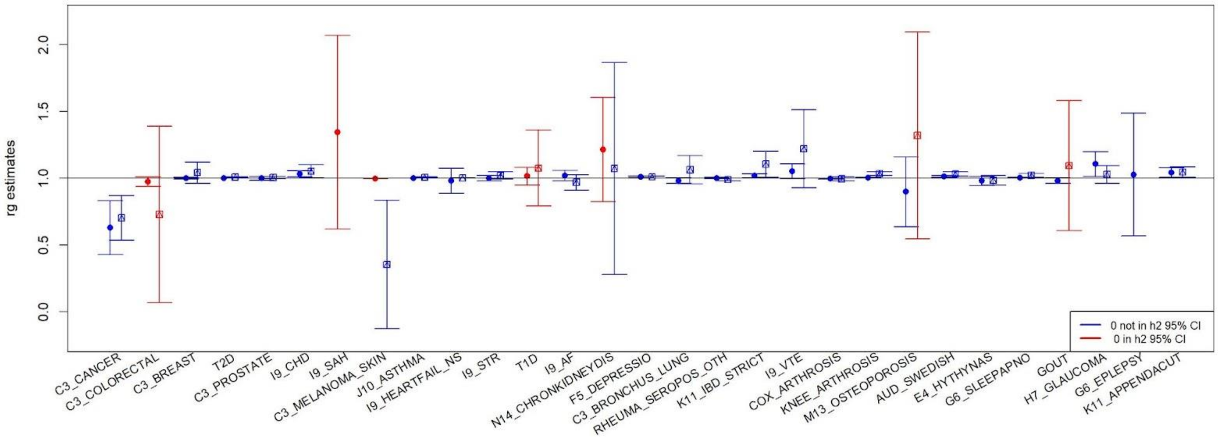


**Figure 5.** Pairwise genetic correlations between the index flagship endpoint and two alternatives (if possible) selected from Finngen's database. Blue colour indicates the heritability estimate ($h^2_{SNP}$) for alternative endpoint is statistically different from 0 (95% CI do not include 0), red indicates that $h^2_{SNP}$ estimate based on GWAS of the alternative endpoints is not significantly different from zero. If $h^2_{SNP}$ is not statistically different from zero, it usually indicates unstable $r_g$ estimates.

We can see that almost all endpoints have an alternative definition which correlates highly with the index phenotype. However, C3_CANCER does not seem to have a very similar alternative and some endpoints have alternatives, which $r_g$ estimates have large variability, therefore less likely to be considered as possible alternative definitions.

All pairwise correlations between index endpoints and two alternative definitions with more details can be found in the Supplementary Table[1]

## Discussion and next steps

Four large biobanks (EstBB, Finngen - including Helsinki Biobank, and UKBB) contributed to the genome-wide association studies bringing total sample size for most of the endpoints around 850,000. All participating cohorts reported GWAS results for European ancestry.

Based on our evaluation of genetic loci identified in GWAS meta-analysis of the tested endpoints, around 5% of genome-wide significant loci exhibit statistically significant heterogeneity. While in some cases, it

[1] https://docs.google.com/spreadsheets/d/1EVTEbMO2fJwr6AzbAb_gGWio8SKgICU1/edit?usp=sharing&ouid=105250623214509639753&rtpof=true&sd=true.

can be attributed to allele frequency differences between analyzed populations, for others the allele frequencies are similar and differences in endpoints seem a plausible candidate to explain the observed heterogeneity. Potential sources for endpoint heterogeneity can include environmental factors, different age distribution in cohorts, sources of ICD codes (inpatient/hospital data or primary health care, whereby hospitalized cases likely reflect more severe cases) or inherent differences in how ICD codes are used by medical systems across countries. In line with this, our comparison of age at diagnosis distribution in analyzed cohorts shows that for endpoints, where age at diagnosis is used for definition, it is important to also consider the source of data and how far back it goes, as in older cohorts, early-onset diagnoses might be missing from the data if electronic health records are incomplete.

Most of the genetic correlations between cohorts are high - median being 1.03. EstBB-FinnGen GWAS results were more strongly correlated that EstBB-UKBB GWAS results. FinnGen-UKBB genetic correlations were overall high. A few endpoints (like VTE and PROSTATE cancer) with statistically significant heterogeneous loci also have $r_g$ estimates lower than 0.8 at least in one cohort-pair analysis. Therefore, it could be worth either to consider alternative definitions for these endpoints (as they exist with high $r_g$ based on FinnGen data, see Figure 5) or inclusion of some threshold on age of first onset in the current definition to try to lessen the heterogeneity between cohorts. However, 95% CI for those two endpoints still include 1, so we cannot make a definite conclusion about $r_g$ being too low to continue with current endpoint definition.

One downside of LDSC is that only for about half of the phenotypes, for which GWAS results were available, heritability estimates in both biobanks under investigation differed significantly from zero. This in combination with warnings from LDSC stating that estimated $h^2_{SNP}$ or GWAS N is too low to calculate genetic correlations means that some of the genetic correlation estimates have quite a large standard error, making $r_g$ estimates unstable and therefore not highly trustworthy to make any conclusion based on them. This is an overall limitation of using genetic methods for estimating endpoint heterogeneity. Genetic methods are sensitive to heritability of an endpoint and overall power of GWAS to reliably detect genetic associations in a given sample size (.../"*GWAS with small effective sample sizes have insufficient power for LDSR to detect polygenic effects, leading to near-zero estimates of heritability*"/[19]). We currently have not done any filtering like UKBB has (based on effective sample size as well as $h^2_{SNP}$ being statistically different from 0[20]). In their post, they say that $N_{eff} < 4500$ (which translates roughly to 1100 cases for EstBB and FinnGen sample size) translates to no confidence in LDSR $h^2_{SNP}$ results and use only phenotypes with at least $N_{eff}$ > 40,000 in their LDRS analysis with some additional filters. In the second stage analysis, different Neff and $h^2_{SNP}$ filters will be tried in our analyses.

We have performed a genetic LDSC analysis between flagship index endpoints and possible alternative definitions, using FinnGen data. For almost all index endpoints, there are alternatives available with high genetic correlations. However, having any cancer diagnosis (C3_CANCER) did not have a good alternative and a few endpoints had alternatives, for which $r_g$ estimates varied significantly. It should also be noted that the alternative definitions we evaluated are also based on the ICD classification system. As a result of the harmonization activities taking place in Deliverable 2.4, in the future it might also be possible to evaluate alternative endpoint definitions based on for example self-reported data or data extracted from additional sources such as epicrises while using specific key-words or -phrases, which will allow for further comparisons to be made.

Our results provide insights for planning further analysis steps. Overall high genetic correlations indicate that selected definitions for endpoints tend to perform similarly across participating cohorts. Based on heterogeneity and genetic correlation analyses, definitions of VTE and prostate cancer seem to perform differently across biobanks and alternative definitions could be considered. Once UKBB results for the rest of the endpoints can be included, a comprehensive list of endpoints showing significant heterogeneity and low $r_g$ estimates can be determined. It is important to note that in further polygenic risk score (PRS) related analysis, careful consideration of how to select individuals into analysis for some endpoints like T1D or K11_APPENDACUT, where age of onset cannot be determined due to possible truncation of diagnosis data in some biobanks, must be performed.

One should note that even though GWAS for some endpoints were under-powered to accurately estimate heritability or genetic correlations between cohorts, similar caveats are not applicable for PRS analyses as:

1. Selected endpoints have been shown to have a genetic component and our input summary statistics selected to construct PRS are more highly powered than individual GWAS from our partners
2. As PRSs tend to have higher effects than any individual common SNP, PRS-phenotype associations can be detected even for low prevalence diseases. So unstable $r_g$ results do not automatically mean that a specific endpoint should be excluded from further analysis.

In the current data freeze, participating cohorts have contributed only with individuals from European ancestry. Therefore, we are unable to address how well are endpoint definitions transferable between cohorts with more diverse genetic backgrounds or how well current endpoint definitions capture flagship diseases in cohorts with different coding systems than ICD-9/10. It is also important to note that the age structure varies in different biobanks - while the UKBB participants are relatively older (aged 40-69 at enrolment), the EstBB for example includes all adult volunteers, so the youngest biobank participants are 18 years old. While age or year of birth is commonly used as a covariate in GWAS analyses, it can still affect the results. We will perform a second stage analysis of genetic correlations once all partners with more diverse data are able to contribute.

## References

1. Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. *Nat Rev Genet*. 2019;20(8):467-484. doi:10.1038/s41576-019-0127-1

2. Zhou W, Nielsen JB, Fritsche LG, et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *bioRxiv*. November 2017:212357. doi:10.1101/212357

3. Zhou W. weizhouUMICH/SAIGE. https://github.com/weizhouUMICH/SAIGE. Accessed June 29, 2022.

4. Mbatchou J, Barnard L, Backman J, et al. Computationally efficient whole-genome regression for quantitative and binary traits. *Nat Genet*. 2021;53(7):1097-1103. doi:10.1038/s41588-021-00870-7

5. Home - regenie. https://rgcgithub.github.io/regenie/. Accessed June 29, 2022.

6. Zhou W, Nielsen JB, Fritsche LG, et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat Genet*. 2018;50(9):1335-1341. doi:10.1038/s41588-018-0184-y

7. Mägi R, Morris AP. GWAMA: software for genome-wide association meta-analysis. *BMC Bioinformatics*. 2010;11(1):288. doi:10.1186/1471-2105-11-288

8. Bulik-Sullivan B, Loh PR, Finucane HK, et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet*. 2015;47(3):291-295. doi:10.1038/ng.3211

9. Bulik-Sullivan B, Finucane HK, Anttila V, et al. An atlas of genetic correlations across human diseases and traits. *Nat Genet*. 2015;47(11):1236-1241. doi:10.1038/ng.3406

10. What Data Are Necessary to Estimate Genetic Correlation? · bulik/ldsc Wiki. https://github.com/bulik/ldsc/wiki/What-Data-Are-Necessary-to-Estimate-Genetic-Correlation%3F. Accessed June 29, 2022.

11. why are the genetic correlation out of the bound of [-1, 1] and have very large se? · Issue #89 · bulik/ldsc. https://github.com/bulik/ldsc/issues/89. Accessed June 29, 2022.

12. Lent S, Deng X, Adrienne Cupples L, Lunetta KL, Liu C, Zhou Y. Imputing rare variants in families using a two-stage approach. In: *BMC Proceedings*. Vol 10. BioMed Central Ltd.; 2016:48. doi:10.1186/s12919-016-0032-y

13. Kreiner-Møller E, Medina-Gomez C, Uitterlinden AG, Rivadeneira F, Estrada K. Improving accuracy of rare variant imputation with a two-step imputation approach. *Eur J Hum Genet*. 2015;23(3):395-400. doi:10.1038/ejhg.2014.91

14. Liu Q, Cirulli ET, Han Y, Yao S, Liu S, Zhu Q. Systematic assessment of imputation performance using the 1000 Genomes reference panels. *Brief Bioinform*. 2015;16(4):549-562. doi:10.1093/bib/bbu035

15. bulik/ldsc: LD Score Regression (LDSC). https://github.com/bulik/ldsc. Accessed June 29, 2022.

16. Heritability and Genetic Correlation · bulik/ldsc Wiki. https://github.com/bulik/ldsc/wiki/Heritability-and-Genetic-Correlation. Accessed June 29, 2022.

17. Kurki MI, Karjalainen J, Palta P, et al. FinnGen: Unique genetic insights from combining isolated

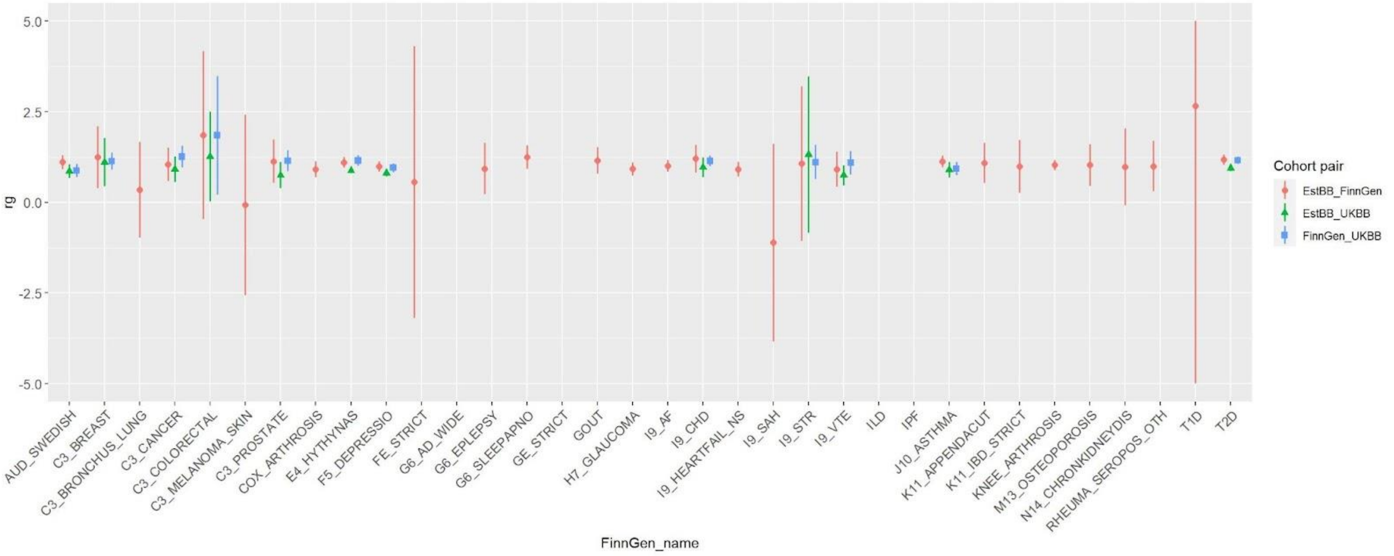
population and national health register data. *medRxiv*. March 2022:2022.03.03.22271360. doi:10.1101/2022.03.03.22271360

18. Allen N, Sudlow C, Downey P, et al. UK Biobank: Current status and what it means for epidemiology. *Heal Policy Technol*. 2012;1(3):123-126. doi:10.1016/j.hlpt.2012.07.003

19. Insights from estimates of SNP-heritability for >2,000 traits and disorders in UK Biobank — Neale lab. https://www.nealelab.is/blog/2017/9/20/insights-from-estimates-of-snp-heritability-for-2000-traits-and-disorders-in-uk-biobank. Accessed June 29, 2022.

20. Defining Confidence Levels for UKB Round 2 LDSR Analyses. https://nealelab.github.io/UKBB_ldsc/confidence.html#summary_of_confidence_ratings. Accessed June 29, 2022.

## Supplementary Figures and Tables



**Supplementary Figure 1.** All pairwise genetic correlations which LDSR provided with 95% CI plotted. For T1D, CI are truncated at (-5,5) to make the graph more informative.

**Supplementary Table 1.** Sample sizes for all evaluated endpoints

| Endpoint name in FinnGen manifest | Partners (n of cases/controls) | | | Total | |
|---|---|---|---|---|---|
| | FinnGen | EstBB | UKBB | cases | controls |
| C3_CANCER | 51271/209134 | 17546/182322 | | | |
| C3_COLORECTAL | 4401/256004 | 1725/198143 | 7269/400837 | 13,395 | 854,984 |
| C3_BREAST | 11573/135488 | 2690/128295 | 13947/ 206669 | 28,210 | 470,452 |
| T2D | 37031/214308 | 12425/187443 | 28668/386926 | 78,124 | 788,677 |
| C3_PROSTATE | 8709/104635 | 2266/66617 | 10825/ 176665 | 21,800 | 347,917 |
| I9_CHD | 25707/234698 | 17594/ 182274 | 26972/347119 | 70,273 | 764,091 |
| I9_SAH | 1620/238926 | 408/199460 | | | |
| C3_MELANOMA_SKIN | 143/260262 | 1383/198485 | | | |
| J10_ASTHMA | 25544/158452 | 23402/ 176466 | | | |
| I9_HEARTFAIL_NS | 30459/229946 | 26203/ 173665 | | | |
| I9_STR | 14351/238854 | 4575/195293 | 10394/ 360479 | 29,320 | 794,626 |
| G6_AD_WIDE | 7329/252879 | 558/199310 | 2160/ 405946 | 10,047 | 858,135 |
| T1D | 3440/214308 | 501/199367 | | | |
| I9_AF | 28670/135821 | 11917/ 187951 | | | |
| N14_CHRONKIDNEYDIS | 4959/252950 | 4263/ 105605 | | | |
| F5_DEPRESSIO | 28098/228817 | 51926/ 147942 | 36838/382918 | 116,862 | 759,677 |
| C3_BRONCHUS_LUNG | 3061/257344 | 879/ 198989 | | | |
| RHEUMA_SEROPOS_OTH | 5793/254548 | 2389/197479 | 985/373007 | 9,167 | 825,034 |
| K11_IBD_STRICT | 4611/249705 | 2105/197763 | | | |
| I9_VTE | 11288/249117 | 13202/ 186666 | 13273/360818 | 37,763 | 796,601 |
| I9_THAORTANEUR | | 924/198944 | | | |
| I9_ABAORTANEUR | | 329/199539 | | | |

| | | | | | |
|---|---|---|---|---|---|
| COX_ARTHROSIS | 13119/203797 | 22352/ 177517 | | | |
| KNEE_ARTHROSIS | 27799/203797 | 38755/ 161113 | | | |
| M13_OSTEOPOROSIS | 3960/249139 | 9714/190154 | 16327/355286 | 30,001 | 794,579 |
| AUD_SWEDISH | 14864/245541 | 10119/ 189749 | 8941/ 365150 | 33,924 | 800,440 |
| E4_HYTHYNAS | 32871/221975 | 14238/ 185630 | 24458/345172 | 71,567 | 752,777 |
| G6_SLEEPAPNO | 20279/239125 | 9049/190819 | | | |
| IPF | 1178/233040 | 117/199751 | | | |
| ILD | 2351/233040 | 546/199322 | | | |
| GOUT | 4502/241230 | 10729/ 189139 | | | |
| H7_GLAUCOMA | 10485/249920 | 15016/ 184852 | | | |
| G6_EPLEPSY | 7224/208845 | 4997/194871 | 5402/ 402704 | 17,623 | 806,420 |
| GE_STRICT | 1751/253181 | 276/199592 | 617/414698 | 2,644 | 867,471 |
| FE_STRICT | 728/253181 | 1875/197993 | 553/ 407553 | 3,156 | 858,727 |
| K11_APPENDACUT | 18798/240075 | 13497/186371 | | | |
| BMI | | 190227 | | | |