Separating the genetics of disease, treatment and treatment response using graphical modeling and large-scale electronic health records.

Malgorzata Borczyk[1,2†], Nick Machnik[1†], Jacek Hajto[2], Ilse Krätschmer[1], Paula Konowalska[2], Bartosz Baszkiewicz[2], Michal Korostynski[2], Matthew R. Robinson[1*]

[1]Institute Of Science and Technology Austria, Am Campus 1, Klosterneuburg, 3400, Austria.
[2]Laboratory of Pharmacogenomics, Maj Institute of Pharmacology PAS, Smetna 12, Krakow, 31-343, Poland.

*Corresponding author(s). E-mail(s): mrobinso@ist.ac.at; ilse.kraetschmer@ist.ac.at;
Contributing authors: gosborcz@if-pan.krakow.pl; nick.machnik@ist.ac.at; hajto@if-pan.krakow.pl; paukon@if-pan.krakow.pl; bar.basz@if-pan.krakow.pl; michkor@if-pan.krakow.pl;
†These authors contributed equally to this work.

Keywords: drug response, complex trait, pharmacogenomics, GWAS

Abstract

Genetic variants affect baseline health and biomarker values, which in turn impact both the therapy selected for an individual and the magnitude of change induced by the medication. Here, we propose an approach for complex longitudinal repeated measures biobank data, which separates genetic effects for disease from the genetic effects for medication usage and those for treatment response. For 211,845 individuals, we construct a pre-post study design from 1,420,443 repeated blood pressure (BP) measurements and 1,117,900 prescription records for common BP influencing drugs, using electronic health records. We model these jointly alongside 8,430,446 imputed single nucleotide polymorphism (SNP) markers and 17,852 whole-exome sequence loss-of-function (LoF) variants, all within a single novel graphical modeling framework. We identify pharmacogenetic candidate SNPs and LoF variants in genes *SLC35F2*, *PKD1* and *KCNIP4*, which are associated with angiotensin receptor blocker therapy and response after controlling for hypertensive disease status. We additionally detect and replicate established clinically relevant variants for statin treatment across multiple biobanks. We find that genetic variation for BP is predominantly shaped prior to the age of 50, but that hundreds of independent loci associate with age-specific BP changes through life. Finally, once post-treatment measures are conditioned on pre-treatment measures and therapy, we find no evidence for genetic variants influencing treatment-response, despite power to detect a genetic variant that contributes $> 0.017\%$ of explained variance. Our graphical modeling and pre-post study design provides a robust way of detecting time-, treatment- and treatment response-specific genetic associations within large-scale biobank studies.

Introduction

Several large-effect pharmacogenomic genes are commonly tested in the clinic to guide therapy administration and dosing, which reduces the incidence of adverse drug reactions [1, 2]. While many recent studies have identified genetic variants associated with medication use [3–5] and shown that baseline biomarker genetic predictors correlate with treatment [3–5], current evidence suggests that for most cardio-metabolic therapies, clinically relevant gene-treatment interactions for treatment response may be unlikely to occur [6–12].

Pharmacogenes have been identified through genome-wide association studies (GWASs), in randomized controlled trials (RCTs), and more recently pharmacogenomic studies within large epidemiological biobank cohorts [1–13]. Typically, drug response phenotypes are often measured in pre- and post-treatment designs in the form of an absolute or relative change in a quantitative trait before and after a particular treatment. In the context of pharmacogenetic effects, the focus is on estimating genetic effects on biomarker change. However, genetic variants also affect baseline values, which in turn impact both the therapy selected for a patient and the magnitude of change induced by the medication [6–12]. This is evidenced by analyzes on cross-sectional and longitudinal data, showing that genetic predisposition to high baseline biomarker value is a reliable indicator of treatment response [3, 5, 11, 12]. Furthermore, follow-up polygenic risk score analyses and studies examining the genetic basis of being placed on particular treatments show that baseline biomarker genetic predictors correlate with the therapy given [5]. This makes separating genetic effects for the disease from the genetic effects for medication usage and those for treatment outcomes challenging. Additionally, interaction studies are susceptible to issues of scale-dependency, endogeneity bias from the relationship of therapy and baseline phenotype [12, 14], and ascertainment bias by sub-setting to individuals with complete pre-post records. These issues currently limit our ability to resolve the genetic basis of treatment and treatment response, especially when using complex large-scale biobank data linked to electronic health records for pharmacogenomic research.

Here, we propose an approach to separate genetic effects on baseline biomarker values from genetic effects for both being placed on a particular therapy and the response of the biomarker to therapy. We show how in a time-aware repeated measures graphical model, genetic variants associated with a particular treatment are identified, conditional on the clinical measure that is the focus of the treatment. Associations for post-treatment response are also identified conditional on both the treatment made and the pre-treatment baseline measure, which separates the disease and the therapy choice from the response to the treatment. We utilize 1,420,443 repeated blood pressure (BP) measurements, annotated with 1,117,900 prescriptions for 72 drugs used in the management of BP grouped into six classes, 47,121 diagnoses of cardiovascular disease, alongside with 8,430,446 imputed SNPs and 17,852 whole exome sequence (WES) data loss of function (LoF) variants, which are all jointly analyzed in a single model (see Figure S1). After conditioning post-treatment measures on pre-treatment measures and therapy, we find no evidence for genetic variants influencing BP treatment response. However, we show that pharmacogenetic candidate SNPs and LoF

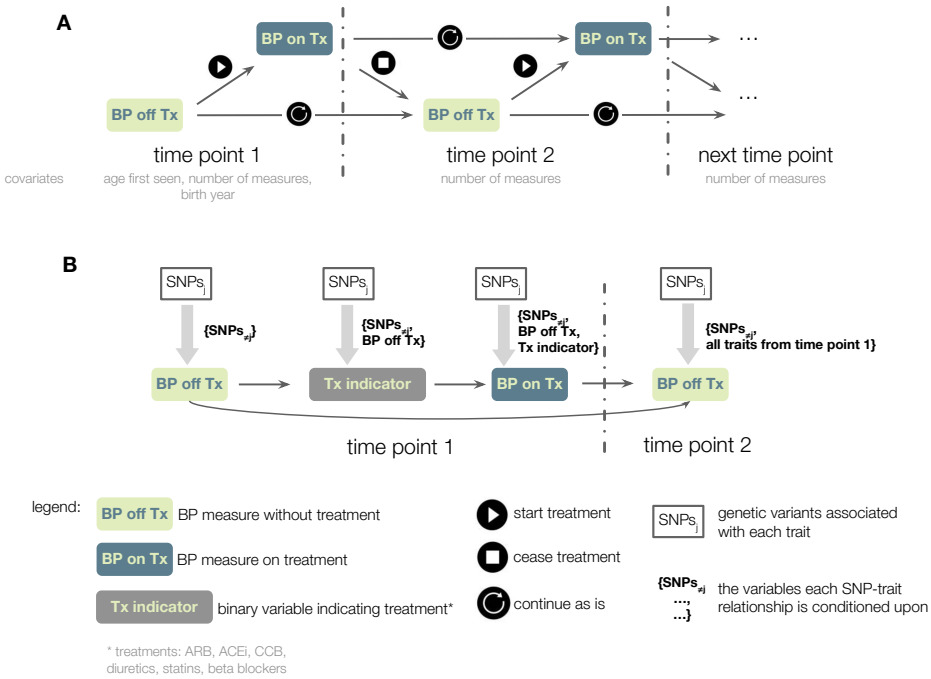

Fig. 1 Overview of the analysis approach. (A) A schematic representation of the way blood pressure (BP) measurements were aggregated and grouped into separate time points (i.e. age groups) for the age-split analysis. For each participant, we determined whether their BP measure was taken when they were not on treatment ("off Tx") or treated with a BP influencing drug ("on Tx"). We include covariates either as baseline values (e.g. birth year) or as time point specific variables (e.g. number of measures, see Methods). (B) An overview of the conditioning scheme in the time series analysis. At the first time point, the association of off-treatment BP with each single nucleotide polymorphism (SNP) is conditioned on all other genetic variants (indicated within the brackets). For all succeeding traits, each SNP association is determined conditionally on all other SNPs as well as on temporally preceding traits, with the conditioning sets given within the brackets. Blood pressure variables are continuously distributed and the six treatment indicator variables are binary (0 for untreated, 1 for treated).

variants for statin and angiotensin receptor blocker therapy and response can be identified after controlling for hypertensive and cardiovascular disease status. Finally, we show that genetic variation for BP is predominantly shaped prior to the age of 50, but that there are hundreds of independent loci associated with age-specific BP changes throughout life.

Results

Pre-post study design

We propose an approach to modeling complex longitudinal biobank data, where we construct a large-scale pre-post study design by extracting data from a combination of electronic health records (EHRs), drug prescription databases, and genomic information (Figure 1A, Figure S1). We stratify clinical blood pressure measurements by age into multiple age groups, and by treatment status into pre- and post-treatment values. In addition, we create binary indicator variables in each age group for each treatment. For each of the measurement and treatment indicator variables, we then create age-group-specific covariates that we wish to study. We also include several additional variables that reflect the structure of the measurements recorded for each individual (measurement source, ages first and last seen in the GP records, year of birth, etc.), for which we wish to adjust (see Methods for full details).

We then apply a novel time-aware graphical inference algorithm to these data that learns the relationships between all observed variables. This is an extension of our recently proposed CI-GWAS algorithm [15]. The algorithm conducts conditional independence (CI) tests between different combinations of variables, which determines whether variables are marginally or conditionally independent of each other (Figure 1B). Since the data are ordered in time, the algorithm uses this prior knowledge when testing for independence between two variables, including only preceding variables or those at the same time point in the conditioning sets.

For example, for the first time point (age group), the CI tests between pre-treatment measures and genetic variants use only the genetic variants and data structure variables as conditioning variables. CI tests between treatment indicators and genetic variants use genetic variants, the data structure variables and the pre-treatment measure as conditioning variables. As a result, SNPs associated with being placed on a particular treatment are those that remain correlated after adjusting for the clinical measure that is the focus of the treatment. Finally, CI tests for the first time point (age group) post-treatment measure condition on the treatment indicator, the pre-treatment measure as well as all other genetic variants and data structure variables. Thus, SNP associations for post-treatment measures are conditional on the treatment made and the pre-treatment measure and can be interpreted as being correlated with the change in the clinical measure post-treatment.

Moving forward in age, measures made later in life are then conditioned on measures at all previous time points. As a result, SNP associations are time-point specific for the treatment indicators and for the pre- and post-treatment clinical values.

All CI tests are conducted simultaneously within a single graph. We propose choosing the cut-off ages between groups so that time points have similar numbers of participants to balance the power to detect associations. In biobank applications, not all individuals will have measures at all time points. Our analysis allows for this, as it is based entirely on summary statistic information, consisting only of the pairwise genetic marker-marker LD correlations, the pairwise genetic marker-variable correlations, and the pairwise variable-variable correlations (Figure S1). It relies on partial correlations vanishing for (conditionally) independent variables, which does not hold

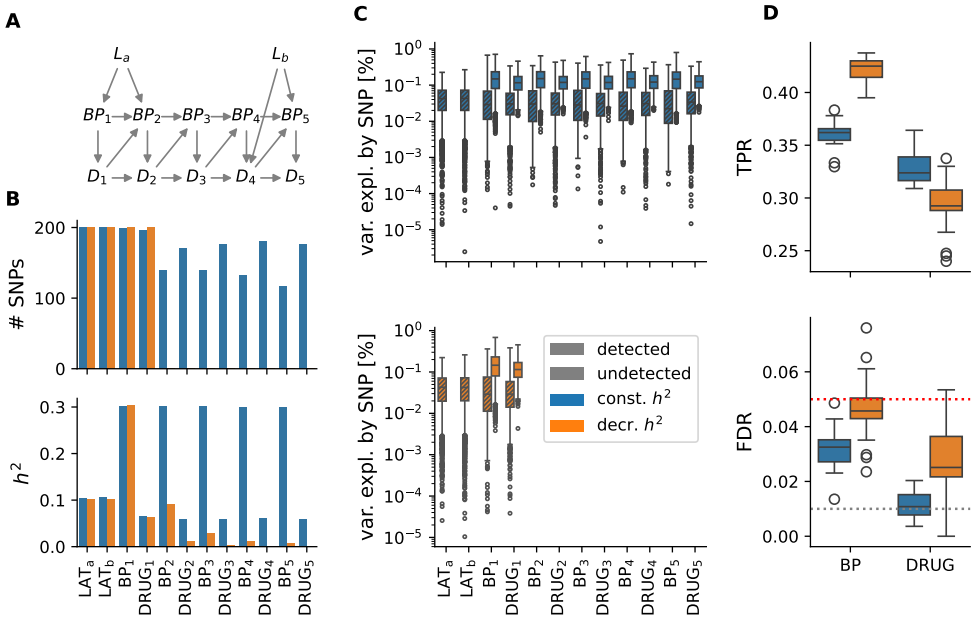

Fig. 2 Detection of time- and trait-specific genetic effects in simulation. We base our simulation on a structural equation model (SEM) that contains genetic and trait variables. The trait variables are latent variables (L, LAT), blood pressure at time t (BP_t), and drug medication status at time t (D_t, $DRUG_t$). We simulated under two scenarios: decreasing and constant heritability (h^2). (A) Graph of trait-time series in the SEM. (B) Numbers of causal single-nucleotide-polymorphisms selected, and heritability for each trait in the SEM. (C) Variance explained by detected and undetected causal genetic variants for each trait. (D) True positive rate (TPR) and false discovery rate (FDR) of the causal genetic variables when applying our marker detection strategy to the simulated data. Only genetic variables that were inferred to be trait- and time-specific were included in the calculations. Boxes in (C, D) indicate interquartile range (IQR), lines in boxes indicate the median and the whisker length is set to $1.5 \times IQR$ from the nearest hinge.

for binary or ordinal variables at low frequencies. To alleviate this issue we use a heterologous correlation matrix, with Pearson correlations between Gaussian, polyserial between Gaussian and ordinal, and polychoric correlations between ordinal variables, as previously proposed [15, 16]. This effectively places binary and ordinal variables on the liability scale and has the additional advantage of allowing comparisons among variables on the same scale, accounting for differences in their prevalence.

Robust identification of age- and trait-specific genetic effects

We first conducted a simulation study where we simulated a time series of continuous and binary variables, together with time-independent latent factors (we refer to these variables as *trait variables* in the following) using a structural equation model (Figure 2A). In addition, the structural equation model (SEM) contains the causal effects of SNPs. Here we distinguish two scenarios for the genetic structure of the variables: (i) *decreasing heritability* (h^2), where SNPs only have causal effects at the first time point,

resulting in the heritability of all variables decreasing throughout the time series, and (ii) *constant heritability*, where there are specific SNPs for each variables at each time point, resulting in constant heritability throughout the time series (Figure 2B). Another way to view the two scenarios is through the lens of genetic correlation: in the scenario of decreasing heritability, the genetic correlations among measurements of the same trait at different ages are stronger and decay much more slowly than is the case in the scenario of continuous heritability, where each time point has its own specific genetic basis (Figure S2).

Across 20 synthetic datasets generated under the model described above, we find that at a sample size of $n = 450,000$ and a significance threshold $\alpha = 10^{-3}$, our procedure has high empirical power to detect genetic variants that explain 0.01% of the variance (Figure 2C, end of lower whiskers for detected variants). This corresponds well to the theoretical limit of detection for a variant at this sample size for our algorithm, which is at $\left(\Phi^{-1}(1-\frac{\alpha}{2})(n-3)^{-\frac{1}{2}}\right)^2 \approx 0.0025\%$ variance explained, where Φ is the cumulative density function of the standard normal distribution. This also corresponds well to the power for this effect size for a standard genome-wide association study, which would be 90% at the genome-wide significance level 5e-8. We note that an age-stratified GWAS would likely detect the same variants, but would not be conditional, and thus markers would be associated across multiple time points, with an assessment of whether their effects are time-specific relying on interaction tests or departures from a consistent effect size [17]. Here, SNPs associated with later ages or treatments are specific, as they are identified conditional on preceding measures.

Overall, in our simulation setting we observe a moderate true positive rate of $30\% - 40\%$, but a well-controlled false discovery rate that lies largely between $1\% - 5\%$ (Figure 2D). The true positive rate and false discovery rate are slightly higher for the continuous traits (i.e. BP), compared to the binary traits (i.e. DRUG), which is due to the fact that placing all binary variables on the liability scale leads to larger standard errors. Furthermore, we infer more markers to be associated with single time points for continuous traits as compared to binary traits. We find $\approx 95\%$ of inferred causal variants to be correctly associated with a single time point, compared to a range of $72\% - 100\%$ for the binary drug traits where there is a slight tendency to assign more than one time point to correctly associated variants (Figure S3). Taken together, our approach accurately selects time- and trait-specific genetic variants accounting for LD down to the theoretical limit of detection given by the sample size and significance threshold, with comparable power to a GWAS.

As time-series data is often subject to data censoring and ascertainment bias, we also explore how structured missing data influences the detection of the simulated causal variants (see Methods and Figures S4, S5 and S6). We find that removal of individuals based on the state of a latent variable mainly decreases TPR for drug-causal variants and has a significant influence on the FPR only when about 50% of all individuals are removed (Figures S4 and S7). Similarly, removal at a rate proportional to the proximity of an individual to the blood-pressure population mean only has a marked influence on the FPR when close to 50% of individuals in the middle of the distribution are removed (Figures S5 and S8). Lastly, fractions of individuals were removed for which the blood pressure value is farther away from the population mean

Table 1 Counts of identified SNPs and whole exome sequencing loss-of-function variants per age group and pre-treatment systolic and diastolic blood pressure (SBP, DBP) traits. To create the LD-pruned input data, SNPs were pre-filtered by setting a maximal correlation threshold of LD $R^2 > 0.8$ within 1MB and selecting a focal variant of highest frequency from each set of highly correlated markers. Post-treatment blood pressure measures are absent from the table as no associations were found.

		all SNPs		LD pruned	
Trait	**Age**	**#SNPs**	**#LoF**	**#SNPs**	**#LoF**
SBP pre-treatment	pooled	311	23	328	23
SBP pre-treatment	< 50	739	30	832	30
SBP pre-treatment	50 - 55	203	5	235	6
SBP pre-treatment	56 - 60	3	0	2	0
SBP pre-treatment	61 - 64	175	8	197	8
DBP pre-treatment	pooled	441	28	471	21
DBP pre-treatment	< 50	849	20	930	21
DBP pre-treatment	50 - 55	265	7	287	7
DBP pre-treatment	56 - 60	3	3	2	0
DBP pre-treatment	61 - 64	1	1	3	0

than a certain threshold, so that they have no record at that and following time points. This also has no significant effect on variant recovery, except for a potential slight downward trend in the recall with increasing number of removed individuals (Figures S6 and S9). For this experiment at most $\approx 15\%$ of people were removed at each time point, which amounts to a removal of 50% of individuals only at the fourth (out of five) time points. It is possible that a more extreme removal would have resulted in similar trends as observed for the above two scenarios.

In summary, our simulations indicate that our graphical modeling approach provides a robust way of detecting time- and trait-specific associations when data are not missing at random with respect to the variables studied. We acknowledge that like with any association study, there may be more complex data missing patterns that we do not consider here. Thus, when applying our approach to the genetics of hypertension as we describe below, we additionally include variables that may correlate with data missingness, with the aim of providing additional control for these factors.

Age-specific genetic effects for blood pressure

We apply our graphical framework to study the genetics of BP and its management through the life-course (for a data overview see Figure S10). From UK Biobank EHRs we extracted: 1,420,443 repeated BP measurements for 211,845 individuals, a series of likely confounding variables (e.g. number of visits) and 1,117,900 prescription records, from which we created treatment indicators for six drug classes commonly used in the management of cardiovascular disease (CVD). For details, see Methods, Supplementary Note 1, Supplementary Data Tables and Figure S10. The six drug classes we analyzed are: Angiotensin-Converting Enzyme Inhibitors (ACEIs), Angiotensin Receptor Blockers (ARBs), Calcium Channel Blocker (CCBs), diuretics used in hypertension management, beta-adrenergic blockers and statins, which all directly or indirectly influence BP (see Figure S11). We include all of these variables,

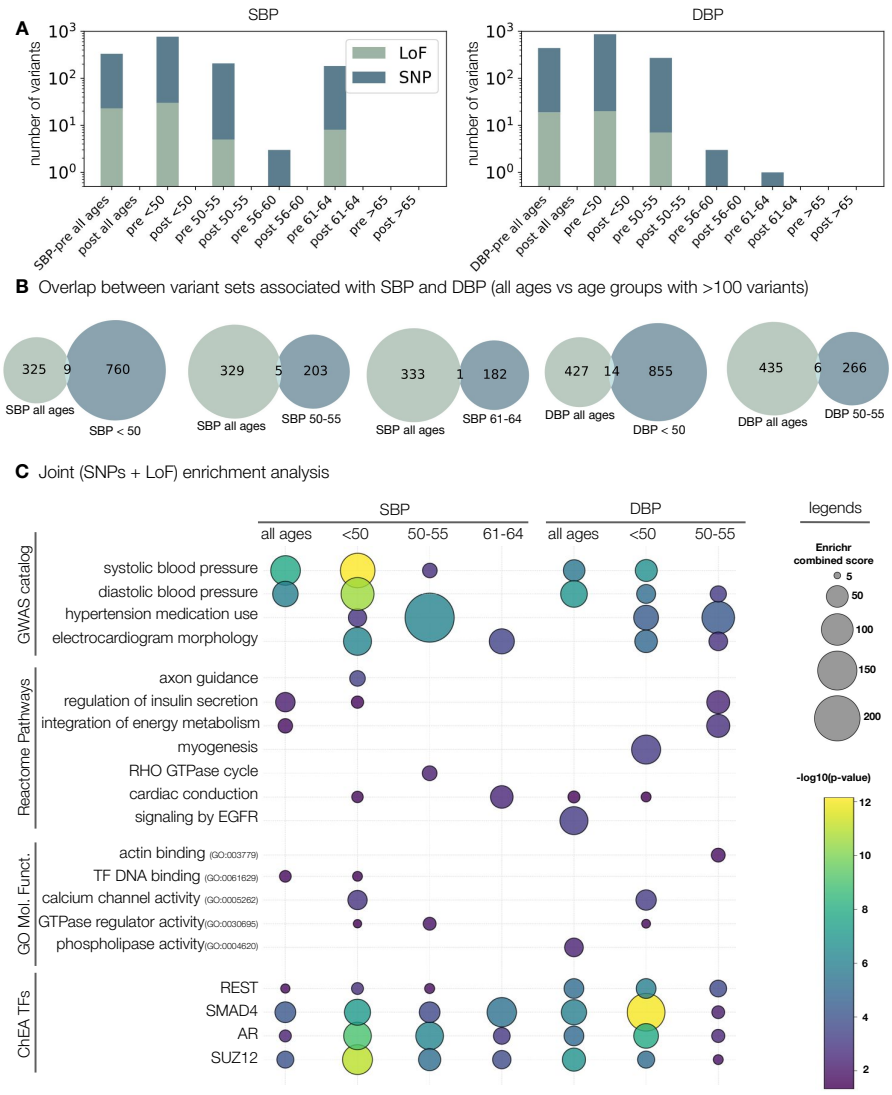

Fig. 3 Genetic effects for systolic and diastolic blood pressure across ages. (A) Stacked barplot showing the numbers of SNPs and whole exome sequencing LoF measures discovered in the age-pooled and age-split models for systolic and diastolic blood pressure (SBP and DBP). (B) Venn diagrams displaying the overlap between variants selected in the age-pooled models (all ages) and each of the age groups with > 100 associated variants. SBP and DBP variants were overlapped separately. (C) Balloon plot of a joint enrichment analysis (full results available in Supplementary Data Tables), showing the top overrepresented terms in four databases. The area of the balloon is directly proportional to the product of the odds ratio and the negative natural log of the p-value. Each balloon is colored by the $-log_{10}$ p-value of the enrichment test. All of the databases listed on the left have been accessed with the EnrichR tool [18]. For the GWAS catalog only cardiovascular traits have been selected, and any terms with less than four overlapping genes or p-value > 0.05 were excluded.

alongside 8,430,446 imputed SNP markers and 17,852 high-confidence loss-of-function (LoF) variants obtained from whole exome sequencing data, within a single graphical model. We compare the results from our model set up in two ways: (i) age-split, where pre- and post-treatment BP records and binary drug treatment indicators are aggregated into five different age groups of similar sample size; and (ii) age-pooled, where all records are averaged into a single set of pre-post measures for each participant (see Methods).

Systolic and diastolic blood pressure (SBP and DBP) are strongly phenotypically correlated ($\rho(SBP_{pre}, DBP_{pre}) = 0.7213$ on average across the 5 age groups, with an average $SE = 0.002$). We also find a consistently high genetic correlation for SBP and DBP within ages of 0.585 (0.137 SE) for early life (< 50) and 0.569 (0.110 SE) for later age (> 65). Thus, we analyzed them both jointly and separately, because we observed that if a SNP has effects on both SBP and DBP of very similar magnitude so that $\rho(SNP, DBP)$ and $\rho(SNP, SBP)$ are very similar and the phenotypic correlation $\rho(SBP, DBP)$ is large, then the partial correlation of the SNP with each blood pressure component becomes insignificant when conditioning on the other (Figure S12). While we find a number of SNPs for each blood pressure component when analyzing them jointly, analyzing them separately allowed us to identify more associated variants and better describe the age-related changes in both variables.

We find hundreds of variants associated with pre-treatment SBP or DBP, but none associated with post-treatment blood pressure at any age (Figure 3A, Table 1, see Supplementary Data Tables for full results). In other words, after conditioning on pre-treatment BP measures, we find no evidence that any SNP or LoF variant is directly associated with the treatment response. We explored if this could be a consequence of our model setup: when analyzing treated and untreated individuals together, the variance in the post-treatment measure of untreated individuals is reduced to 0 after conditioning on pre-treatment values, leaving remaining variance only for the treated individuals, which is desired. However, the post-treatment distribution is mainly governed by the untreated individuals, because the treatment prevalence is low (see Supplementary Data Tables). This reduces the observed strength of genetic associations with the post-treatment measure and thereby influences our ability to detect treatment response associations. To explore this further, we conducted a follow-up simulation study exploring our detection power in a setting directly modeled after the age-pooled analysis. We confirm that variants associated with the post-treatment measure become easier to detect when focusing the analysis on treated individuals (Figure S13). This leads to a lower detection threshold of 0.16% variance explained in the analysis focused on treated individuals, compared to 1% in the analysis with all individuals. Based on this observation, we confirmed our results by reanalyzing only the subset of treated individuals in the age-pooled data, where we have the highest sample size and therefore the highest power to detect such variants (see Supplementary Data Tables for sample sizes). Again, we do not find any SNP or LoF associations with post-treatment SBP or DBP.

We find evidence for a small number of age-specific genetic associations with pre-treatment SBP and DBP in later age groups, while conditioning on earlier BP measurements, as well as many other control variables for data structure (Figure 3,

Table 1 and Methods). We find that there is only limited potential for age-specific associations as the genetic correlation (estimated in an LD-pruned subset of the 8M SNPs) between the first and last age group is 0.929 (0.108 SE) for SBP and at 0.984 (0.116 SE) for DBP. Additionally, we find little evidence that the SNP heritability varies with age, being 0.184 (0.023 SE) and 0.180 (0.023 SE) for the first age group and 0.190 (0.019 SE) and 0.179 (0.019 SE) for the last age group, for SBP and DBP respectively.

In summary, we find evidence for genetic variation for BP that is predominantly shaped prior to the age of 50, evidence for a smaller number of loci associated with age-specific phenotypic change, but no evidence for genetic effects underlying treatment response. Associations selected by our algorithm are those that remain correlated when conditioning on all other markers and traits. Conditioning on the entire genetic background of all markers is likely beneficial for accurately detecting rare to moderate frequency variant associations. However, a marker in high LD with many other markers may be more likely to be removed unless an effect can be definitively assigned to it, simply because little variation is left after conditioning on all correlated markers. We contrasted the age-pooled and age-split analyses for SBP, where we observe that the variant sets selected are largely non-overlapping for SBP measured prior to age 50 and age-pooled SBP measures (Figure 3B). Overall there is a strong positive relationship between the marginal correlations of SNPs with the first age group SBP measures and of SNPs with age-pooled SBP (Figure S14). However, common variants with strong correlation with both the first age group SBP and age-pooled SBP appear to be less likely to be detected in both models, as they are in strong LD with other markers (Figure S14). Thus, to ensure that common SNP associations are not overlooked, we also re-run the age-pooled and age-split analyses for SBP pre-filtering the markers by setting a maximal correlation threshold of LD $R^2 > 0.8$ within 1MB and selecting a focal variant of highest frequency from each set of highly correlated markers. This serves to keep the majority of the SNP data unchanged and facilitates the selection of sets of highly correlated variables. From these data, we find a larger number of common variant sets for BP across ages, but we again find no evidence for SNPs associated with post-treatment SBP or DBP measures (Table 1).

We assessed the replicability of our BP associations by estimating the regression slope of SNP correlations estimated in an independent sample of UK Biobank participants onto the SNP correlations (corrected for Winnner's curse effects) used in our analyses. A similar approach to quantify replicability has been applied in other studies [19]. As we analyze a subset of 211,845 UK Biobank individuals for whom repeated drug prescription and general practitioner measures and genotype data were available, this leaves a subset of 246,902 individuals who had only a single baseline BP measure recorded at different ages along with knowledge of whether they were on hypertension or cholesterol modifying medication at the time of measurement. We found significant regression slopes for pre-treatment SBP in age group 1 (0.82, SE 0.01, adjusted $R^2 = 0.73$), pre-treatment DBP in age group 1 (0.80, SE 0.01, adjusted $R^2 = 0.78$), pre-treatment age-pooled SBP (0.89, SE 0.01, adjusted $R^2 = 0.87$), and pre-treatment age-pooled DBP (0.89, SE 0.01, adjusted $R^2 = 0.92$). These are all close to one, with

substantial variance explained, and therefore suggest a high level of replicability of our findings.

We also analyzed the p-value enrichment of the SNPs discovered in our analysis in a recent large-scale BP GWAS [20], finding an over-representation of lower than expected p-values (Figure S15). We performed gene level enrichment analyses that included both SNPs and the genes with LoF variants with the GWAS Catalog to look for known associations with BP and other cardiovascular traits at the gene level (Figure 3C, Supplementary Note 2, Supplementary Data Tables). From all of the cardiovascular-health related traits within the GWAS catalog, the top overrepresented terms are systolic blood pressure, diastolic blood pressure, electrocardiogram morphology and medication use for hypertension (Figure 3C). As an example, from the 195 SBP-associated genes in the age-pooled analysis, 32 (16,4%) have been previously associated with SBP in GWAS (enrichment p-value $= 1.50e-8$, OR $= 3.54$). For DBP, we find significant over-representation of genes previously associated with DBP in GWAS for all three DBP gene sets (age-pooled p-value $= 1.19e-07$, OR $= 3.22$; age < 50 p-value $= 8.89e-6$, OR $= 2.23$; age 50-55 p-value 0.00078, OR $= 2.67$). For more details see Supplementary Note 2. Taken together, this suggest our results are highly replicable and that the genetic variation for BP is shaped prior to mid-life, with some evidence for SNPs whose effects contribute to BP variation later in life.

Conditional on these SNP associations and all other variants genome-wide, we also discover many genes with LoF variants associated with the pre-treatment measures (Table 1). Examples of LoF variant associations (with three or more carriers within the UK Biobank) include *SMAD2*, detected in the pre-treatment SBP from all ages. This gene produces a signaling protein involved in artery remodeling [21, 22]. For DBP, in the < 50 age group, we replicate the previously reported association with *PDE3A*. This gene encodes a phosphodiesterase and its gain of function mutations cause brachydactyly with severe hypertension [23, 24]. Thus, we replicate known genetic syndromes associated with hypertension as well as identified novel LoF variants conditional on common variant background. In Supplementary Note 2 we provide a full discussion of our findings, with all LoF associations listed in the Supplementary Data Tables.

Finally, we note that we find hundreds of associations with potential confounders, especially with the number of observed measurement points (~ 200 SNPs and LoF variants in the age-pooled models, and ~ 1000 SNPs and LoF variants across the five age groups in the age-split models). These associations are described in detail in the Supplementary Data Tables and Supplementary Note 2. For CVD diagnoses, after conditioning on all likely confounders, ongoing drug treatments, and pre-treatment BP values, we find no associated SNPs or LoF variants.

Treatment-associated pharmacogenomic variants

Our graphical model allows for the discovery of pharmacogenetic variants that have a significant association with treatment variables after controlling for the disease state. In our analyses, we find a limited number of such variants associated with the use of statins and ARBs (Figure 4A). We discover two variants for ARB-taking (both LoF) and five associated with the statin use (four SNPs, one LoF). When we specifically

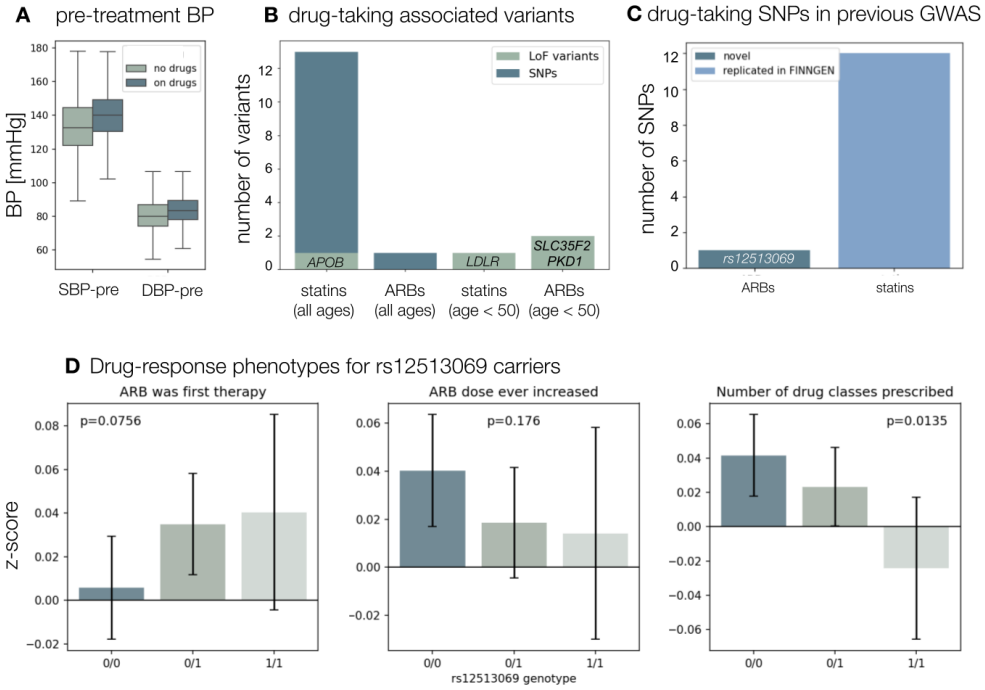

Fig. 4 Treatment associated genetic variants discovered in the UK Biobank. (A) Comparison of the pre-treatment systolic (SBP) and diastolic BP (DBP) measurements for UK Biobank participants with no prescriptions for any antihypertensive drugs (no drugs) versus those that were given drugs at later time points (on drugs). (B) Stacked barplot showing the total number of single nucleotide polymorphism (SNP) or whole exome sequence loss-of-function (LoF) variants discovered for each drug class and age groups (those with no discoveries are omitted, findings from general and on-drugs population are joined). ARB - angiotensin receptor blocker. (C) Stacked barplot indicating novel and previously found associations for SNPs. Variants that additionally replicate an external cohort (FinnGen) are marked separately, (D) Barplots (mean +/- 95% CI) showing the z-score values of phenotypes associated with drug response for carriers of the SNP discovered as associated with taking ARBs. Each phenotype is adjusted for covariates and standardized to a z-score. Nominal generalized linear model p-values are noted in each of the graphs. Three phenotypes with the lowest p-values are displayed and full results are available in the Supplementary Data Tables.

select hypertensive individuals by sub-setting our dataset to participants who received drugs at at least one time-point, we uncover an additional LoF variant and eight SNPs for statins and one SNP for ARBs. In the age-split model, we discover variants only in the youngest age-group (< 50) and all of these are LoF variants, while in the age-pooled analysis the majority of our findings are SNPs. For a summary of these findings, see Table 2 and Figure 4B.

Examples of associations with statin-taking include established clinically relevant variants such as SNP rs7412 in the *APOE* gene, which is associated with statin efficacy, among other phenotypes [25]. We also identify SNP rs11591147, a missense (R46L)

Table 2 Genetic variants associated with treatment in the UK Biobank blood pressure management dataset. For each detected whole exome sequence LoF variant, the gene symbol is given. For each SNP, the rsID/gene symbol column gives the SNP rsID and the gene symbols in brackets if the SNP falls inside a protein-coding gene. The liability scale squared correlation (r^2) between the variant and the phenotype is given as well as the odds ratio (OR) for variant carriers (both homo- and heterozygotes) being on a drug as well as the lower and upper bounds of the 95% confidence interval (CI). A star indicates variants detected only in the on-drugs subset of the UKB participants (see Methods). MAF - minor allele frequency. ARB - angiotensin receptor blocker.

age group	trait	variant class	rsID / gene symbol	MAF	liability scale r^2	OR	95% CI lower	95% CI upper
< 50	ARB	LoF	*SLC35F2*	0.000031	0.140	72.3	12.058	433.491
			PKD1	0.000068	0.147	46.555	12.009	180.477
	statin		*LDLR*	0.000051	0.106	NA (all carriers on statins)		
all ages	ARB	SNP	rs12513069* (*KCNIP4*)	0.363	-0.0359	0.89	0.856	0.926
	statin	LoF	*APOB**	0.000417	-0.0299	0.715	0.669	0.764
		SNP	rs11591147 (*PCSK9*)	0.0174	-0.0553	0.17	0.075	0.383
			rs118039278* (*LPA*)	0.0777	0.0402	1.158	1.124	1.193
			rs56393506*	0.164	0.0367	1.106	1.08	1.133
			rs186696265	0.0149	0.0244	1.272	1.197	1.352
			rs11648003* (*DHODH*)	0.220	0.0368	1.075	1.051	1.099
			rs8110479* (*SLC44A2*)	0.0443	-0.0347	0.925	0.888	0.962
			rs72658867 (*LDLR*)	0.0125	-0.0410	0.751	0.696	0.811
			rs5742911* (*LDLR*)	0.306	-0.0345	0.937	0.917	0.958
			rs12608822 (*ZNF14*)	0.0463	-0.0308	0.907	0.872	0.943
			rs62120566*	0.0248	-0.0329	0.87	0.825	0.917
			rs11881756*	0.114	-0.0297	0.937	0.912	0.962
			rs7412 (*APOE*)	0.0802	-0.0883	0.731	0.707	0.755

variant in *PCSK9* with a strong reducing effect on cardiovascular risk [26]. *PCSK9* has also recently been found to be a genetic mediator of statin effects [27]. We find that carriers of the above SNPs are less likely to be treated with statins after correcting for their hypertensive status (Table 2). Amongst individuals in the < 50 age group who take statin, we identify *LDLR* LoF carriers, which serves as a positive control as *LDLR* LoF is the most common cause of familial hypercholesterolaemia [28].

For each of the discovered SNPs, we searched for existing associations (at 0.8 LD value and 10e-6 p-value threshold) in the Open Targets and GWAS Catalog databases. All statin-taking variants that we found have previously been reported to be associated with statin-taking or response to statins, confirming the validity of our proposed approach. All twelve statin SNPs replicate in the publicly available summary statistics

(from 500,000 individuals) of the Finnish cohort, FinnGen (see the Supplementary Data Tables and Figure 4C).

For ARB treatment, we find an association with Polycystin-1 (*PKD1*) LoF variants. Deleterious variants in *PKD1* cause autosomal dominant polycystic kidney disease (ADPKD). ADPKD is associated with early hypertension and loss of *PKD1* has been reported to directly impair cardiomyocyte activity and contractile function [29]. Furthermore, we find an association of LoF variants in the *SLC35F2* gene with ARB treatment. Other members of the solute carrier (SLC) gene family are well-established pharmacogenes [30]. *SLC35F2* is known to be up-regulated in tumors [31, 32]. In an animal model with a renin–angiotensin–aldosterone system (RAAS)-affecting mutation, *SLC35F2* expression was down-regulated in kidneys [33]. Notably, a closely related gene, *SLC35F3*, transports thiamine and variants in it have long been established as associated with hypertension [34]. To our knowledge, there are no reports of consequences of LoF variants in *SLC35F2* neither in humans nor in animals, and no known associations with cardiovascular traits, making it a candidate for further study. We find nine additional LoF variants for ARB-taking in the youngest age group, but exclude them from our report, as they only had a single carrier in the subset of the youngest age group that was given ARBs. See Supplementary Data Tables for full results.

Finally, in the age-pooled analysis of hypertensive participants, we find an association between the SNP rs12513069 and ARB-taking. It is a common (minor allele frequency MAF=0.363) intronic variant of *KCNIP4*. At a p-value threshold of $1e-6$ we find no existing associations in the GWAS Catalog nor in the Open Targets databases, neither for this SNP nor for any other SNP in > 0.8 LD with it. One GWAS study identified another SNP in *KCNIP4* (rs145489027), in LD $r^2 = 0.40$ with rs12513069, as positively associated with ACEi-induced cough [35]. We find that carriers of rs12513069 are less likely to be treated with ARBs conditioned on their pre-treatment BP (Table 2). This additionally makes *KCNIP4* a candidate for further study.

These treatment indicator associations are conditioned on pre-treatment BP, all control phenotypes, all other SNPs and LoF scores and all other drug classes. Thus, we interpret these as SNPs associated with individuals being placed onto a particular treatment as opposed to others, conditional on their BP readings. For these binary treatment indicators, we are limited in our detection ability to SNPs that explain on average more than 0.35% of the liability-scale variance of the indicator, depending on age and trait (see Table S1). As an example, in the analysis of the age stratified data, we have on average 77,120 individuals with ARB prescription indications in each age group. To have the same power for ARB as for a continuous trait with normal distribution like SBP, for which we have about 60,000 individuals with records in each age group, we would need records for 200,000 people instead, given its current low prescription rate of about 0.08 (Figure S16). However, despite this, we discover a limited number of associations with drug taking that are likely to be clinically relevant, as they have large effect sizes or high frequency in the population.

For a pharmacogenomic interpretation of the discovered variants, we performed a detailed prescription data analysis focusing on ARBs and the variants that we

find to be associated with individuals being placed on these drugs as opposed to others, conditional on their pre-treatment SBP. Our dataset consists of indirect ARB-response phenotypes standardized and adjusted for covariates and includes traits such as switches to other drug classes, dose adjustments and therapy augmentations (see Figure S17, details in Supplementary Note 1). For each of the three discovered variants associated with ARB taking (rs12513069, LoF in *PKD1*, LoF in *SLC35F2*) we performed single-variant association tests (see full results in the Supplementary Data Tables), assuming an additive model. We find that for rs12513069, non-reference-allele carriers are, on average, prescribed drugs from less drug classes than reference-allele carriers (regression coefficient, beta = -0.029, p-value = 0.013, Figure 4D). For the LoF variant carriers we found associations with average doses of ARBs: Irbesartan for *PKD1* LoF carriers (beta = 1.18, p-value = 0.0025) and Candesartan cilexetil for *SLC35F2* LoF carriers (beta = 2.60, p-value = 0.0021).

Finally, since the treatment indicators are conditioned on initial blood pressure, we also investigated whether running the model without BP measurements would increase the number of findings. In this analysis, only one more variant passes the detection threshold for likely causal association for the binary treatment indicators, namely rs7675300 - another SBP in *KCNIP4* with a positive liability scale squared correlation ($R^2 = 0.035$) with ARB treatment. We also investigated if the SBP-associated variants were correlated with any of the indirect ARB-response phenotypes. We analyzed the extracted ARB-response variables using multiple regression models and found that, after adjusting for the number of predictors, genetic principal components and geographical coordinates, SBP SNPs had no explanatory power regardless of whether SBP was included in the model or not (see the Supplementary Data Tables).

Thus, we conclude that there is evidence for pharmacogenetic variants that have a significant association with treatment variables even when controlling for the disease state. Despite having the power to detect clinically actionable variants, as indicated by our detection of the statin associations that are currently used in clinical practice to screen patients, we find that there is likely only an extremely limited number of such variants. However, for the discovered ARB variants, we provide evidence supporting that they represent candidates for screening in terms of drug-response associations for further pharmacogenomic studies.

Discussion

Here, we show how genetic effects on baseline biomarker values can be modeled, together with genetic effects, for both: *(i)* being placed on a particular therapy and *(ii)* the response of the biomarker to therapy. Through the analysis of a pre-post treatment study design with a time-aware repeated measures graphical model, we show that pharmacogenetic candidate SNPs and LoF variants for statin and angiotensin receptor blocker therapy and response can be identified in the UK Biobank and replicated in other studies. We replicate a large number of known statin therapy and efficacy associated variants [6, 25, 36, 37], which is a way of validating our approach. Additionally, we discover a novel variant in *KCNIP4* that is associated with both ARB treatment and proxy ARB treatment response traits. As this is acommon variant it

could represent a clinically relevant biomarker. Another variant in *KCNIP4* is known to increase the risk of adverse drug reaction to another antihypertensive [35]. *KCNIP4* is also a potentially functionally relevant gene, associated with cardiac conduction (in Reactome Pathways) with its main function being the regulation of Kv4 potassium channels. It is hypothetically connected to ACEi-induced cough via its role in sensory nerve endings in the lungs [35].Importantly, the associations we detect for with treatment and treatment response are those that remain after controlling for hypertensive and cardiovascular disease status, alongside a host of variables correlated with data missingness, ascertainment and data structure.

We find no evidence for genetic variants directly influencing BP in response to treatment. Recent studies suggest that the influence of common and rare genetic variants on cardiometabolic drug response is relatively low [11, 12]. Our results suggest that, in fact, there is little possibility of large genetic effects when conditioned on pre-treatment measures and therapy. Previous studies have faced issues separating genetic effects for disease from genetic effects for medication usage. For example, adjusting drug response phenotypes or change scores for baseline biomarker levels has previous been widespread, but has recently been shown to introduce bias [7, 10, 11]. We note that we avoid these issues here because we do not calculate a response of change phenotype. Rather we simply examine the variation in post-treatment values conditional on both the pre-treatment measures and the therapy allocated. We additionally show that genetic variation for BP is predominantly shaped prior to the age of 50. While we identify a small number of independent loci associated with age-specific BP changes through life, genetic correlations and SNP-heritability estimates were high and close to 1 across mid- to late-life, suggesting that there is little potential for age-specific genetic effects for BP. Taken together, our results support an emerging consistent picture of a genetic basis of baseline cardiometabolic biomarker values that is shaped in early-life, which then impacts both the therapy selected for a patient and the magnitude of change induced by the medication.

There remain a number of limitations. First, larger sample sizes are required, which can be achieved by combining drug response GWAS from observational, EHR and RCT data as well as studies in more diverse ancestries. Our framework can easily accommodate meta-analyzed summary statistics across studies, or the model can be run for different ancestry groups. In this work, we have focused on describing a data analysis design and statistical approach, with the aim of providing a novel way in which biobanks coupled with EHRs can contribute to pharmacogenomics research with larger sample sizes than RCTs. For systematic full meta-analyses to occur, a coordinated international effort is now required to explore and deal with the likely substantial heterogeneity between biobanks in their record linkage, prescription profiles, and ascertainment biases. Here, we have avoided the issue of heterogeneity by seeking replication in a large hold-out set of UK Biobank individuals.

Second, this is an observational study that is both prospective and retrospective. Despite, *(i)* showing that our framework is robust to moderate data missingness, *(ii)* including a large number of variables that we hope control for missing data structure and ascertainment biases, and *(iii)* replicating our results, we cannot be certain that any of our associations are not driven by some unaccounted for confounding. Third, we

do not observe drug responses directly and can only associate SNPs and LoFs with differences between individuals in the therapy they receive and with proxy drug response phenotypes. In our pharmacogenomics follow-up study, we see that rs12513069 carriers are generally prescribed drugs from fewer classes than non-carriers and we can suggest that patients are often prescribed drugs from multiple classes if they are non-responders, but this is only conjecture. Fourth, our approach detects genetic variants whose effects remain after conditioning on all other SNPs and may miss highly correlated common genetic variants unless these are grouped into sets prior to analysis. This is handled by simply running the model on the full and grouped data and comparing the results, which is facilitated by the computational efficiency of our algorithm.

In summary, our approach is unique for the analysis of repeated measures and can be applied to very large-scale data efficiently. We show that the power to detect variants in our algorithm is similar to a standard genome-wide association study, while controlling the FDR. Thus, we provide a robust way of detecting time-, treatment- and treatment response-specific genetic associations within large-scale biobank studies.

Methods

Data processing and quality control

A detailed description of the extraction and quality control of all blood pressure measurements (1,420,443 in the raw dataset), associated prescriptions (1,117,900 in the raw dataset) and other related phenotypes from the UK Biobank electronic health records is available in Supplementary Note 1. All data were extracted from the UK Biobank Research Access platform with pyspark (v 3.2.3) and then processed with Hail (see Code Availability). Prescription, diagnosis and measurement codes were all recoded into unified systems and the final dataset included all blood pressure measurements from GP and assessment centers annotated with: date, source, cardiovascular disease diagnosis and prescription information. We grouped the prescriptions into six drug categories, as given in Table S2). We considered the measurement to be associated with a drug if a prescription had been issued within 60 days prior to the measurement. Measurements were hard-filtered for highly improbable values and separated into pre- and post-drug treatment measures. BP measures were aggregated into five age group in a way that minimized the standard deviation of the group size. For ARB response phenotypes, we also extracted dose information and prescription dates. Numerical variables were adjusted for covariates (sex, 20 genetic principal components, assessment center, genotyping batch, coordinates) and standardized.

While it is impossible to account for all potential confounding factors in an observational study, we also collect the following data structure variables that are commonly available:

- Year of birth: Treatment patterns, health status and data collection vary over time.
- Age first and last seen in the GP records: Younger age participants have less follow-up, older participants may have missing earlier measures.
- Number of measures at each time point: Underlying health status, behavior, and clinical variation influence the patient record.

- Participant location: Treatment pattern, record collection and health status show regional variation within a study.
- Median age at measurement: This measure is used to standardize longitudinal records across time points.

We include these variables because biobank data are both prospective and retrospective: individuals enter the study at different ages and are followed over time, with information about their past being inferred from linked records. Because follow-up may be inaccurate and record linkage is imperfect, the data collected are incomplete. Both the presence/absence of an individual at specific time points and the amount of data recorded about them will reflect their underlying health status, unobserved confounding, missing data structure and time trends.

Whole-exome sequencing data processing

The UK Biobank final release dataset of population level exome variant calls files were analyzed on the UK Biobank Research Analysis Platform (RAP) with DXJupyterLab Spark Cluster App (v. 2.1.1). Only biallelic sites and high quality variants were retained according to the following criteria: variant missingness <10%, Hardy Weinberg Equilibrium (HWE) p-value $> 10^{-15}$, minimum read coverage depth of 7, at least one sample per site passing the allele balance threshold >0.15 (as in [38]). Genomic variants in Ensembl canonical, protein coding transcripts were annotated with the Ensembl Variant Effect Predictor (VEP) tool (see Code Availability). High-confidence loss-of-function (LoF) variants were identified with the LOFTEE plugin (v1.0.4_GRCh38). For each gene, homozygous or multiple heterozygous individuals for LoF variants received a score of 2, those with a single heterozygous LoF variant 1, and the rest 0. As genotyping data were prepared with GRCh37, position of each gene was located with the Ensembl 87 GRCh37 gtf file. Coordinates for 402 genes which were not located were lifted over from the Ensembl 110 GRCh38 gtf file with UCSC liftover. This resulted in 17,852 genes annotated with LoF variants.

Imputed genotyping data processing

We restricted our analysis to a sample of European-ancestry UK Biobank individuals. To infer ancestry, we used both self-reported ethnic background (UK Biobank field 21000-0), selecting coding 1, and genetic ethnicity (UK Biobank field 22006-0), selecting coding 1. We projected the 488,377 genotyped participants onto the first two genotypic principal components (PC) calculated from 2,504 individuals of the 1,000 Genomes project. Using the obtained PC loadings, we then assigned each participant to the closest 1,000 Genomes project population, selecting individuals with PC1 projection $\leq |4|$ and PC2 projection $\leq |3|$.

Samples were also excluded based on UK Biobank quality control procedures with individuals removed of (i) extreme heterozygosity and missing genotype outliers; (ii) a genetically inferred gender that did not match the self-reported gender; (iii) putative sex chromosome aneuploidy; (iv) exclusion from kinship inference; (v) withdrawn consent. We used genotype probabilities from version 3 of the imputed autosomal

genotype data provided by the UK Biobank to hard-call the single nucleotide polymorphism (SNP) genotypes for variants with an imputation quality score above 0.3. The hard-call-threshold was 0.1, setting the genotypes with probability ≤ 0.9 as missing. From the good quality markers (with missingness less than 5% and p-value for HWE test larger than 10^{-6}, as determined in the set of unrelated Europeans), we selected those with minor allele frequency (MAF) ≥ 0.0002 and rs identifier, in the set of European-ancestry participants. This resulted in data set with 458,747 individuals and 8,430,446 markers (denoted as 8M dataset in the following). We apply CI-GWAS to these data as described below and we use these data for our simulation study as described below.

In addition, we produced an LD-pruned set of SNPs by removing markers in very high LD using the clumping approach of plink [39] (version 1.90, see Code Availability), where we ranked SNPs by minor allele frequency and then selected the highest-MAF SNPs from any set of markers with LD $R^2 \leq 0.8$ within a 1-Mb window. This results in the selection of a tagging set of variants, with only variants in very high LD with the tag SNPs removed. These filters resulted in a data set with 2,174,071 markers (denoted as 2M dataset in the following).

CI-GWAS

In this work, we use a modified version of CI-GWAS [15] (see Code Availability). In short, CI-GWAS learns a causal graphical model of genetic and trait variables. It performs variable selection on the genetic variables, removing variables which are not adjacent to any trait in the learned graph, or, in other words, do not exhibit any direct effects on traits in the learned model. The graph is learned through a series of conditional independence (CI) tests, by which an edge between a pair of variables is removed if the partial correlation of the two variables is insignificant given some other set of adjacent variables.

Here, we introduce the use of time indices of the variables in the model to CI-GWAS. The user can provide the order of the variables' measurements in time (the time index). For an edge between two variables X, Y that have time indices t_X, t_Y, CI-GWAS will not include any variable Z in the conditioning set for which $t_Z > t_X$ or $t_Z > t_Y$. The rationale is that events in the future cannot have causal effects on events in the past. Therefore, the model should not use variables for which it is known that their measurement lies in the future to explain variation in variables in the present. Tables 3 and 4 show the time indices used in CI-GWAS for the analysis of the simulated and real data, respectively.

The input data to CI-GWAS were prepared as follows: First, we selected the intersection of individuals with records in the extracted phenotypic data and records in the UK Biobank. We then joined the LoF variants together with either the genetic variant data from the 8M or the 2M datasets. For these two sets of variants we computed LD-blocks, as described previously [15], using:

```
$ ci-gwas block \
  --max-block-size 11000 \
  --corr-width 2000
```

Table 3 Time indices of variables in simulation analysis, as provided to CI-GWAS.

Variable	Time Index
SNPs	0
BP_t	t
D_t	t

Table 4 Time indices of variables in real data analysis, as provided to CI-GWAS.

Variable	Time Index
SNPs	0
year of birth	1
age first seen	1
age last seen	1
BP pre-treatment in age group t	$3t-2$
BP post-treatment in age group t	$3t$
treatment in age group t	$3t-1$
mean number of measurements in age group t	$3t-2$
median age in age group t	$3t-2$

For each of the resulting blocks, we calculated the correlation between the genetic variables using plink with:

```
$ plink --r triangle bin4
```

This constitutes the ‘mxm‘ (correlations between genotypic variables) input data for CI-GWAS.

For the ‘mxp‘ portion of the CI-GWAS input (correlations between genotypic variables and traits), we calculated, for every genetic variant, the correlation with each phenotype. Here we used either the biserial correlation coefficient [40], for binary phenotypes, and the Pearson correlation coefficient for continuous phenotypes, together with the corresponding standard errors. For each correlation value, we used only the subset of individuals with records for the corresponding genetic variant and phenotype.

For the ‘pxp‘ portion of the CI-GWAS input (correlations between traits), we calculated the correlation between all pairs of phenotypes. Again, for each pair, we used only the subset of individuals with records in both phenotypes. We calculated the biserial correlation coefficient for pairs of continuous and binary traits, the tetrachoric correlation coefficient [41] for pairs of binary traits, and the Pearson correlation coefficient for pairs of continuous traits, with the corresponding standard errors. Unless stated explicitly otherwise, we set the p-value acceptance threshold for the individual CI-GWAS tests to $\alpha = 10^{-3}$.

Table 5 Trait effect size sampling parameters

Edge	Distribution
$LAT \rightarrow$ any	$\mathcal{U}([-0.1, -0.05] \cup [0.05, 0.1])$
$BP \rightarrow BP$	$\mathcal{U}([0.5, 0.7])$
$BP \rightarrow DRUG$	$\mathcal{U}([0.05, 0.2])$
$DRUG \rightarrow DRUG$	$\mathcal{U}([0.5, 0.7])$
$DRUG \rightarrow BP$	$\mathcal{U}([-0.2, -0.05])$

Simulation studies

In our simulation study, we set up a structural equation model of a time-series of blood pressure readings and medication status, measured at five consecutive time points, and two latent variables (Figure 2A). We set the model up such that blood pressure at time t has a positive effect on blood pressure at time $t+1$, as well as a positive effect on the medication status at time t. The medication status at time t has a positive effect on the medication status at time $t+1$ and a negative effect on the blood pressure at time $t+1$. Latent variables have positive or negative effects on random subsets of observed trait variables. We choose the effect sizes between traits from uniform distributions with parameters given in Table 5.

In addition to these traits, the model includes genetic variables. We simulate two scenarios for the genetic architecture:

1. **Decreasing heritability:** The heritability of blood pressure and medication status decreases over time. The idea here is that genetics influence blood pressure only early in life (e.g., through influence on heart development, metabolism, etc.) and that further exogenous influences are environmental. Thus, the heritability is maximal at the first time point and decreases at subsequent time points at a rate depending on the magnitude of the causal effects between traits. In this scenario, we sample 200 genetic variables each for the two latent variables at time point 1, $DRUG_1$ and BP_1. We then assign effect sizes $\beta(V_i)$ to the genetic variables of each trait V_i by first sampling them from a standard normal and then scaling them so that the sum of their squares matches the desired trait heritability:

$$\begin{aligned} &\tilde{\beta} \sim \mathcal{N}(0,1), \\ &\beta = sgn(\tilde{\beta})\sqrt{|\tilde{\beta}| h_t^2 \|\tilde{\beta}\|_1^{-1}}, \end{aligned} \tag{1}$$

   where h_t^2 is the target trait heritability (0.1 for latent traits, 0.2 for $DRUG_1$ and 0.3 for BP_1).
2. **Constant heritability:** The heritability of blood pressure and drug taking is the same at each time point, and there are genetic markers specific to time points. We start as in the decreasing heritability scenario described above by sampling the markers and effects for the latent variables, $DRUG_1$ and BP_1. For all other traits

we compute a transmitted heritability h^2_{trans}:

$$h^2_{trans}(V_i) = \sum_{V_j \in anc(V_i)} (\beta(V_j)\Psi_{ji})^2,$$

where Ψ_{ji} is the total causal effect of V_j onto V_i, the product of the direct causal effects on each directed path from V_j to V_i, summed over all such paths, and $anc(V_i)$ is the set of ancestors of V_i in the graphical model. Ψ_{ji} is computed as

$$\Psi = \sum_{l=1}^{p} B^l,$$

where B is the matrix of direct causal effects such that B_{kl} is the direct causal effect of $V_k \rightarrow V_l$ and p is the number of trait variables in the model. We then define a local heritability

$$h^2_{loc}(V_i) = h^2_t(V_i) - h^2_{trans}(V_i).$$

In order to keep the distribution of $|\beta|$ the same for all traits, we sample only $200 \times \frac{h^2_{loc}(V_i)}{h^2_t(V_i)}$ variables per trait. Using $h^2_{loc}(V_i)$ in place of $h^2_t(V_i)$, we then sample the genetic effects $\beta(V_i)$ as described in Equation 1.

In both scenarios, the genetic variables are randomly picked from 169,903 markers on human chromosome 1 in the 2M dataset of UK Biobank markers. We sample without replacement, such that there is no overlap between the sets of causal variants. Thus, there is no "horizontal" or "true" pleiotropy in this simulation. Given all direct effects in the structural equation model we then generate data v_i corresponding to each each variable V_i. We standardize the genetic data to unit variance and zero mean, (denoted as X below), and add noise to a target variance of 1 for each trait variable:

$$\begin{aligned}
v'_i &= v^T B_{,i} + X_i \beta(V_i), \\
\epsilon_i &\sim \mathcal{N}\left(0, \sqrt{1 - Var(v_i)}\right), \\
v_i &= v'_i + \epsilon_i.
\end{aligned}$$

If v_i is a binary variable (i.e. a medication status) in the model, then we interpret $v'_i + \epsilon_i$ as the trait on the the liability scale, from which we binarize:

$$v_i = \begin{cases} 1, \text{if } v'_i + \epsilon_i \geq \Phi^{-1}(1-K) \\ 0, \text{else} \end{cases},$$

where K is the prevalence and Φ is the cumulative density function of the standard normal distribution. We used a prevalence of 0.08 in this simulation. After that, we standardize v_i to unit variance.

We generated data for 20 sampled models in each genetics scenario and ran CI-GWAS at $\alpha = 10^{-3}$ to select genetic variables in the cuda-skeleton step. To assess the quality of the marker recovery, we computed the TPR and FDR of the recovered markers. A marker is only counted as true positive if it was inferred as a cause for the correct variable at the correct time point, and false otherwise. We are therefore assessing the ability of CI-GWAS to recover age specific effects.

Censoring based on a latent selection variable. To obtain a censored dataset, we first repeat the simulation procedure described above. We then define a threshold $frac$ and remove all individuals from the dataset entirely for which $LAT_a \geq frac$. We generated and analyzed datasets with $frac = 0.25$ and $frac = 0.5$.

Censoring due to decreased incentive of healthy people to go to medical check-ups. Again, to obtain a censored dataset, we first repeat the simulation procedure described above. At each time point t we then examine the values of BP_t of all individuals. We remove the entries of individuals for which $|BP_t| \leq 1.2$ from the data of BP_t and $DRUG_t$ with a probability $P = \gamma\phi(BP_t)$, where ϕ is the Gaussian probability density function and γ is a proportionality factor (0.5 or 1.0).

Censoring due to death. Again, to obtain a censored dataset, we first repeat the simulation procedure described above. At each time point t, we examine the values of BP_t of all individuals and remove the entries of individuals for which $BP_t \geq thr$ from the data of BP_τ and $DRUG_\tau$ for $\tau \geq t$. We generated and analyzed datasets for $thr = 1$ and $thr = 2$.

Detection limit calculations

We present theoretical limits of the percentage of variance in a trait explained by a marker, below which CI-GWAS labels associations as insignificant. We calculate these limits based on the conditional independence test in CI-GWAS, which tests the significance of the sample correlation between two variables A and B, r_{AB}, and rejects the null hypothesis $H_0 : \rho_{AB} = 0$ (ρ_{AB} being the unobserved population correlation coefficient) if

$$|z(r_{AB})| \geq \frac{\Phi^{-1}(1 - \frac{\alpha}{2})}{\sqrt{n-3}}, \tag{2}$$

where z is Fisher's z-transform of the correlation, n is the sample size, and Φ is the cumulative density function of the standard normal distribution.

CI-GWAS calculates n from the standard errors of the correlations provided, s_r, using

$$s_r = \frac{1 - r^2}{\sqrt{n-3}} \iff \tag{3}$$

$$n = \left(\frac{1-r^2}{s_r}\right)^2 + 3 \tag{4}$$

To calculate detection limits, we also use the standard errors in the CI-GWAS input, but make the approximation $n = \frac{1}{s_r^2}$, since the correlations considered are very small, such that $1-r^2 \approx 1$, and $n-3 \approx n$ for large n. We then calculate the r^2_{AB} that satisfies the equality in the above equation, and interpret it as as the variance explained in B by A at the detection limit.

To calculate the standard errors for biserial correlations, r_b, (used for correlations between markers and binary traits), we used Hunter's and Schmidt's approximation [42]:

$$s_{r_b} = \frac{\sqrt{pq}(1-r^2)}{\phi(\Phi^{-1}(p))\sqrt{n-3}}, \tag{5}$$

where ϕ is the probability density function of the standard normal distribution, p is the prevalence of the binary trait and $q = 1-p$.

Since CI-GWAS calculates the sample size from the standard error as stated above (using the standard error to sample size relation of the product-moment correlation), the resulting estimated sample size $\hat{n}$ is strictly smaller than n in Equation 5. This reduction in the effective sample size is a consequence of the information loss resulting from the dichotomization of the liability-scale distribution of the binary trait, and has been shown to be effective in controlling the number of false positive test results caused by treating binary variable with low prevalence as Gaussian [43].

To calculate the n needed to get an effective sample size $\hat{n}$ (and the corresponding detection limit in CI-GWAS) for a binary trait with prevalence p, we set $s_{r_b} = s_r$ using Equations 3 and 5 and the approximations $1-r^2 \approx 1$ and $n-3 \approx n$, and solve for n:

$$\frac{\sqrt{pq}}{\phi(\Phi^{-1}(p))\sqrt{n}} = \frac{1}{\sqrt{\hat{n}}} \iff \tag{6}$$

$$n = \hat{n}\frac{pq}{\phi(\Phi^{-1}(p))^2}. \tag{7}$$

Simulation to find detection limits of genetic effects on post-treatment blood pressure

We conducted a simulation based on the parameters observed in the analysis of the age-pooled dataset to explore the effect of including all individuals in the analysis, as opposed to focusing only on those who received treatment. We simulated data for $n = 180,000$ individuals, which is approximately the number of individuals with non-missing records in both ARB and BP_{pre} in the age-pooled analysis. We simulated a Gaussian BP_{pre} and, based on that, a drug indicator on the liability scale

$$D_l = \beta_B BP_{pre} + \epsilon_D, \tag{8}$$

$$\epsilon_D \sim \mathcal{N}(0, 1-\beta^2) \tag{9}$$

where $\beta_B = 0.3$, which we then dichotomized to generate a binary variable D with prevalence $k = 0.05$. We then generated a variable,

$$BP_{post} = BP_{pre} + \beta_G G \odot D + \beta_D D + \epsilon D, \tag{10}$$

representing blood pressure after treatment, which only differs from BP_{pre} for people who have been treated (for who $D = 1$), by on average the difference $\beta_D = -0.25$ plus the effect of a Gaussian genetic variable G of size β_G that we varied in the range $[10^{-3}, 2 \times 10^{-1}]$. $\odot$ represents element-wise multiplication. For each value of β_G we created 100 replicates. We then calculated the partial correlations ($\hat{\rho}$)

$$\hat{\rho}_{\text{all}} := \hat{\rho}(G, BP_{post} | \{BP_{pre}, D\}), \tag{11}$$

$$\hat{\rho}_{\text{treated}} := \hat{\rho}(G^t, BP^t_{post} | \{BP^t_{pre}\}), \tag{12}$$

where $X^t = \{X_i | D_i = 1 \forall i \in [1..n]\}$, and assessed their absolute value as a function of β_G.

Replication and enrichment analyses

We replicated our results within a UK Biobank hold-out set of data. In our main CIGWAS analysis, we used a subset of 211,845 UK Biobank individuals where repeated drug prescription and general practitioner measures and genotype data were available. This left a subset of 246,902 individuals who had only a single baseline BP measure recorded at different ages alongside knowledge of whether they were on hypertension or cholesterol modifying medication at the time of measurement, which we used for replication. We assessed the replicability of our BP associations by estimating the regression slope of SNP correlations estimated in the independent sample of UK Biobank participants onto the SNP correlations (corrected for winnner's curse effects) from our main analysis. A similar approach to quantify replicability has been applied in other studies [19]. We controlled for Winner's curse using a conditional likelihood method for obtaining bias-reduced association estimates [44].

We additionally mapped all of our discovered SNPs to other SNPs in high LD ($R^2 > 0.8$) with them using the available UKB LD matrices from the imputed genotype data (see Code Availability). We downloaded summary statistics from SBP/DBP GWAS [20] and for each of our SNPs assigned the lowest p-value from all its LD partners. We also parsed all of the Open Targets (including FinnGen results) and GWAS catalog records (access date: March 2025) to extract any existing associations at the p-value threshold of 10^{-6} for all of the discovered SNPs and their LD partners (see Code Availability). For the enrichment analyses, SNPs were mapped to genes based on their rsIDs with Ensembl VEP (Ensembl version 113) and joined with genes with found LoF variants into unified gene lists. All of the enrichment analyses were performed with the Enrichr tool [18] and included the following databases: GWAS catalog (gene-level), Reactome Pathways Database, Gene Ontology (GO) Molecular Function and ChEA and ENCODE consensus TFs. For the main analysis and visualization, we only included traits from the GWAS catalog that contained any of the following strings: coronary, systolic, diastolic, hypertension, medication, cardio, pressure, heart,

myocardial. Full results are available in the Supplementary Data Tables. Traits with p-values < 0.05 and less than four overlapping genes were excluded. For visualisation we present a set of terms from each database where for each of the gene lists we selected a term with the lowest p-value (maximum of 7 terms - one for each list).

SNP heritability and genetic correlation estimation

We used GCTA (see Code Availability) to estimate the proportion of variance attributable to a relationship matrix calculated from the 2M SNP data subset for SBP and DBP at each age group. We also used bivariate models to estimate the covariance (correlation) of the SNP effects of SBP and DBP across ages.

Associations with drug response phenotypes

The extracted drug-response phenotypes (see Supplementary Note 1) were analyzed with linear models for their association with selected variants. Each of the drug response phenotypes was first adjusted for covariates (sex, geographical coordinates and genetic PCs) and standardized. For each of the discovered variants associated with ARB taking, we performed single-variant association tests using ordinary least squares regression. To evaluate any association of the drug-response traits with the SBP/DBP associated variants we ran multiple linear regression, where missing variant information was mean-imputed. The R^2 values were adjusted for the number of predictors.

Supplementary information. Supplementary Information accompanies this work.

Acknowledgments. We thank members of the Robinson group at ISTA for their comments. This work was funded by an SNSF Eccellenza Grant to MRR (PCEGP3-181181), by core funding from the Institute of Science and Technology Austria, and by the SONATINA 5 National Science Center Poland grant awarded to MB (2021/40/C/NZ2/00218). We also gratefully acknowledge Poland's high-performance infrastructure PLGrid ACK Cyfronet AGH, for providing computer facilities and support within computational grant no PLG/2021/015079. We would like to acknowledge the participants and investigators of the UK Biobank study. High-performance computing was supported by the Scientific Service Units (SSU) of IST Austria through resources provided by Scientific Computing (SciComp).

Declarations

Author contributions

MRR, NM and MB conceived and designed the study, conducted the analysis and wrote the paper. MRR and MK provided study oversight. JH, BB, PK, IK contributed to the analyses and the writing of the manuscript. All authors approved the final manuscript prior to submission.

Author competing interests

MRR has received research funding from Boehringer Ingelheim through a research collaboration agreement for the research that created the CI-GWAS algorithm. All other authors declare no competing interests.

Data availability

This project uses the UK Biobank data under project number 35520. This work uses data provided by patients and collected by the NHS as part of their care and support. UK Biobank genotypic, phenotypic and linked healthcare data is available through a formal request at (http://www.ukbiobank.ac.uk).

UK Biobank Linkage Disequilibrium Matrices are available publicly https://registry.opendata.aws/ukbb-ld.

All summary statistic estimates are released publicly on Dryad: https://doi.org/xx.xxxx/dryad.xxxxxxxxx.

Code availability

The CI-GWAS code is fully open source and available at https://github.com/medical-genomics-group/ci-gwas. All of the code used to prepare and analyze the data in this paper is available in a dedicated repository https://github.com/medical-genomics-group/bp-rep. The prescription and diagnoses recoding dictionary is available as json file within the projects' repository https://github.com/medical-genomics-group/bp-rep/blob/main/bp-data-extraction/recoding_dict.json.

Scripts that take a list of SNPs with their GRCh37 coordinates, extract their LD partners, and look for existing associations in GWAS Catalog and Open Targets databases) at the user-defined p-value threshold are published as a containerised stand-alone software https://github.com/ippas/ld-assoc-explorer.

R version 4.2.1 is available at https://www.r-project.org/.

Plink version 1.9 is available at https://www.cog-genomics.org/plink/1.9/.

Hail 0.2.104. is available at https://github.com/hail-is/hail.

Ensembl Variant Effect Predictor (VEP) tool is available at https://www.ensembl.org/Tools/VEP.

GCTA is available at https://yanglab.westlake.edu.cn/software/gcta/index.html.

References

[1] Pirmohamed, M. Pharmacogenomics: current status and future perspectives. *Nature Reviews Genetics* **24**, 350–362 (2023).

[2] McInnes, G., Yee, S. W., Pershad, Y. & Altman, R. B. Genomewide association studies in pharmacogenomics. *Clin. Pharmacol. Ther.* **110**, 637–648 (2021).

[3] Kiiskinen, T. *et al.* Genetic predictors of lifelong medication-use patterns in cardiometabolic diseases. *Nature Medicine* **29**, 209–218 (2023).

[4] Cordioli, M. *et al.* Socio-demographic and genetic risk factors for drug adherence and persistence: a retrospective nationwide and biobank study across 5 medication classes and 1 845 665 individuals. *medRxiv* (2024).

[5] Wu, Y. *et al.* Genome-wide association study of medication-use and associated disease in the uk biobank. *Nature Communications* **10**, 1891 (2019).

[6] Postmus, I. *et al.* Pharmacogenetic meta-analysis of genome-wide association studies of ldl cholesterol response to statins. *Nature Communications* **5**, 5068 (2014).

[7] Oni-Orisan, A. *et al.* The impact of adjusting for baseline in pharmacogenomic genome-wide association studies of quantitative change. *NPJ Genom. Med.* **5**, 1 (2020).

[8] Salvi, E. *et al.* Genome-wide and gene-based meta-analyses identify novel loci influencing blood pressure response to hydrochlorothiazide. *Hypertension* **69**, 51–59 (2017).

[9] McDonough, C. W. *et al.* Adverse cardiovascular outcomes and antihypertensive treatment: A genome-wide interaction meta-analysis in the international consortium for antihypertensive pharmacogenomics studies. *Clinical Pharmacology & Therapeutics* **110**, 723–732 (2021).

[10] Zhang, H., Chhibber, A., Shaw, P. M., Mehrotra, D. V. & Shen, J. A statistical perspective on baseline adjustment in pharmacogenomic genome-wide association studies of quantitative change. *npj Genomic Medicine* **7**, 33 (2022).

[11] Sadler, M. C. *et al.* Leveraging large-scale biobank ehrs to enhance pharmacogenetics of cardiometabolic disease medications. *Nature Communications* **16**, 2913 (2025).

[12] Sadowski, M. *et al.* Characterizing the genetic architecture of drug response using gene-context interaction methods. *Cell Genomics* **4** (2024).

[13] Zhou, K. *et al.* Variation in the glucose transporter gene slc2a2 is associated with glycemic response to metformin. *Nature Genetics* **48**, 1055–1059 (2016).

[14] Robinson, M. R. *et al.* Genotype–covariate interaction effects and the heritability of adult body mass index. *Nat Genet* **49**, 1174–1181 (2017). Publisher: Nature Publishing Group.

[15] Machnik, N. *et al.* Causal inference for multiple risk factors and diseases from genomics data (2025). BioRxiv 2023.12.06.570392.

[16] Cui, R., Groot, P. & Heskes, T. Learning causal structure from mixed data with missing values using Gaussian copula models. *Stat Comput* **29**, 311–333 (2019).

[17] Ojavee, S. E. *et al.* Genetic insights into the age-specific biological mechanisms governing human ovarian aging. *The American Journal of Human Genetics* **110**, 1549–1563 (2023).

[18] Xie, Z. *et al.* Gene set knowledge discovery with enrichr. *Current protocols* **1**, e90 (2021).

[19] Yengo, L. *et al.* Meta-analysis of genome-wide association studies for height and body mass index in ≈700000 individuals of european ancestry. *Human Molecular Genetics* **27**, 3641–3649 (2018).

[20] Keaton, J. M. *et al.* Genome-wide analysis in over 1 million individuals of european ancestry yields improved polygenic risk scores for blood pressure traits. *Nature Genetics* **56**, 778–791 (2024).

[21] Deng, H. *et al.* Activation of Smad2/3 signaling by low fluid shear stress mediates artery inward remodeling. *Proceedings of the National Academy of Sciences* **118**, e2105339118 (2021). Publisher: Proceedings of the National Academy of Sciences.

[22] Bjørnstad, J. L. *et al.* Inhibition of SMAD2 phosphorylation preserves cardiac function during pressure overload. *Cardiovascular Research* **93**, 100–110 (2012).

[23] Gallegos, F. *et al.* Decoding Monogenic Hypertension: A Review of Rare Hypertension Disorders. *American Journal of Hypertension* hpaf005 (2025).

[24] Chung, Y. W. *et al.* Targeted disruption of PDE3B, but not PDE3A, protects murine heart from ischemia/reperfusion injury. *Proceedings of the National Academy of Sciences* **112**, E2253–E2262 (2015). Publisher: Proceedings of the National Academy of Sciences.

[25] Xiang, Q. *et al.* The association between the slco1b1, apolipoprotein e, and cyp2c9 genes and lipid response to fluvastatin: a meta-analysis. *Pharmacogenetics and genomics* **28**, 261–267 (2018).

[26] Ference, B. A. *et al.* Effect of long-term exposure to lower low-density lipoprotein cholesterol beginning early in life on the risk of coronary heart disease: a mendelian randomization analysis. *Journal of the American College of Cardiology* **60**, 2631–2639 (2012).

[27] Sadowski, M. *et al.* Characterizing the genetic architecture of drug response using gene-context interaction methods. *Cell Genomics* **4** (2024).

[28] Defesche, J. C. *et al.* Familial hypercholesterolaemia. *Nature reviews Disease primers* **3**, 1–20 (2017).

[29] Altamirano, F. *et al.* Polycystin-1 assembles with kv channels to govern cardiomyocyte repolarization and contractility. *Circulation* **140**, 921–936 (2019).

[30] Schaller, L. & Lauschke, V. M. The genetic landscape of the human solute carrier (slc) transporter superfamily. *Human genetics* **138**, 1359–1377 (2019).

[31] He, J. *et al.* Solute carrier family 35 member f2 is indispensable for papillary thyroid carcinoma progression through activation of transforming growth factor-β type i receptor/apoptosis signal-regulating kinase 1/mitogen-activated protein kinase signaling axis. *Cancer Science* **109**, 642–655 (2018).

[32] Che, B. *et al.* Slc35f2–syvn1–trim59 axis critically regulates ferroptosis of pancreatic cancer cells by inhibiting endogenous p53. *Oncogene* **42**, 3260–3273 (2023).

[33] Rodriguez, S. *et al.* Aging and chronic high-fat feeding negatively affect kidney size, function, and gene expression in ctrp1-deficient mice. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology* **320**, R19–R35 (2021).

[34] Zhang, K. *et al.* Genetic implication of a novel thiamine transporter in human hypertension. *Journal of the American College of Cardiology* **63**, 1542–1555 (2014).

[35] Mosley, J. D. *et al.* A genome-wide association study identifies variants in kcnip4 associated with ace inhibitor-induced cough. *The pharmacogenomics journal* **16**, 231–237 (2016).

[36] Lavertu, A., McInnes, G., Tanigawa, Y., Altman, R. B. & Rivas, M. A. LPA and APOE are associated with statin selection in the UK Biobank (2020). 2020.08.28.272765.

[37] Guan, Z.-W. *et al.* Pharmacogenetics of statins treatment: Efficacy and safety. *Journal of Clinical Pharmacy and Therapeutics* **44**, 858–867 (2019).

[38] Szustakowski, J. D. *et al.* Advancing human genetics research and drug discovery through exome sequencing of the uk biobank. *Nature genetics* **53**, 942–948 (2021).

[39] Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, s13742–015–0047–8 (2015).

[40] Pearson, K. On a New Method of Determining Correlation Between a Measured Character A, and a Character B, of which Only the Percentage of Cases Wherein B Exceeds (or Falls Short of) a Given Intensity is Recorded for Each Grade of A. *Biometrika* **7**, 96–105 (1909).

[41] Pearson, K. Mathematical Contributions to the Theory of Evolution. VIII. On the Correlation of Characters Not Quantitatively Measurable. [Abstract]. *Proceedings of the Royal Society of London* **66**, 241–244 (1899).

[42] Hunter, J. E. & Schmidt, F. L. Dichotomization of continuous variables: The implications for meta-analysis. *Journal of Applied Psychology* **75**, 334–349 (1990). Place: US Publisher: American Psychological Association.

[43] Cui, R., Groot, P. & Heskes, T. L. Learning causal structure from mixed data with missing values using gaussian copula models. *Stat Comput* **29**, 311–333 (2019).

[44] Zhong, H. & Prentice, R. L. Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. *Biostatistics* **9**, 621–634 (2008).

[45] Vandenwijngaert, S. *et al.* Blood pressure-associated genetic variants in the natriuretic peptide receptor 1 gene modulate guanylate cyclase activity. *Circ. Genom. Precis. Med.* **12**, e002472 (2019).

[46] Chittani, M. *et al.* TET2 and CSMD1genes affect SBP response to hydrochlorothiazide in never-treated essential hypertensives. *J Hypertens* **33**, 1301–1309 (2015).

[47] The International Consortium for Blood Pressure Genome-Wide Association Studies. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature* **478**, 103–109 (2011).

[48] Kubota, Y. *et al.* Cooperative interaction of Angiopoietin-like proteins 1 and 2 in zebrafish vascular development. *Proc Natl Acad Sci U S A* **102**, 13502–13507 (2005).

[49] Motahari, Z., Moody, S. A., Maynard, T. M. & LaMantia, A.-S. In the line-up: deleted genes associated with digeorge/22q11. 2 deletion syndrome: are they all suspects? *Journal of neurodevelopmental disorders* **11**, 1–28 (2019).

[50] Kim, Y. J. *et al.* The contribution of common and rare genetic variants to variation in metabolic traits in 288,137 East Asians. *Nat Commun* **13**, 6642 (2022).

[51] Williams, J. *et al.* Integrating Common and Rare Variants Improves Polygenic Risk Prediction Across Diverse Populations (2024). MedrXiv, 2024.11.05.24316779.

[52] Cheong, A. *et al.* Downregulated rest transcription factor is a switch enabling critical potassium channel expression and cell proliferation. *Molecular cell* **20**, 45–52 (2005).

[53] Lin, A. E. *et al.* Gain-of-function mutations in smad4 cause a distinctive repertoire of cardiovascular phenotypes in patients with myhre syndrome. *American journal of medical genetics Part A* **170**, 2617–2631 (2016).

[54] Davis, G. K. *et al.* Androgen receptor blockade differentially regulates blood pressure in growth-restricted versus ovarian deficient rats. *Hypertension* **74**, 975–982 (2019).

[55] Wu, C.-C. *et al.* Androgen-dependent hypertension is mediated by 20-hydroxy-5, 8, 11, 14-eicosatetraenoic acid–induced vascular dysfunction: role of inhibitor of κb kinase. *Hypertension* **57**, 788–794 (2011).

[56] Bisserier, M. *et al.* Regulation of the methylation and expression levels of the bmpr2 gene by sin3a as a novel therapeutic mechanism in pulmonary arterial hypertension. *Circulation* **144**, 52–73 (2021).

[57] Yu, J. *et al.* Long noncoding rna ahit protects against cardiac hypertrophy through suz12 (suppressor of zeste 12 protein homolog)-mediated downregulation of mef2a (myocyte enhancer factor 2a). *Circulation: Heart Failure* **13**, e006525 (2020).

[58] del Bosque-Plata, L., Martínez-Martínez, E., Espinoza-Camacho, M. Á. & Gragnoli, C. The role of tcf7l2 in type 2 diabetes. *Diabetes* **70**, 1220–1228 (2021).

[59] Chadha, P. S. *et al.* Reduced kcnq4-encoded voltage-dependent potassium channel activity underlies impaired β-adrenoceptor–mediated relaxation of renal arteries in hypertension. *Hypertension* **59**, 877–884 (2012).

[60] Staruschenko, A., Palygin, O., Ilatovskaya, D. V. & Pavlov, T. S. Epidermal growth factors in the kidney and relationship to hypertension. *American Journal of Physiology-Renal Physiology* **305**, F12–F20 (2013).

Supplementary Information

Separating the genetics of disease, treatment and treatment response using graphical modeling and large-scale electronic health records.

Malgorzata Borczyk, Nick Machnik, Jacek Hajto, Ilse Krätschmer, Paula Konowalska, Bartosz Baszkiewicz, Michal Korostynski, Matthew Robinson

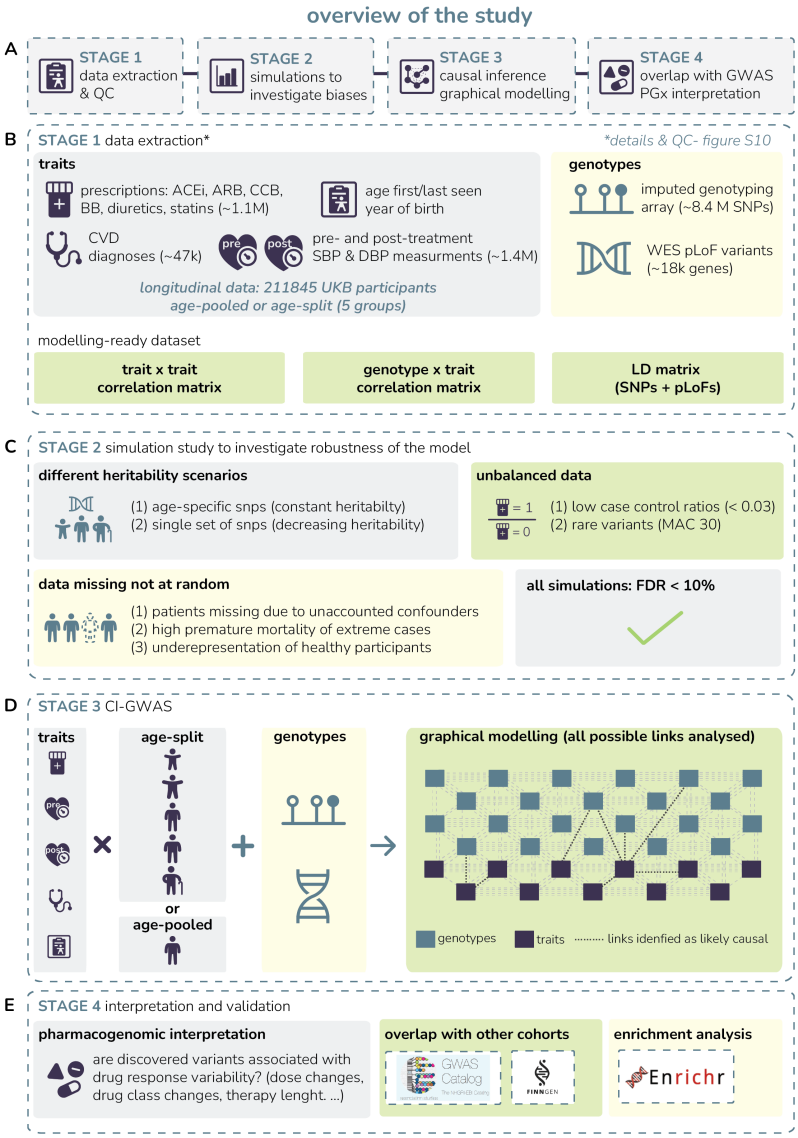

Fig. S1 General overview of the study. (A) Stages of the study (B) We first selected traits and processed the data to create a pre-post design and datasets ready for modelling. (C) We conducted a simulation study that included examining the effects of different missing data patterns. (D) We then applied our graphical model to the UK Biobank data. (E) We then conducted a pharmacogenetic follow-up study to aid interpretation, alongside replication and enrichment analyses.

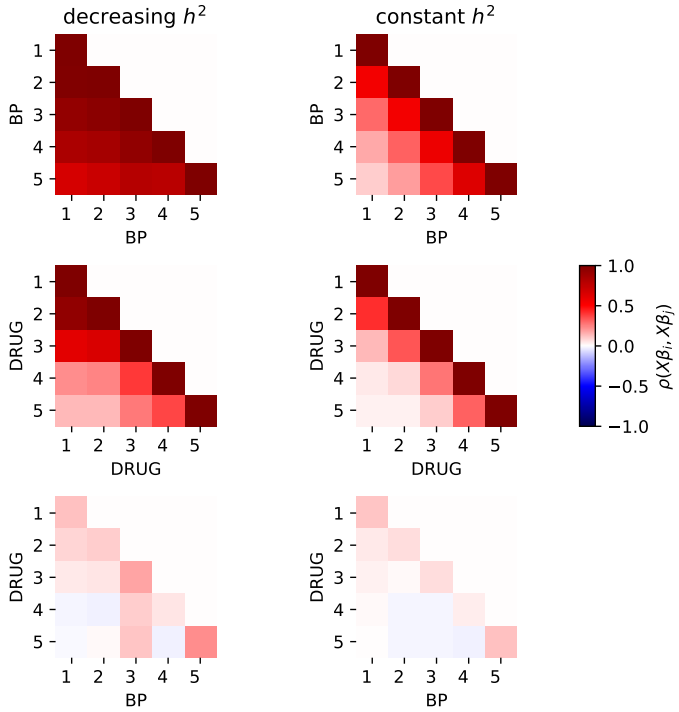

Fig. S2 Genetic correlations between traits in simulation studies with decreasing and constant heritability, h^2. The strength of the genetic correlations, $\rho(X\beta_i, X\beta_j)$, between pairs of traits i and j at the five time-points simulated from the structural equation model are shown as colors. X represents the model matrix containing the data of all causal genetic variables and β_k is the vector of regression coefficients in a model $trait_k = X\beta_k + \epsilon_k$. For more details, see Methods.

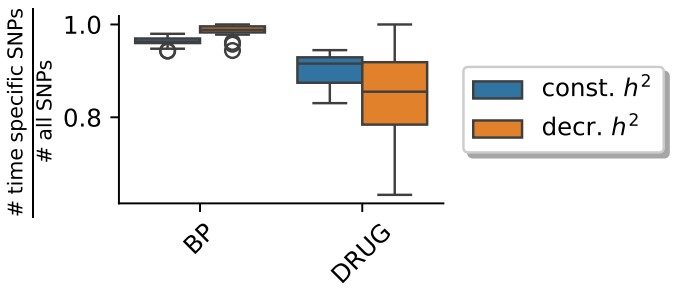

Fig. S3 Time specificity of recovered variants in simulations with constant and decreasing heritability, h^2. The figure shows the fraction of SNPs inferred to be time-point specific, out of all recovered SNPs, for blood pressure (BP, continuous trait) and drug taking (DRUG, binary trait), assuming constant h^2, where there are true causal markers for each time-point, or decreasing h^2, where there are true causal markers only at the first time-point).

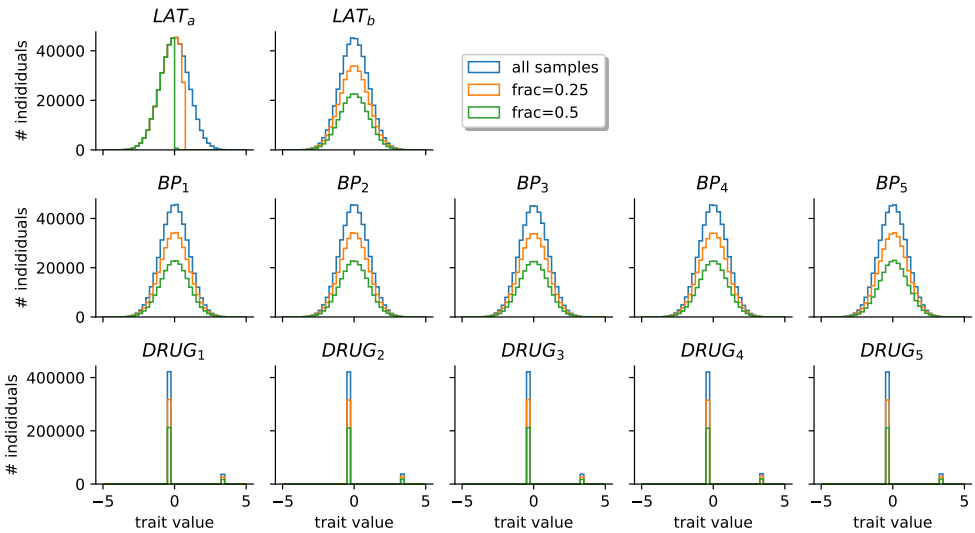

Fig. S4 Distributions of traits in simulation with censoring based on a latent selection variable. In the simulation, individuals were removed if their value in the trait LAT_a exceeded a threshold $\Phi^{-1}(1 - frac)$, where Φ is the cumulative distribution function of the Standard Normal distribution. Distributions are shown for one representative replicate out of 20.

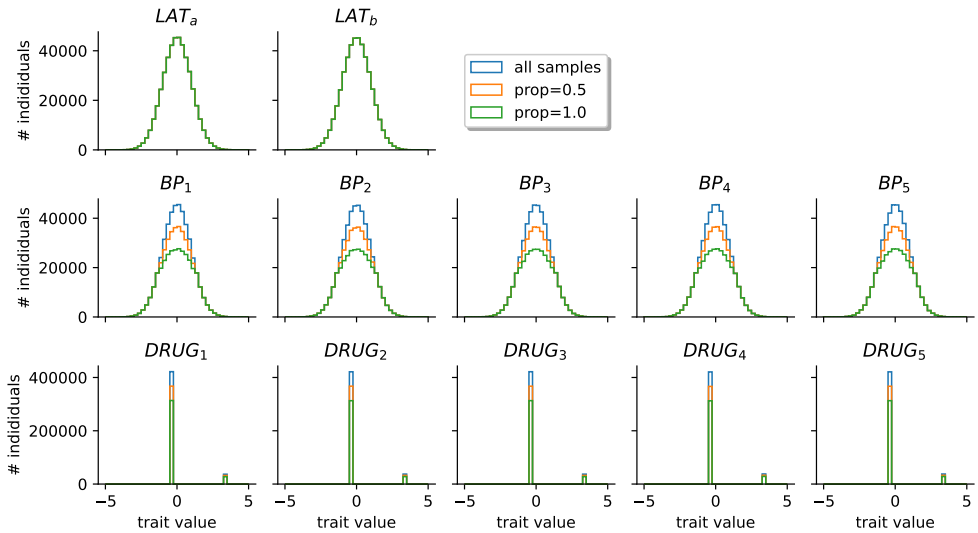

Fig. S5 Distributions of traits in simulation with censoring due to decreased incentive of healthy people to go to medical check-ups. In the simulation, individuals were removed only from that specific time-point at random at a rate $\phi(BP_t) \times prop$ if their value $|BP_t| \leq 1.2$, where ϕ is the probability density function of the Standard Normal distribution. Distributions are shown for one representative replicate out of 20.

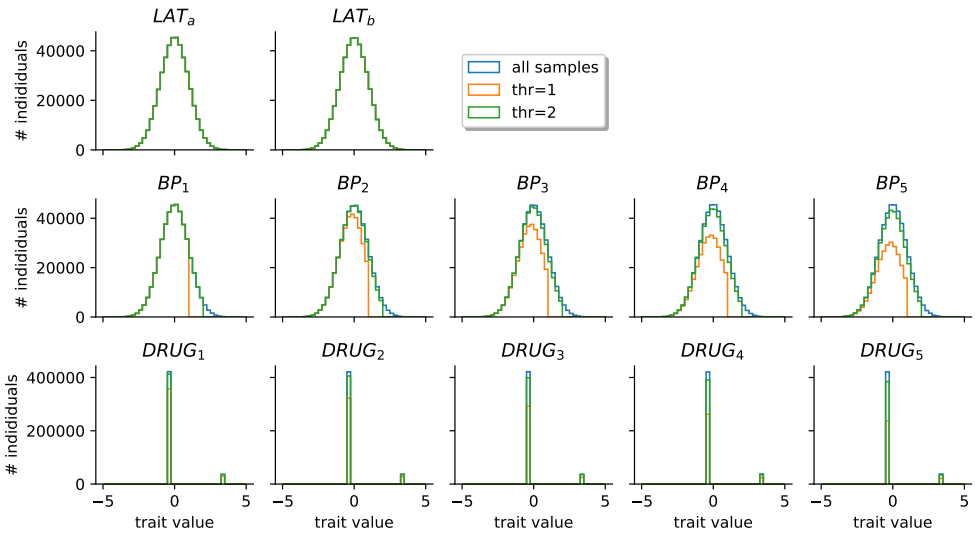

Fig. S6 Distributions of traits in simulation with censoring due to death. In the simulation, individuals were removed from all time-points $t' \geq t$ if their value $BP_t > thr$. Distributions are shown for one representative replicate out of 20.

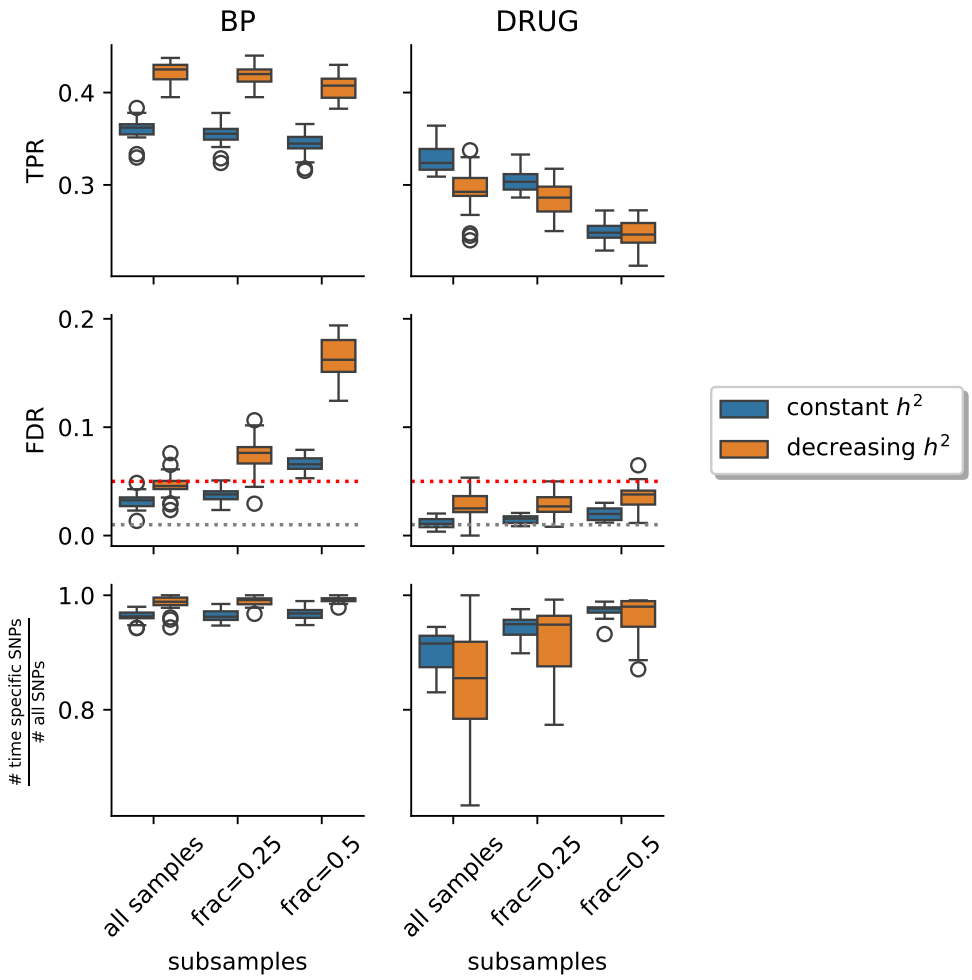

Fig. S7 Marker recovery performance in a simulation with censoring based on a latent selection variable. In the simulation, individuals were removed if their value in the trait LAT_a exceeded a threshold $\Phi^{-1}(1 - frac)$, where Φ is the cumulative distribution function of the Standard Normal distribution. We show the true positive rate (TPR) and false discovery rate (FDR) of the causal genetic variables when applying our marker detection strategy to the simulated data. Only genetic variables that were inferred to be trait- and time-specific were included in the calculations. Numbers of markers inferred to be time-point specific as a fraction of all recovered markers are shown below. Box indicates interquartile range (IQR), line in box indicates the median, whisker length is set to $1.5 \times IQR$ from the nearest hinge. Red line in FPR plot is at 5% FDR, grey line 1% FDR. All results shown across 20 replicates.

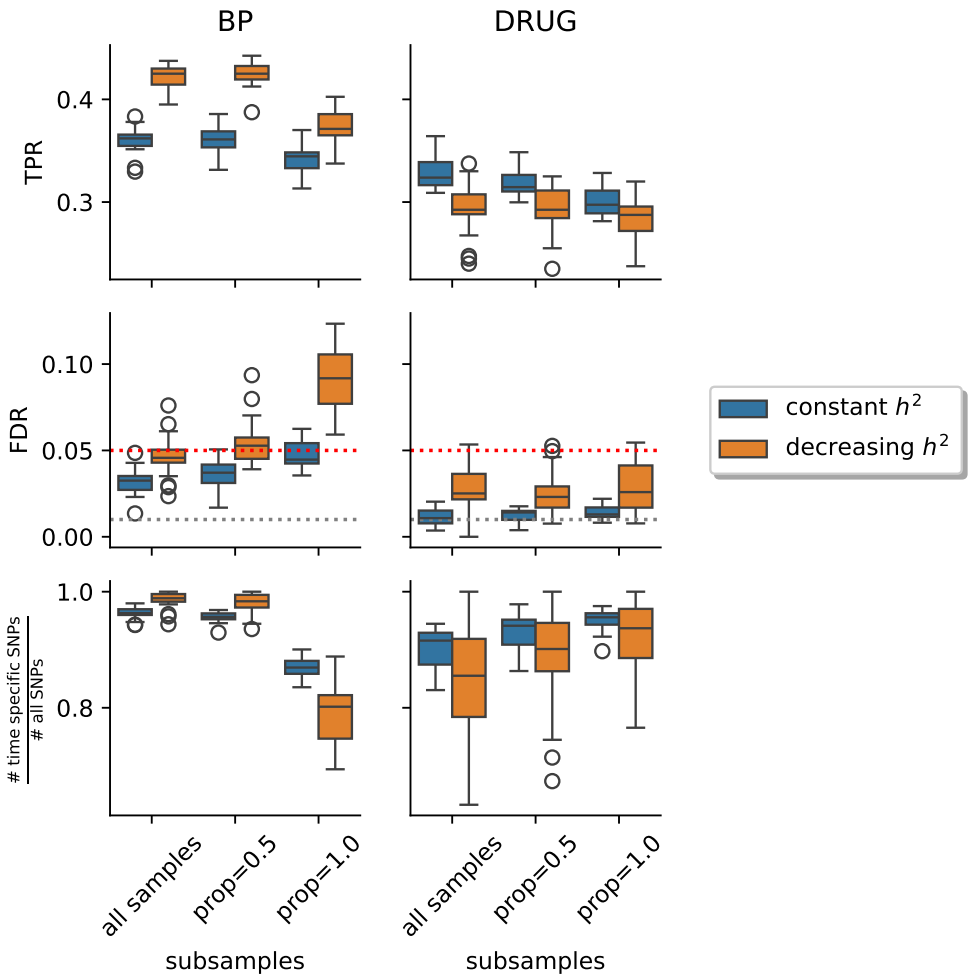

Fig. S8 Marker recovery performance in a simulation with censoring due to decreased incentive of healthy people to go to medical check-ups. In the simulation, individuals were removed only from that specific time-point at random at a rate $\phi(BP_t) \times prop$ if their value $|BP_t| \leq 1.2$, where ϕ is the probability density function of the Standard Normal distribution. We show the true positive rate (TPR) and false discovery rate (FDR) of the causal genetic variables when applying our marker detection strategy to the simulated data. Only genetic variables that were inferred to be trait- and time-specific were included in the calculations. Numbers of markers inferred to be time-point specific as a fraction of all recovered markers are shown below. Box indicates interquartile range (IQR), line in box indicates the median, whisker length is set to $1.5 \times IQR$ from the nearest hinge. Red line in FPR plot is at 5% FDR, grey line 1% FDR. All results shown across 20 replicates.

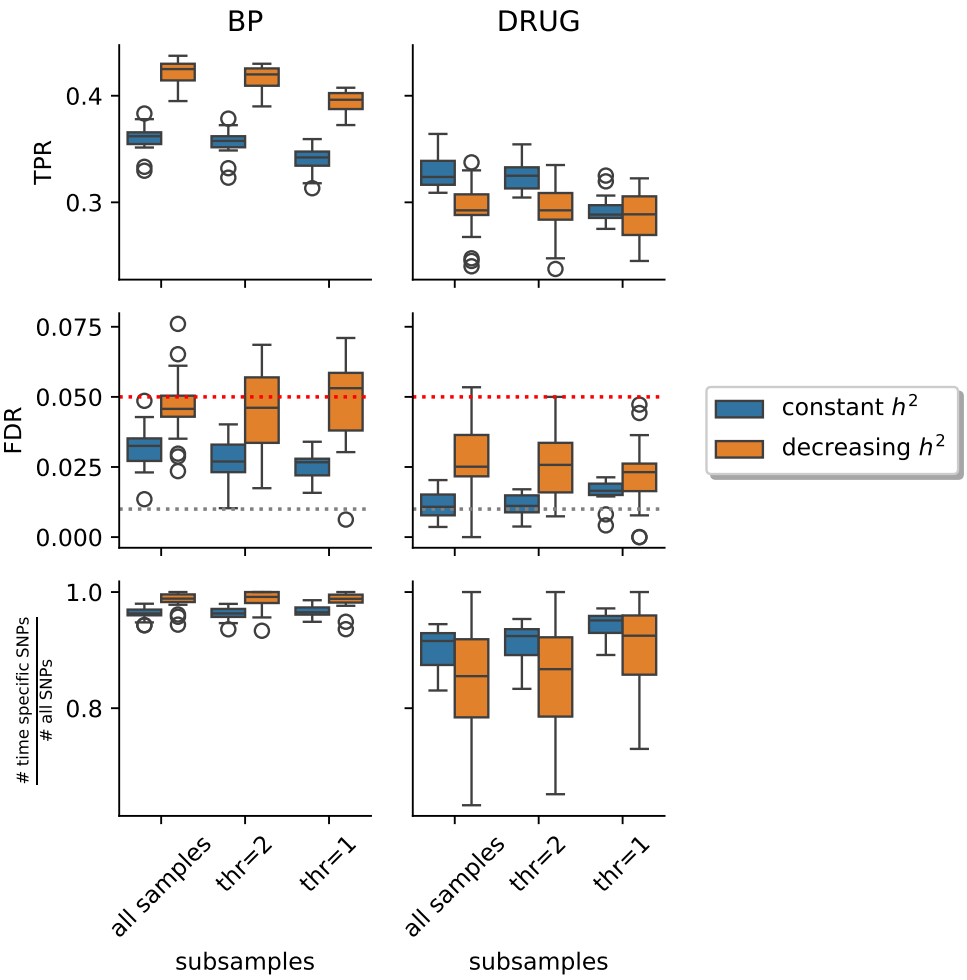

Fig. S9 Marker recovery performance in a simulation with censoring due to death. In the simulation, individuals were removed from all time-points $t' \geq t$ if their value $BP_t > thr$. We show the true positive rate (TPR) and false discovery rate (FDR) of the causal genetic variables when applying our marker detection strategy to the simulated data. Only genetic variables that were inferred to be trait- and time-specific were included in the calculations. Numbers of markers inferred to be time-point specific as a fraction of all recovered markers are shown below. Box indicates interquartile range (IQR), line in box indicates the median, whisker length is set to $1.5 \times IQR$ from the nearest hinge. Red line in FPR plot is at 5% FDR, grey line 1% FDR. All results shown across 20 replicates.

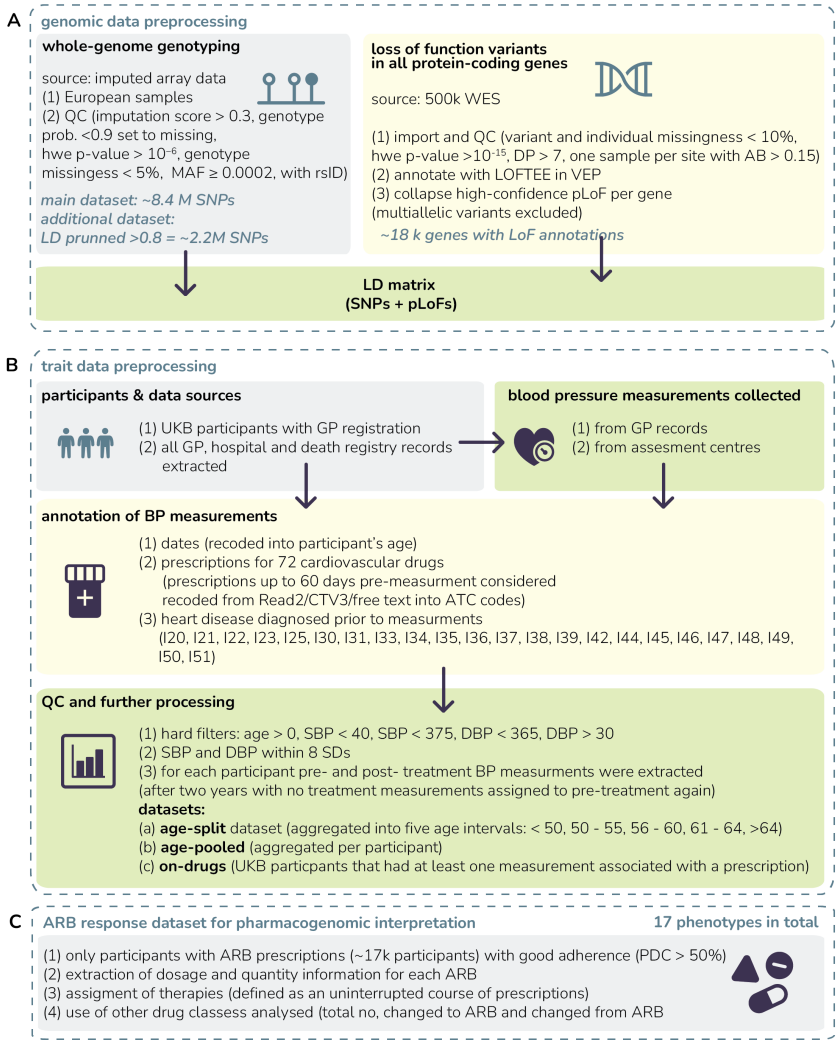

Fig. S10 Data extraction and quality control process. (A) Imputed SNPs were included alongside whole-exome sequence loss of function variants. (B) Phenotypic values were obtained from general practitioner electronic health records and drug prescription databases, annotated and filtered. (C) A follow-up pharmacogenetic study was also conducted.

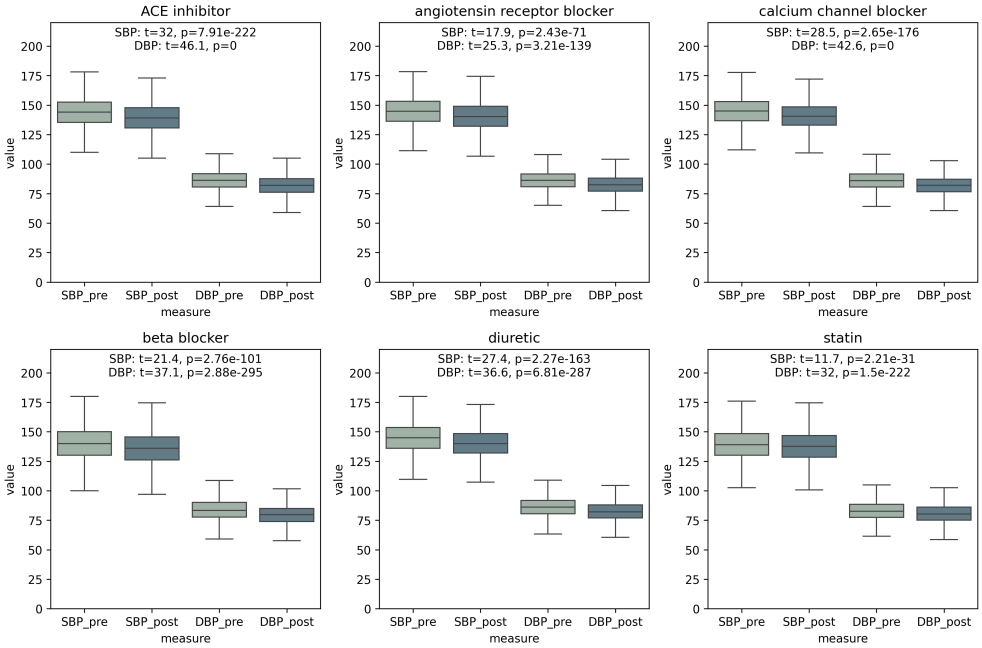

Fig. S11 Pre- and post- treatment BP values for each of the analysed drug classes. Boxplots are plotted for all UKB participants with GP records that received drugs from each class. Each of the pre- and post- treatment pairs were compared with t-tests t statistic and p-values are given on each subplot for SBP and DBP.

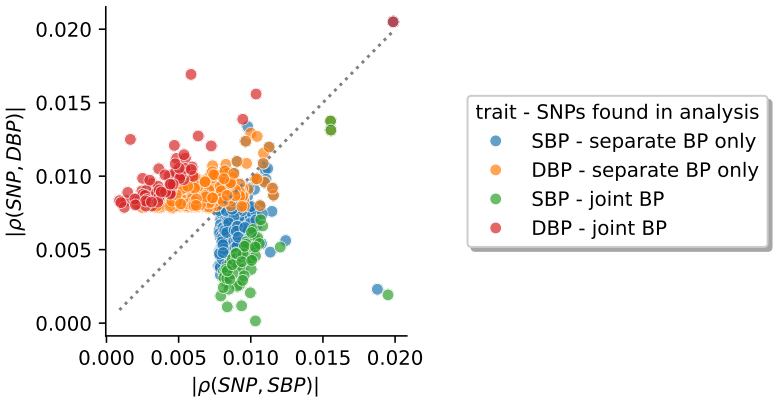

Fig. S12 Correlations of diastolic and systolic blood pressure with markers recovered in joint and separate analyses. We inferred causal markers for systolic (SBP) and diastolic blood pressure (DBP) in three analyses: (i) including only SBP, drugs and controls, (ii) including DBP, drugs and controls and (iii) including DBP, SBP, drugs and controls. Here we show that markers found to be associated with SBP (blue, green) or DBP (orange, red) are found in the joint analysis only when their correlation with SBP is not too similar to their correlation with DBP, or else they are found only in the separate analyses.

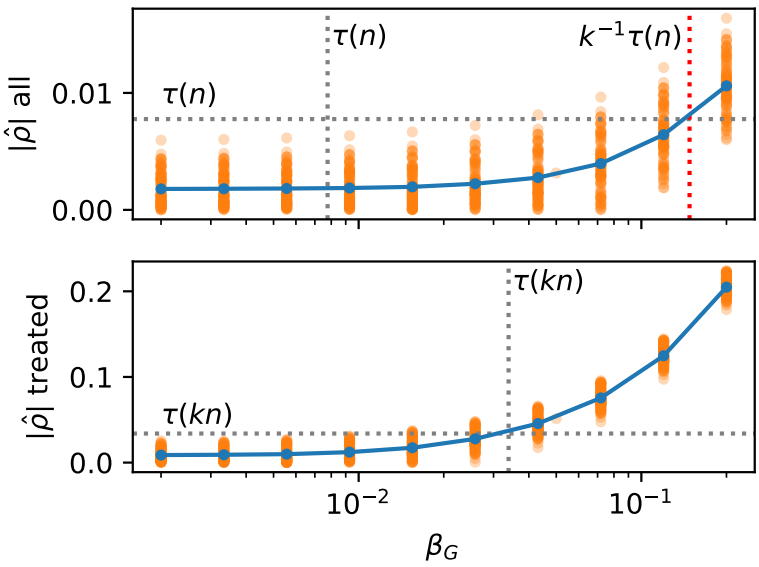

Fig. S13 Detection thresholds of variants associated with BP_{post} measures. We present the results of a simulation designed to match the correlations estimated in the analysis of real data with aged-pooled phenotypes, with $n = 180,000$ individuals. We simulated a blood pressure trait before treatment BP_{pre}, a dichotomous treatment indicator D with prevalence k, blood pressure after treatment BP_{post}, and a genetic variable G with an effect β_G on blood pressure after treatment, using a structural equation model with parameters estimated from our real observed data (see Methods). We generated data 100 replicates for each value of β_G. We then computed the partial correlations $\hat{\rho}\text{all} := \hat{\rho}(G, BP_{post}|\{BP_{pre}, D\})$ and $\hat{\rho}\text{treated} := \hat{\rho}(G^t, BP^t_{post}|\{BP^t_{pre}\})$, where $X^t = \{X_i|D_i = 1 \forall i \in [1..n]\}$. We find that, when testing $\hat{\rho}$treated, it becomes significant very closely to the marginal correlation threshold $\beta_G = \tau(kn)$, whereas $\hat{\rho}$all becomes significant at $\beta_G = k^{-1}\tau(n)$, where $\tau(x) = \Phi^{-1}(1-\alpha/2)/\sqrt{x-3}$ with $\alpha = 1e-03$. Orange dots show individual values of $\hat{\rho}$, blue dots show the arithmetic mean across the 100 replicates.

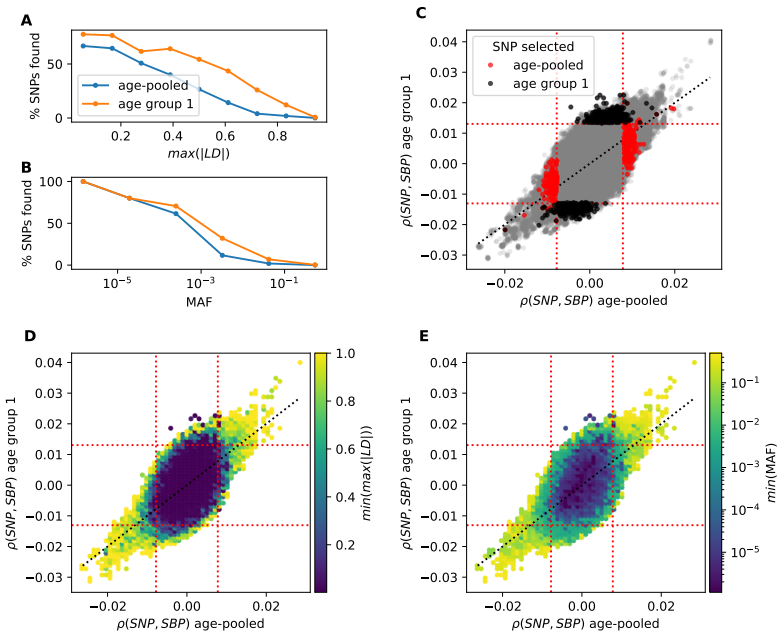

Fig. S14 Effects of linkage disequilibrium (LD) and minor allele frequency (MAF) on the discovery of marker-trait associations with CI-GWAS and the overlap of discovered markers between the age-pooled and age-split analyses. (A) Percentage of markers with marginal correlations above the detection limits of 0.008 (age-pooled) and 0.013 (age group 1) that were recovered by CI-GWAS for SBP_{pre}, calculated in ten bins in the interval $[0, 1]$ of $max(|LD|)$, the maximum absolute value of LD that a marker is in with any other marker. (B) Percentage as in A, but calculated in bins of MAF. (C) Marginal correlations of all 8M markers with SBP_{pre} in age group 1 and SBP_{pre} in the age-pooled dataset, with SNPs selected in each analysis highlighted in black (age group 1) and red (age-pooled), non-selected markers in grey. (D) We binned the 8M markers by correlation into 80 bins in the interval $[-0.04, 0.04]$, then in each bin we calculated the $min(max(|LD|))$ over all markers. (E) Same binning as in D, but with $min(MAF)$ (minor allele frequency) calculated over all markers in the bin. C, D, E: Red dotted lines indicate the detection limits in each analysis. Black dotted line shows the diagonal.

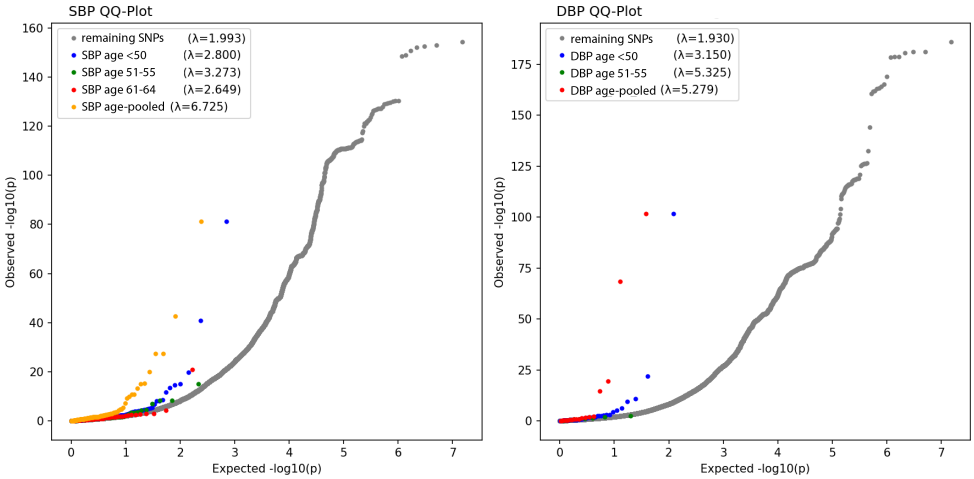

Fig. S15 Replication of SNPs discovered in our study in a recent large-scale GWAS of blood pressure For SNPs discovered in our analysis as well as their LD-partners (LD $r^2 > 0.8$) we extracted p-values from a large-scale BP GWAS [20] and selected one representing SNP from each LD block (with the lowest p-value). The expected versus the observed $-log_{10}(p)$ values are plotted for the SNPs reported in [20] and discovered in our analysis, stratified by groups. Remaining SNPs (grey dots) include all GWAS variants after excluding our lead SNPs and their LD partners. For each list we provide the inflation factor λ which gives the ratio of the median chi-square statistic observed to the expected values with one degree of freedom. Higher values of λ indicate p-values lower than expected by chance p-values.

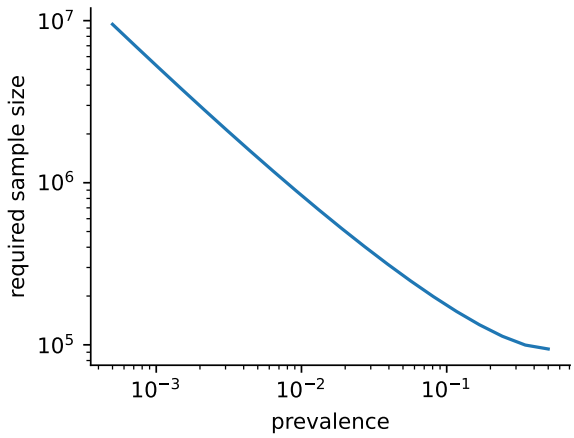

Fig. S16 Sample size required for dichotomous trait to get same power as for continuous trait with 60k observations. Shows the the sample size required as a function of the prevalence of the dichotomous trait. See Methods for details about calculation.

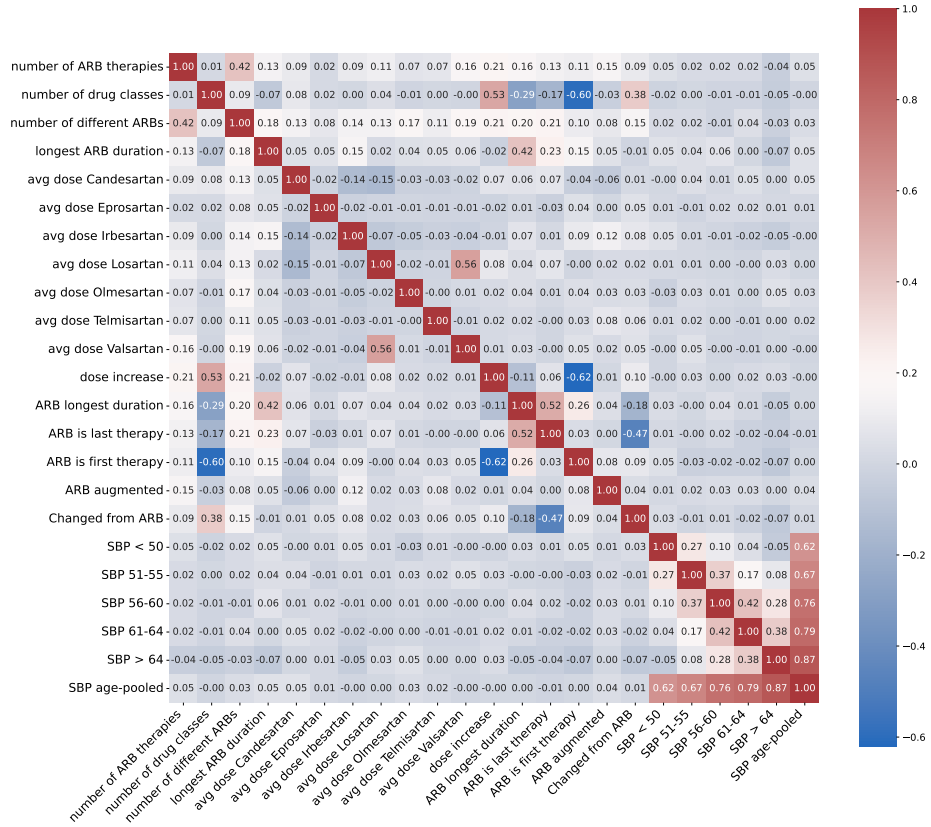

Fig. S17 Correlation heatmap of extracted ARB-related phenotypes. See Supplementary Note 1 for detais on the phenotypes. Pearson correlation values are given in each of the heatmap rectangles and colored accordingly. Pretreatment SBP values from the age-pooled and age-split datasets are also included for comparison. Each phenotype was adjusted for covariates and standardised.

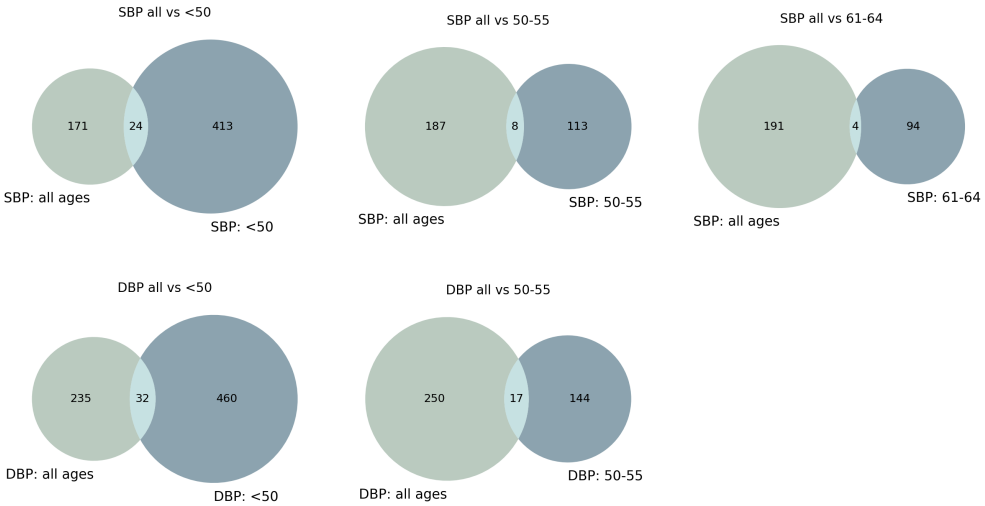

Fig. S18 Gene-level overlap between discovered variants. SNPs were mapped to genes with Ensembl VEP and joined with genes with LoF variants. The Venn diagrams show the overlap between the age-pooled (SBP all) and age-split gene lists for DBP and SBP separately. Only lists with over a 100 of variants have been included.

Table S1 Detection limit. Minimum percentage of variance explained by SNPs that can be detected in our analysis. See Methods for details about the calculations.

trait	age-group	detection-limit [% var. expl.]
ACE inhibitor	1	0.535
ACE inhibitor	2	0.241
ACE inhibitor	3	0.166
ACE inhibitor	4	0.14
ACE inhibitor	5	0.096
beta blocker	1	0.625
beta blocker	2	0.325
beta blocker	3	0.22
beta blocker	4	0.181
beta blocker	5	0.116
CCB	1	0.897
CCB	2	0.385
CCB	3	0.222
CCB	4	0.172
CCB	5	0.102
cvd	1	0.861
cvd	2	0.397
cvd	3	0.232
cvd	4	0.169
cvd	5	0.1
diuretic	1	0.806
diuretic	2	0.348
diuretic	3	0.212
diuretic	4	0.166
diuretic	5	0.112
statin	1	0.624
statin	2	0.234
statin	3	0.132
statin	4	0.095
statin	5	0.067
ARB	1	1.793
ARB	2	0.737
ARB	3	0.426
ARB	4	0.31
ARB	5	0.19
SBP pre	1	0.017
SBP pre	2	0.017
SBP pre	3	0.018
SBP pre	4	0.021
SBP pre	5	0.022
SBP post	1	0.017
SBP post	2	0.015
SBP post	3	0.015
SBP post	4	0.015
SBP post	5	0.015
DBP pre	1	0.018
DBP pre	2	0.018
DBP pre	3	0.019
DBP pre	4	0.022
DBP pre	5	0.023
DBP post	1	0.017
DBP post	2	0.016
DBP post	3	0.016
DBP post	4	0.016
DBP post	5	0.016

Table S2 Drug substances and their classifications.

ACE inhibitors		Angiot. II receptor blockers		Calcium channel blockers		Beta blockers		Diuretics		Statins	
ATC code	substance	ATC code	substance	ATC code	substance	ATC code	substance	ATC code	substance	ATC code	substance
C09AA01	captopril	C09CA01	losartan	C08CA01	amlodipine	C07AA01	alprenolol	C03AA01	bendroflumethiazide	C10AA01	simvastatin
C09AA02	enalapril	C09CA02	eprosartan	C08CA02	felodipine	C07AA02	oxprenolol	C03AA02	hydroflumethiazide	C10AA02	lovastatin
C09AA03	lisinopril	C09CA02	eprosartan	C08CA03	isradipine	C07AA03	pindolol	C03AA03	hydrochlorothiazide	C10AA03	pravastatin
C09AA04	perindopril	C09CA03	valsartan	C08CA04	nicardipine	C07AA05	propranolol	C03AA04	chlorothiazide	C10AA04	fluvastatin
C09AA05	ramipril	C09CA04	irbesartan	C08CA05	nifedipine	C07AA06	timolol	C03AA05	polythiazide	C10AA05	atorvastatin
C09AA06	quinapril	C09CA05	tasosartan	C08CA06	nimodipine	C07AA07	sotalol	C03AA06	trichlormethiazide	C10AA06	cerivastatin
C09AA07	benazepril	C09CA06	candesartan	C08CA07	nisoldipine	C07AA12	nadolol	C03AA07	cyclopenthiazide	C10AA07	rosuvastatin
C09AA08	cilazapril	C09CA07	telmisartan	C08CA08	nitrendipine	C07AA14	mepindolol	C03AA08	methyclothiazide	C10AA08	pitavastatin
C09AA09	fosinopril	C09CA08	olmesartan medoxomil	C08CA09	lacidipine	C07AA15	carteolol	C03AA09	cyclothiazide		
C09AA10	trandolapril	C09CA09	azilsartan medoxomil	C08CA10	nilvadipine	C07AA16	tertatolol	C03AA13	mebutizide		
C09AA11	spirapril	C09CA10	fimasartan	C08CA11	manidipine	C07AA17	bopindolol	C03BA02	quinethazone		
C09AA12	delapril			C08CA12	barnidipine	C07AA19	bupranolol	C03BA03	clopamide		
C09AA13	moexipril			C08CA13	lercanidipine	C07AA23	penbutolol	C03BA04	chlortalidone		
C09AA14	temocapril			C08DB01	diltiazem	C07AA27	cloranolol	C03BA05	mefruside		
C09AA15	zofenopril			C08CX01	mibefradil	C07AB01	practolol	C03BA07	clofenamide		
C09AA16	imidapril			C08CA14	cilnidipine	C07AB02	metoprolol	C03BA08	metolazone		
				C08CA15	benidipine	C07AB03	atenolol	C03BA09	meticrane		
				C08CA16	clevidipine	C07AB04	acebutolol	C03BA10	xipamide		
				C08CA17	levamlodipine	C07AB05	betaxolol	C03BA11	indapamide		
				C08CA51	amlodipine celecoxib	C07AB06	bevantolol	C03BA12	clorexolone		
				C08CA55	nifedipine combinations	C07AB07	bisoprolol	C03BA13	fenquizone		
						C07AB08	celiprolol	C03BC01	mersalyl		
				C08DA01	verapamil	C07AB09	esmolol	C03BD01	theobromine		
				C08DA02	gallopamil	C07AB10	epanolol	C03DA01	spironolactone		
						C07AB11	s-atenolol	C03DA02	potassium canrenoate		
				C08EA01	fendiline	C07AB12	nebivolol	C03DA03	canrenone		
				C08EA02	bepridil	C07AB13	talinolol	C03DA04	eplerenone		
				C08EX01	lidoflazine	C07AB14	landiolol	C03DA05	finerenone		
				C08EX02	perhexiline	C07AG01	labetalol	C03DB01	amiloride		
						C07AG02	carvedilol	C03DB02	triamterene		
								C03CA01	furosemide		
								C03CA02	bumetanide		
								C03CA03	piretanide		
								C03CA04	torasemide		
								C03CC01	etacrynic acid		
								C03CC02	tienilic acid		
								C03CD01	muzolimine		
								C03CX01	etozolin		
								C03XA01	tolvaptan		
								C03XA02	conivaptan		
								C03BX03	cicletanine		

Supplementary Note 1

Extraction of blood pressure measurements, diagnoses and prescription data from the UK Biobank

Blood pressure measurements and prescription data were extracted from the UK Biobank Research Access platform with pyspark [v 3.2.3] and then processed with Hail (see Code Availability). All of the UK Biobank participants with GP (General Practice) registration were selected and then all GP, hospital and death registry records were extracted for these participants with accompanying dates. For prescription extraction, a dictionary of medications used in blood pressure management was prepared and codes in each of the prescription coding systems used in the UK Biobank GP data were translated into Anatomical Therapeutic Chemical (ATC) codes. The free text fields were also parsed for medication names.

For blood pressure (BP) measurements relevant Read2/CTV3 codes were extracted together with the measurement value. Measurements performed in the UK Biobank Assessment Center were also added to the dataset. For BP measurements in the GP data with two numerical values, the higher value was assigned to systolic blood pressure (SBP) and the lower to diastolic blood pressure (DBP). For single-value measurements, the value has been assigned to DBP if 'diastolic' was present in the procedure code description, otherwise to SBP. The following hard filters were applied to remove impossible measurements: age > 0, SBP > 40, SBP < 375, DBP < 365 and DBP > 30. Heart disease diagnoses that included codes from the ICD10 chapters I20, I21, I22, I23, I25, I30, I31, I33, I34, I35, I36, I37, I38, I39, I42, I44, I45, I46, I47, I48, I49, I50, I51 together with equivalent Read2/CTV3 and ICD9 codes were also re-coded and extracted. A full json dictionary of all of the above items can be found in the Code Availability section. Finally, we prepared a dataset that included all blood pressure measurements annotated with date, source, cardiovascular disease (CVD) diagnosis and prescription information. CVD has been assigned 1 if any relevant diagnosis has been made prior to the measurement and 0 otherwise. Prescriptions have been grouped into six drug categories (Table S2). Each measurement was assigned 1 if a prescription for a drug in a category has been issued within 60 days prior to the measurement. Measurements were separated into pre- and post-treatment in the following procedure: Measurements were accepted as pre-treatment if there was no prescription record for any drug from the investigated classes in two years prior. Measurements associated with prescription for any of the investigated drugs were considered post-treatment. Measurements for each participant were then aggregated. In the age-pooled approach, we averaged pre- and post- treatement SBP and DBP values and the maximum value from the remaining variables (drug classes, measurement source, CVD diagnosis) was used in the aggregation. If a person had no treatment records, we copied the pre-treatment BP values into the post-treatment values. In the age-split approximation, measurements for each participant were first aggregated into five age intervals averaging pre- and post-treatment and choosing the maximum value for other variables. The intervals were chosen as to minimize the standard deviation of group size. If a participant did not have on-drug measurements within an interval, the pre-treatment values were copied. All variables were

then adjusted for covariates (sex, 20 genetic principal components, assessment center, genotyping batch, coordinates) and standardized.

Extraction of data to evaluate individual drug responses

Major classes of cardiological drugs were defined based on the British National Formulary (BNF) classification system. Drugs were identified by filtering lookup tables from the UK Biobank resource (see biobank.ndph.ox.ac.uk/ukb/ukb/auxdata/primarycare_codings.zip) according to the appropriate BNF paragraphs, sections and chapters.

- Diuretics: Defined as drugs listed under section 2.2 ("Diuretics").
- Statins: Defined as drugs listed under section 2.12 ("Lipid-Regulating Drugs").
- Beta-Blockers: Defined as drugs listed under section 2.4 ("Beta-Adrenoceptor Blocking Drugs").
- Calcium-Channel Blockers (CCBs): Defined as drugs from paragraph 2.6.2 ("Calcium-Channel Blockers").
- Angiotensin-Converting Enzyme Inhibitors (ACE Inhibitors): Defined as drugs from subparagraph 2.5.5.1 ("Angiotensin-Converting Enzyme Inhibitors").
- Angiotensin II Receptor Antagonists: Defined as drugs from subparagraph 2.5.5.2 ("Angiotensin II Receptor Antagonists").
- Renin Inhibitors: Defined as drugs from subparagraph 2.5.5.3. ("Renin Inhibitors").

For each group, active substances were matched with corresponding brand names, using the classification data to associate these substances with entries in different drug coding systems (Read2, CTV3, DMD). This yielded a set of codes with which the electronic health records from UK Biobank was then filtered.

In addition, dosage and quantity information was extracted for each drug code. Where available, dosage and quantity values were directly retrieved from the structured data fields associated with the codes. In cases where such structured information was missing, quantity values were extracted from the textual descriptions within the coding system using regular expression techniques (e.g., identifying patterns such as "mg", "ml", "mg/ml"). This two-step approach ensured that dosage data was included wherever possible, providing a more complete picture of each drug's use and characteristics (see Code Availability).

In the case of drug combinations, cases where drugs were given together with diuretics or calcium channel blockers were included. When the combination was not explicitly stated in the data (e.g. 'with diuretics' or 'with a calcium channel blocker'), the information was noted and the drugs were treated as drugs of the relevant group (diuretic or CCB) without specifying the exact substance.

To ensure the integrity and quality of the dataset, records with missing values were removed. In particular, the following data were removed:

- Missing Dose Information: A total of 0.86% of records were found to be missing dose information.
- Missing Quantity Information: 13.85% of records were missing information on the quantity.

- Missing Dates: A small proportion (0.002%) of records had missing date information.

Combined, these filters removed 14.10% of all prescriptions due to missing data. In addition, individuals with no ARB prescriptions were excluded from the analysis. As a result, the final dataset included 19,114 individuals.

We then defined therapies and calculated the proportion of days covered (PDC) for ARBs. To ensure the correct classification of therapies, we identified the intervals between prescriptions for each patient and drug, which were added to the database. We defined a therapy for each patient as a set of prescriptions involving the same drug and a similar expected daily dose.

Prescriptions were grouped into separate therapies based on the interval between adjacent prescriptions. If the interval between two consecutive prescriptions exceeded 90 days, adjusted for the number of tablets prescribed, they were considered separate therapies. In addition, the algorithm considered the median mean daily dose over 11 consecutive prescriptions. It compared consecutive median values and considered that a significant dose change occurred when the dose was changed at least 2 times. When such a change occurred, treatment was divided into separate treatment periods.

For ARB therapy, the percentage of days on treatment (PDC), an indicator used to assess medication adherence, was also calculated. The calculation was based on the estimated daily dose of the drug based on the fact that the minimum dose is no less than 1 tablet.

Finally, we then applied specific inclusion criteria by removing therapies with a PDC below 50% or exceeding 200%. This filtering step eliminated approximately 1,200 therapies, ensuring that only reliable adherence data remained. After the PDC filtering, the final dataset contained approximately 17,000 individuals who were prescribed ARBs.

ARB therapy data was then aggregated for each participant to extract drug response proxy phenotypes. This included collection of counts of different ARB substances prescribed, total number of different drug classes prescribed to ARB-takers, duration of longest ARB therapy, average doses of most common ARBs and binary traits indicating if ARB was first or last therapy for a participant, if it was ever augmented with another drug, if the dose was ever increase or if an ARB was discontinued and a drug from another class was introduced. This process resulted in 17 ARB response traits.

Supplementary Note 2

Overlap between the result sets

Between the age-pooled and age-split analyses the variant sets that we discover are largely non-overlapping (Figure S11B). In the age-split analysis, there is no overlap between variants in any of the age groups for DBP and one SNP for SBP that overlaps between the < 50 and 50-55 age groups. This SNP is rs35479618 which is a missense variant (E967K) in the Natriuretic Peptide Receptor 1 (*NPR1*) known to be associated with elevated SBP [45]. This variant is also found in the age-pooled analysis.

For each of the SBP and DBP-associated variant sets that exceeded 100 variants, we mapped SNPs to genes using Ensemble's VEP. In total 2733 SNPs were mapped to 2038 genes (see Supplementary Data Tables) and joined with the genes with associated LoF variants. First, we observe that the gene-level overlap between age-split and age-pooled analyses for SBP and DBP is higher than the SNP-level overlap (Figure S18). As an example, only one SNP is found in the overlap between SBP-associated variants in the age-pooled model and SBP-associated variants in the age group 60-64. This variant is rs146210632, an intergenic variant on chromosome 2. At the gene level however, we find common mappings to 4 protein-coding genes (*FBXL17*, *WWOX*, *LDB2* and *COL12A1*). We note that the directions of the SNP effect sizes within the same genes, but for different ages may be opposite. The rs186089134 intron variant in *FBXL17* (gnomAD MAF = 0.0099) has a Pearson correlation coefficient of -0.010 for SBP in the age-pooled analysis, while the distinct intron variant rs141714549 in the same gene has a correlation of 0.017 (gnomAD MAF = 0.0065). Variants in *FBXL17* have been previously associated with BMI and adipose tissue quantity, according to the GWAS catalog, as well as with antihypertensive medication response [46]. The example of these two variants shows how CI-GWAS is able to refine known associations and discover markers in the same gene but with different effects depending on the age of the person.

Further description of loss-of-function variants

Conditional on the SNP associations and all other variants genome-wide, we also discover many genes with loss-of-function (LoF) variants associated with the pre-treatment measures (Table 1). Examples of LoF variant associations (with 3 or more carriers within the UKB) include *SMAD2* detected in the pre-treatment SBP from all ages. This gene produces a signaling protein involved in artery remodeling [21, 22]. For DBP in all ages, we report associations with LoFs in *BAG6*. Common variants in this gene have been previously associated with BP [47]. In the youngest age group, for SBP, we detect association with LoF variants in *ANGPTL1* - a gene known for regulation of angiogenesis [48]. For DBP, in the same age group, we replicate association with *PDE3A*. This gene encodes a phosphodiesterase and its gain of function mutations cause brachydactyly with severe hypertension [23, 24]. In the age group 60-4, LoFs in *ARVCF* are associated with SBP. Common variants in this gene have previously been found in relation to pulse pressure (GWAS catalog: GCST004278, GCST007705). Furthermore, deletion of this catenin, among other protein-coding genes, occurs in the autosomal dominant Velo-Cardio-Facial syndrome (VCFS) that is associated with

abnormal cardiovascular system morphology [49]. In summary, in our WES analysis, conditioned on common variant background, we replicate known genetic syndromes associated with hypertension as well as provide novel findings. All LoF associations are listed in Supplementary Data Table D.

Enrichment analysis

As rare and common variants act together to influence complex traits, we join genes with LoF variants discovered in our model together with the genes annotated to SNPs in a gene-level enrichment analysis [50, 51]. We performed gene level enrichment analyses with the (1) Reactome Pathways Database for pathway enrichment; (2) Gene Ontology (GO) Molecular Function to integrate the biological function of discovered genes; (3) ChEA and ENCODE transcription factor database to search for possible common regulators of the discovered genes (Figure 3C, Supplementary Data Tables).

From all of the cardiovascular-health related traits within the GWAS catalog the top overrepresented terms are Systolic Blood Pressure, Diastolic Blood Pressure, Electrocardiogram Morphology (Amplitude At Temporal Datapoints) and Medication Use For Hypertension (Number Of Purchases). As an example, 32 out of 195 SBP-associated genes in the age-pooled analysis (16.4%) have been previously associated with SBP in GWAS (enrichment p-value = 1.50e-8, odds ratio = 3.54). In the age-split analysis, in the youngest age group (< 50), 63 out of the 437 SBP-associated genes (14.4%) are listed in the GWAS catalog for SBP (p-value 7.07e-13, OR = 3.09). In the 50-55 years-old age group, 15 out of 121 genes (12.4%) are also in the SBP GWAS catalog term (p-value 0.0021, OR = 2.52). There is no significant enrichment for SBP GWAS genes in the 98 genes we discover in the 61-64 age group, suggesting that these are novel findings (Figure 3C). For DBP, we find a significant (max p-value 0.00078) overrepresentation of genes previously associated with DBP in GWAS for three sets of DBP genes (all ages, < 50, 50-55). Notably, the highest overlap with the genes associated with medication use for hypertension is observed in the 50-55 years-old age group (8 out of the 121 SBP genes, p-value = 7.23e-7, OR = 12.38). Taken together, our results are highly replicable and suggest that the genetic variation for BP is shaped prior to mid-life, with limited evidence for SNPs whose effects contribute to BP variation later in life.

In a functional enrichment analysis, we find terms only in the youngest age groups (< 50) that are associated with organism development: axon guidance for SBP (p-value = 0.00039, OR = 2.24) and myogenesis for DBP (p-value = 0.00064, OR = 8.34) (Figure 3C). Cardiac conduction, on the other hand, has the strongest overrepresentation in genes associated with SBP in the 61-64 age group (p-value = 0.0038, OR = 6.68). In terms of molecular function, in the youngest age groups, we find calcium channel activity enrichment in both the SBP (p-value = 0.0013, OR = 4.08) and DBP (p-value = 0.00064, OR = 4.11) genes. Based on ChIP-seq data, the discovered genes share common regulatory pathways across the age groups. However, the enrichment for the top Transcritpion Factor (TF) binding sites is most pronounced in the youngest age groups for both SBP and DBP-associated gene (Figure 3C). The TFs that we discover are REST (RE1-silencing transcription factor), SMAD4 (a central $TGF-\beta$ pathway mediator), AR (androgen receptor), and SUZ12 (a Polycomb group

protein); all with crucial functions in human development. REST affects potassium channel expression and vascular smooth muscle cell proliferation [52]. SMAD4 regulates vascular and cardiac development and remodeling. Gain of funcion mutations in SMAD4 cause hypertension in about 15% of patients [53]. In animal studies, AR signaling has been linked to hypertension in both males and females [54] [55]. SUZ12 belongs to a complex that epigenetically silences gene expression and mediates heart development and its abberant signalling has been previously implicated in pulmonary hypertension [56] [57].

Associations with possible confounders

Across the analyses we discover multiple SNP and LoF variants associated with likely confounders. We note that we find no variants associated with the year of birth, one SNP associated with age at first record in the dataset (rs187729359) and three SNPs associated with age at last record in the dataset (rs147750638, rs115024119, rs78651908). All four are discovered in the age-pooled analysis and only rs147750638 is located within a protein-coding gene locus: *MARCHF4*, a membrane-bound E3 ubiquitin ligase, variants in which have been previously associated with height, smoking initiation and acute myeloid leukemia in the GWAS catalog. In the age-pooled model, we find 234 variants associated with number of measurements and none with median age at measurement. Among these 234 variants, we find SNPs in genes previously associated with risk of type 2 diabetes (*TCF7L2* and *HNF4A*). In *TCF7L2*, which is the most potent risk locus for type 2 diabetes risk [58], we find a previously unreported SNP associated positively with number of measurements (Pearson $r^2 = 0.008$). We therefore hypothesize that the inclusion of the "number of measurements" phenotype may adjust for comorbidities as patients with accompanying diseases are more often measured.

In the age-split models, we find variants associated with number of measurements for each age group, but just as with SBP/DBP these numbers tend to decrease for older participants (age < 50 - 1058 variants, age $50 - 55$ - 18 variants, age $56 - 60$ - 4 variants, age $61 - 64$ - 7 variants, age > 64 - 6 variants). We also find 689 variants associated with the median age at measurement in the youngest age group. This is expected as patients with genetically-determined syndromes should have more BP measurements early in life. Accordingly, 45 of these variants are LoF and include genes like *KCNQ4* associated with hearing loss in humans and known for its regulation of vascular tone [59] and Epidermal Growth Factor (*EGF*) which expression determines appropriate kidney function [60]. A pooled enrichment analysis of all of the SNPs and LoF variants (863 genes) associated with the possible confounders in the age-split approach points to multiple cardiovascular phenotypes. According to the PhenGenI Association 2021 database (accessed via Enrichr), 114 genes (13%) are associated with HLD cholesterol levels (adjusted p-value 1.09e-7). Other examples of overrepresented terms include 65 genes associated with asthma (adjusted p-value 1.85e-4), 69 with myocardial infraction (1.85e-4), 80 with stroke (3.89e-4) and 162 with Pervasive Child Development Disorders (4.61e-4). In summary, we show that the possible confounding factors included in our model are associated with variants linked to cardiovascular disease outcomes.

Description of Supplementary Data Tables

In all Supplementary Data Tables genomic coordinates are reported for GRCh37.

Supplementary Data Table A Sample size for all traits in all models. In addition, we report trait prevalence for binary traits.

Supplementary Data Table B All SNPs associated with pretreatment SBP and DBP in both the age-pooled and the age-split model.

Supplementary Data Table C Gene level enrichment analyses that includes both SNPs and the genes with LoF variants with the GWAS Catalog. This table includes any results with adjusted p-value < 0.05 as reported by Enrichr.

Supplementary Data Table D All LoF variatms associated with SBP/DBP in the age-pooled and age-split analyses.

Supplementary Data Table F Associations with potential confounders. Gene symbols are provided for LoF variants and rsIDs for SNPs.

Supplementary Data Table G Associations of drug-response SNPs in Open Targets Database including Finngen. This table includes replication information for discovered SNPs and their LD partners in multiple datasets, but only up to r6 of FinnGen. Associations for rs12608822 and rs186696265 have been searched and confirmed in the r12 FinnGen release (see https://r12.finngen.fi/variant/19:19730943-C-T and https://r12.finngen.fi/variant/6:160690668-C-T)

Supplementary Data Table H ARB-taking SNPs and LoFs findings.

Supplementary Data Table I Results for ARB-response variables association with ARB-taking variants.

Supplementary Data Table J ARB-response variable multiple regression with SBP and DBP SNPs.

Supplementary Data Table K VEP results for a joined list of SBP- and DBP-associated SNPs across the age-split and age-pooled analyses.