

## Adjusting for medication status in genome-wide association studies

Amanda H.W. Chong<sup>\*1</sup>, Christopher Kintu<sup>2</sup>, Yoonsu Cho<sup>1</sup>, Segun Fatumo<sup>2,3</sup>, Jason Torres<sup>4</sup>, George Davey Smith<sup>1,4</sup>, Tom R. Gaunt<sup>1,5¶</sup> and Gibran Hemani<sup>1,5¶</sup>

<sup>1</sup> MRC Integrative Epidemiology Unit, Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, United Kingdom

<sup>2</sup> MRC/UVRI and LSHTM Uganda Research Unit, Entebbe, Uganda

<sup>3</sup> Department of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, UK

<sup>4</sup> Medical Research Council Population Health Research Unit, Clinical Trial Service Unit and Epidemiological Studies Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK

<sup>5</sup> NIHR Bristol Biomedical Research Centre, University Hospitals Bristol and Weston NHS Foundation Trust and University of Bristol

\* Corresponding author: Amanda H.W. Chong ([a.chong@bristol.ac.uk](mailto:a.chong@bristol.ac.uk))

¶ Equal contributors

Keywords: Medication; GWAS; Bias

## Abstract

When conducting genome-wide association studies, improper handling of medication status that is relevant to the trait of interest can induce biases by opening up different pathways that distort estimates of the true effect. Here, we propose the genetic empirical medication reduction adjustment (GEMRA) method which uses a heuristic search for an empirical adjustment to be applied to phenotypic values of participants reporting medication use. Through simulations we show that the direct genetic effect estimates in the GEMRA approach exhibited less bias and greater statistical power than either restricting the sample to unmedicated users, or including all samples without adjustment. We then applied the GEMRA approach to estimate statin medication adjustment for analysis of LDL cholesterol levels, using multi ancestry data from UK Biobank and the Uganda Genome Resource. We found that a relative rather than an absolute adjustment better modelled the effect of medication on LDL cholesterol, with an effect of 40% reduction appearing to be consistent across ancestral groups. These findings are consistent with the current clinical guidelines.

## Introduction

Inclusion of participants taking medication in genome-wide association studies (GWAS) of a phenotype affected by that medication is challenging as it introduces heterogeneity in the analyses. This is due to medicated participants having an observed phenotype value that is not reflective of their true, unmedicated phenotype value included in the analyses with unmedicated participants, resulting in findings that don't represent either group (1-3). For example, participants taking LDL-lowering medication will have a lower observed LDL cholesterol (LDLc) value due to the effect of medication use. This issue has been previously explored by Bell et al. 2022(4) and Fang et al. 2022(5), with both findings showing the misleading apparent effects of BMI on lowering LDLc when the effects of statin use is not addressed. However, when medication use is considered, these results can be attributable to people with higher BMI being more likely to be given statins.

To date, there is no approach to correct for medication status that would result in an adjustment that reliably restores the medication-adjusted phenotype value to the unmedicated value. In previous literature, approximations of the unmedicated phenotypic value have been applied for phenotypes such as blood pressure(6-8), lipid traits(9-12), and type 2 diabetes(7, 9, 13) with studies accounting for treatment effect using a plausible constant, censored normal regression, or a non-parametric method(6, 14). However, for LDLc, studies have commonly excluded participants based on medication status (15-23), and there are several concerns that can occur due to this approach. As shown in Figure 1, in the example involving LDL-lowering medication and LDLc, the exclusion or inclusion of participants based on medication status or stratifying by bi-medication status opens a causal pathway through medication status to the observed phenotype value. This can introduce bias to the estimated causal effect (i.e. collider bias)(24). Furthermore, additional collider bias can also be

introduced as prescription of medication may be influenced by other factors other than high LDLc such as a cardiovascular event or a family history of high cholesterol (i.e., genetic variants associated with other risk factors in Figure 1). Therefore, selection by medication status could induce an association between genetic variants associated with non-LDLc risk factors and LDLc. In addition, by excluding participants by medication status, this approach could be ignoring participants that could better inform analyses of potential genetic variants associated with dyslipidemia(25). Exclusion would also decrease the sample size, reducing the statistical power of the GWAS which, in some instances, could be substantial. Without adequate statistical power, analyses may be unable to detect genetic signals at the genome-wide threshold(26). In contrast, not adjusting for medication status would introduce heterogeneity into the analyses as the observed value would be different from the unmedicated value. This could then result in a bias that either overestimates or underestimates the true effect, or opens a path for genetic variants affecting other factors influencing medication use.

Therefore, to account for medication status, a relative or absolute adjustment of medication reduction could be applied. We propose the genetic empirical medication reduction adjustment (GEMRA) method that would heuristically search for an optimal adjustment value to approximately correct the medication-biased phenotype measurements to a value that better reflects the unmedicated phenotypic value in participants taking medication. *Relative adjustment* would be a percentage reduction in medication effect that is applied to the observed value. *Absolute adjustment* would be a specific medication reduction value that is applied to the observed value. In a cross-ancestry setting, differences in prevalence of medication use and phenotype distribution may contribute to differences in the optimum adjustment for each population.

In this study, we applied the GEMRA approach to LDLc and LDL-lowering medication status. We first employed a simulation framework to examine the impact of this approach on direct and indirect genetic associations for LDLc. We then conducted the analyses using real-world data from UK Biobank (UKB) and the Uganda Genome Resource (UGR) to evaluate the adjustment approach in different populations.

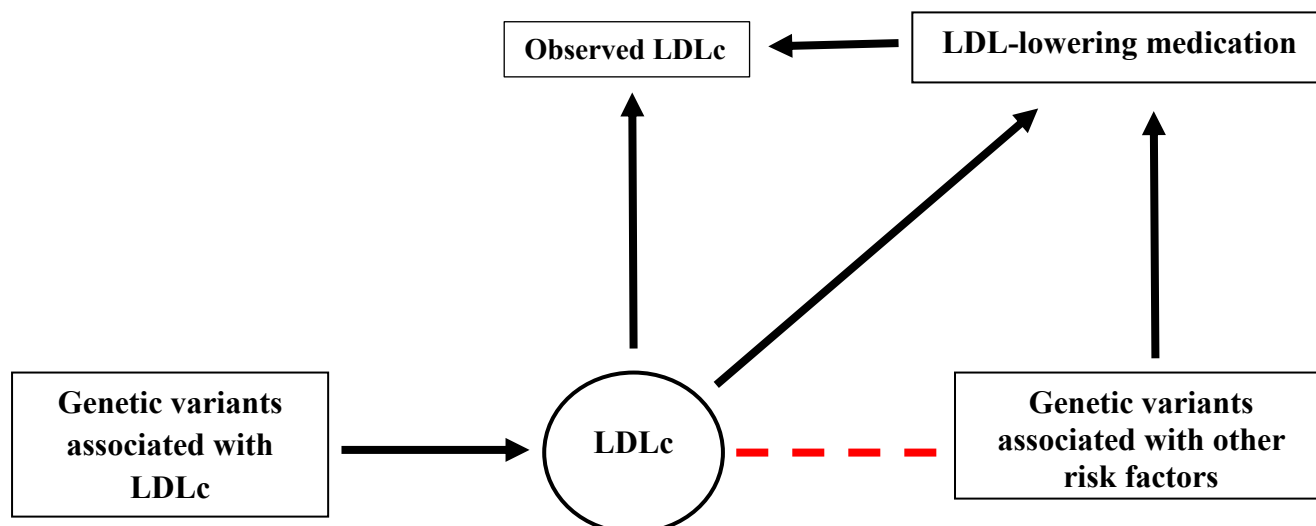


Figure 1. Directed acyclic graph (DAG) illustrating how selection of samples based on LDL-lowering medication use can introduce collider bias (indicated by the dashed red line), as conditioning on or adjusting for medication use can induce an association between non-LDLc genetic factors and LDLc. This is because non-genetic factors that influence LDLc will become related to genetic variants when adjusted and stratified for these non-genetic factors or medication status. Similarly, using a complete sample, including both users and non-users of medication, effectively introduces the same collider bias path by lowering LDLc levels amongst medication users only.

## Methods

### GEMRA

We developed the genetic empirical medication reduction adjustment (GEMRA) approach to adjust the observed LDLc phenotype value in participants taking statins and restoring it to an LDLc value that is more reflective of the underlying, unmedicated value. This approach applies medication reduction adjustments (i.e., absolute and relative) to the observed phenotypic values of medicated participants. Genome-wide association analyses is conducted using unmedicated and medicated participants adjusted for medication status for each model, and  $R^2$  is then calculated as a measure of goodness of fit for each independent, genome-wide SNP identified in the model.  $R^2$  is calculated by: 1) Calculating the t-statistic using the equation,  $t_{stat} = \beta^2 / \beta_{SE}^2$ , where  $\beta$  is the beta effect estimate of the SNP and  $\beta_{SE}$  is the standard error of the beta effect estimate of the SNP; 2) Then calculating  $R^2$  using the equation,  $R^2 = t_{stat}^2 / (t_{stat}^2 + (n+1))$ , where n is the sample size. The model is determined to be the ‘best’ model for adjusting for the effect of statins on LDLc if the summed  $R^2$  value is the highest.

### Simulation study

We implemented a simulation study on 100,000 individuals with genetic effects associated with LDLc ( $G_{\text{direct}}$ ) and genetic effects associated with non-LDLc factors ( $G_{\text{indirect}}$ ). The LDLc phenotype distribution was modelled on measured LDLc values in UK Biobank (UKB) with a phenotypic mean of 3.55 mmol/L and a standard deviation of 0.87 mmol/L. Furthermore, we applied a medication use prevalence of 14%, as previously reported in UKB(27), and a medication effect of 0.8 (i.e., a relative LDL-lowering medication reduction of 20%). We then repeated the simulations 1,000 times with parameters of the direct and indirect genetic effects on LDLc and the beta effect estimate for LDLc resampled in each simulation.

For each simulation, we applied five approaches: 1) True effect; 2) Unadjusted observation approach (No medication status adjustment); 3) Medication only approach (Restricted to individuals taking statin medication); 4) Medication excluded approach (Restricted to individuals not taking any statin medication); 5) Approximate medication adjustment approach (GEMRA).

### Empirical analysis: UK Biobank and Ugandan Genome Resource

#### *Cohort profiles*

UKB is a population-based health research resource consisting of approximately 500,000 people, aged between 38 years and 73 years, who were recruited between the years 2006 and 2010 from across the UK(28). Particularly focused on identifying determinants of human diseases in middle aged and older individuals, participants provided a range of information (such as demographics, health status, lifestyle measures, cognitive testing, personality self-report, and physical and mental health measures) via questionnaires and interviews; anthropometric measures, blood pressure readings and samples of blood, urine and saliva were also taken (data available at [www.ukbiobank.ac.uk](http://www.ukbiobank.ac.uk)). A full description of the study design, participants and quality control (QC) methods have been described in detail previously(29). UK Biobank received ethical approval from the Research Ethics Committee (REC reference for UK Biobank is 11/NW/0382).

The Ugandan Genome Resource (UGR) is a population-based cohort which comprises of approximately 22,000 individuals from 12 ethno-linguistic groups recruited from rural south-western Uganda(30). The initial purpose of the cohort was to study human immunodeficiency viruses in the general population. However, since 2011, this cohort has extended its scope to also focus on understanding the epidemiology and genetics of communicable and non-communicable diseases with a wide range of phenotypic data collected (such as anthropometric measurements, demographic and lifestyle measurements through self-reported questionnaires, as well as blood specimens)(30). The Ugandan Genome Resource was approved by the Science and Ethics Committee of the UVRI Research and Ethics Committee (UVRI-REC #HS 1978), the Uganda National Council for Science

and Technology (UNCST #SS 4283), and the East of England-Cambridge South (formerly Cambridgeshire 4) NHS Research Ethics Committee UK.

#### *Genotyping, imputation and quality control*

A full description of the genotype generation, quality control and imputation methods have been previously reported in detail for UKB(31) and UGR(30).

In brief, for UKB, 488,377 participants were successfully genotyped with 438,427 participants genotyped using the Affymetrix UK Biobank Axiom Array and 49,950 participants genotyped using the UK BiLEVE Axiom Array. Multiallelic single-nucleotide polymorphisms (SNPs), SNPs with a minor allele frequency less than or equal to 1%, and samples that were classified as outliers for missing rate and heterozygosity were removed prior to phasing. Phasing was conducted using a modified version of the SHAPEIT2 algorithm(32), and genotype imputation to a reference set consisting of the Haplotype Reference Consortium reference panel merged with the UK10K and 1000 Genomes phase 3 reference panels(33) was performed using IMPUTE2 algorithms(34).

For UGR, the 3.5M Illumina chip array was used to genotype 5,000 participants at the Wellcome Trust Sanger Institute. Pre-phasing protocols removed samples due to sex mismatch, high relatedness, call rate, and heterozygosity. Phasing was implemented using SHAPEIT2(35). Genotype imputation was conducted using IMPUTE2(36) with a reference set combining Uganda 2000 Genomes reference panel and the 1000 Genomes phase 3 reference panel(30).

#### *Phenotype data*

In UKB, venous blood samples were collected from participants in a non-fasting state at their baseline clinical visit between 2006 and 2010. Circulating LDLc was directly measured using an enzymatic selective protection method on a Beckman Coulter AU5800. Further details of the phenotype collection and analysis have been previously reported ([https://biobank.ndph.ox.ac.uk/showcase/ukb/docs/serum\\_biochemistry.pdf](https://biobank.ndph.ox.ac.uk/showcase/ukb/docs/serum_biochemistry.pdf)). In addition, medication data (Data Field: 20003; Data coding 4) were collected through verbal interviews conducted by trained members of staff during the baseline assessment visit and subsequently during follow-up visits. Duration, dosage and compliance of medication treatments were not recorded.

In UGR, participants provided venous blood samples in a non-fasting state with circulating LDLc measured using homogenous enzymatic colorimetric assays on a Cobas Integra 400 Plus Chemistry analyser (Roche Diagnostics)(37). Protocol describing the biochemical analyses has been previously reported elsewhere(38, 39). Medication data for UGR was obtained from participants during enrolment through an interview. The participants would be asked if they had taken any medication for high cholesterol as prescribed by a professional health worker in the last two weeks. The responses would then be coded as per study interview codes.

## Statistical analyses

GWAS were performed in UGR and all ancestries in UKB using an ancestry-specific approach. GWAS was conducted using a linear mixed model association method in BOLT-LMM(v2.3)(40) for participants of European ancestry in UKB and GENESIS(v3.17)(41) for all other ancestries and cohorts. All analyses were conducted using an additive genetic model and were adjusted for age (defined as the year of the survey minus participant year of birth), sex, and the first 20 principal components. Analyses were limited to genetic variants that were genotyped and imputed with a minor allele frequency >0.1% and an INFO score >0.8.

To evaluate the performance of each model, we firstly filtered the genetic variants in the GWAS output using the genome-wide threshold ( $p < 5 \times 10^{-8}$ ) and conducted linkage disequilibrium (LD) clumping ( $r^2 < 0.001$ ) to identify independent genetic variants for each model using PLINK(v1.90) ([www.cog-genomics.org/plink/1.9/](http://www.cog-genomics.org/plink/1.9/))(42). We then determined the best model by calculating the  $R^2$  for each SNP as a metric for goodness of fit for the model. We then summed the  $R^2$  values for each model to evaluate which model was the ‘best’ by identifying the model with the highest summed  $R^2$  value.

## **Results**

### Simulation study

We conducted 1,000 simulations to evaluate the direct and indirect genetic effects associated with LDLc. We used 100,000 individuals with simulated phenotypic data based on the LDLc distribution and prevalence of LDL-lowering medication use in UKB, and a medication effect of 0.8 which corresponds to a relative LDL-lowering medication reduction of 20%. In each simulation, we applied and compared five approaches: 1) True effect; 2) Unadjusted observational approach, which did not adjustment for medication status; 3) Medication only approach, which was restricted on individuals taking statin medication; 4) Medication excluded approach, which was restricted to individuals not taking any statin medication; and 5) Approximate medication adjustment approach using the GEMRA method.

For the direct effects, we observed stronger attenuated genetic effect estimates for the medication-only and observational methods compared to the methods adjusting for and excluding medication status (Figure 2). The genetic effect estimates for the methods excluding medication status and adjusting for medication status (GEMRA) were strengthened and closer to the true effect estimates (Figure 2). Furthermore, with the exception of the medication-only group, power across the methods were consistent ( $p \leq 7.1 \times 10^{-190}$ ).

For the indirect effects, the GEMRA method to approximately adjust for medication showed low false positive rate ( $p > 0.01$ ) similar to the true effect estimate results in comparison to the other methods (Figure 2). In addition, we observed that collider bias induced a negative association between the

unadjusted LDLc phenotype and non-LDLc genetic factors for the observational, medication-only and medication-excluded methods.

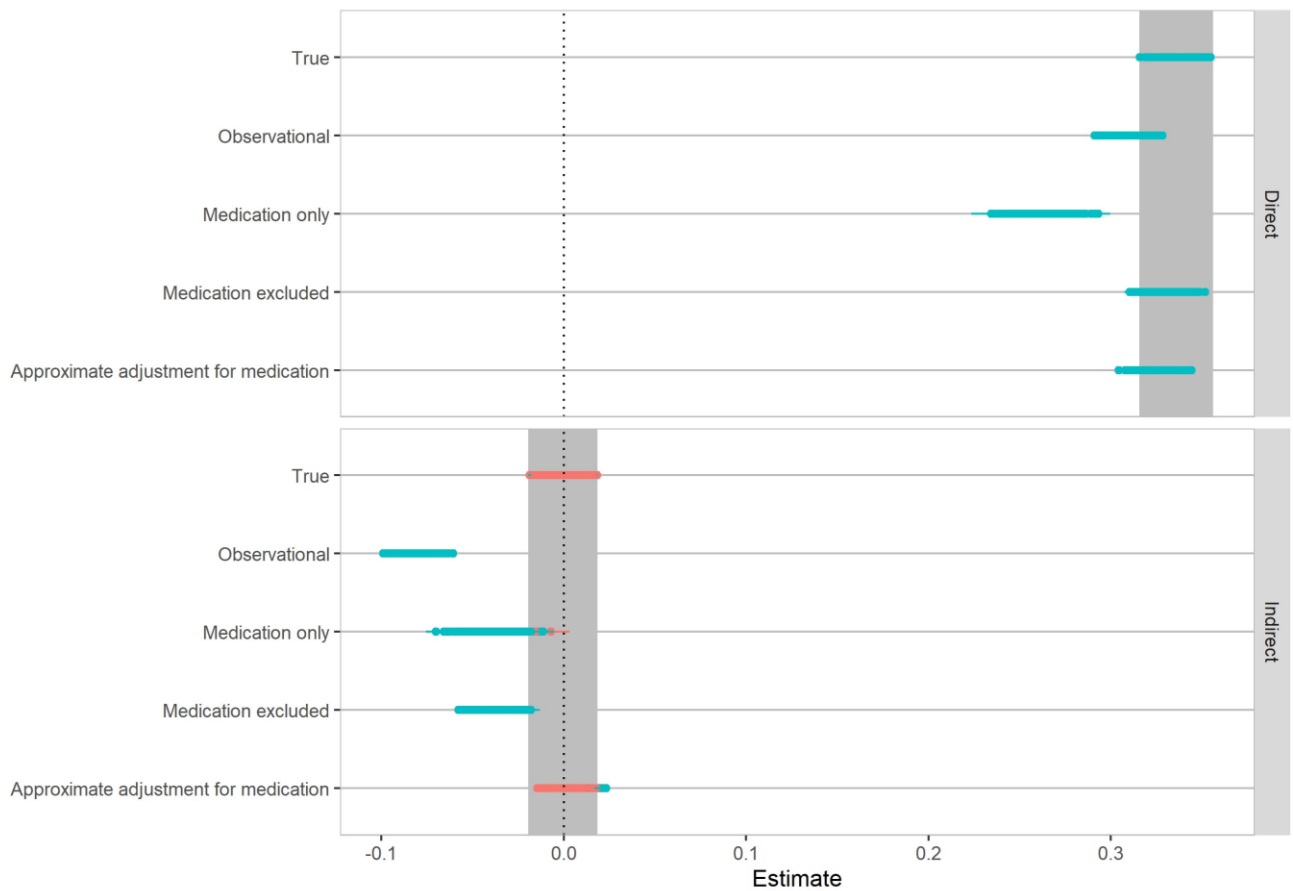


Figure 2. A comparison of the true genetic effect estimate precisely adjusting for medication status compared to the genetic effect estimates of the direct and indirect genotype-LDLc association when: (1) Medication status is not accounted for (Observational); (2) Restricted to individuals taking medication (Medication-only); (3) Restricted to individuals without medication (Medication-excluded); (4) Medication use is adjusted for (Approximate adjustment for medication). For the indirect effect, the blue points denote a  $p$ -value threshold  $< 0.01$  indicating false-positive results and the red points indicate a  $p$ -value threshold  $> 0.01$  indicating no evidence of false-positive results.

#### Empirical analysis: UK Biobank and Ugandan Genome Resource

We then applied these approaches to real-world data from two data sources (UKB and UGR) to evaluate the best model in each population. This was determined by the highest summed  $R^2$  value when applying different thresholds to adjust for medication reduction in observed LDL values. Descriptive characteristics of LDLc and statin use in UKB across ancestries and UGR are shown in Table 1.



Table 1. Description of LDL cholesterol and statin use across different ancestries in UK Biobank and Ugandan Genome Resource.

Data source	Ancestry	LDLc (Median (IQR)) mmol/L	Statin use (Medicated/Unmedicated; %)
UKB	African	3.2 (2.7-3.8)	573/3,231; 18%
	East Asian	3.4 (2.9-3.9)	171/1,415; 12%
	European	3.5 (2.9-4.1)	31,547/219,151; 14%
	South Asian	3.3 (2.7-3.9)	273/1,057; 26%
UGR	Ugandan	2.0 (1.5-2.5)	24/6,144; 0.4%

For the relative medication adjustment approach, findings were consistent amongst all populations with the highest summed  $R^2$  value at the threshold adjusting for either 30% or 40% reduction in observed LDL due to medication (Figure 3A). Overall, the distribution in summed  $R^2$  values were similar across populations.

For the absolute medication adjustment approach, the best models were similar amongst the East Asian, European and South Asian populations with the summed  $R^2$  values increasing until they peaked at either the 1.6 or 1.8 mmol/L adjustment threshold (Figure 3B). In comparison, the African population showed an undulating distribution in the summed  $R^2$  values with the best model at the 1.2 mmol/L adjustment threshold (Figure 3B). Furthermore, the Ugandan population showed a decreasing distribution in summed  $R^2$  values with the best model at the 0.3 mmol/L adjustment threshold (Figure 3B).

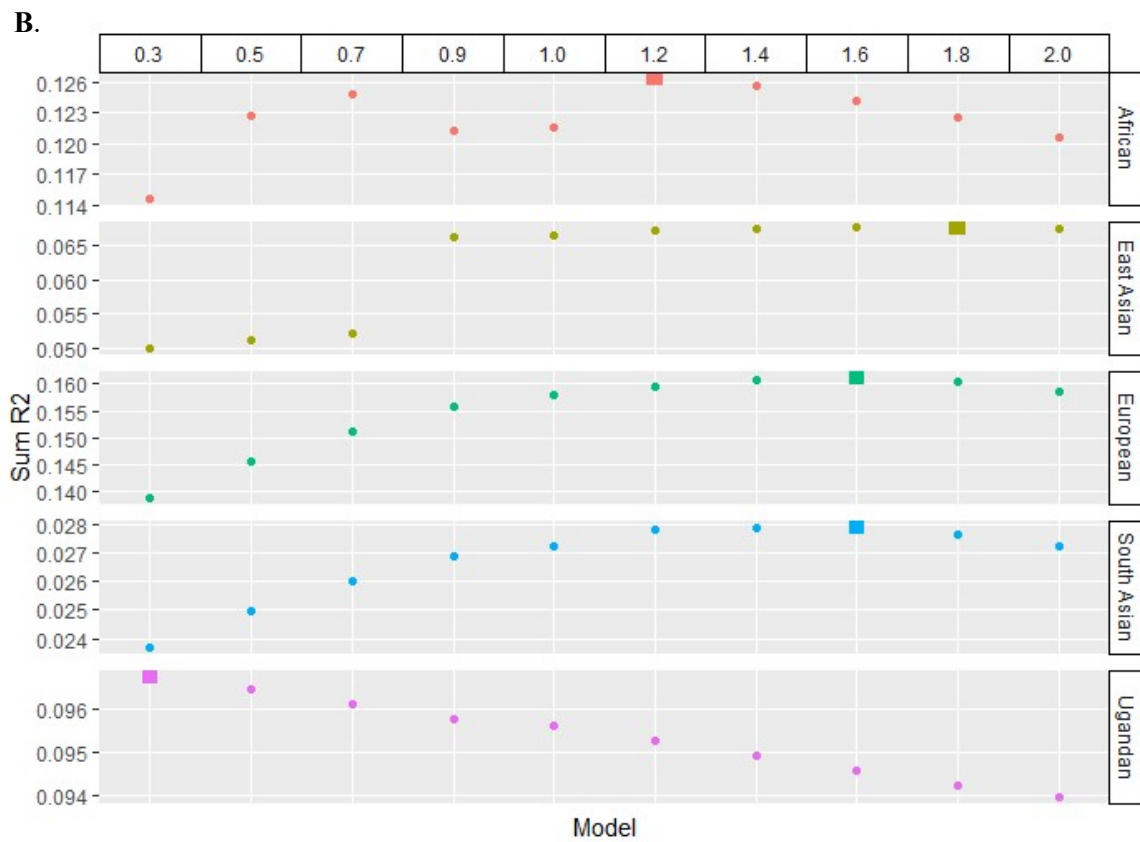
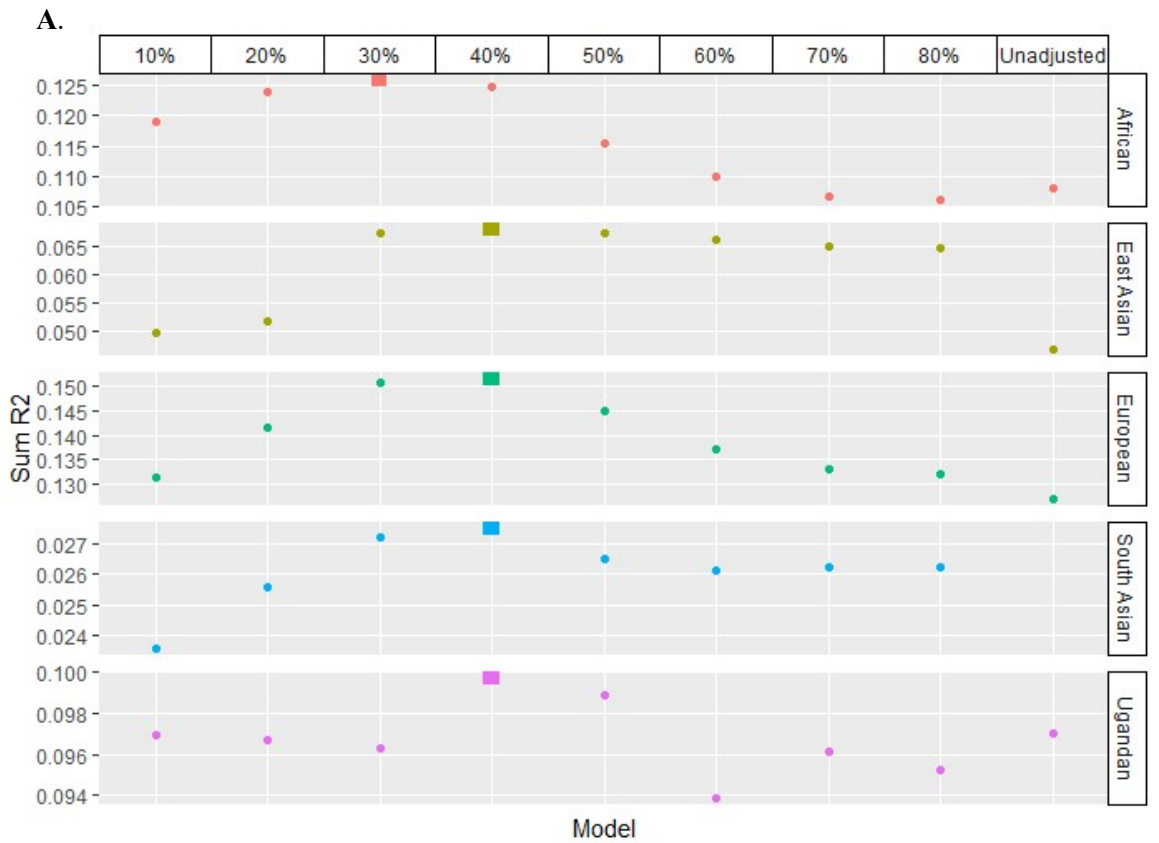


Figure 3. Comparison of relative (A) and absolute (B) adjustment of statins (mmol/L) in different populations using the GEMRA approach. In UKB (i.e., African, East Asian, European, and South Asian) and UGR (i.e., Ugandan), for each ancestry, the points represent the summed  $R^2$  value for each model, with the large square highlighting the model with the highest summed  $R^2$  value.

## Discussion

In this study, we aimed to evaluate the effects of adjusting for medication status when conducting genome-wide association analyses using the GEMRA approach. When performing simulations, our findings suggested that the direct effects of adjusting for medication use and excluding medication use strengthened the effect estimates compared to the unadjusted, observational approach and were closer to the true effect estimates. Furthermore, our simulations illustrated that the indirect effects of approximately adjusting for medication status resulted in low false positive results (similar to the true effect) and results for this method were shown to, in part, resolve the collider bias issue, although some bias remains.

In addition, we used real-world data to evaluate the best model (determined by the largest summed  $R^2$  value) when applying different thresholds to adjust observed LDLc values affected by medication use across different populations. Overall, the best model for the relative adjustment approach was consistent across all populations (30-40% threshold). In contrast, the best model in the absolute adjustment approach was less consistent with only the East Asian, European and South Asian populations showing similar thresholds as the best model (1.6-1.8 mmol/L threshold). For the African and Ugandan population, the best models were attenuated with an adjustment threshold of 1.2 mmol/L and 0.3 mmol/L, respectively.

Difference in medication adjustments required across populations may be a result of the difference in age distribution of the individuals. This could have contributed to the heterogeneity in the absolute best model thresholds as individuals of European descent in UKB had a median age of 72 years (Interquartile range (IQR): 64-77), whilst individuals in the UGR cohort had a median age of 30 years (IQR: 18-46). As a result, the mean observed LDLc would be low due to the age distribution of the UGR cohort being younger. Therefore, a 40% reduction in medication effect of a small value is a smaller absolute change, compared to a 40% reduction of a larger value. This difference can also be seen when comparing the number of individuals taking LDLc-lowering medication (UKB EUR: 14%; UGR: <1%) (Table 1).

In addition, the type of LDLc-lowering therapy prescribed to individuals may have contributed to the heterogeneity of best models. Randomized controlled trials (RCTs) have been conducted to evaluate the efficacy of these drug regimens with studies showing varying degrees of serum LDLc reduction(43-48). In a meta-analysis of RCTs, statin treatment was associated with a 28% reduction in LDLc levels(43) which is similar to the anticipated 30-50% reduction in LDLc levels for moderate-

intensity statin therapy outlined by the 2018 US blood cholesterol treatment guidelines(49). This is also comparable to the best model thresholds identified in the relative approach across populations. Furthermore, when statin therapy is used in combination with other non-statin drugs (e.g., Ezetimibe, Evacetrapib or Evolocumab), RCTs showed a decrease in LDLc levels between 0.9-2.4 mmol/L(46-48), which is within the range of best model thresholds identified when applying the absolute approach to adjust for medication status, with the exception of the Ugandan population finding. However, as UKB and UGR did not report the dose, duration of treatment and level of compliance towards the statin treatment, the GEMRA approach is liable to over- or under adjusting for the stain effect on LDLc without this additional context. As a result, some heterogeneity can be introduced into the analyses.

In addition, this study was limited by the sample size of the data sources applied in the empirical analysis. The non-European demographic groups in UKB were limited by their sample size compared to the European group (Non-European:  $N < 6,200$ ; Europeans:  $N = 219,151$ ). As a result, the statistical power to detect genetic variants associated with LDLc in these groups, and subsequently to evaluate the best model using the summed  $R^2$  values, may have affected the precision of the findings. This can be observed when using absolute values to adjust for statin use (Figure 3B), which suggest that the average statin reduces LDLc to a greater extent in East Asians than in Europeans. However, this finding is likely a result of imprecision due to large confidence intervals surrounding the  $R^2$  values, as the mean population levels of LDLc tend to be notably lower in East Asian populations than in European populations.

Overall, our study has illustrated the technical challenges that can arise when analysing phenotypes altered by medication status in genome-wide association analyses, and we have proposed the GEMRA approach to approximately adjust the observed values using LDLc and statin use. Findings using this approach suggest that statins have a relative effect of 40% on LDLc values which is consistent across ancestries.

## **Acknowledgements**

Quality control filtering of the UK Biobank data was conducted by R.Mitchell, G.Hemani, T.Dudding, L.Corbin, S.Harrison, L.Paternoster as described in the published protocol (doi: 10.5523/bris.1ovaaau5sxunp2cv8rcy88688v).

## **Funding**

This research has been conducted using the UK Biobank Resource under Application No 15825. This project was supported by the UK Medical Research Council (MRC) as part of a network grant awarded to the MRC Integrative Epidemiology Unit at the University of Bristol (MC\_UU\_00011/4, MC\_UU\_00011/1, MC\_UU\_00032/01, MC\_UU\_00032/03), core funding to the MRC/Uganda Virus

Research Institute and London School of Hygiene & Tropical Medicine Uganda Research Unit under the MRC/UK Department of International Development (DFID) Concordat agreement with Segun Fatumo supported by the Wellcome Trust (220740/Z/20/Z), and the MRC Population Health Research Unit at the University of Oxford (MC\_UU\_00017/2).

This study was also supported by the NIHR Biomedical Research Centre at the University Hospitals Bristol and Weston NHS Foundation Trust and the University of Bristol. This publication is the work of the authors who will serve as guarantors for the contents of this paper. The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the National Institute for Health Research.

### **Competing interests**

The authors declare that they have no conflict of interest.

### **Data availability**

All scripts used to perform these analyses are available at [https://github.com/MRCIEU/bp-drug-mr/tree/main/analysis/gwas/ldlc\\_adj\\_med\\_analysis](https://github.com/MRCIEU/bp-drug-mr/tree/main/analysis/gwas/ldlc_adj_med_analysis).

### **References**

1. Cole SR, Platt RW, Schisterman EF, Chu H, Westreich D, Richardson D, et al. Illustrating bias due to conditioning on a collider. *International journal of epidemiology*. 2010;39(2):417-20.
2. Greenland S. Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology*. 2003:300-6.
3. Griffith GJ, Morris TT, Tudball MJ, Herbert A, Mancano G, Pike L, et al. Collider bias undermines our understanding of COVID-19 disease risk and severity. *Nature communications*. 2020;11(1):5749.
4. Bell JA, Richardson TG, Wang Q, Sanderson E, Palmer T, Walker V, et al. Effects of general and central adiposity on circulating lipoprotein, lipid, and metabolite levels in UK Biobank: a multivariable Mendelian randomization study. *The Lancet Regional Health–Europe*. 2022;21.
5. Fang S, Holmes MV, Gaunt TR, Smith GD, Richardson TG. Constructing an atlas of associations between polygenic scores from across the human phenome and circulating metabolic biomarkers. *Elife*. 2022;11:e73951.
6. Tobin MD, Sheehan NA, Scurrah KJ, Burton PR. Adjusting for treatment effects in studies of quantitative traits: antihypertensive therapy and systolic blood pressure. *Statistics in medicine*. 2005;24(19):2911-35.

7. Norris T, Cole TJ, Bann D, Hamer M, Hardy R, Li L, et al. Duration of obesity exposure between ages 10 and 40 years and its relationship with cardiometabolic disease risk factors: a cohort study. *PLoS medicine*. 2020;17(12):e1003387.
8. Stitzel NO, Khera AV, Wang X, Bierhals AJ, Vourakis AC, Sperry AE, et al. ANGPTL3 deficiency and protection against coronary artery disease. *Journal of the American College of Cardiology*. 2017;69(16):2054-63.
9. Pinto Pereira SM, Ki M, Power C. Sedentary behaviour and biomarkers for cardiovascular disease and diabetes in mid-life: the role of television-viewing and sitting at work. *PloS one*. 2012;7(2):e31132.
10. Pencina KM, Thanassoulis G, Wilkins JT, Vasan RS, Navar AM, Peterson ED, et al. Trajectories of non-HDL cholesterol across midlife: implications for cardiovascular prevention. *Journal of the American College of Cardiology*. 2019;74(1):70-9.
11. Friedewald WT, Levy RI, Fredrickson DS. Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. *Clinical chemistry*. 1972;18(6):499-502.
12. Pencina MJ, Navar AM, Wojdyla D, Sanchez RJ, Khan I, Ellassal J, et al. Quantifying importance of major risk factors for coronary heart disease. *Circulation*. 2019;139(13):1603-11.
13. Bennett WL, Maruthur NM, Singh S, Segal JB, Wilson LM, Chatterjee R, et al. Comparative effectiveness and safety of medications for type 2 diabetes: an update including new drugs and 2-drug combinations. *Annals of internal medicine*. 2011;154(9):602-13.
14. White IR, Koupilova I, Carpenter J. The use of regression models for medians when observed outcomes may be modified by interventions. *Statistics in medicine*. 2003;22(7):1083-96.
15. Kullo IJ, Ding K, Jouni H, Smith CY, Chute CG. A genome-wide association study of red blood cell traits using the electronic medical record. *PloS one*. 2010;5(9):e13011.
16. Kim YJ, Go MJ, Hu C, Hong CB, Kim YK, Lee JY, et al. Large-scale genome-wide association studies in East Asians identify new genetic loci influencing metabolic traits. *Nature genetics*. 2011;43(10):990-5.
17. Cho YS, Go MJ, Kim YJ, Heo JY, Oh JH, Ban H-J, et al. A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. *Nature genetics*. 2009;41(5):527-34.
18. Wallace C, Newhouse SJ, Braund P, Zhang F, Tobin M, Falchi M, et al. Genome-wide association study identifies genes for biomarkers of cardiovascular disease: serum urate and dyslipidemia. *The American Journal of Human Genetics*. 2008;82(1):139-49.

19. Sabatti C, Service SK, Hartikainen A-L, Pouta A, Ripatti S, Brodsky J, et al. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nature genetics*. 2009;41(1):35-46.
20. Rasmussen-Torvik LJ, Pacheco JA, Wilke RA, Thompson WK, Ritchie MD, Kho AN, et al. High density GWAS for LDL cholesterol in African Americans using electronic medical records reveals a strong protective variant in APOE. *Clinical and translational science*. 2012;5(5):394-9.
21. Kamatani Y, Matsuda K, Okada Y, Kubo M, Hosono N, Daigo Y, et al. Genome-wide association study of hematological and biochemical traits in a Japanese population. *Nature genetics*. 2010;42(3):210-5.
22. Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, Clarke R, et al. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nature genetics*. 2008;40(2):161-9.
23. Kathiresan S, Melander O, Guiducci C, Surti A, Burt NP, Rieder MJ, et al. Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nature genetics*. 2008;40(2):189-97.
24. Munafò MR, Tilling K, Taylor AE, Evans DM, Davey Smith G. Collider scope: when selection bias can substantially influence observed associations. *International journal of epidemiology*. 2018;47(1):226-35.
25. Wu J, Province MA, Coon H, Hunt SC, Eckfeldt JH, Arnett DK, et al. An investigation of the effects of lipid-lowering medications: genome-wide linkage analysis of lipids in the HyperGEN study. *BMC genetics*. 2007;8(1):1-9.
26. Hong EP, Park JW. Sample size and statistical power calculation in genetic association studies. *Genomics & informatics*. 2012;10(2):117.
27. Mitchell RE, Hartley AE, Walker VM, Gkatzionis A, Yarmolinsky J, Bell JA, et al. Strategies to investigate and mitigate collider bias in genetic and Mendelian randomisation studies of disease progression. *PLoS Genetics*. 2023;19(2):e1010596.
28. Allen NE, Sudlow C, Peakman T, Collins R, biobank U. UK biobank data: come and get it. *American Association for the Advancement of Science*; 2014. p. 224ed4-ed4.
29. Collins R. What makes UK Biobank special? *The Lancet*. 2012;379(9822):1173-4.
30. Fatumo S, Mugisha J, Soremekun OS, Kalungi A, Mayanja R, Kintu C, et al. Uganda Genome Resource: A rich research database for genomic studies of communicable and non-communicable diseases in Africa. *Cell Genomics*. 2022;2(11).
31. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*. 2018;562(7726):203-9.
32. O'Connell J, Sharp K, Shrine N, Wain L, Hall I, Tobin M, et al. Haplotype estimation for biobank-scale data sets. *Nature genetics*. 2016;48(7):817-20.

33. Huang J, Howie B, McCarthy S, Memari Y, Walter K, Min JL, et al. Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nature communications*. 2015;6(1):8111.
34. Howie B, Marchini J, Stephens M. Genotype imputation with thousands of genomes. *G3: Genes| Genomes| Genetics*. 2011;1(6):457-70.
35. Delaneau O, Coulonges C, Zagury J-F. Shape-IT: new rapid and accurate algorithm for haplotype inference. *BMC bioinformatics*. 2008;9(1):1-14.
36. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics*. 2009;5(6):e1000529.
37. Gurdasani D, Carstensen T, Fatumo S, Chen G, Franklin CS, Prado-Martinez J, et al. Uganda genome resource enables insights into population history and genomic discovery in Africa. *Cell*. 2019;179(4):984-1002. e36.
38. Sugiuchi H, Uji Y, Okabe H, Irie T, Uekama K, Kayahara N, et al. Direct measurement of high-density lipoprotein cholesterol in serum with polyethylene glycol-modified enzymes and sulfated alpha-cyclodextrin. *Clinical chemistry*. 1995;41(5):717-23.
39. Sugiuchi H, Irie T, Uji Y, Ueno T, Chaen T, Uekama K, et al. Homogeneous assay for measuring low-density lipoprotein cholesterol in serum with triblock copolymer and  $\alpha$ -cyclodextrin sulfate. *Clinical chemistry*. 1998;44(3):522-31.
40. Loh P-R, Tucker G, Bulik-Sullivan BK, Vilhjálmsson BJ, Finucane HK, Salem RM, et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature genetics*. 2015;47(3):284-90.
41. Gogarten SM, Sofer T, Chen H, Yu C, Brody JA, Thornton TA, et al. Genetic association testing using the GENESIS R/Bioconductor package. *Bioinformatics*. 2019;35(24):5346-8.
42. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*. 2015;4(1):s13742-015-0047-8.
43. LaRosa JC, He J, Vupputuri S. Effect of statins on risk of coronary disease: a meta-analysis of randomized controlled trials. *Jama*. 1999;282(24):2340-6.
44. Naci H, Bruggs JJ, Fleurence R, Ades AE. Dose-comparative effects of different statins on serum lipid levels: a network meta-analysis of 256,827 individuals in 181 randomized controlled trials. *European journal of preventive cardiology*. 2013;20(4):658-70.
45. Josan K, Majumdar SR, McAlister FA. The efficacy and safety of intensive statin therapy: a meta-analysis of randomized trials. *Cmaj*. 2008;178(5):576-84.
46. Robinson JG, Nedergaard BS, Rogers WJ, Fialkow J, Neutel JM, Ramstad D, et al. Effect of evolocumab or ezetimibe added to moderate-or high-intensity statin therapy on LDL-C lowering in patients with hypercholesterolemia: the LAPLACE-2 randomized clinical trial. *Jama*. 2014;311(18):1870-83.



47. Nicholls SJ, Brewer HB, Kastelein JJ, Krueger KA, Wang M-D, Shao M, et al. Effects of the CETP inhibitor evacetrapib administered as monotherapy or in combination with statins on HDL and LDL cholesterol: a randomized controlled trial. *Jama*. 2011;306(19):2099-109.
48. Nicholls SJ, Puri R, Anderson T, Ballantyne CM, Cho L, Kastelein JJ, et al. Effect of evolocumab on progression of coronary disease in statin-treated patients: the GLAGOV randomized clinical trial. *Jama*. 2016;316(22):2373-84.
49. Grundy SM, Stone NJ, Bailey AL, Beam C, Birtcher KK, Blumenthal RS, et al. 2018 AHA/ACC/AACVPR/AAPA/ABC/ACPM/ADA/AGS/APhA/ASPC/NLA/PCNA guideline on the management of blood cholesterol: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Circulation*. 2019;139(25):e1082-e143.