Université de Montréal

La génétique humaine pour l'étude de cibles pharmacologiques

Par
Marc-André Legault

Département de biochimie et médecine moléculaire, Faculté de Médecine

Thèse présentée en vue de l'obtention du grade de Philosophiæ doctor (Ph.D.)
en Bio-informatique

Mars 2021

Université de Montréal

Département de biochimie et médecine moléculaire, Faculté de Médecine

---

Cette thèse intitulée

La génétique humaine pour la validation de cibles pharmacologiques

Présentée par
Marc-André Legault

A été évaluée par un jury composé des personnes suivantes

Pavel Hamet
Président Rapporteur

Marie-Pierre Dubé
Directrice de recherche

Sébastien Lemieux
Codirecteur

Alexandre Bureau
Membre du jury

Brent Richards
Examinateur externe

Gregor Andelfinger
Représentant du Doyen

Résumé

En étudiant les variations génétiques au sein d'une population, il est possible d'identifier des polymorphismes génétiques qui confèrent une protection naturelle contre la maladie. Si l'on parvient à comprendre le mécanisme moléculaire qui sous-tend cette protection, par exemple en reliant la variation génétique à la perturbation d'une protéine bien précise, il pourrait être possible de développer des thérapies pharmacologiques qui agissent sur la même cible biologique. Cette relation entre les médicaments et les variations génétiques est une des prémisses centrales de la validation génétique de cibles pharmacologiques qui est un facteur de réussite dans le développement de médicaments.

Dans cette thèse, nous utiliserons un modèle génétique pour prédire les effets bénéfiques et indésirables de l'ivabradine, un médicament utilisé afin de réduire la fréquence cardiaque. L'ivabradine est un inhibiteur du canal ionique *potassium/sodium hyperpolarization-activated cyclic nucleotide-gated channel 4*, encodé par le gène *HCN4*, dont les bénéfices sont hétérogènes chez différentes populations de patients. Ce médicament est efficace pour le traitement de l'angine et de l'insuffisance cardiaque, mais s'est avéré inefficace en prévention secondaire chez des patients coronariens stables sans dysfonction systolique. La caractérisation des effets de l'ivabradine s'est échelonnée sur une période de 6 ans et trois grands essais de phase III ont été menées. Nous étudierons la possibilité d'avoir prédit ou accéléré ce processus à l'aide de modèles génétiques et nous contrasterons les effets spécifiques à l'ivabradine des effets généraux de la réduction de la fréquence cardiaque par une approche de randomisation mendélienne.

Deuxièmement, une approche génétique sera utilisée pour évaluer l'effet de l'inhibition de la *cholesteryl ester tranfer protein* (CETP), une enzyme responsable du transfert des cholestérols estérifiés et des triglycérides entre différentes lipoprotéines ainsi qu'une cible pharmacologique largement étudiée pour le traitement de la maladie coronarienne. Les études génétiques prédisent un bénéfice à l'inhibition de CETP, mais les essais randomisés ont eu des résultats hétérogènes et décevants. Nous utiliserons un modèle génétique d'inhibition de la CETP pour identifier des variables qui peuvent moduler l'effet de l'inhibition de la

CETP sur des biomarqueurs et la maladie ischémique. Les biomarqueurs pris en compte comprennent les taux de cholestérol à lipoprotéines de basse et haute densité, mais aussi la capacité du plasma à absorber le cholestérol, une mesure fonctionnelle importante et sous-étudiée. Le sexe et l'indice de masse corporelle se sont avérés être deux variables qui modifient fortement les effets d'une réduction génétiquement prédite de la concentration de CETP sur les paramètres étudiés. Notre modèle prédit un bénéfice plus important de l'inhibition de la CETP pour les femmes et les individus ayant un indice de masse corporelle normal sur le profil lipidique, mais nous n'avons pas pu démontrer une modulation de l'effet sur la maladie ischémique. Cette étude reste importante sur le plan méthodologique, car elle soulève la possibilité d'utiliser des modèles génétiques de cibles pharmacologiques pour prédire l'hétérogénéité dans la réponse au médicament, une lacune des essais randomisés classiques.

Enfin, nous avons adopté une approche centrée sur les gènes pour caractériser l'effet de 19 114 protéines humaines sur 1 210 phénotypes de la UK Biobank. Les résultats de cette étude sont accessibles au public (exphewas.statgen.org) et constituent une ressource précieuse pour cerner rapidement les conséquences phénotypiques associées à un locus. Dans le contexte de validation de cibles pharmacologiques, cette plate-forme web peut aider à rapidement identifier les problèmes de sécurité potentiels ou à découvrir des possibilités de repositionnement du médicament. Un exemple d'utilisation de cette plate-forme est présenté où nous identifions le gène de la myotiline comme un nouvel acteur potentiel dans la pathogénèse de la fibrillation auriculaire.

Mots clés : cible pharmacologique, randomisation mendélienne, PheWAS, ivabradine, CETP, fibrillation auriculaire, génétique humaine, maladie coronarienne

Abstract

Using population-level data, it is possible to identify genetic polymorphisms that confer natural protection against disease. If the molecular mechanism underlying this protection can be understood, for example by linking variants to the disruption of a particular protein, it may be possible to develop drugs that act on the same biological target. This link between drugs and variants is a central premise of genetic drug target validation.

In this work, a genetic model is used to predict the beneficial and adverse effects of ivabradine, a drug used to lower heart rate. Ivabradine is an inhibitor of the ion channel *potassium/sodium hyperpolarization-activated cyclic nucleotide-gated channel 4*, encoded by the *HCN4* gene, with heterogeneous benefits in different patient populations. This drug is effective in the treatment of angina and heart failure but it is ineffective in patients with stable coronary artery disease without systolic dysfunction. Characterization of the effect of ivabradine has occurred over a 6-year period and three large phase III trials have been conducted. We will investigate whether this process could have been streamlined using genetic models and contrast the ivabradine-specific effect with the general effect of heart rate reduction using a Mendelian Randomization approach.

Second, a genetic approach is used to study the effect of inhibiting *cholesteryl ester tranfer protein* (CETP), an enzyme responsible for the transfer of cholestery esters and triglycerides between different lipoproteins and a widely studied drug target for the treatment of coronary artery disease. Genetic studies predict a benefit of CETP inhibition, but randomized trials yielded heterogeneous and disappointing results. We will use a genetic model of CETP inhibition to identify variables that may modulate the effect of CETP inhibition on biomarkers and ischemic disease. The biomarkers we considered included low- and high-density lipoprotein cholesterol levels but also the plasma cholesterol efflux capacity, an important and understudied functional measure of high density lipoproteins. Sex and body mass index strongly modulated the effect of a genetically predicted lower CETP concentration on the lipid profile. Our model predicts a greater benefit of CETP inhibition in women and individuals with normal body mass index on the lipid profile, but these observations

did not translate to changes in the effect on cardiovascular outcomes. This study remains methodologically important because it demonstrates the possibility of using genetic models of drug targets to predict heterogeneity in drug response, a shortcoming of conventional randomized trials.

Finally, we adopted a gene-centric approach to characterize the effect of 19,114 human protein-coding genes on 1,210 UK Biobank phenotypes. The results of this study are publicly available (exphewas.statgen.org) and provide a valuable resource to rapidly screen the phenotypic consequences associated with a gene. In the context of drug target validation, this platform can help quickly identify potential safety issues or discover drug repurposing opportunities. An example of the use of this platform is presented where we identify the myotilin gene as a potential atrial fibrillation gene.

Keywords: drug target, mendelian randomization, PheWAS, ivabradine, CETP, atrial fibrillation, human genetics, coronary artery disease

Table des matières

Liste des figures

Liste des tableaux

Liste des abréviations

ACCELERATE *Assessment of Clinical Effects of Cholesteryl Ester Transfer Protein Inhibition with Evacetrapib in Patients at a High Risk for Vascular Outcomes*

API interface de programmation d'application (*Application Programming Interface*)

apoB apolipoprotéine B

apoE apolipoprotéine E

AVC Accident Vasculaire Cérébral

BEAUTIFUL *morBidity-mortality EvAlUaTion of the I_f inhibitor ivabradine in patients with coronary disease and left-ventricULar dysfunction*

DIAGRAM *DIAbetes Genetics Replication And Meta-analysis Consortium*

FDA *U.S. Food & Drug Administration*

FOURIER *Further Cardiovascular Outcomes Research with PCSK9 Inhibition in Subjects with Elevated Risk*

GoF gain de fonction (*Gain of Function*)

GRS score de risque génétique (*Genetic Risk Score*)

GWAS étude d'association pangénomique (*Genome-Wide Association Study*)

HDL lipoprotéines à haute densité (*High Density Lipoprotein*)

HR rapport des risques instantanés (*Hazard Ratio*)

IC intervalle de confiance

IMC indice de masse corporelle

InSIDE *Instrument Strength Independant of Direct Effect*

IVW estimés par ratio pondérés par l'inverse de leur variance (*Inverse Variance Weighted*)

JUPITER *Justification for the Use of Statins in Prevention: an Intervention Trial Evaluating Rosuvastatin*

LD déséquilibre de liaison (*Linkage Disequilibrium*)

LDL lipoprotéines à faible densité (*Low Density Lipoprotein*)

LoF perte de fonction (*Loss of Function*)

lp(a) lipoprotéine(a)

MR randomisation mendélienne (*Mendelian Randomization*)

OMIM *Online Mendelian Inheritance in Man*

OR rapport des cotes (*Odds Ratio*)

PCA analyse en composantes principales (*Principal Component Analysis*)

PheWAS études d'association panphénomique (*Phenome-Wide Association Study*)

pQTL *locus* associé au niveau protéique (*protein Quantitative Trait Locus*)

PRS score de risque polygénique (*Polygenic Risk Score*)

REVEAL *Evaluation of the Effects of Anacetrapib through Lipid Modification*

SHIFT *Systolic Heart failure treatment with the I_f inhibitor ivabradine Trial*

SIGNIFY *Study assessInG the morbidity-mortality beNefits of the I_f inhibitor ivabradine in patients with coronarY artery disease*

SKAT *SNP-set Kernel Association Test*

SNP polymorphisme d'un seul nucléotide (*Single Nucleotide Polymorphism*)

VLDL lipoprotéines à très faible densité (*Very Low Density Lipoprotein*)

À Flo et Adri

Remerciements

J'aimerais d'abord remercier Dre Marie-Pierre Dubé qui m'a accepté dans son laboratoire de recherche à l'été 2012 alors que j'étais en première année du programme de baccalauréat en bio-informatique à l'Université de Montréal. Au fil de ces 9 années, j'ai développé plusieurs habiletés techniques et scientifiques et j'ai pu participer à une grande diversité de projets. Je la remercie de m'avoir transmis son approche rigoureuse, diligente et intelligente à la science que je porterai tout au long de ma carrière. Toutes ces expériences m'ont donné le goût de poursuivre une carrière dans le monde de la recherche.

Je remercie aussi les membres de mon laboratoire d'accueil de m'avoir supporté, conseillé et aidé pendant tout ce temps. Je remercie Amina pour les longues discussions sur les différentes approches statistiques. Je remercie Louis-Philippe qui a été mon premier mentor à StatGen et un acteur important dans ma formation en bio-informatique. Je remercie aussi Michel, Géraldine, Daniel, Johanna et Audrey. Vous avez fait de notre sous-sol sombre un endroit lumineux et chaleureux.

De l'Institut de Cardiologie, je remercie aussi les collaborateurs de mes projets de recherche en particulier Dr Jean-Claude Tardif, Simon de Denus et Julie Hussin dont les commentaires ont grandement amélioré mes travaux. Merci aussi à Guillaume Lettre pour son implication dans la vie académique de l'Institut de Cardiologie de Montréal et pour avoir siégé sur mon comité de thèse.

Merci à Sébastien Lemieux, mon co-directeur, qui m'a offert un mentorat essentiel et un regard différent quand j'en avais besoin. Il a toujours été très généreux de son temps et son approche agnostique m'a souvent forcé à pousser ma réflexion et à reconsidérer les dogmes.

J'aimerais aussi remercier les membres du comité d'évaluation de ma thèse de me prêter votre temps précieux. Je suis reconnaissant à Sylvie Mader d'avoir présidé mon comité de thèse et à Élaine Meunier pour son aide.

Sur une note plus personnelle, ce travail aurait été impossible sans le support de mes parents Diane Tassé et Michel Legault qui ont toujours valorisé l'éducation, le travail, la

rigueur et le dépassement de soi. Ils m'ont certainement offert un environnement qui m'a permis de compléter mes études doctorales et je leur en suis extrêmement reconnaissant. Merci à mon frère François pour son amour et son appui.

Je remercie de tout cœur ma conjointe Florence pour son support infini, son oreille attentive, ses conseils judicieux et ses encouragements. J'ai l'immense chance de partager mon quotidien avec elle depuis mon adolescence. Je remercie aussi Adrien, mon petit complice, d'embellir mes journées. Je suis très reconnaissant à ma belle famille, à Sophie et Denis de m'avoir encouragé pendant toute cette aventure. Votre aide a souvent été déterminante. Merci à Barbara et Myriam pour votre amitié et votre générosité.

Finalement, j'aimerais souligner le support financier de la faculté de Médecine et de la faculté des Études Supérieures et Postdoctorales de l'Université de Montréal, de l'Institut de Cardiologie de Montréal, des Fonds de Recherche du Québec en Santé et des Institut de Recherche en Santé du Canada.

CHAPITRE 1

Introduction

Les variations génétiques sont à l'origine de notre identité biologique et contribuent à nos caractéristiques observables. La taille, la pression artérielle, la réponse aux médicaments et le risque d'infarctus, par exemple, ont des composantes à la fois environnementales et génétiques. L'étude des variations génétiques qui sous-tendent ces caractéristiques a plusieurs buts, notamment d'améliorer notre compréhension de la base moléculaire de ces caractéristiques et de permettre leur prédiction. Dans cette thèse, nous nous concentrerons sur une application particulière de l'étude des variations génétiques, à savoir leur utilisation pour prédire l'effet de certains médicaments et pour découvrir de nouvelles cibles pharmacologiques.

1.1 La génétique des cibles pharmacologiques

Les médicaments agissent généralement en modulant une cible pharmacologique, soit une enzyme ou un récepteur impliqué dans la physiopathologie de la maladie. Par exemple, les statines, des médicaments utilisés pour abaisser le taux de cholestérol associé aux lipoprotéines à faible densité (*Low Density Lipoprotein*, LDL), inhibent une enzyme responsable d'une réaction limitante de la biosynthèse du cholestérol [1]. Des mutations génétiques de cette enzyme, c'est-à-dire la HMG CoA reductase, entraînent également une réduction de la concentration plasmatique de cholestérol LDL, offrant ainsi une protection naturelle contre la maladie coronarienne aux individus qui en sont porteurs [2, 3]. Cet exemple particulier ne fait pas exception, car le soutien génétique de cibles pharmacologiques est un facteur prédictif de succès dans le développement de médicaments. AstraZeneca a publié une revue de son portfolio des médicaments en développement entre 2005 et 2010, qui comprenaient

des projets de la phase préclinique à la phase II [4]. Le taux d'échec des projets pour lesquels la cible pharmacologique était appuyée par des preuves génétiques était de 27%, contre 57% pour les médicaments qui n'avaient pas de telles données. L'importance d'avoir de bons biomarqueurs d'efficacité et de cibler adéquatement la population de patients susceptibles de bénéficier d'une intervention pharmacologique était aussi évoquée, deux aspects pour lesquels une approche génétique peut être utile [5]. Dans une étude plus large, Nelson *et coll.* ont montré que les gènes codants de cibles de médicaments approuvés sont enrichis parmi les gènes identifiés dans les études d'association pan-génomiques (*Genome-Wide Association Studies*, GWAS) et répertoriés dans la base de données *GWASdb* (rapport des cotes (*Odds Ratio*, OR) = 2.0, $P = 1.3 \times 10^{-14}$). Cet enrichissement était encore plus fort pour les gènes identifiés dans la base de données *Online Mendelian Inheritance in Man* (OMIM) pour leur association avec des maladies mendéliennes (OR = 7.2, $P = 2.9 \times 10^{-10}$). De plus, les indications des médicaments coïncidaient fréquemment avec les traits associés dans les études génétiques de leurs cibles. Cette concordance grandissait à travers les phases du développement de médicaments [6]. Les constats de cette étude ont ajouté à l'engouement existant face à l'utilisation de la génétique humaine pour valider, de façon naturelle, l'hypothèse thérapeutique [7, 8]. Quatre ans plus tard, une étude de réplication a été menée bénéficiant d'un nombre considérable de nouvelles découvertes dans les bases de données d'associations génétiques [9]. La progression à travers les étapes de développement du médicament était encore plus fortement prédite par les gènes associés à des maladies mendéliennes dans la base de données OMIM par rapport à l'étude originale. Pour les gènes appuyés par des associations issues de GWAS, l'effet était grandement atténué, possiblement à cause de l'ajout de variations à effet plus faible issues de GWAS de très grande taille. Ce résultat est compatible avec l'hypothèse de l'architecture omnigénique des traits complexes selon laquelle leur héritabilité est composée à la fois de contributions majeures de quelques gènes *centraux* et d'un grand nombre de très faibles contributions de gènes *périphériques* [10]. Selon ce modèle, les gènes périphériques agissent en perturbant de façon subtile le réseau densément interconnecté d'expression génique de la cellule et leurs effets se propagent jusqu'aux gènes centraux. L'appui d'une cible thérapeutique par une évidence issue d'association génétique est donc important, mais insuffisant en soi pour prédire l'efficacité et la sécurité de la modulation pharmacologique d'une cible thérapeutique.

Cette thèse se concentre sur les approches génétiques permettant d'améliorer la validation de cibles pharmacologiques et d'inférer l'effet causal de la modulation des produits géniques (protéines, enzymes et biomarqueurs) ciblés par les médicaments. Ces méthodes promettent d'optimiser l'utilisation de données génétiques dans le cadre du développement de nouvelles thérapies efficaces et sécuritaires ainsi que de discerner les interventions futiles.

1.1.1 L'histoire à succès de la PCSK9

Un exemple de succès d'une découverte génétique ayant mené au développement d'un nouveau médicament est sans doute la *proprotein convertase subtilisin/kexin type 9* (PCSK9). Cette protéine a été découverte et caractérisée à Montréal en 2003 par l'équipe du Dr Seidah [11], puis rapidement associée à l'hypercholestérolémie familiale, une maladie mendélienne caractérisée par de très hauts taux de cholestérol LDL et le développement précoce de la maladie coronarienne [12]. Les mutations initialement identifiées par liaison génétique étaient des mutations faux-sens entraînant un gain de fonction (*Gain of Function*, GoF) de la PCSK9, expliquant le mécanisme de transmission autosomique dominante. La nature GoF de ces mutations a aussi été confirmée dans des modèles murins où la surexpression de l'orthologue *Pcsk9* augmente le cholestérol non associé aux lipoprotéines à haute densité (*High Density Lipoprotein*, HDL) par 500% [13].

Afin d'identifier des mutations entraînant la perte de fonction de la PCSK9, une étude de séquençage a été réalisée dans la *Dallas Heart Study*, une cohorte populationnelle multiethnique [14]. Cette étude a d'abord ciblé les participants dont les taux de cholestérol LDL étaient les plus bas, menant à la découverte de mutations non-sens de *PCSK9* dont la prévalence a par la suite été mesurée dans la population complète. Les deux mutations non-sens identifiées étaient spécifiques aux Afro-américains et leur fréquence combinée était d'environ 2% (*vs* $< 0.1\%$ chez les Européens). Les effets protecteurs de ces mutations sur l'athérosclérose mesurée par imagerie de la carotide et sur la maladie coronarienne ont par la suite été confirmés par génotypage ciblé dans la cohorte *Atherosclerosis Risk in Communities* (ARIC) [15]. Dans cette étude, 1 seul des 85 (1.2%) porteurs Afro-américains des mutations non-sens $PCSK9^{142X}$ et $PCSK9^{679X}$ a développé la maladie coronarienne *vs* 319 des 3,278 (9.7%) non-porteurs ($P = 0.008$). Une méta-analyse de plus grande envergure incluant 11 339 cas de maladie cardiaque ischémique et 55 359 témoins a démontré l'effet protecteur de la mutation $PCSK9^{R46L}$. L'effet de cette variation génétique sur la maladie coronarienne rapporté selon le OR dans la méta-analyse à effets fixes était de 0.72 (intervalle de confiance (IC) 95% 0.62, 0.84) [16].

L'identification des mutations non-sens de *PCSK9* témoigne d'une certaine tolérance à la perte de fonction de ce gène. Il demeure intéressant d'évaluer les conséquences indésirables associées à ces mutations afin d'anticiper les effets possibles d'une thérapie. À l'aide d'une approche de randomisation mendélienne (*Mendelian Randomization*, MR), des variations génétiques de *PCSK9* ont été associées à une augmentation du risque de diabète de type 2 [17, 18]. Le OR pour une réduction de 10 mg/dl du cholestérol LDL par un score allélique de PCSK9 sur le diabète de type 2 était de 1.11 (IC 95% 1.04, 1.19), mais le bénéfice pour les

évènements cardiovasculaires demeurait probant avec un OR de 0.81 (IC 95% 0.74, 0.89).

L'ensemble de ces découvertes génétiques ont permis de cerner les bienfaits de l'inhibition de la PCSK9 et de repérer de possibles effets indésirables. Par la suite, plusieurs programmes de développement pharmaceutique d'inhibiteurs de PCSK9 ont vu le jour et l'étude *Further Cardiovascular Outcomes Research with PCSK9 Inhibition in Subjects with Elevated Risk* (FOURIER), publiée en 2017, a été la première étude de phase 3 à démontrer un bénéfice de cette classe thérapeutique dans la prévention d'évènements cardiovasculaires chez des patients à risque [19]. L'étude FOURIER portait sur l'évolocumab, un anticorps monoclonal ciblant la PCSK9, et le critère de jugement [1] primaire était une variable composite de décès pour cause cardiovasculaire, infarctus du myocarde, accident vasculaire cérébral, hospitalisation pour angine instable ou revascularisation coronarienne. Le rapport des risques instantanés (*Hazard Ratio*, HR) estimé dans l'étude pour l'issue primaire était de 0.85 (IC 95% 0.79, 0.92) en faveur du traitement.

Cet exemple démontre comment l'observation de variations génétiques perturbant une protéine peut prédire les bienfaits et les risques potentiels d'une inhibition pharmacologique. D'un point de vue génétique, l'identification des variations GoF et perte de fonction (*Loss of Function*, LoF) a permis d'établir une relation dose-réponse de la conséquence d'une modulation de la PCSK9. L'étude de variations plus communes au sein de populations a aussi aidé à élucider la relation avec les issues cliniques ainsi que la conduite d'études de MR. Le développement des inhibiteurs de la PCSK9 en un temps record est, par la suite, devenu une histoire à succès qui a renouvelé l'intérêt des approches génétiques pour appuyer l'identification de cibles pharmacologiques [20].

1.1.2 Le courant *funny* et l'ivabradine

Dans cette thèse, nous utiliserons une approche basée sur la génétique de cibles pharmacologiques pour étudier le cas de l'ivabradine. Ce médicament est un inhibiteur du canal ionique responsable du courant *«funny»* (I_f) et encodé par le gène *potassium/sodium hyperpolarization-activated cyclic nucleotide-gated channel 4* (*HCN4*). Ce canal est impliqué dans la dépolarisation spontanée des cardiomyocytes du nœud sinusal amorçant la contraction autonome du cœur [21]. L'inhibition du HCN4 par l'ivabradine réduit la fréquence cardiaque, ce qui peut entraîner un effet cardioprotecteur en abaissant la demande énergétique du myocarde, le stress oxydatif, la dysfonction endothéliale et en réduisant le risque de rup-

1. Dans le reste de la thèse, le terme «issue» sera utilisé pour faire référence au critère de jugement d'un essai randomisé ou à la variable d'intérêt dans une étude de MR. Le terme anglais utilisé pour désigner ce terme dans la littérature est «outcome».

ture de plaques d'athérome [22]. De plus, l'ivabradine, contrairement à d'autres médicaments abaissant la fréquence cardiaque, agit de façon spécifique sur la fréquence cardiaque et non sur d'autres paramètres comme la contractilité ou la pression artérielle [23]. Ce mécanisme d'action bien caractérisé permet donc de valider l'hypothèse thérapeutique des bénéfices de la réduction de la fréquence cardiaque.

L'ivabradine a initialement été évalué pour traiter l'angine stable. Cette maladie est causée par la formation de plaque d'athérome qui entrave la perfusion du myocarde entraînant des symptômes comme la dyspnée ou la douleur thoracique. Dans ce contexte ischémique, la réduction de la fréquence cardiaque par l'ivabradine devait améliorer la perfusion en augmentant le temps de diastole tout en réduisant la demande du myocarde en oxygène. Ces bénéfices escomptés ont été démontrés dans un essai randomisé où l'ivabradine augmentait le temps avant l'apparition de symptômes ischémique durant l'activité physique et la tolérance à l'exercice [24].

Par la suite, l'ivabradine a été testé dans un contexte de maladie coronarienne stable avec dysfonction systolique dans l'étude *morBidity-mortality EvAlUaTion of the I_f inhibitor ivabradine in patients with coronary disease and left-ventricULar dysfunction* (BEAUTIFUL) [23]. Comme les β-bloquants réduisent la fréquence cardiaque et qu'ils sont efficaces pour le traitement de la dysfonction systolique, l'ivabradine était pressentie efficace dans ce contexte. De plus, comme l'efficacité de l'ivabradine pour le traitement de l'angine est établie, la population clinique de BEAUTIFUL représentait un choix prometteur. Malgré une réduction de la fréquence cardiaque de 6 battements par minute après 12 mois de traitement, l'étude n'a pas démontré l'efficacité de l'ivabradine pour son issue primaire d'infarctus, de décès cardiovasculaire ou d'hospitalisation pour insuffisance cardiaque [23]. Des analyses de sous-groupe suggèrent cependant un bénéfice sur les évènements ischémiques chez les patients dont la fréquence cardiaque était élevée ($\geq$ 70 battements par minute) à l'entrée de l'étude. Dans ce sous-groupe, l'ivabradine était efficace pour prévenir l'hospitalisation en raison d'un infarctus aigu et la réduction relative des évènements était de 36% (P=0.001). Le résultat négatif sur le critère de jugement primaire a donc été attribué en partie à la fréquence cardiaque trop basse à l'entrée de l'étude.

Pour donner suite à ces résultats, l'étude *Systolic Heart failure treatment with the I_f inhibitor ivabradine Trial* (SHIFT) a été conçue pour évaluer l'efficacité de l'ivabradine chez des patients avec insuffisance cardiaque chronique avec dysfonction systolique présentant une fréquence cardiaque $\geq$ 70 battements par minute à l'entrée de l'étude [25]. Cette étude a démontré un bénéfice sur son issue primaire, une variable composite de décès cardiovasculaire ou hospitalisation pour détérioration de l'insuffisance cardiaque (HR = 0.82, IC 95%

0.75, 0.90). Ce résultat confirme l'hypothèse thérapeutique du bénéfice de la réduction de la fréquence cardiaque pour traiter l'insuffisance cardiaque. Cet effet était particulièrement probant chez les participants dont la fréquence cardiaque était supérieure à la médiane ($\geq$ 77 battements par minute) à l'entrée de l'étude.

Finalement, une troisième étude randomisée en double insu a évalué le bénéfice de l'ivabradine chez des patients coronariens stables sans dysfonction systolique présentant une fréquence cardiaque élevée ($\geq$ 70 battements par minute). L'étude *Study assessInG the morbidity-mortality beNefits of the I_f inhibitor ivabradine in patients with coronarY artery disease* (SIGNIFY) s'appuyait sur le résultat positif dans le sous-groupe de patients coronariens avec fréquence cardiaque élevée de l'étude BEAUTIFUL [26]. Étonnamment, cette étude s'avéra négative et l'incidence de l'issue primaire était de 6.8% dans le bras de traitement contre 6.4% dans le bras placébo [27]. Les auteurs ont donc conclu que, chez les patients coronariens aigus, une fréquence cardiaque élevée est un marqueur de risque et non un facteur causal des évènements coronariens.

Du point de vue de son innocuité, l'ivabradine est un médicament bien toléré et la bradycardie est l'effet indésirable le plus souvent rapporté dans les études. Cependant, une augmentation du risque de fibrillation auriculaire, une arythmie caractérisée par une contraction désorganisée des oreillettes, a été rapportée dans deux des études. Dans l'étude SHIFT, cet évènement a été observé chez 9% des individus du bras ivabradine contre 8% dans le bras placébo ($P = 0.012$). Dans l'étude SIGNIFY, 5.3% des patients ont développé la fibrillation auriculaire dans le bras ivabradine et 3.8% dans le bras placébo. Dans une méta-analyse de 11 essais randomisés en double insu incluant 21,571 participants, le risque relatif de développer la fibrillation auriculaire chez les utilisateurs d'ivabradine comparé aux utilisateurs du placébo était de 1.15 (IC 95% 1.07, 1.24 ; $P = 0.0027$) [28]. Les auteurs ont cependant estimé que le nombre d'individus à traiter pour mener à l'observation d'un effet indésirable (*number needed to harm*) est de 208 (IC 95% 122, 667) par année de traitement, ce qui peut représenter un risque acceptable considérant le bénéfice important de l'ivabradine en insuffisance cardiaque [28]. D'un point de vue épidémiologique, la fibrillation auriculaire et l'insuffisance cardiaque sont souvent observées conjointement et la relation de cause à effet est difficile à établir. Cette comorbidité s'explique par des causes communes (variables confondantes) telles l'hypertension, les maladies valvulaires ou ischémiques. Il existe également plusieurs hypothèses suggérant une causalité mutuelle des deux maladies [29].

Sur une période de 6 ans BEAUTIFUL, SHIFT et SIGNIFY ont évalué l'efficacité de l'ivabradine dans différentes populations de patients avec un succès variable. Face à cet historique complexe, le 2^{e} chapitre de cette thèse évaluera comment une approche génétique

aurait pu être utilisée pour guider ou prédire le résultat de ces essais randomisés. Bien que l'intérêt d'une approche en amont des essais cliniques soit indéniable, mener ces études en aval permet de bien cerner les avantages et inconvénients d'une approche génétique tout en permettant le développement de la méthodologie statistique. Le cas de l'ivabradine est un bon exemple où l'effet sur la fibrillation auriculaire est surestimé et dissimule l'effet bénéfique sur l'insuffisance cardiaque lorsqu'on utilise des modèles statistiques conventionnels. Pour modéliser ce scénario complexe, nous avons employé des modèles de risques en compétition. Nous avons aussi utilisé une approche de randomisation mendélienne pour valider l'hypothèse thérapeutique derrière l'ivabradine, c'est-à-dire que la réduction de la fréquence cardiaque entraîne un effet bénéfique chez des patients souffrant d'insuffisance cardiaque, mais pas chez des patients avec la maladie coronarienne.

1.1.3 L'exemple complexe de la CETP

La *cholesteryl ester transfer protein* (CETP) est une enzyme ayant la forme d'un tunnel et qui permet l'échange de cholestérols estérifiés contre des triglycérides entre les HDL et les lipoprotéines contenant de l'apolipoprotéine B (apoB) [30]. Des mutations génétiques de la CETP augmentent fortement le cholestérol HDL. Comme des données épidémiologiques suggéraient qu'une augmentation du cholestérol HDL pourrait protéger contre la maladie coronarienne, la découverte de cette enzyme a suscité beaucoup d'intérêt et mené au développement d'inhibiteurs de la CETP : le torcetrapib, l'anacetrapib, le dalcetrapib et l'evacetrapib [31-33]. Le premier essai de phase 3 évaluant un inhibiteur de la CETP, le torcetrapib, a été suspendu à cause de préoccupations liées à la sécurité du médicament qui augmente la pression artérielle et les évènements cardiovasculaires [34]. Ces effets ont été attribués à des effets de la molécule indépendants de l'inhibition de la CETP (*off-target*). Par la suite, les essais cliniques dal-OUTCOMES et *Assessment of Clinical Effects of Cholesteryl Ester Transfer Protein Inhibition with Evacetrapib in Patients at a High Risk for Vascular Outcomes* (ACCELERATE), évaluant respectivement le dalcetrapib et l'evacetrapib, ont été arrêtés pour cause de futilité. C'est seulement le dernier essai clinique, *Evaluation of the Effects of Anacetrapib through Lipid Modification* (REVEAL), qui a démontré un bénéfice en faveur de l'anacetrapib [35]. Cet essai clinique était aussi celui de plus grande taille incluant plus de 30 000 patients suivis durant 50 mois. Le mécanisme derrière le bénéfice de ce médicament est toujours débattu, car les inhibiteurs de la CETP ont plusieurs effets possiblement athéroprotecteurs. L'hypothèse avancée par plusieurs est que le bénéfice est attribuable à la réduction du cholestérol non HDL ou à la diminution de l'apoB. Dans l'étude REVEAL, l'anacetrapib a réduit le cholestérol non-HDL par 17 mg/dl et le bienfait prédit par une telle

réduction du cholestérol par les statines est cohérent avec l'effet protecteur observé dans l'étude. Cependant, ces raisonnements sont basés sur des calculs approximatifs et d'autres mécanismes sous-jacents pourraient aussi être impliqués.

Le transport inverse du cholestérol est un autre mécanisme par lequel les HDL peuvent exercer un effet athéroprotecteur. Le cholestérol est excrété vers les HDL au niveau de tissus périphériques pour être ensuite transporté vers le foie ou des organes endocrines où il sera excrété sous forme d'acides biliaires ou d'hormones stéroïdiennes, respectivement. La première étape de ce processus, l'efflux de cholestérol des tissus périphériques, a un effet athéroprotecteur, car il permet d'éliminer le cholestérol accumulé dans des cellules spumeuses (*foam cells*) au niveau de la plaque d'athérome [36]. Dans une cohorte de 2 924 participants à la *Dallas Heart Study* sans maladie coronarienne, l'efflux de cholestérol était associé avec une réduction de 67% de l'incidence des évènements coronariens lorsque les groupes formés à partir des 1er et 4^{e} quartiles d'efflux étaient comparés, alors que le cholestérol HDL n'avait pas d'effet [37]. Ces résultats ont été confirmés dans d'autres études et sont cohérents avec l'effet de l'efflux de cholestérol observé sur l'épaisseur entre la media et l'intima de la carotide, une mesure échographique de l'athérosclérose [38].

La capacité du plasma (et de ses HDL) à absorber du cholestérol est donc une mesure *in vitro* de la fonction des HDL qui est plus représentative de leur caractère athéroprotecteur que le cholestérol HDL [39]. Outre leur rôle dans le transport inverse du cholestérol, les HDL ont aussi des effets reconnus sur la fonction de l'endothélium [40], la production de lp(a) [41] et l'activation de plaquettes [42] qui pourraient contribuer à leur rôle athéroprotecteur. Les HDL forment un ensemble hétérogène de particules dont la taille et la composition en lipides et en apolipoprotéines diffèrent. À ce jour, nous ne connaissons pas l'ensemble des conséquences de l'inhibition de la CETP sur ces paramètres [43, 44].

Plusieurs études ont tenté de mieux comprendre les effets de la CETP par l'effet des variations génétiques de son gène. Les premières études se sont concentrées sur l'étude de variations communes de la CETP comme TaqIB (rs708272), une variation au premier intron du gène. TaqIB est fortement associée avec le cholestérol HDL, la concentration plasmatique de CETP, l'activité enzymatique de CETP, l'athérosclérose et la maladie coronarienne [45, 46]. Dans une méta-analyse publiée en 2008 incluant jusqu'à 92 études et 113 833 participants, Thompson *et coll.* ont décrit de façon exhaustive l'effet de 3 variations communes (TaqIB [rs708272], I405V [rs5882], -629C>A [rs1800775]]) et 3 variations plus rares de CETP (D442G [rs2303790], -631C>A [rs1800776] et R451Q [rs1800777]) sur un grand nombre de paramètres lipidiques ainsi que sur les évènements coronariens [47]. Il a été démontré qu'en général, les allèles diminuant l'activité de la CETP augmentaient le cholestérol HDL et dimi-

nuaient le niveau de cholestérol LDL et de triglycérides. Plusieurs de ces variations avaient un effet protecteur faible sur la maladie coronarienne (*p. ex.* $OR = 0.95$; IC 95% 0.91, 1.00 pour *CETP* -629C>A). La plupart des études génétiques subséquentes basées sur une approche prospective [48] ou de MR [49, 50] ont pu confirmer la réduction du risque d'évènements cardiovasculaires. Le mécanisme par lequel l'inhibition de la CETP exerce un effet cardioprotecteur demeure controversé. Dans une étude de MR, l'effet protecteur de CETP sur la maladie coronarienne, exprimé pour une réduction de 10 mg/dl de cholestérol LDL, était comparable à celui estimé par des variations génétiques de cibles pharmacologiques modulant principalement le cholestérol LDL : *HMGCR*, *PCSK9* et *NPC1L1* [51]. Ces résultats étaient aussi cohérents lorsqu'exprimés en fonction de la réduction d'apoB. Les auteurs ont donc proposé que le bienfait de l'inhibition de la CETP soit principalement dû aux effets sur les niveaux d'apoB. Ces estimations doivent cependant être interprétées avec précaution, car elles nécessitent une grande extrapolation. Par exemple, selon le score présenté au Chapitre 3 et construit selon des paramètres semblables à l'étude en question, la différence d'apoB entre le 1% le plus haut et le plus bas du score de CETP était de 3.4 mg/dl en moyenne. Une extrapolation substantielle est donc nécessaire pour arriver à exprimer un effet correspondant à 10 mg/dl. Des effets de cet ordre de grandeur ont été observés à l'aide de variations ayant des effets sévères sur la protéine CETP, mais leur effet sur la maladie coronarienne est estimé avec peu de précision à cause de leur faible prévalence [52]. L'émergence d'études utilisant la résonance magnétique nucléaire a permis d'évaluer l'effet de la CETP sur les fractions lipidiques avec une plus grande précision. La réduction génétique de la CETP a, en réalité, peu d'effet sur le cholestérol LDL, mais elle réduit le cholestérol des lipoprotéines à très faible densité (*Very Low Density Lipoprotein*, VLDL) et des restes de chylomicrons en plus d'augmenter la composition en triglycérides des HDL [53, 54]. En résumé, les analyses génétiques appuient fortement le rôle protecteur attribuable à une réduction génétique de la concentration de CETP sur les évènements coronariens. Une portion substantielle de cet effet protecteur est possiblement due à la réduction de cholestérol associé aux lipoprotéines contenant de l'apoB, mais d'autres mécanismes pourraient aussi y contribuer.

Dans une sous-étude pharmacogénomique de l'essai clinique dal-OUTCOMES, les individus traités au dalcetrapib et portant le génotype «AA» de la variation rs1967309 (représentant environ 17% de la population) avaient une réduction de 39% des évènements coronariens en comparaison avec les individus du bras placébo [55]. Suite à cette découverte, un modèle murin inactivé pour l'*Adcy9* a été développé et ces souris développaient moins d'athérosclérose, avaient une meilleure fonction endothéliale et prenaient plus de poids que les souris de génotype sauvage [56]. De façon concordante avec l'étude pharmacogénomique du dalcetrapib, ces effets disparaissaient chez les souris transgéniques pour la CETP malgré l'inactiva-

tion de l'*Adcy9*. En revisitant les données des études dal-OUTCOMES et dal-PLAQUE2, un bénéfice spécifique aux individus portant le génotype rs1967309-AA a été observé. Contrairement aux porteurs des autres génotypes, les individus homozygotes AA ne présentaient pas d'augmentation de la protéine C réactive, un marqueur d'inflammation systémique élevé chez les individus AG et GG, et les AA présentaient une augmentation de l'efflux du cholestérol comparé aux AG et GG [57]. Un essai randomisé ciblant les patients avec le génotype AA à la variation génétique rs1967309 est en cours, dal-GenE (NCT02525939), et permettra d'évaluer de façon définitive le bénéfice du dalcetrapib comparé au placébo dans la sous-population génétique. L'effet pharmacogénomique a aussi été évalué pour les autres inhibiteurs de la CETP. Dans une sous-étude de ACCELERATE, rs1967309 avait un effet concordant, mais non statistiquement significatif avec l'observation de dal-OUTCOMES. La valeur-p de l'interaction entre la variation génétique et le médicament était de 0.17 lorsqu'évaluée avec un modèle génotypique, et de 0.06 pour un modèle additif (test de tendance) [58]. Dans une sous-étude de REVEAL, il n'y avait cependant aucun indice d'interaction entre rs1967309 et l'anacetrapib ($P = 0.96$) [59].

1.2 Les études d'association

1.2.1 Les études d'association pangénomiques (GWAS)

Les GWAS représentent une approche agnostique pour identifier des variations génétiques corrélées à un phénotype d'intérêt. Une grande panoplie de phénotypes humains ont été étudiés par les GWAS [60], incluant, pour n'en nommer que quelques-uns, la taille, le risque de maladie coronarienne ou encore la réponse au médicament [61-63]. Dans les GWAS, chaque variation génétique est évaluée séquentiellement par un test statistique pour détecter l'association avec un phénotype. Les variations génétiques d'intérêt dans ces études sont fréquemment des polymorphismes d'un seul nucléotide (*Single Nucleotide Polymorphisms*, SNPs) ou de courtes insertions ou délétions relativement communes au sein de la population à l'étude (*p. ex.* fréquence allélique de plus de 5%). Au sens statistique, ceci signifie que les variations sont testées selon une approche marginale et non en considérant la distribution jointe des variations génétiques. Dans les études GWAS contemporaines, un très grand nombre de variations sont considérées grâce aux technologies permettant l'imputation de génotypes qui ne sont pas directement interrogés par les technologies de génotypage. Par exemple, en utilisant les génomes séquencés dans le cadre du projet TOPMed en guise de référence et la plate-forme publique d'imputation qui s'y rattache, un GWAS moderne peut ainsi tester l'association d'environ 6 millions de variations génétiques dont la fréquence

allélique est supérieure à 5% [2] [64, 65].

La visée initiale des études GWAS était d'élucider les causes génétiques des maladies humaines [66]. Cette tâche s'avère cependant complexe, car établir le mécanisme génétique derrière un signal d'association statistique est une tâche difficile qui requiert habituellement des études de validation fonctionnelle ciblées. Plusieurs considérations statistiques comme la correction pour les tests multiples ou le risque de biais par des variables confondantes ajoutent aussi à la complexité d'interpréter les associations identifiées dans les GWAS. Bien qu'une meilleure compréhension des voies biologiques demeure un objectif majeur des GWAS, ces études sont maintenant utilisées dans plusieurs contextes et servent aussi d'outil en génétique statistique.

Un phénomène ayant contribué à faire des GWAS des outils de génétique statistique est la grande popularité de la publication des résultats des GWAS. Notamment, les effets statistiques des variations génétiques peuvent être diffusés publiquement sans enfreindre la confidentialité des participants de recherche. Des statistiques sommaires obtenues des GWAS comme le coefficient de régression et son erreur type peuvent être publiées, car elles ne permettent pas la réidentification des données individuelles des participants. La plupart des grands consortiums étudiant des maladies génétiquement complexes adhèrent à cette approche et publient les statistiques sommaires de GWAS, et une grande partie des travaux présentés dans cette thèse utilisent de telles données.

Les statistiques sommaires de GWAS peuvent être utilisées dans une multitude de contextes et le développement d'approches statistiques adaptées à leur utilisation est un domaine de recherche important. Par exemple, elles sont utilisées pour estimer l'héritabilité (h^2_{SNP}) d'un phénotype, soit la fraction de la variance totale d'un phénotype attribuable aux effets génétiques additifs [67]. Elles sont aussi utilisées pour construire des scores de risque polygénique (*Polygenic Risk Scores*, PRS) qui peuvent être utilisés pour prédire le risque génétique de développer une maladie [68, 69]. L'estimation d'effets génétiques conditionnels est aussi possible à partir de statistiques sommaires. Par exemple, pour estimer l'effet d'une variation génétique sur un phénotype tout en ajustant pour une covariable, la méthode «multi-trait-based conditional & joint analysis» (mtCOJO) peut être utilisée [70]. Cette méthode est protégée du biais de collisionneur et permet un ajustement dans l'estimation de l'effet génétique sur le phénotype en y soustrayant l'effet médié par la covariable. Finalement, ces statistiques peuvent être utilisées pour inférer la relation de causalité entre deux variables ayant une composante génétique en utilisant l'approche de la MR, et qui sera

2. En considérant aussi un filtre sur la qualité de l'imputation de $r^2 \geq 0.6$, la plate-forme de génotypage Illumina Infinium Global Screening Array MD v3 et la référence du génome GRCh38

abordée à la section 1.4.

1.2.2 Les études d'association panphénomiques (PheWAS)

Les dossiers médicaux électroniques contenant des codes diagnostiques et des résultats de tests médicaux représentent d'imposantes banques des données. Bien souvent, ces dossiers sont rattachés à un système de santé et incluent des données sur des centaines de milliers d'individus. Une des premières implémentations à grande échelle d'un tel système est *bioVU*, la biobanque de l'Université de Vanderbilt, qui utilise un modèle *opt-out* où les participants sont inclus dans la biobanque à moins de retirer leur consentement à participer au projet lors de la signature du consentement aux soins [71]. Les données des dossiers médicaux sont, par la suite, anonymisées et l'excédent de sang collecté pour les analyses de laboratoire de routine est utilisé afin d'en extraire l'ADN et de faire le génotypage. Grâce aux données génétiques et aux codes diagnostiques enregistrés pour fins de facturation, cette biobanque a permis la conduite de la première étude études d'association panphénomique (*Phenome-Wide Association Study*, PheWAS). Dans cette première mouture, les phénotypes étaient des combinaisons de codes de facturation ICD9 (*Phecodes*) permettant de représenter l'ensemble du phénome d'intérêt clinique [72]. Des groupes contrôles étaient aussi automatiquement sélectionnés en excluant des individus atteints de maladies reliées. Dans cette démonstration de faisabilité, 7 associations précédemment identifiées par l'approche GWAS ont été testées avec tous les *Phecodes*. Quatre de ces variations ont vu leur association répliquée par PheWAS et 19 nouvelles associations potentielles ont été identifiées [72].

L'approche PheWAS a plusieurs avantages dans le contexte de l'étude des cibles pharmacologiques, notamment pour le repositionnement de médicaments ou pour la prédiction des effets indésirables d'un médicament. Par repositionnement de médicament, nous entendons la découverte d'une nouvelle indication pour un médicament ayant déjà été approuvé pour une indication thérapeutique existante. Comme les paramètres pharmacocinétiques et l'innocuité de tels médicaments sont déjà établis, seule une démonstration d'efficacité pour la nouvelle indication thérapeutique est nécessaire [73]. Par exemple, une étude portant sur l'arthrite rhumatoïde a utilisé une approche bio-informatique incluant des GWAS et des PheWAS pour établir la relation entre les régions identifiés et les protéines affectées ainsi que leurs partenaires d'interaction [74]. Les protéines identifiées étaient enrichies pour des cibles pharmacologiques de médicaments approuvés pour le traitement de l'arthrite rhumatoïde par un facteur de 3.7x [74]. Certains des gènes identifiés encodaient des cibles pharmacologiques pour d'autres maladies incluant certains cancers et le psoriasis, révélant des opportunités de repositionnement de médicaments. Malgré tout, les repositionnements réussis demeurent

limités et les approches GWAS et PheWAS doivent être intégrées dans des approches computationnelles plus élaborées prenant en considération, notamment l'analyse de causalité [75, 76].

La deuxième application pour laquelle l'approche PheWAS est avantageuse est la prédiction d'effets indésirables lors de la validation de cibles pharmacologiques. Si une variation génétique de cible pharmacologique est associée à un bienfait thérapeutique, l'approche PheWAS peut être utilisée pour découvrir des conséquences inattendues de l'usage du médicament. Par exemple, une variation bénéfique dans le gène *PCSK9* associée avec une diminution du risque d'hyperlipidémie (OR = 0.63), de maladie coronarienne (OR = 0.73) et d'AVC ischémique (OR = 0.61) a été utilisée dans un PheWAS afin de déceler des effets indésirables dus à la modulation de la PCSK9 (effet *«on target»*) [77]. Ce criblage a révélé une augmentation du risque de diabète de type 2 (OR = 1.24) après ajustement pour l'usage de médicaments contre l'hypercholestérolémie. Aucun effet sur le risque de cataracte, d'insuffisance cardiaque, de fibrillation auriculaire ou de dysfonction cognitive n'a été observé. Bien qu'une augmentation de l'incidence du diabète soit un effet indésirable connu des statines, les essais randomisés d'inhibiteurs de la PCSK9 n'ont pas révélé d'effet sur la glycémie ou sur le diabète [78]. Cette discordance avec les études génétiques pourrait cependant être due à la différence dans la durée de l'exposition et des essais cliniques à long terme seront nécessaires pour élucider la question. Les observations du PheWAS sur la fonction cognitive sont cohérentes avec les résultats de l'essai clinique EBBINGHAUS démontrant que l'évolocumab, un inhibiteur de la PCSK9, n'a pas d'effet sur cette dernière [79].

Bien entendu, l'approche PheWAS n'est pas adéquate pour prédire des effets non spécifiques des médicaments (effet *«off target»*). Cette approche est aussi limitée dans le cas de situations plus complexes, par exemple, si des risques en compétition sont en jeu ou si la médication interfère avec les effets génétiques. Les PheWAS sont tout de même communes en génétique et ils sont souvent utilisés pour déterminer le spectre des conséquences phénotypiques de variations génétiques nouvellement identifiées.

1.2.3 Modèles d'association basés sur les gènes

L'étude des associations génétiques se fait habituellement au niveau des variations génétiques. Dans certains cas, il peut cependant être avantageux de tester l'apport combiné des variations dans un test d'association gène-centrique. D'abord, cette approche peut faciliter l'interprétation des résultats en priorisant l'implication du gène à l'étude dans le mécanisme génétique qui sous-tend l'association. La prudence reste de mise, car il est possible que l'effet soit, en réalité, modulé par un élément génétique partageant le même *locus*, ou agissant à dis-

tance (effet en *trans*). Néanmoins, l'approche gène-centrique permet de faciliter les analyses bio-informatiques comme l'enrichissement d'annotations de gènes ou de voies biologiques qui utilisent des listes des gènes associés. Un autre avantage de ces approches est de réduire le fardeau des tests multiples. La correction utilisée pour les GWAS suppose fréquemment un million de tests indépendants. C'est pour cette raison que le seuil de signification de $5 \times 10^{-8} = 0.05/10^6$ est un choix populaire [80]. En considérant que le génome humain compte environ 20 000 gènes, une correction de Bonferroni pour une étude gène-centrique serait environ 100 fois plus permissive qu'une correction GWAS, ce qui représente une opportunité intéressante pour développer des méthodes d'associations plus puissantes tout en contrôlant adéquatement le taux de faux positifs. Ces approches ont aussi un avantage particulier pour l'étude des variations rares qui ne peuvent être étudiées individuellement à cause de leur faible fréquence dans la population. Des méthodes comme l'analyse de fardeau («*burden test*») permettent alors de tester la différence de fréquence des allèles rares à un *locus* entre les individus atteints et non atteints [81]. La combinaison de variations génétiques peut aussi entraîner un gain de puissance statistique dans le cas des variations communes s'il existe plus d'une variation causale. Comme l'étude des variations rares n'est pas abordée dans cette thèse, le reste de la section se limitera aux approches statistiques pour les tests gène-centriques adaptés aux variations communes (un tableau résumé est présenté à l'annexe C.1).

Plusieurs approches combinent les statistiques d'association des variations génétiques individuelles (*p. ex.* les statistiques χ^2) afin d'obtenir une statistique de test omnibus. Une des premières méthodes proposées, le test d'ensemble de variations implémenté dans le logiciel généraliste *Plink*, utilise une approche de regroupement par déséquilibre de liaison (*Linkage Disequilibrium*, LD) et d'un filtre sur les valeurs-p pour cibler des variations susceptibles de contribuer de façon indépendante au phénotype [82]. Cet algorithme de sélection des variations est décrit de façon plus détaillée à la section 1.3. Une fois les variations indépendantes sélectionnées, la moyenne de leurs statistiques d'association χ^2 est utilisée comme statistique omnibus. La valeur-p empirique est ensuite calculée en permutant les phénotypes afin d'estimer la distribution des statistiques sous l'hypothèse nulle. Cette approche est conceptuellement intéressante, mais l'utilisation du test de permutation pour le calcul de la valeur-p entraîne un temps de calcul considérable. Au lieu d'utiliser la moyenne des statistiques χ^2, la méthode *«VErsatile Gene-based Association Study»* (VEGAS) utilise la somme des χ^2, ce qui peut s'avérer plus puissant si les variations ont toutes de petites tailles d'effet. Cette méthode utilise la corrélation entre les variations génétiques (matrice de LD) afin de simuler des statistiques d'association sous l'hypothèse nulle et dérive une valeur-p empirique. L'utilisation du modèle de simulation s'avère toujours coûteuse sur le plan computationnel

[83]. Une stratégie permettant d'éviter le recours aux permutations ou aux simulations est l'adaptation du test de Simes développé comme correction pour les tests multiples et adapté aux analyses gène-centriques dans l'outil *«Gene-based Association Test that uses Extended Simes procedure»* (GATES) [84]. Cet outil estime le nombre de variations génétiques indépendantes à l'aide de la matrice de corrélation des valeurs-p pour l'association des variations génétiques. La valeur-p du test gène-centrique correspond ensuite à la plus petite valeur-p corrigée par la procédure de Simes. L'hypothèse nulle de ce test correspond au scénario où aucune variation génétique de l'ensemble n'est associée au phénotype. Comme la procédure de Simes est analogue à l'approche de contrôle du taux de fausses découvertes proposé par Benjamini et Hochberg, on peut voir ce test comme une approche visant à contrôler le taux de fausses découvertes à même l'ensemble de variations génétiques considérées [85]. Dans un modèle où toutes les variations génétiques du gène ont de petites tailles d'effet, cette approche risque d'être moins puissante que les alternatives. Une implémentation optimisée de tests basés sur la moyenne des χ^2 ou sur la somme des χ^2 a ensuite été proposée dans l'outil *«PAthway Scoring ALgorithm»* (PASCAL) qui utilise une simulation de Monte Carlo ou un algorithme d'intégration numérique pour le calcul efficace des valeurs-p [86]. Comme ces différentes façons de combiner les statistiques d'association ont toutes certains avantages et inconvénients, il peut être souhaitable de combiner leurs résultats pour obtenir un meilleur prédicteur. L'outil *«COMbined gene-based Association Test»* (COMBAT) offre cette possibilité en utilisant la procédure de Simes pour combiner les valeurs-p des différents modèles d'association dans un paradigme d'apprentissage ensembliste [87].

Une autre façon de combiner les variations génétiques à un *locus* en contrôlant pour le LD est d'utiliser l'analyse en composantes principales des génotypes. Cette technique permet de décomposer la matrice de covariance des génotypes en composantes orthogonales qui expliquent une proportion décroissante de la variance. Même si le nombre de variations génétiques considérées est grand, cette décomposition permettra de générer de nouvelles variables continues (les valeurs propres) retenant la majorité de la variance génétique en tirant profit du LD. Les composantes principales dérivées des valeurs propres peuvent ensuite être utilisées dans un modèle de régression sur phénotype conventionnel et l'association peut être testée en utilisant une approche comme le ratio des vraisemblances [88, 89]. Un avantage important de l'approche par analyse en composantes principales est qu'elle ne nécessite pas de tests individuels des variations génétiques ni de simulations pour calculer des valeurs-p empiriques, ce qui en fait une alternative efficace en matière de temps de calcul. Le nombre de composantes principales incluses au modèle est un hyperparamètre et peut être déterminé selon des analyses de sensibilité ou sélectionné par validation croisée. De plus, comme les composantes principales sont des combinaisons linéaires de variations

génétiques, il est possible d'établir *a posteriori* les variations génétiques contribuant au signal d'association.

Les approches à noyau forment une autre famille de méthodes utilisées pour les associations avec des ensembles de variations génétiques. Ces approches utilisent une fonction de similarité, nommée fonction de noyau, pour calculer la similarité génétique de chaque paire d'individus. Différents choix de noyaux ont été utilisés, mais le noyau d'identité par l'état (*identity by state*) qui est proportionnel à la somme pondérée des variations génétiques partagées entre deux individus est un choix populaire [90]. L'estimation statistique peut ensuite se faire facilement, car il existe un lien entre l'approche à noyau et les modèles linéaires mixtes ce qui permet l'utilisation de procédures statistiques conventionnelles. La méthode *SNP-set Kernel Association Test* (SKAT) est un exemple d'une approche à noyau et permet de tester l'effet combiné de variations rares et communes [91, 92]. Contrairement à l'approche des méthodes par fardeau génétique, le SKAT est puissant même si des allèles rares ont des effets de direction opposée sur un phénotype, ou si plusieurs variations considérées sont neutres. Cependant, l'approche de fardeau est plus puissante si la majorité des variations sont causales et ont des effets dans la même direction. Pour concilier leurs avantages, le modèle SKAT-O a été développé et correspond à un mélange paramétrable des statistiques d'association de ces deux approches [93].

1.2.4 Bases de données d'associations

Les approches GWAS et PheWAS tentent toutes les deux d'identifier des associations entre des variations génétiques et des phénotypes avec des perspectives différentes. Plusieurs bases de données ont émergé pour tenter de coordonner et de centraliser les résultats issus de ces études (Table 1.1). La première de ces initiatives est le *GWAS Catalog* qui recense les associations statistiques significatives publiées et les statistiques sommaires complètes de certaines études [60]. Au moment de l'écriture, ce catalogue contenait 4 865 publications et 247 051 associations génétiques. Plus récemment, la base de données *PhenoScanner* a été développée et étend considérablement le spectre des associations considérées [97, 100]. Cette ressource intègre plus de 65 milliards d'associations portant sur l'expression des gènes, les métabolites, les niveaux protéiques, l'épigénétique et les maladies. Elle permet aussi la recherche par variation corrélée (*proxy*) en utilisant une population de référence issue du *1000 Genomes Project* pour estimer le LD [101]. *PhenoScanner* peut aussi être interrogé de façon automatique en utilisant le langage de programmation *R* ou un outil en ligne de commande. Un autre portail d'accès aux données d'associations est le *Open Targets Genetics* qui s'inscrit dans une plate-forme plus large d'intégration des données *omiques* nommée

Open Targets Platform et qui se spécialise dans la validation de cibles pharmacologiques [95, 102]. En plus de recenser les associations SNP–phénotype, le portail *Open Target Genetics* ajoute des outils de priorisation du gène causal basé sur la distance, la colocalisation avec l'expression génique, la pathogénicité prédite par des outils informatiques et l'interaction avec la chromatine. Lorsqu'on recherche de l'information sur une variation génétique, ces ressources sont maintenant le point de départ, car elles permettent de rapidement identifier des associations déjà connues, et ce à travers un spectre très large de phénotypes.

En plus de ces ressources centralisées, d'excellents outils ont été développés pour faciliter l'exploration interactive et la visualisation des données d'association. *PheWeb* permet le déploiement d'instances indépendantes et l'intégration de statistiques sommaires sur plusieurs phénotypes [94]. Une interface web permet de parcourir les résultats par variant, par *locus* ou par phénotype. Il est aussi possible de générer les graphiques communément utilisés dans le contexte d'études d'association comme le diagramme de Manhattan pour visualiser l'association statistique à travers le génome et le diagramme quantile-quantile qui permet d'observer l'inflation de la statistique d'association en comparant la distribution des millions de statistiques observées à la distribution attendue sous l'hypothèse nulle. Le diagramme de type *locus* (*Locus Plot*) est aussi implémenté et permet de visualiser l'association statistique à petite échelle et d'observer la structure de LD et les gènes présents dans la région d'intérêt. Ce graphique est généré à l'aide d'un outil développé de façon indépendante et nommé *LocusZoom* qui permet lui aussi la présentation de statistiques sommaires [103]. Certaines instances publiques notables de *PheWeb* qui permettent l'interrogation d'études d'association sont décrites à la Table 1.2. Les sous-études pharmacogénomiques des essais randomisés COLCOT et COLCORONA auxquels j'ai participé dans le contexte de ma thèse sont rattachés à des instances publiques de PheWeb : statgen.org/pheweb/colcot et statgen.org/pheweb/colcorona [104, 105].

Divers portails de connaissances ont aussi été développés dans le cadre de l'initiative *Accelerating Medecines Partnership* (AMP) financée par le National Institute of Health (NIH), la *U.S. Food & Drug Administration* (FDA) et 10 compagnies privées du secteur biopharmaceutique. Ces portails, coordonnés par le *Knowledge Portal Network*, sont organisés par regroupement de maladies incluant les maladies métaboliques communes, la sclérose latérale amyotrophique et les maladies musculo-squelettiques (hugeamp.org). Ces ressources font l'agrégation de résultats issus de grandes études liant la génétique humaine aux maladies complexes et permettent le téléchargement, l'exploration et la visualisation des résultats. En contraste avec l'outil *PheWeb*, ces portails permettent aussi la recherche des jeux de données en fonction des maladies et des types de données génétiques utilisées. Au moment de l'écri-

ture, le portail des maladies métaboliques communes recensait 277 jeux de données portant sur 332 traits.

Certaines bases de données de résultats d'associations ont une vocation plus spécialisée. C'est le cas par exemple de *MR-BASE* qui inclut des statistiques sommaires de plusieurs GWAS dans le but de permettre les études de MR à deux échantillons [108]. En plus de permettre l'estimation des effets causaux, la plate-forme offre plusieurs options pour effectuer des analyses de sensibilité et étudier la validité des suppositions de la MR. La recherche de variations corrélées (*proxy*) et l'harmonisation des allèles se font aussi automatiquement facilitant les analyses et réduisant le risque d'erreurs. Le *Polygenic Score Catalog* est un autre exemple de base de données spécialisée dont la vocation est d'harmoniser la représentation des PRS de façon à favoriser la reproductibilité et l'évaluation de nouveaux scores [109].

Ces ressources bio-informatiques offrent un appui important à la recherche et l'avancée des connaissances reliant les variations génétiques à des conséquences phénotypiques et fonctionnelles. Néanmoins, elles sont limitées aux résultats issus d'approches conventionnelles considérant habituellement l'effet d'une seule variation génétique individuelle obtenu à l'aide de modèles linéaires et additifs. Il est encore rare d'y trouver des résultats d'analyses stratifiées par sexe ou pour d'autres sous-groupes d'intérêt clinique, ce qui peut restreindre l'utilité et la validité des résultats dans certains contextes. De façon plus générale, la contribution des interactions gène-gène, gène-environnement ou des effets dominants est habituellement perdue dans ces bases de données.

TABLE 1.1 – Description de bases de données d'associations génétiques sélectionnées.

Ressource	Description	URL
PheWeb [94]	– Outil d'exploration et visualisation de résultats d'études d'association – Permet le déploiement d'instances personnalisées	pheweb.sph.umich.edu
Finngen	– Résultats d'associations de l'étude Finngen, une cohorte populationnelle basée en Finlande – Données sommaires de GWAS disponibles après embargo d'un an – Version adaptée de PheWeb	r4.finngen.fi
AMP Knowledge Portal Network	– Agrégation de données issues de plusieurs études – Permet la recherche de jeux de données – Accès centralisé aux statistiques sommaires – Outils d'analyse (*p. ex.* priorisation de gène)	hugeamp.org
Open Targets Genetics [95]	– Approche centrée sur la validation de cibles pharmacologiques – Références croisées avec bases de données de voies biologiques, d'animaux modèles et de cibles pharmacologiques – Priorisation de gène et de variation génétique – Intégration de données épigénétiques, d'expression des gènes et de niveaux protéiques	genetics.opentargets.org
GWAS Catalog [60]	– Effort de recensement de toutes les associations génétiques issues de GWAS – Inclut aussi plusieurs GWAS de plus petite envergure	ebi.ac.uk/gwas
PheWAS Catalog	– Portail spécialisé pour les études PheWAS – Inclut les dictionnaires Phecodes utilisés pour l'agrégation de codes diagnostiques	phewascatalog.org
Polygenic Score Catalog [96]	– Base de données spécialisée dans les scores de risque génétique	pgscatalog.org
PhenoScanner [97, 98]	– Intégration de plusieurs sources de données d'association – Recherche par variation corrélée (*proxy*) – Outils pour accès automatisé avec le langage de programmation R	phenoscanner.medschl.cam.ac.uk
Online Mendelian Inheritance In Man (OMIM) et ClinVar [99]	– Bases de données spécialisée dans les maladies et mutations rares – Intégration et révision de mutations observées dans des cliniques génétiques	omim.org et ncbi.nlm.nih.gov/clinvar

TABLE 1.2 – Exemples sélectionnés d'instances publiques de PheWeb permettant la consultation en ligne de résultats d'études d'association.

Étude	n participants	n variants	n phén.	URL
Finngen	176 899	17M	2 444	r4.finngen.fi
UK Biobank (TOPMed) [106]	400 000	57M	1 400	pheweb.org/UKB-TOPMed
UK Biobank (Neale lab v1)	337 000	11M	2 400	pheweb.org/UKB-Neale
Michigan Genomics Initiative (MGI)	40 000	24M	1 700	pheweb.org/MGI-freeze2
Méta-analyse données de laboratoire (MGI + BioVU) [107]	–	0.8M	70	pheweb.org/MGI-BioVU
METSIM	1 000 – 6 000	21M	1 400	pheweb.org/metsim-metab
BioBank Japan	$\leq$ 200 000	–	218	pheweb.jp

1.3 Les scores de risque génétique

Les scores de risque génétique (*Genetic Risk Scores*, GRS) représentent une mesure de l'effet combiné de variations génétiques en lien avec un trait ou une maladie. Ils peuvent inclure l'effet d'un nombre restreint de variations génétiques au sein d'un gène d'intérêt ou encore combiner l'effet de millions de variations à travers le génome. On utilise parfois le terme «score de risque polygénique (*Polygenic Risk Score*, PRS)» pour désigner ces derniers, mais pas exclusivement.

Les applications des GRS sont diverses. D'abord, ils peuvent être utilisés pour prédire un trait ou l'incidence d'une maladie [110]. Les GRS pangénomiques contemporains ont un pouvoir prédictif considérable qui peut même égaler celui de mutations rares causant des maladies mendéliennes. Par exemple, les mutations causant l'hypercholestérolémie familiale ont une prévalence d'environ 0.4% dans la population et elles triplent le risque de maladie coronarienne chez les individus porteurs de ces mutations. Dans une étude portant sur les GRS pangénomiques, Khera *et coll.* ont construit un score basé sur 6.6 millions de variations génétiques et ont démontré que les individus dont le score était supérieur au 92^{e} percentile avaient un risque de maladie coronarienne équivalent aux porteurs de mutations rares causant l'hypercholestérolémie familiale [111]. Selon ce seuil, pour un même risque génétique, la proportion d'individus ayant un risque polygénique accru dans la population (8%) est largement supérieure à celle des porteurs de mutations rares (0.4%), et s'ajoute à celles-ci. Les auteurs ont aussi montré que ces individus n'auraient pu être dépistés en fonction de facteurs de risque conventionnels comme l'hypertension ou l'hypercholestérolémie. En comparant le pouvoir prédictif d'un score de maladie coronarienne aux facteurs de risque conventionnels, Inouye *et coll.* ont montré que le risque polygénique est un prédicteur au moins aussi fort que le tabagisme, l'obésité, l'historique familial de maladie coronarienne, l'hypertension ou l'hypercholestérolémie [69]. De plus, l'ajout du score polygénique à un modèle multivariable incluant ces facteurs de risque augmentait le pouvoir prédictif de façon significative. Contrairement aux autres facteurs de risques, les GRS ne changent pas dans le temps et peuvent être mesurés tôt dans la vie. Leur effet n'est toutefois pas déterministe et une modification des habitudes de vie peut contrebalancer de façon substantielle le risque génétique [112]. Ces scores peuvent aussi être utilisés pour prioriser le traitement pharmacologique. Par exemple, les individus ayant un plus grand GRS de maladie coronarienne bénéficient plus de la prise d'une statine en prévention primaire, c'est-à-dire avant un premier évènement coronarien. Dans une étude de l'essai randomisé *Justification for the Use of Statins in Prevention: an Intervention Trial Evaluating Rosuvastatin* (JUPITER), la statistique du nombre nécessaire à traiter (NNT) avec une statine sur une période de 10 ans pour éviter un évènement car-

diovasculaire était de 66 dans le groupe à faible risque génétique versus 25 dans le groupe à haut risque [113]. L'application de GRS dans un contexte clinique ou prédictif est appuyée d'évidence tirée de résultats préliminaires, mais la démonstration d'un bénéfice réel n'a pas encore été faite dans un essai randomisé. Les GRS peuvent aussi être utilisés comme variables instrumentales dans les études de MR et c'est l'utilisation principale qui en est faite dans cette thèse (voir la section 1.4.3).

Nous aborderons maintenant les aspects techniques liés à la construction et au calcul de GRS. Mathématiquement, le GRS d'un individu correspond à :

$$GRS_i = \sum_{j=1}^{m} w_j G_{ij}$$

Où G_{ij} correspond au génotype du $i^{\text{ème}}$ individu à la $j^{\text{ème}}$ variation génétique et w_j le poids associé à cette variation génétique. Le génotype suit habituellement un encodage additif et prend comme valeur le nombre d'allèles mineurs portés par l'individu ($G_{ij} \in \{0, 1, 2\}$). Afin de refléter l'incertitude lors de l'imputation des génotypes, une valeur continue entre 0 et 2 représentant un dosage génotypique peut être utilisée. Les poids w_j correspondent souvent aux effets alléliques estimés dans des GWAS (c'est-à-dire que $w_j = \hat{\beta}_j$, où $\hat{\beta}_j$ est le coefficient de régression pour l'effet de la variation génétique). Autrement, quand de telles statistiques ne sont pas disponibles ou qu'elles sont susceptibles d'induire un biais, on peut calculer un GRS non pondéré. Cette approche est un cas spécial où les poids $w_j = 1$ et où le GRS correspond à la somme des allèles portés par un individu. On prend cependant soin d'exprimer toutes les variations génétiques en fonction de l'allèle dont l'effet va dans la même direction.

Les principaux paramètres à déterminer lors de la conception d'un score sont le choix des variations génétiques à inclure et leur poids. L'approche classique pour la sélection des variations à inclure est la méthode de tri par valeur-p et d'agrégation par LD (approche C+T pour *clumping and thresholding*) [114]. Cette approche correspond à un algorithme vorace de sélection de variables. D'abord, on trie la liste de variations candidates en ordre croissant de la valeur-p d'association avec le trait d'intérêt. Ensuite, on inclut itérativement la première variation, puis on retire les variations corrélées de la liste de variations candidates. Cette étape est répétée jusqu'à ce qu'il ne reste aucune variation candidate ayant une valeur-p inférieure à un certain seuil. Cet algorithme très simple fait usage de deux hyperparamètres, soient le seuil de valeur-p et le seuil utilisé pour l'agrégation de variations par LD (seuil de r^2). La détermination de ces paramètres peut se faire en utilisant les méthodes habituelles telles que la validation croisée. Dans le cas où le GRS d'intérêt doive être limité à un gène, par

exemple dans le contexte de la création d'une variable instrumentale pour une étude de cis-MR, on restreint simplement la liste de variations candidates au *locus* d'intérêt. Si l'accès aux données individuelles est possible, on peut aussi adopter une approche de conditionnement pas-à-pas. L'analyse d'association est alors répétée en incluant la plus forte association de l'itération précédente jusqu'à ce qu'aucun signal ne soit détecté. Cette approche a l'avantage de contrôler pour le LD implicitement et de permettre l'intégration de variations corrélées, mais dont l'effet est partiellement indépendant. Il faudra cependant prendre soin d'éviter le surapprentissage en utilisant un jeu de données indépendant ou la validation croisée. Au lieu de sélectionner des variations à inclure dans le score, une autre approche consiste à inclure toutes les variations génétiques disponibles. Cette approche mène à la construction de scores incluant plusieurs millions de variations génétiques qui sont, la plupart du temps, de meilleurs prédicteurs que les scores plus restrictifs [111]. Bien qu'ils soient adéquats pour des tâches prédictives, ces scores sont souvent pléiotropiques et moins adaptés à des approches de MR.

Plusieurs approches statistiques ont été développées pour optimiser les poids des GRS. L'approche *LDPred* adopte un modèle bayésien permettant d'imposer un *a priori* sur les poids afin de tenir compte du LD et de la distribution attendue des poids selon l'héritabilité et la fraction des variations qui sont causales [115]. Le LD et l'héritabilité peuvent être estimés automatiquement par l'outil, mais la fraction de variations causales est un hyperparamètre à déterminer par l'utilisateur. La plus récente version de l'outil, *LDPred2*, permet l'estimation automatique de ce paramètre et offre aussi la possibilité d'estimer des poids épars où certains poids sont fixés à 0 [116]. L'autre famille de méthodes pour induire un rétrécissement (*shrinkage*) des paramètres est la régression pénalisée. Ces modèles imposent des contraintes sur la taille totale des poids ce qui a pour conséquence de réduire la contribution des variations dont l'effet est nul. Des implémentations adaptées aux PRS ont été développées, incluant l'outil *lassosum* [117].

Le calcul de GRS peut se faire à l'aide de plusieurs outils bio-informatiques. La plupart des outils discutés pour l'optimisation des poids permettent le calcul subséquent des GRS. Autrement, le logiciel *Plink* peut être utilisé pour calculer ces scores, mais d'autres outils plus spécialisés sont parfois mieux adaptés. Dans le cadre de la création et du calcul de scores par l'approche C+T par exemple, il est souhaitable d'utiliser un logiciel comme *PRSice-2* qui permet de tester plusieurs seuils de valeur-p afin de déterminer le seuil optimal. L'outil facilite aussi les tâches de contrôle de qualité comme l'arrimage des allèles en fonction du poids et du génotype encodé et permet la visualisation de pouvoir prédictif du score [118, 119]. J'ai aussi développé une suite d'outils pour la création, le calcul et l'évaluation de GRS qui sera présentée à la section 5.2.

1.4 La randomisation mendélienne

Les variations génétiques sont héritées à la naissance et demeurent généralement fixes tout au long de la vie. Ce phénomène a de puissantes implications, car l'immuabilité des variations génétiques les rend moins sensibles aux biais. Dans une courte lettre à l'éditeur de 1986 et portant sur la relation causale entre le cholestérol sérique et le cancer, Dr Martijn B. Katan écrit : « *Unlike most other indices of lipid metabolism, apolipoprotein aminoacid sequences are not disturbed by disease, and the apo E phenotype found in a patient will have been present since birth.* » [120]. L'apolipoprotéine E (apoE) étant impliquée dans l'élimination du cholestérol du plasma, les isoformes de l'apoE sont associées au cholestérol plasmatique. Pour élucider la relation entre le risque de cancer et le cholestérol, le Dr Katan recommande de comparer la fréquence des isoformes de l'apoE entre des individus avec et sans cancer. Cette lettre à l'éditeur, formulée avant même le séquençage du génome humain, est aujourd'hui considérée comme la première mention du concept de la MR [121].

1.4.1 Estimateurs et approches statistiques

La MR permet d'estimer l'effet d'une exposition (X) sur une issue (Y) en utilisant des variations génétiques (G) associées à l'exposition (Figure 1.1). Cette approche propose l'utilisation de variations génétiques à titre de variables instrumentales dans une analyse en inférence causale. L'estimation par variable instrumentale est une méthode fréquemment utilisée pour l'inférence causale en économétrie et les premiers estimateurs utilisés en MR sont issus de ce domaine. Avant de parler des estimateurs, abordons les suppositions nécessaires à leur validité. Dans la littérature des variables instrumentales, les trois suppositions sont les suivantes [122, 123] :

1. **Pertinence**. La variable instrumentale est associée avec l'exposition d'intérêt.
2. ***«Exclusion restriction»***. La variable instrumentale est conditionnellement indépendante de l'issue sachant l'exposition.
3. **Indépendance**. La variable instrumentale ne doit pas avoir de cause commune avec l'issue.

La première condition exige que l'instrument génétique soit fortement associé à l'exposition d'intérêt. En pratique, cette hypothèse peut être validée aisément en utilisant des mesures d'association statistique (*p. ex.* statistique F partielle) et de taille d'effet (*p. ex.* R^2) [124]. La 2^{e} supposition stipule que l'effet de l'instrument génétique est complètement médié par l'exposition. Cette condition pourrait se révéler fausse si les variations génétiques

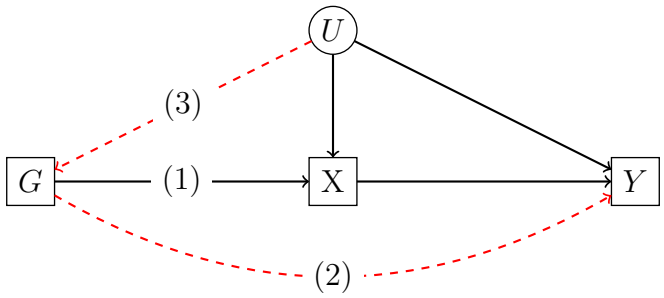
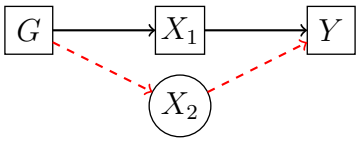


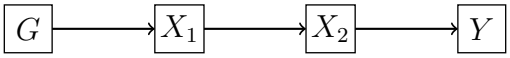
FIGURE 1.1 – **Un graphe acyclique dirigé représentant les conditions de validité d'une variable instrumentale.** Les arêtes pointillées représentent les effets qui ne doivent pas exister pour que la variable G soit une variable instrumentale valide de X. Les numéros portés par les arêtes représentent les conditions des variables instrumentales. Les noeuds encadrés représentent des variables observées et les noeuds encerclés des variables latentes.

utilisées sont en LD avec un autre *locus* ayant un effet sur Y (1.2c). Des outils de colocalisation peuvent être utilisés afin de réduire le risque de confusion par le LD [125, 126]. L'analyse de colocalisation permet de confirmer que les variations derrière l'association avec l'exposition sont les mêmes que celles qui sous-tendent l'association avec l'issue. Les effets pléiotropiques pourraient aussi invalider la 2^{e} supposition. On parle de *pléiotropie horizontale* ou de *pléiotropie biologique* lorsqu'un *locus* a des effets sur deux voies biologiques distinctes [127] (Figure 1.2a). Comme ces effets sont fréquents en génétique, il est important d'adresser le risque de biais en évaluant de façon critique le choix des variables instrumentales, en effectuant des analyses de sensibilité et en utilisant des méthodes de MR robustes à l'inclusion d'instruments invalides. La *pléiotropie verticale*, aussi appelée *pléiotropie médiée*, représente un scénario où deux variables sur la même voie causale sont affectées par la variation génétique (Figure 1.2b). Ce type de pléiotropie ne pose pas de problème en MR et correspond simplement à la succession d'effets associés à une modulation génétique. Finalement, la 3^{e} hypothèse exige que l'instrument génétique n'ait pas de cause commune avec l'issue. Comme les variations génétiques sont fixes, la variable la plus susceptible d'invalider cette supposition est la structure de population attribuable à la diversité d'origine génétique dans une population étudiée (Figure 1.2d). Les fréquences alléliques varient d'une population à l'autre selon l'origine géographique et génétique des populations. Bien souvent, l'origine ethnique est aussi corrélée avec des facteurs environnementaux comme l'alimentation ou les mœurs et habitudes de vie qui peuvent avoir un effet sur les issues étudiées en MR.

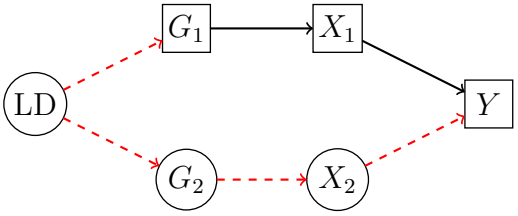
Si ces trois critères sont respectés, l'estimation de l'effet causal entre l'exposition et l'issue est possible (voir [128] pour une revue plus approfondie des estimateurs). L'estimateur par ratio est populaire en MR, car il peut être calculé à partir de statistiques sommaires.



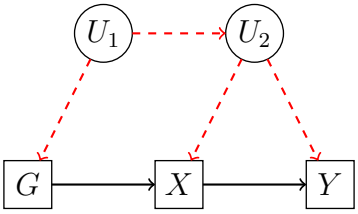
(a) Biais causé par la pléiotropie biologique ou pléiotropie horizontale.



(b) Pléiotropie verticale (pas de biais).



(c) Biais causé par le LD.



(d) Biais causé par l'origine ethnique ancestrale. Ici, U_1 représente l'effet de l'origine géographique et génétique sur les fréquences alléliques. Cette variable est corrélée avec U_2 qui représente des facteurs environnementaux (*p. ex.* l'alimentation).

FIGURE 1.2 – **Graphes acycliques dirigés représentant différentes structures causales plausibles dans les études de MR.** Les vertices encadrés représentent des variables typiquement observées et les vertices encerclée les variables non-observées. Lorsqu'un chemin causal est source de biais dans le contexte d'études de MR, les arêtes sont pointillées. Les arêtes pleines représentent la chemin causal utilisé pour l'inférence MR. Suivant la notation précédemment décrite, les variables G représentent les instruments génétiques ou les variations génétiques non-observées. Les variables X représentent des biomarqueurs complexes et les variables Y représentent l'issue d'intérêt dans le contexte de l'étude MR.

Il s'agit simplement du ratio des coefficients de régression $\hat{\beta}_{MR} = \hat{\beta}_{Y|G}/\hat{\beta}_{X|G}$ estimés par une régression de l'issue sur l'instrument au numérateur et de l'exposition sur l'instrument au dénominateur. La variance de cet estimateur peut être calculée de façon analytique en utilisant l'approximation normale [128] :

$$Var(\hat{\beta}_{MR}) = \sqrt{\frac{\text{se}\left(\hat{\beta}_{Y|G}\right)^2}{\hat{\beta}^2_{X|G}} + \frac{\hat{\beta}^2_{Y|G}\,\text{se}\left(\hat{\beta}_{X|G}\right)^2}{\hat{\beta}^4_{X|G}} - \frac{2\hat{\beta}_{Y|G}\,\text{cov}\left(\hat{\beta}_{X|G},\ \hat{\beta}_{Y|G}\right)^2}{\hat{\beta}^3_{X|G}}}$$

Une autre possibilité, qui ne suppose pas la normalité de l'estimateur, mais qui nécessite un plus grand temps de calcul, est le rééchantillonnage *bootstrap* [129].

Dans une approche dite de MR à deux échantillons (*two sample* MR), les estimateurs par

ratio peuvent être issus de deux jeux de données indépendants. Ceci permet les études de MR à partir de statistiques sommaires publiées par de grands consortiums de GWAS [130]. Dans ce contexte, la variance de l'estimateur par ratio est réduite tel que discuté dans le cadre de l'approche *Summary data-based Mendelian Randomization* (SMR) [131]. Si plusieurs variations génétiques indépendantes sont utilisées en guise de variables instrumentales, un estimateur combiné peut être obtenu en calculant la moyenne des estimés par ratio pondérés par l'inverse de leur variance (*Inverse Variance Weighted*, IVW). Cette méthode est identique à une méta-analyse à effets fixes des estimateurs par ratio. Autrement, dans l'approche GSMR, les estimateurs par ratios sont modélisés comme étant issus d'une variable normale multivariée dont l'espérance est partagée et égale à l'effet causal [70].

Plusieurs méthodes ont été développées permettant d'assouplir les hypothèses imposées aux instruments et d'augmenter la robustesse à l'inclusion d'instruments invalides. L'approche MR-Egger utilise la supposition *Instrument Strength Independant of Direct Effect* (InSIDE) qui exige que les effets directs ($G \to Y$) soient indépendants de l'effet sur l'exposition ($G \to X$) [123]. Cette nouvelle condition est une façon d'assouplir la supposition #2 des variables instrumentales, l'*exclusion restriction*. Un avantage de l'approche MR-Egger est aussi qu'elle permet d'effectuer un test statistique de pléiotropie directionnelle, c'est-à-dire qu'elle permet l'estimation de l'effet direct (effet de G sur Y non médié par l'exposition) moyen de toutes les variations génétiques incluses comme variables instrumentales. En pratique, la supposition InSIDE peut s'avérer stricte et difficile à vérifier et il est possible qu'une telle pléiotropie directionnelle soit rare [132]. De plus l'approche MR-Egger est particulièrement susceptible au biais dans un contexte de MR à deux échantillons [132, 133]. Une autre approche pour combiner plusieurs variables instrumentales est l'approche de la médiane ou de la médiane pondérée [134]. En prenant la médiane des estimateurs par ratio, pondérée par leur précision, l'estimateur obtenu est robuste à l'inclusion de 50% (exclusivement) de variables instrumentales invalides.

De façon plus pragmatique, plusieurs approches tentent d'identifier et de retirer les variations génétiques exhibant un comportement différent des autres. Ces approches s'inspirent de la détection des données aberrantes et supposent que le plus grand sous-groupe de variations génétiques ayant un effet homogène représente l'effet causal. Le mécanisme *HEIDI-outlier*, implémenté dans l'outil GSMR, teste la différence entre l'estimateur causal d'un instrument et un instrument de référence. Si cette différence est trop grande, l'instrument ne participe pas à l'estimation de l'effet causal [70]. De façon similaire, l'approche MR-PRESSO utilise une approche d'exclusion itérative (*leave one out*) et calcule les estimateurs IVW en excluant chaque variation tour à tour [135]. En exploitant cette approche, l'outil MR-PRESSO

propose un test pour la détection de la pléiotropie horizontale, un test de valeur aberrante pour chaque variation génétique et un test de distorsion qui compare l'estimateur omnibus à l'estimateur excluant les variations aberrantes. Une approche connexe est le *contamination mixture model* qui utilise un mélange de modèles à deux composantes [136]. La première composante correspond à l'effet causal et est modélisée par une distribution normale centrée à la vraie valeur de l'effet causal. L'autre composante est utilisée pour modéliser les instruments invalides et la probabilité d'appartenance aux deux composantes peut être utilisée pour identifier les instruments invalides. Finalement, l'approche *Latent Causal Variable* (LCV) offre une approche complémentaire à la MR formelle [137]. Dans cette approche, une variable causale latente est utilisée pour modéliser la corrélation génétique entre deux issues (Y_1 et Y_2). Comme une relation de cause à effet entre Y_1 et Y_2 implique que les variations génétiques influençant Y_1 auront un effet proportionnel sur Y_2 et non l'inverse, l'approche LCV peut distinguer la causalité génétique de la corrélation génétique. Cependant, contrairement aux approches de MR, cette méthode nécessite l'estimation de la corrélation génétique à partir de données pangénomiques ce qui la rend moins propice à l'étude de gènes ou de voies biologiques spécifiques. L'orientation de l'effet causal des effets de causalité entre deux variables peut aussi être déterminée dans un contexte de MR en utilisant la MR bidirectionnelle qui consiste à mener l'étude MR en alternant l'exposition et l'issue. Les variables instrumentales sont alors sélectionnées en fonction de l'exposition considérée [138].

1.4.2 MR des fractions lipidiques et MR multivariable

L'effet des diverses fractions lipidiques (*p. ex.* le cholestérol LDL et le cholestérol HDL) sur le risque de maladie coronarienne est un sujet qui a beaucoup été étudié par les approches de MR. L'effet causal du cholestérol LDL sur la maladie coronarienne a été démontré à la fois dans les études de MR et les essais cliniques [113, 139, 140]. L'étude du cholestérol HDL, par exemple, s'avère plus complexe et varie selon la sélection de l'instrument et de la méthode utilisée. Dans une des premières études sur le sujet, Voight *et coll.* ont remarqué que des variations de la lipase endothéliale (encodée par le gène, *LIPG*) n'étaient pas associées à l'infarctus du myocarde et il en était de même pour un score formé de 14 variations génétiques associées exclusivement avec le cholestérol HDL [141]. Comme il en a été question plus tôt (section 1.1.3), les études de cis-MR ont démontré que la modulation de différentes voies biologiques ayant aussi un impact sur les taux de cholestérol HDL, comme par l'inhibition de la CETP, peut diminuer le risque de maladie coronarienne.

la MR multivariable est une approche permettant de considérer plusieurs expositions candidates conjointement, comme les différentes fractions lipidiques. Elle permet de modéliser

explicitement la pléiotropie horizontale, mais elle est limitée aux expositions connues et observables. Les suppositions des variables instrumentales sont adaptées au contexte multivariable de sorte que les instruments doivent être associés avec au moins un des facteurs de risque (*pertinence*), ils ne doivent pas être susceptibles à des variables de confusion (*indépendance*) et ils doivent être indépendants de l'issue conditionnellement à l'ensemble des expositions (*exclusion restriction*) [142]. Les effets causaux peuvent ensuite être estimés en utilisant une approche adaptée de moindre carrés à deux étapes (*two-stage least squares*) si des données individuelles sont disponibles ou en utilisant une approche basée sur la vraisemblance si des données sommaires sont utilisées. En appliquant l'approche de MR multivariable par vraisemblance à l'étude du cholestérol LDL, du cholestérol HDL et des triglycérides, Burgess *et coll.* ont identifié un effet causal entre le cholestérol LDL et les triglycérides et la maladie coronarienne, tandis que le cholestérol HDL n'avait pas d'effet causal. Cependant, une étude du même groupe a obtenu des résultats divergents en utilisant 185 variations génétiques associées avec le cholestérol LDL, le cholestérol HDL ou les triglycérides dans le GWAS du *Global Lipid Genetics Consortium*. L'effet causal multivariable du cholestérol HDL sur la maladie coronarienne devenait alors significatif ($P = 0.008$) et robuste à l'exclusion de variations associées à la pression artérielle ou à l'indice de masse corporel ($P = 0.027$) [143].

Ce sujet est encore un important secteur de recherche et le choix de la méthode et des expositions concurrentes ont un grand impact sur les résultats des études de MR [144]. En utilisant des modèles de simulation, Burgess *et coll.* ont souligné que la méthode de MR multivariable permet d'estimer l'effet direct des facteurs de risque sur l'issue, et non l'effet total qui inclut les effets médiés par la structure de causalité entre les facteurs de risque. Ainsi, l'effet estimé par MR multivariable correspond à l'effet d'une intervention sur l'exposition en maintenant les autres expositions constantes, soit l'effet direct contrôlé (*controlled direct effect*).

1.4.3 MR des cibles pharmacologiques

Dans la section précédente, nous nous sommes concentrés sur l'application de la MR à des questions de recherche portant sur l'estimation de l'effet d'une exposition. Dans le cas de l'étude de cibles pharmacologiques, on s'intéresse cependant à une façon bien précise de modifier l'exposition. Par exemple, une approche MR pour prédire l'effet causal de réduire le cholestérol LDL par l'inhibition de la PCSK9 pourrait restreindre les instruments utilisés à des variations génétiques situées seulement dans la région du gène en question. Ou encore, l'exposition d'intérêt peut être limitée à l'expression du gène en question, ou à la concentration plasmatique de son produit. Quand l'exposition est intimement liée au gène d'intérêt,

et que l'instrument utilisé est restreint au *locus* de ce dernier, on parle de randomisation mendélienne en *cis* (cis-MR) [145]. La Figure 1.3, tirée de Holmes *et coll.* [5], illustre la distinction entre une étude de MR qui vise à établir l'effet d'un biomarqueur complexe et une étude de cis-MR qui vise à établir l'effet spécifique à une cible pharmacologique.

En général, les approches cis-MR utilisent les mêmes estimateurs et méthodes que la MR conventionnelle. Cependant, comme les variations génétiques sont restreintes à la région d'un seul gène d'intérêt, il est habituellement souhaitable d'utiliser des méthodes adaptées à l'inclusion de variations génétiques en LD [146]. L'utilisation de ces méthodes nécessite tout de même une sélection des variations génétiques pour s'assurer qu'elles ont un effet fort sur l'exposition et pour éviter la multicolinéarité. Shmidt *et coll.* ont comparé des modèles restrictifs incluant des variations génétiques sur la base d'annotations fonctionnelles à des modèles libéraux qui incluaient plusieurs variations en LD [145]. Pour éviter la multicolinéarité dans ces modèles libéraux, une étape d'agrégation par LD était utilisée comme unique mécanisme de sélection des variations. Le seuil de $r^2 = 0.5$ donnait de bons résultats pour les gènes étudiés. Les estimateurs ponctuels étaient en général similaires entre les deux approches, mais les estimateurs des modèles plus inclusifs étaient plus précis. Il est cependant plausible que ces résultats dépendent de l'architecture génétique sous-jacente, notamment du nombre de variations causales dans la région génétique ciblée et de la structure de LD. Une autre autre approche consiste à utiliser une PCA de la matrice de covariance pondérée des génotypes afin de combiner les variations génétiques en fonction à la fois du LD et de la taille de l'effet sur l'exposition [146, 147]. L'estimateur PCA-IVW a été conçu pour estimer l'effet causal dans ce contexte. Cette méthode a une puissance comparable aux approches conventionnelles tout en étant plus robuste au choix de paramètres (*p. ex.* le seuil de r^2 dans l'approche de regroupement) [146, 147]. Il est aussi possible de combiner des variations génétiques en un score qu'on utilise comme variable instrumentale [128, 148]. L'inconvénient de cette approche est qu'il devient impossible de détecter des variations pléiotropiques dont l'effet est aberrant, mais ce risque est réduit dans un contexte de cis-MR.

Lorsqu'une cis-MR s'intéresse à l'effet d'une cible pharmacologique, on parle alors de MR de cible pharmacologique. Un exemple d'une telle analyse est celui de l'étude de *HMGCR*, la cible des statines et de *NPC1L1* la cible de l'ezetimibe menée par Ference *et coll.* [3]. Les statines inhibent la HMG CoA reductase, une enzyme responsable d'une étape limitante de la biosynthèse des cholestérols, et l'ezetimibe inhibe la *Niemann-Pick C1-Like 1 protein* qui est responsable de l'absorption intestinale des cholestérols. Les instruments génétiques utilisés dans l'étude étaient des scores composés de variations génétiques aux *loci* de *HMGCR* et *NPC1L1* et pondérés pour leur effet sur le cholestérol LDL. La sélection des variations était

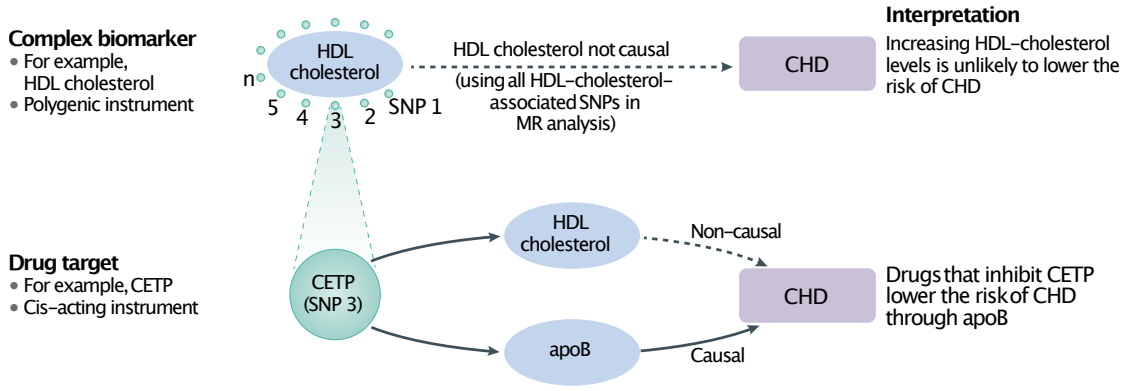


FIGURE 1.3 – **Illustration comparant une étude de MR visant à estimer l'effet du cholestérol HDL sur la maladie coronarienne à une étude de cis-MR de la cible pharmacologique CETP (Holmes *et coll.* [5])** [3]. Bien qu'illustrant bien le contraste ente une étude de cis-MR et une étude de MR axées sur l'étude de biomarqueurs, cette illustration ne reflète pas l'ensemble de la complexité de la modulation de la CETP. La Section 1.1.3 et le Chapitre 3 s'intéressent à cette question en profondeur.

basée sur l'approche de seuil de valeur-p et regroupement par LD. Au final, 5 variations ont été sélectionnées pour le score de *NPC1L1* contre 3 variations pour le score de *HGMCR*. La population de l'étude a ensuite été divisée en quatre groupes formés en séparant les scores au niveau de la médiane. Ces groupes ont été utilisés pour représenter des modèles génétiques d'utilisateurs de statines, d'ezetimibe ou de la combinaison des deux (Table 1.4). Le choix de séparer les groupes au niveau de la médiane est intuitif et permet une analogie facile avec la structure d'un essai randomisé, mais cette méthode n'est pas la plus puissante et il est préférable de conserver les scores sur l'échelle continue [149]. Comparativement au groupe de référence, le groupe du modèle génétique d'utilisateurs de statines avait une réduction moyenne de 2.9 (IC 95% 2.4, 3.4) mg/dl de cholestérol LDL. Pour une même réduction du cholestérol LDL, l'effet était similaire chez les individus du groupe du modèle génétique de l'ezetimibe. Dans le groupe correspondant à la combinaison des médicaments (les deux scores supérieurs à la moyenne), la réduction de cholestérol LDL était de 5.8 (IC 95% 5.3, 6.3) mg/dl ce qui est cohérent avec un effet additif des deux scores. L'effet sur la maladie coronarienne était concordant et proportionnel à la réduction de cholestérol LDL avec un OR = 0.89 (IC 95% 0.85, 0.93) dans le groupe représentant la combinaison statine + ezetimibe.

Cette étude a été publiée très peu de temps après le dévoilement des résultats de l'essai randomisé *The Improved Reduction of Outcomes : Vytorin Efficacy International Trial*

3. Reprinted by permission from Springer Nature Customer Service Centre GmbH : Springer Nature, Nature Reviews Cardiology. *Integrating genomics with biomarkers and therapeutic targets to invigorate cardiovascular drug development*, MV Holmes, TG Richardson, BA Ference, NM Davies and GD Smith. Copyright 2021

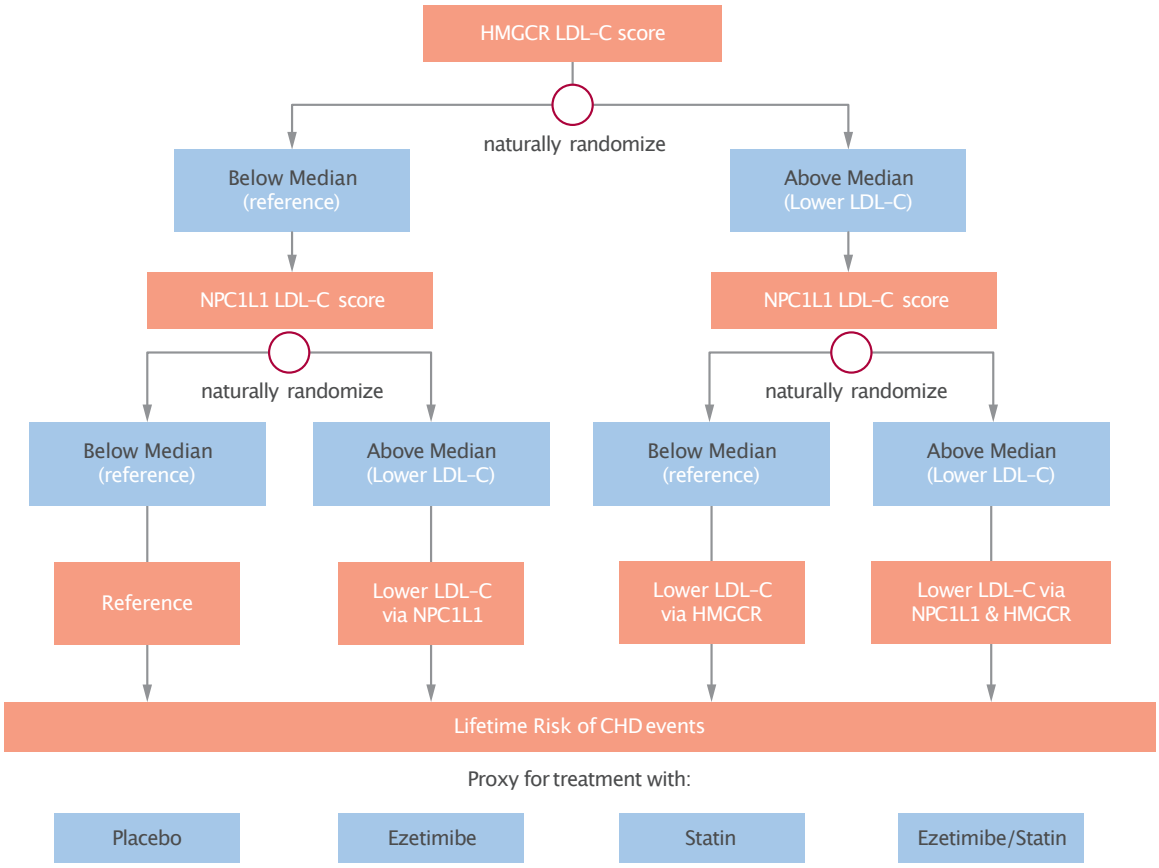


FIGURE 1.4 – **Groupes de comparaison suivant le devis de l'étude de randomisation mendélienne factorielle** 2×2 **(Ference *et coll.* [3]).** Le but de l'étude était de prédire l'effet de l'ezetimibe, des statines et de leur combinaison sur le cholestérol LDL et les évènements coronariens. Les scores sont des scores génétiques composés de 3 variations pour *HMGCR* (modèle pour les statines) et 5 variations pour *NPC1L1* (modèle pour l'ezetimibe). Ils sont pondérés pour l'effet des variations sur le cholestérol LDL et en encodant l'allèle réduisant le cholestérol LDL de sorte qu'une augmentation du score prédit une réduction du cholestérol LDL. [4]

(IMPROVE-IT) évaluant le bénéfice de l'ajout d'ezetimibe aux statines pour prévenir les évènements coronariens et dont les résultats sont concordants [150]. Le groupe du Dr Ference a mené plusieurs études de validation de cibles pharmacologiques en utilisant une approche de MR incluant les gènes *CETP*, *PCSK9*, *ACLY* et *LPA* [18, 51, 151-153]. Les études de MR de cibles pharmacologiques ne sont pas directement comparables avec les essais cliniques à cause de différences au niveau de la durée de l'exposition (toute la vie dans le cas de la MR) et de la nature des populations à l'étude. Par exemple, les essais randomisés d'inhibiteurs de la PCSK9 portent souvent sur des individus atteints de la maladie coronarienne et avec un

4. Reprinted from JACC, 65 / 15, BA. Ference, F. Majeed, R. Penumetcha, JM. Flack, RD. Brook, Effect of Naturally Random Allocation to Lower Low-Density Lipoprotein Cholesterol on the Risk of Coronary Heart Disease Mediated by Polymorphisms in NPC1L1, HMGCR, or Both A 2 × 2 Factorial Mendelian Randomization Study, 1552-1561, Copyright (2015), with permission from Elsevier.

taux de LDL-c élevé, ce qui diffère des études populationnelles souvent utilisées en génétique [18, 19]. Néanmoins, la direction des effets est habituellement concordante avec les observations cliniques malgré ces différences. Certains auteurs ont d'ailleurs rapporté la relation de proportionnalité entre les estimations issus d'essais randomisés et de MR afin de quantifier le bénéfice clinique attendu de médicaments [3, 5, 8, 151]. De nouvelles méthodes d'inférence causale étudient aussi la possibilité de combiner des données observationnelles et expérimentales afin d'accroître la précision des estimés causaux, ce qui pourrait éventuellement permettre une plus grande réconciliation des effets de MR [154].

L'approche de MR peut aussi être utilisée pour identifier de nouvelles cibles pharmacologiques et l'émergence de plate-formes de protéomique à haut débit favorise ces études. Dans une étude phare, Sun *et coll.* ont quantifié l'abondance de 3 622 protéines plasmatiques chez 3 301 participants sains de l'étude INTERVAL [155]. En testant l'association entre les variations génétiques et les niveaux plasmatiques de protéines, les auteurs ont identifié plusieurs *loci* associés aux niveaux protéiques (*protein Quantitative Trait Loci*, pQTLs). Les pQTLs qui modifiaient les niveaux de la protéine RANK étaient aussi associés avec la maladie de Paget, une rare maladie des os, ce qui suggère que cette protéine pourrait être une cible pharmacologique pour le traitement de cette maladie. Dans une approche de MR à deux échantillons, l'étude de Zheng *et coll.* a par la suite combiné les statistiques sommaires de 5 GWAS de pQTLs afin d'évaluer de façon systématique l'effet de 1 002 protéines plasmatiques sur 153 maladies et 72 facteurs de risque [156]. Cette étude a révélé 238 relations causales entre des protéines plasmatiques et des traits humains. Les protéines identifiées incluent la *urokinase-type plasminogen activator* qui diminue le risque de maladie inflammatoire de l'intestin avec MR OR = 0.75 (IC 95% 0.69, 0.83). Une formulation injectable de cette enzyme, kinlytic (urokinase), a initialement été développée pour la thrombolyse lors d'infarctus ou d'Accident Vasculaire Cérébral (AVC) et il reste à déterminer si un repositionnement sera possible en fonction de ses paramètres pharmacocinétiques et du ratio risque-bénéfice. L'étude de pQTL situés près de leur protéine d'intérêt pose un moins grand risque de pléiotropie horizontale ce qui favorise les études de MR, et de telles approches ont un grand potentiel pour la découverte de cibles pharmacologiques.

1.4.4 Le biais de l'instrument de faible

Le biais de l'instrument faible est un problème important dans les études de MR [157]. Lorsque l'association entre la variable instrumentale et l'issue est faible, l'effet estimé par MR sera biaisé en direction de l'association observationnelle entre l'exposition et l'issue. Cette situation est facile à reconnaître en pratique, mais il existe une manifestation plus

pernicieuse si plusieurs variables instrumentales sont utilisées. Le biais peut alors survenir si l'estimation de l'effet des variations génétiques sur l'exposition se fait dans un contexte de MR à un échantillon (*one sample* MR), c'est-à-dire si tous les effets sont estimés au sein du même jeu de données. Intuitivement, ce phénomène est dû au risque accru que les facteurs de confusion, par hasard, aient une distribution inégale à travers les niveaux des différentes variables instrumentales. Autrement dit, l'ajout de variations génétiques de faible effet au modèle peut induire un surapprentissage (*overfitting*) au niveau de l'association avec l'exposition à cause de l'estimation dans un échantillon statistique de taille finie [148, 157]. Il faut donc être prudent face à ce risque de surapprentissage et s'assurer de la force des variables instrumentales utilisées.

1.5 Introduction de la thèse

L'objectif global de cette thèse est d'améliorer notre capacité à prédire l'effet de la modulation de cibles pharmacologiques en utilisant des techniques modernes en génétique humaine. Le domaine de la validation génétique de cibles pharmacologiques n'est pas nouveau [158], mais il connaît un essor récent et un intérêt renouvelé dû à de nouveaux développements méthodologiques et de récents succès dans le développement de médicaments appuyés par des données génétiques [5].

L'objectif de l'étude présentée au Chapitre 2 est de réconcilier les modèles génétiques de l'inhibition du canal ionique cardiaque HCN4 (cible de l'ivabradine) avec les résultats des études cliniques de cette molécule. Ces études ont révélé un profil clinique complexe caractérisé par un risque accru de fibrillation auriculaire combiné à un bénéfice sur les symptômes angineux et sur l'insuffisance cardiaque. Notre approche génétique permet de récapituler ces résultats en plus de mettre en lumière les limites d'une approche génétique en illustrant des différences entre l'effet pharmacologique de l'ivabradine et l'effet de mutations de la cible.

Le 3^{e} chapitre a pour objectif d'identifier des modificateurs de l'effet des inhibiteurs de la CETP à l'aide de modèles génétiques. Au sens plus large, cette approche vise à développer une nouvelle application des modèles de validation des cibles pharmacologiques pour prédire les déterminants cliniques de l'effet des médicaments. Peu d'études se sont penchées sur cette question importante qui apporte un outil prometteur pour étudier la validité externe d'essais randomisés. Ce chapitre représente une contribution importante à la fois pour la caractérisation d'interactions dans le contexte de l'étude de cibles pharmacologiques et au niveau de notre compréhension des effets de la CETP.

Le 4^{e} chapitre de cette thèse décrit une approche statistique et une importante base de

données d'associations entre des gènes encodant des protéines et une grande diversité de phénotypes. Bien que l'intérêt de ce portail dépasse l'étude des cibles pharmacologiques, la présentation des résultats illustre plusieurs exemples d'associations coïncidant avec des gènes de cibles pharmacologiques. Le portail permet aussi de visualiser l'enrichissement de gènes associés avec un phénotype et les cibles pharmacologiques par classe de médicaments. Cette fonctionnalité pourrait être utile pour guider le repositionnement de médicaments. En contraste avec les deux chapitres précédents qui étudient en profondeur des cibles pharmacologiques bien précises, cette contribution vise à appuyer le développement de médicament en démocratisant l'utilisation de bases de données génétiques pour la validation de cibles pharmacologiques.

Finalement, le 5^{e} chapitre de la thèse décrit des contributions scientifiques additionnelles découlant de mes études doctorales. Notamment, il est question d'un article de revue sur la pharmacogénomique des médicaments ciblant les lipides plasmatiques. Cette revue décrit à la fois nos connaissances actuelles sur la génétique de la réponse à ces médicaments et de nouvelles données en lien avec l'utilisation de scores polygéniques pour définir des sous-groupes d'individus susceptibles de bénéficier de façon accrue de la prise de médicaments. Le 5^{e} chapitre présente aussi une contribution bio-informatique d'intérêt général pour la création de scores de risque génétique. Cet outil a été utilisé abondamment pour mener les études présentées dans cette thèse afin de créer des modèles génétiques de l'inhibition de cibles pharmacologiques.

En conclusion, cette thèse aborde plusieurs aspects de l'utilisation de modèles génétiques pour la validation de cibles pharmacologiques. Elle considère à la fois des exemples précis avec l'ivabradine et les inhibiteurs de la CETP ainsi que des approches générales et d'intérêt plus large qui pourront être utilisées pour la validation de cibles émergentes.

CHAPITRE 2

A genetic model of ivabradine recapitulates results from randomized controlled trials

L'ivabradine est un médicament efficace pour atténuer les symptômes angineux et pour prévenir les événements cardiovasculaires chez les patients avec insuffisance cardiaque. Ce médicament inhibite le canal ionique HCN4 qui permet la dépolarisation spontanée du cœur et sa contraction rythmique. L'inhibition de ce canal par l'ivabradine réduit donc la fréquence cardiaque de façon spécifique sans affecter d'autres paramètres cardiovasculaires comme la tension artérielle ou la contractilité cardiaque. Sous l'hypothèse qu'une réduction de la fréquence cardiaque réduit les besoins énergétiques du myocarde tout en favorisant la perfusion durant la diastole, l'ivabradine était pressentie pour prévenir les événements ischémiques chez les patients coronariens stables sans dysfonction systolique. Cette hypothèse a été évaluée dans l'essai randomisé SIGNIFY, une grande étude de phase 3 incluant 19 102 patients. Malgré le bénéfice du médicament pour traiter les patients souffrant d'angine et d'insuffisance cardiaque, aucun bénéfice n'a été observé dans la population de patients avec maladie coronarienne. De plus, il a été noté dans plusieurs essais randomisés que les utilisateurs de l'ivabradine présentaient une légère augmentation du risque de fibrillation auriculaire comparé aux groupes placébos. Les données cliniques suggèrent cependant que ce risque n'a pas de conséquences sur la sécurité de l'ivabradine et n'augmente pas le risque d'accidents vasculaires cérébraux.

À la lumière de ces résultats, nous avons construit un modèle génétique de l'ivabradine que nous avons tenté de réconcilier avec les données des essais randomisés. Comme le gène *HCN4* est court et très intolérant aux mutations, il y avait peu de variations génétiques disponibles pour construire la variable instrumentale. Nous avons choisi une seule variation

génétique (rs8038766), sélectionnée pour sa forte association avec la fréquence cardiaque, un phénotype *proxy* de l'activité du canal HCN4. Bien que son effet soit modeste (réduction de 0.57 battement par minute), l'association statistique était forte ($P = 2.76 \times 10^{-66}$) grâce à la grande taille de l'étude qui incluait 413 083 participants de la UK Biobank. Notre étude a démontré qu'une diminution de la fréquence cardiaque par la réduction prédite de l'activité de HCN4 était associée à une augmentation du risque de fibrillation auriculaire et d'accident vasculaire cérébral ischémique récapitulant les principales préoccupations de sécurité soulevées dans les essais randomisés. Nous avons noté toutefois qu'il est possible que la fibrillation auriculaire augmente aussi le risque de développer l'insuffisance cardiaque. Nous avons estimé l'effet de rs8038766 sur l'insuffisance cardiaque indépendamment de l'augmentation attendue du risque médiée par la fibrillation auriculaire. L'estimation de cet effet conditionnel est pertinente afin de réconcilier le modèle génétique avec les résultats des essais randomisés où le risque de fibrillation auriculaire, pour une même réduction de fréquence cardiaque, était plus faible. De plus, plusieurs stratégies peuvent être utilisées cliniquement afin de contrôler le risque de fibrillation auriculaire et pourraient être envisagées si un bénéfice sur l'insuffisance cardiaque est attendu. Pour modéliser cet effet, nous avons opté pour l'utilisation de modèles de risque en compétition. L'approche utilisée modélise l'incidence de l'insuffisance cardiaque chez les participants de la UK Biobank qui ne souffraient ni d'insuffisance cardiaque, ni de la fibrillation auriculaire au recrutement dans l'étude. L'utilisation d'un modèle prospectif permet d'estimer l'effet de rs8038766 sur l'insuffisance cardiaque avant qu'une médiation par la fibrillation auriculaire ne soit possible (c'est-à-dire avant le développement de cette maladie). Il est improbable que cette approche soit aussi susceptible au biais du collisionneur qu'une étude de sous-groupe dans un devis transversal, mais une étude de simulation formelle serait nécessaire pour mieux quantifier le biais associé aux deux méthodes. L'approche mtCOJO, introduite à la Section 1.2.1 et qui est robuste au biais du collisionneur, a aussi été utilisée et a permis de corroborer l'effet conditionnel protecteur de rs8038766 sur l'insuffisance cardiaque qui n'était pas apparent dans les études conventionnelles [70]. Nous avons aussi utilisé une approche de MR bidirectionnelle qui démontre que la fibrillation auriculaire et l'insuffisance cardiaque ont une relation de causalité mutuelle. Un bénéfice robuste de la réduction génétiquement prédite de la fréquence cardiaque par l'ivabradine n'a pas été observé pour la maladie coronarienne en utilisant une approche similaire. Ensemble, ces résultats récapitulent les résultats principaux observés dans les essais cliniques SHIFT et SIGNIFY. Notre étude est la première à utiliser des modèles de risques en compétition dans les études génétiques visant à valider des cibles pharmacologiques et démontre l'intérêt d'utiliser une méthodologie adaptée et sensible aux comorbidités des populations de patients ciblés.

Contributions : La concept de cette étude a été développé par Marc-André Legault,

Simon de Denus, Benoit Tyl, Jean-Claude Tardif et Marie-Pierre Dubé. Johanna Sandoval et Sylvie Provost ont contribué à préparer les données et au contrôle de la qualité des données génétiques et cliniques de la UK Biobank. Marc-André Legault, Jean-Claude Tardif et Marie-Pierre Dubé ont contribué au développement et à l'implémentation des algorithmes diagnostics utilisés pour définir les issues cardiovasculaires dans la UK Biobank. La méthodologie a été développée par Marc-André Legault, Sylvie Provost, Amina Barhdadi, Louis-Philippe Lemieux Perreault et Marie-Pierre Dubé. Louis-Philippe Lemieux Perreault et Marc-André Legault ont contribué au développement des logiciels utilisés. La production de graphiques et les analyses statistiques ont été réalisées par Marc-André Legault. La supervision de ce projet a été faite par Marie-Pierre Dubé. Tous les auteurs ont lu et contribué à l'écriture du manuscrit et les contributions majeures ont été faites par Marc-André Legault, Marie-Pierre Dubé et Jean-Claude Tardif.

A genetic model of ivabradine recapitulates results from randomized controlled trials

PloS One 2020, 15 (7): e0236193. doi:10.1371/journal.pone.0236193

Marc-André Legault[1,2,3], Johanna Sandoval[1,3], Sylvie Provost[1,3], Amina Barhdadi[1,3], Louis-Philippe Lemieux Perreault[1,3], Sonia Shah[4,5], R.Thomas Lumbers[6,7,8], Simon de Denus[1,9], Benoit Tyl[10], Jean-Claude Tardif[1,11], Marie-Pierre Dubé[1,3,11]

1. Montreal Heart Institute, Montreal, Canada;
2. Université de Montréal, Department of biochemistry and molecular medicine, Montreal, Canada
3. Université de Montréal Beaulieu-Saucier Pharmacogenomics Centre, Montreal, Canada;
4. Institute for Molecular Bioscience, The University of Queensland, Brisbane, Queensland, 4072, Australia
5. Institute of Cardiovascular Science, University College London, UK
6. Institute of Health Informatics, University College London, UK
7. Health Data Research UK London, University College London, UK
8. Bart's Heart Centre, St. Bartholomew's Hospital, London, UK
9. Faculty of Pharmacy, Université de Montréal, Montreal, Canada
10. Cardiovascular Center for Therapeutic Innovation, Institut de Recherches Internationales Servier, Suresnes, France
11. Department of medicine, Université de Montréal, Montreal, Canada.

2.1 Abstract

Background. Naturally occurring human genetic variants provide a valuable tool to identify drug targets and guide drug prioritization and clinical trial design. Ivabradine is a heart rate lowering drug with protective effects on heart failure despite increasing the risk of atrial fibrillation. In patients with coronary artery disease without heart failure, the drug does not protect against major cardiovascular adverse events prompting questions about the ability of genetics to have predicted those effects. This study evaluates the effect of a variant in HCN4, ivabradine's drug target, on safety and efficacy endpoints.

Methods. We used genetic association testing and Mendelian randomization to predict the effect of ivabradine and heart rate lowering on cardiovascular outcomes.

Results. Using data from the UK Biobank and large GWAS consortia, we evaluated the effect of a heart rate-reducing genetic variant at the *HCN4* locus encoding ivabradine's drug target. These genetic association analyses showed increases in risk for atrial fibrillation (OR 1.09, 95% CI: 1.06-1.13, P=9.3×10^{-9}) in the UK Biobank. In a cause-specific competing risk model to account for the increased risk of atrial fibrillation, the *HCN4* variant reduced incident heart failure in participants that did not develop atrial fibrillation (HR 0.90, 95% CI: 0.83-0.98, P=0.013). In contrast, the same heart rate reducing *HCN4* variant did not prevent a composite endpoint of myocardial infarction or cardiovascular death (OR 0.99, 95% CI: 0.93-1.04, P=0.61).

Conclusion. Genetic modelling of ivabradine recapitulates its benefits in heart failure, promotion of atrial fibrillation, and neutral effect on myocardial infarction.

Keywords: Mendelian randomization; HCN4; SHIFT; SIGNIFY

2.2 Condensed Abstract

The effects of drugs can sometimes be predicted from the effects of mutations in genes encoding drug targets. We tested the effect of a heart rate reducing allele at the *HCN4* locus encoding ivabradine's drug target and found results coherent with the SHIFT and SIGNIFY clinical trials of ivabradine. The genetic variant increased the risk of atrial fibrillation and cardioembolic stroke and protected against heart failure in a competing risk model accounting for the increased risk of atrial fibrillation. The variant had a neutral effect on a composite of myocardial infarction and cardiovascular death.

2.3 Introduction

Human genetics can be a powerful tool to guide drug development. The identification of mutations in important coronary artery disease associated genes has led to the development of new drugs and the approach of Mendelian randomization (MR) is widely used to predict the effect of interventions on biomarkers [159], to validate drug targets and to predict the effect of drug combinations [3]. There are limitations, however, in the value of human genetics to predict the effects of drugs. The main problems are caused by pleiotropic effects of genetic variants [160], the difference between a lifelong exposure to a risk factor and interventions that are administered after disease onset [3] and the generalizability of the results to specific patient populations and to different ethnic populations [124].

Here, we investigate whether human genetics can reproduce the diverging results obtained on different clinical outcomes in randomized clinical trials of ivabradine. This heart-rate lowering drug was demonstrated to reduce the composite of cardiovascular death and hospitalization for worsening heart failure in patients with symptomatic heart failure and a heart rate above 70 bpm (beats per minute) at baseline in the SHIFT trial [25]. In this study, there was a placebo-adjusted reduction in heart rate of 10.9 (10.4, 11.4) bpm after 28 days on treatment with ivabradine and the hazard ratio (HR) for the cardiovascular composite endpoint was 0.82 (95% CI: 0.75-0.90, p<0.0001). In contrast, in the SIGNIFY trial, ivabradine did not reduce the composite of cardiovascular death or myocardial infarction in patients with stable coronary artery disease (CAD) without heart failure and with a heart rate > 70 bpm at baseline (HR 1.08, 95% confidence interval (CI): 0.96-1.20, P=0.20) [27]. In both studies, there was an increase in the risk of atrial fibrillation in patients randomized to ivabradine. The incidence of atrial fibrillation was 9% and 8% in the ivabradine and placebo arms in SHIFT (P=0.012) respectively, whereas it was 5.3% and 3.8% in SIGNIFY (Supplementary Table A.1). The heart rate reduction induced by ivabradine is due to the inhibition of the "funny" current (I_f), which is important for cardiac depolarization during phase 4 of the action potential in the sino-atrial node [161]. The hyperpolarization-activated cyclic nucleotide-gated channel 4 encoded by the *HCN4* gene is responsible for this current [162]. Here, we study naturally occurring variants at this locus as a genetic model of ivabradine therapy to predict the effects of the drug on heart failure, atrial fibrillation and CAD.

2.4 Methods

2.4.1 Data sources

The UK Biobank is a prospective population cohort of over 500,000 individuals aged between 40 and 69 at recruitment, and has been previously described [106]. We used hospitalization data between the beginning of the Health Episode Statistics (HES) linkage (April 1st 1997) and the last available date for the current data release (March 1st 2016). Codes used to define the clinical variables are presented in Supplementary Table A.2. All UK Biobank participants were previously genotyped. We applied genetic quality control leaving 413,083 individuals for analysis (Supplementary Methods A.1.1). All reported genomic positions are reported with respect to build GRCh37. We also used the largest available meta-analysis of genome-wide association studies (GWAS) with summary statistics reporting the effect of the *HCN4* rs8038766 variant on stroke, atrial fibrillation, CAD, myocardial infarction and heart failure (Supplementary Methods A.1.1). [62, 163–165]. All participants

of the UK Biobank gave their informed consent and the present study was approved by the institutional ethics review board of the Montreal Heart Institute.

2.4.2 Statistical analyses

To identify independent variants at the *HCN4* locus (chr15:73,612,200-73,661,605 ± 200kb) associated with resting heart rate at baseline in the UK Biobank dataset, we used forward stepwise linear regression with additive allele coding and a genome-wide significance threshold ($p \leq 5.0 \times 10^{-8}$). Association between *HCN4* variant rs8038766 and clinical endpoints was assessed using multivariable logistic regression. All models were adjusted for age, sex and the first 10 principal components. For the prospective and competing risk analyses we used Cox proportional-hazards regression. For the competing risk analyses, we estimated the cause-specific hazards where individuals are censored at the time of occurrence of the competing risk if it occurred prior or if it was reported at the same time as the event of interest [166]. We used time from the first baseline visit in years as the timescale and the censure was the date of death or end of follow-up period. For the construction of the heart rate Genetic Risk Score (GRS), we used 64 previously reported genome-wide significant heart rate associated SNPs (with $r^2 < 0.1$) [167]. We split the participants based on the GRS quintiles with the group formed by the 5^{th} GRS quintile (and above) corresponding to the higher heart rate group and the odds ratio for CAD, heart failure and atrial fibrillation were obtained by comparing the first 4 groups individually to the 5^{th} group used as reference in logistic regression. All analyses were performed using the R (v.3.5.2) programming language unless otherwise specified.

2.4.3 Mendelian randomization

We used the inverse variance weighted (IVW) [168], MR-Egger [123], contamination mixture [136] and Mendelian Randomization Pleiotropy RESidual Sum and Outlier (MR-PRESSO) methods [135]. We present results of all four methods in order to outweigh the drawbacks of individual approaches and help guide conclusions, and we report all causal effect estimates for a standard deviation reduction in heart rate as measured in the UK Biobank (11.1 bpm). Coincidentally, this heart rate reduction is similar to the effect of ivabradine in randomized clinical trials (*e.g.* 10.9 bpm in the SHIFT trial) and is comparable in magnitude to the heart rate reduction by ivabradine. For the heart rate GRS, we used the two-stage method with individual level data and the effect estimates are for a 11.1 bpm reduction in heart rate as well [128]. Analyses were performed with the "MendelianRandomization" R package and MR-PRESSO [135]. Refer to Supplementary Methods A.1.1 for additional

details.

2.5 Results

2.5.1 Genetic model of ivabradine

To construct a genetic model of ivabradine treatment, we tested the association between variants at the *HCN4* locus (defined as the gene boundaries ± 200 kb) and heart rate in the UK Biobank by stepwise forward regression analysis. Two independent signals were identified, the first was led by rs8038766 with every copy of the "G" allele reducing heart rate by 0.57 bpm (95% CI 0.51-0.64, $P = 2.76 \times 10^{-66}$). This variant is in high linkage disequilibrium (LD) with variants previously associated with resting heart rate, heart rate variability traits and atrial fibrillation (Supplementary Table A.3). The second variant was rs3743496 with the "T" allele reducing heart rate by 0.30 bpm (95% CI 0.25-0.35, $P = 3.96 \times 10^{-30}$). The region spanned by the association signal led by this variant was wide and overlapped more of the neighbouring gene (*NEO1*), than *HCN4* (Supplementary Figure A.1). Furthermore, the lead variant was in linkage equilibrium with rs8038766 ($D' = 0.43$ and χ^2 test of independence p-value < 0.0001 in 1000 Genomes phase III Europeans) suggesting that the secondary association signal could in fact not be independent of the first. For these reasons, we were not confident that rs3743496 could be used as a specific and independent genetic instrument of *HCN4* activity and excluded it from the genetic model of ivabradine. Additionally, *HCN4* is a short gene (49,405 bases) that is intolerant to loss-of-function mutations (probability of being loss of function intolerant of 1 in the gnomAD database) [169] which could explain the scarcity of functional variants to be used as genetic instruments.

2.5.2 Genetically predicted effect of ivabradine on safety endpoints

We tested the association between the heart rate reducing allele of the *HCN4* variant rs8038766 and atrial fibrillation and stroke, a common and well-known consequence of atrial fibrillation. *HCN4* has previously been implicated in atrial fibrillation, and we replicated these results using rs8038766 [23]. In the UK Biobank, rs8038766 was strongly associated with atrial fibrillation (OR 1.09, 95% CI 1.06-1.13; $P = 9.3 \times 10^{-9}$) but not with any stroke or ischemic stroke (Figure 2.1). The association between rs8038766 and atrial fibrillation was also observed in summary statistics from previously published GWAS of atrial fibrillation with OR=1.11 ($P = 1.8 \times 10^{-26}$), and 1.12 ($P = 5.4 \times 10^{-35}$) for Roselli *et al.* and Nielsen

et al. respectively (Figure 2.1) [164, 170]. Previous epidemiologic studies have shown that chronic atrial fibrillation leads to a five-fold increase in the risk of stroke [171]. We did not find a significant association between rs8038766 and stroke in the UK Biobank, potentially because of the low number of cases (4,158 cases for ischemic stroke). Summary results from the MEGASTROKE consortium show an association between rs8038766 and cardioembolic (OR=1.08, 95% CI 1.03-1.13, $P = 1.54 \times 10^{-3}$) and ischemic stroke (OR=1.03, 95% CI 1.01-1.05, P=0.0152) (Figure 2.1) [172].

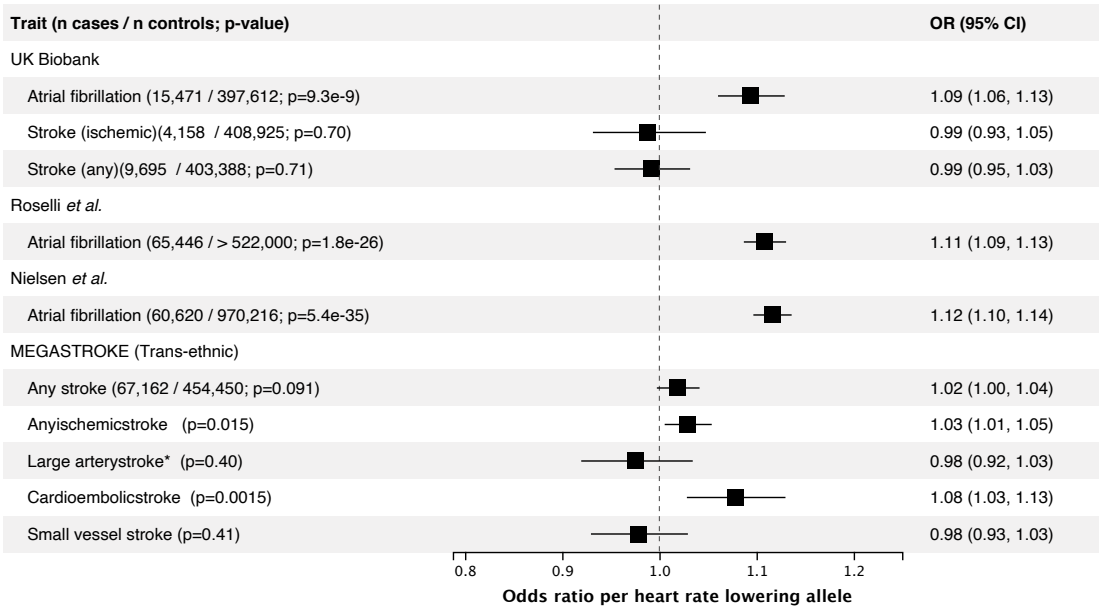


Figure 2.1 – **Association between the heart rate lowering allele (G) of the *HCN4* variant rs8038766 and safety outcomes in the UK Biobank and in published GWAS summary statistics from large consortia.** For the UK Biobank, reporting results from logistic regression comparing the combined prevalent and incident cases to non-cases. References: Roselli *et al.* [170], Nielsen *et al.* [164], MEGASTROKE [172].
* rs7174098 (LD $r^2 = 1$ in 1000 Genomes Europeans) was used instead of rs8038766.

2.5.3 Genetically predicted effect of ivabradine on efficacy endpoints

We tested for association of the heart rate-reducing allele at the *HCN4* variant rs8038766 with combined prevalent and incident heart failure in the UK Biobank and found a non-significant trend for a protective effect (OR = 0.96, 95% CI 0.91-1.00, p=0.071) (Figure 2.2). However, because atrial fibrillation is an important risk factor for heart failure [173], it is a possible that the increased risk of atrial fibrillation attenuates a possible protective

association with heart failure. Indeed, after adjustment for any prevalent or incident atrial fibrillation, the association of rs8038766 with heart failure was OR=0.91, 95% CI 0.87-0.96 ($P = 6.7 \times 10^{-4}$). In a model including the interaction term between rs8038766 and atrial fibrillation, the estimated OR of the variant on heart failure was OR=0.87, 95% CI 0.81-0.94, ($P = 0.00011$) and the interaction term OR was 1.12, 95% CI 1.00-1.24 ($P = 0.04$). However, these associations could be biased if both the SNP and the outcome increase atrial fibrillation risk resulting in a possible collider bias. To account for this, we used a cause-specific hazards model for the incidence of heart failure and atrial fibrillation separately using 404,767 UK Biobank participants that were free of both diseases at baseline. In this group, there were 3,385 incident heart failure cases and the HCN4 variant rs8038766 showed a non-significant trend for a protective effect (HR = 0.96, 95% CI 0.89-1.02; $P = 0.177$) (Table 2.1). However, in a competing risk model accounting for incident occurrences of atrial fibrillation, the protective effect of the heart rate-reducing variant on heart failure was brought to focus with HR=0.90, 95% CI 0.83-0.98 ($P = 0.013$) (Table 2.1). We conducted a similar analysis using incident myocardial infarction or cardiovascular death corresponding to the primary endpoint in the SIGNIFY trial, which was also potentially exposed to the opposing effects of the heart rate-reducing variant on atrial fibrillation and myocardial infarction. There was no detectable association of the heart rate-reducing variant with myocardial infarction or cardiovascular death in the simple Cox proportional-hazards model (HR=0.99, 95% CI 0.94-1.05) or in the cause-specific competing risk model (HR=0.99, 95% CI 0.93-1.04) (Table 2.1). We did see, however, an association between rs8038766 and prevalent or incident cases of unstable angina (OR 0.92 95% CI 0.86-0.98, $P = 0.0056$) (Figure 2.2).

In the HERMES case-control consortium, the heart rate reducing allele of rs8038766 was only weakly associated with heart failure (OR 0.98 95% CI 0.96-1.00, $P = 0.079$) (Figure 2.2), but when using the mtCOJO method to adjust for atrial fibrillation using summary statistics [70], the protective effect was increased with a conditional OR = 0.96 95% CI 0.94-0.98 ($P = 9.7 \times 10^{-4}$) [165]. In the CARDIoGRAMplusC4D consortium, there was no association between rs8038766 and CAD or myocardial infarction (Figure 2.2).

2.5.4 Bi-directional MR

Bi-directional MR supports a causal effect of atrial fibrillation on heart failure with a causal OR estimate of 1.22 and ranging up to 1.25 according to different MR models (Table 2.2, Supplementary Table A.7), and supports a causal effect of heart failure on atrial fibrillation with OR ranging from 1.21-1.94 (excluding the contamination mixture model estimate of 6.82 which is an outlier among the other methods). These results are concordant

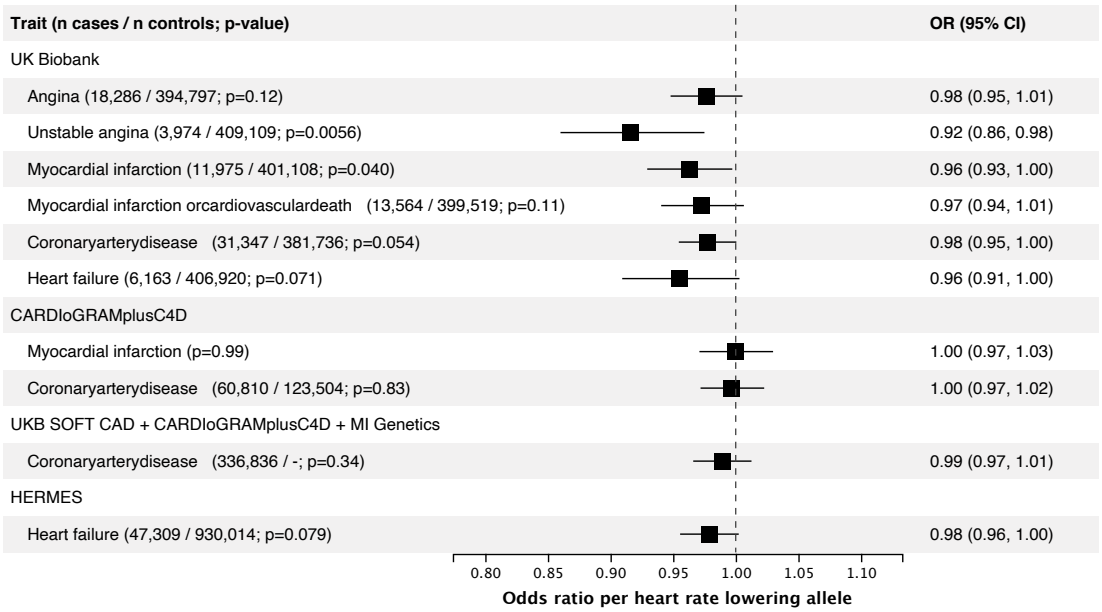


Figure 2.2 – **Association between the heart rate lowering allele (G) of the *HCN4* variant rs8038766 and efficacy outcomes in the UK Biobank and in GWAS summary statistics from large consortia.** For the UK Biobank, reporting results from logistic regression comparing the combined prevalent and incident cases to non-cases. References: CARDIoGRAMplusC4D [62], UKB SOFT CAD + CARDIoGRAMplusC4D + MI Genetics [163], HERMES [165].

with observational longitudinal studies that have observed an increased incidence of atrial fibrillation in new heart failure patients and vice versa and where both diseases are often diagnosed on the same day [173]. Bi-directional MR also supported a causal effect of CAD and myocardial infarction on atrial fibrillation, but not the opposite. The point estimates ranged from OR: 1.12-1.17 for the effect of CAD on atrial fibrillation and OR: 1.11-1.22 for the effect of myocardial infarction on atrial fibrillation (Table 2.2, Supplementary Table A.7).

2.5.5 Effect of heart rate on cardiovascular outcomes

To compare our observations of heart rate reduction attributable to the *HCN4* variant to that of polygenic origin, we constructed a genetic risk score (GRS) with 64 variants previously associated with heart rate (Supplementary Table A.4) [167]. An increase of 1 standard deviation in the heart rate GRS was associated with a 1.76 (1.73, 1.80) bpm increase in heart rate explaining 2.5% of the variance in the UK Biobank data (Supplementary Figure A.2). The two stage method causal estimate scaled for a 11.1 bpm (corresponding to 1 standard deviation of heart rate in the UK Biobank and concordant with the effect of ivabradine in

Table 2.1 – Association of the *HCN4* variant with outcomes in the UK Biobank using prospective and cause-specific hazard competing risk analyses.

Outcome	Model	N total	N events	HR (95% CI)*	P-value
Genetic model for SHIFT. *Using participants without a history of atrial fibrillation or heart failure at recruitment*					
Heart failure	Cox proportional-hazards	404,767	3,385	0.96 (0.89, 1.02)	0.18
Atrial fibrillation	Cox proportional-hazards	404,767	8,461	1.08 (1.04, 1.13)	9.4×10^{-5}
Heart failure	Competing risk (atrial fibrillation)	404,767	2,380	0.90 (0.83, 0.98)	0.013
Atrial fibrillation	Competing risk (heart failure)	404,767	7,663	1.08 (1.04, 1.13)	3.2×10^{-4}
Genetic model for SIGNIFY *Using participants without a history of atrial fibrillation or MI at recruitment*					
MI or CV Death	Cox proportional-hazards	397,008	4,976	0.99 (0.94, 1.05)	0.84
Atrial fibrillation	Cox proportional-hazards	397,008	7,880	1.08 (1.04, 1.13)	3.1×10^{-4}
MI or CV Death	Competing risk (atrial fibrillation)	397,008	4,534	0.99 (0.93, 1.04)	0.61
Atrial fibrillation	Competing risk (MI or CV death)	397,008	7,482	1.09 (1.04, 1.13)	1.4×10^{-4}

* Reporting the effect of the heart rate reducing allele of rs8038766 at the *HCN4* gene. All models were adjusted for age, sex and the first 10 principal components. In the Cox proportional-hazards models, individuals were censored at the time of death or end of follow up; in the competing risk models, individuals were censored at the time of occurrence of the competing event, death, or end of follow up. CV, cardiovascular; HR, hazard ratio; MI, myocardial infarction.

Table 2.2 – Bi-directional Mendelian randomization estimates.

Exposure	Outcome	MR Causal OR* (95% CI)	P-value
Atrial fibrillation (152 variants)	Heart failure	1.23 (1.20, 1.27)	3.7×10^{-52}
Atrial fibrillation (152 variants)	Coronary artery disease	1.00 (0.98, 1.03)	0.76
Atrial fibrillation (152 variants)	Myocardial infarction	0.98 (0.95, 1.02)	0.30
Heart failure (11 variants)	Atrial Fibrillation	1.45 (1.11, 1.90)	0.0067
Coronary artery disease (68 variants)	Atrial Fibrillation	1.15 (1.11, 1.21)	1.7×10^{-10}
Myocardial infarction (31 variants)	Atrial Fibrillation	1.11 (1.06, 1.16)	1.3×10^{-5}

Summary statistics for atrial fibrillation taken from Nielsen *et al.* [164] for myocardial infarction and CAD from CARDIoGRAMplusC4D and CARDIoGRAMplusC4D + UKB SOFT + MiGen [62, 163], for heart failure from HERMES [165]. * IVW MR model. For MR results using MR-Egger, the contamination mixture model and MR-PRESSO, see Supplementary Table A.7. Causal ORs relate the odds of the outcome in exposed individuals vs non-exposed. IVW, inverse-variance weighted; MR, Mendelian randomization; OR, odds ratio.

clinical trials) genetic reduction in heart rate was OR=1.25 (95% CI: 1.13-1.39) for atrial fibrillation, OR=1.03 (95% CI: 0.88-1.21) for heart failure and 1.03 (95% CI: 0.96-1.11) for CAD. To account for the presence of pleiotropy, we also used MR-Egger, contamination mixture model and the MR-PRESSO methods (Supplementary Table A.5), and saw an increase in the risk for atrial fibrillation associated with heart rate reduction (OR 1.54, $P = 1.3 \times 10^{-7}$ for MR-PRESSO), but not with heart failure or CAD, although these did not take into account the possible competing effect of atrial fibrillation. MR analyses using effect estimates derived from larger GWAS consortia using the same set of 64 heart rate

variants supported results observed with the UK Biobank data (Supplementary Table A.6).

2.6 Discussion

In the present study, we used genetics to infer the causal effect of ivabradine on safety and efficacy outcomes in an attempt to reproduce observations from randomized clinical trials, and to assess the value of genetic approaches to support drug targets and trial design issues such as target patient population and clinical outcomes.

2.6.1 Effect of *HCN4* on atrial fibrillation

Genetically predicted heart rate reduction from the HCN4 gene variant rs8038766 was associated with an increase in risk of atrial fibrillation, recapitulating the observations from the SIGNIFY and SHIFT trials. In our MR analyses using methods robust to the inclusion of invalid instruments, we observed that a genetically predicted reduction in heart rate of approximately 11 bpm conferred an increased risk of atrial fibrillation with an OR of 1.54 in the UK Biobank and OR ranging from 1.36 to 1.56 using summary statistics from previous large GWAS. These results are also coherent with a recent MR study reporting a protective effect of increased heart rate for atrial fibrillation and cardioembolic stroke [174]. Atrial fibrillation is known to increase the risk of stroke by 3 to 5-fold [175]. In clinical trials of ivabradine, there was no treatment association with stroke, but the small number of incident atrial fibrillation events would have made such an observation unlikely [176]. The estimated effects of heart rate on atrial fibrillation are smaller than the effect predicted from the *HCN4* variant alone, whose scaled OR for a comparable 11 bpm reduction would be greater than 5. This may be explained partly by the inaccuracy of extrapolation of the OR estimate derived from a single genetic variant, and also possibly by an effect of HCN4 on atrial fibrillation that may be specific to modulation of the I_f current or other structural consequences of *HCN4* mutations. For example, genetic mutations in *HCN4* have been associated to Brugada syndrome and sick sinus syndrome as well as left ventricular noncompaction and it is possible that common polymorphisms in the *HCN4* gene have more subtle effects on myocardium structure or conduction parameters that may be independent of heart rate [177, 178]. Additionally, altering HCN4 function or levels during embryogenesis in other species have been shown to structurally alter heart development which could explain effects beyond heart rate modulation alone [179]. The MR estimates from the UK Biobank are also based on mostly healthy individuals with a low heart rate (mean of 69 bpm) possibly limiting clinical interpretation [180]. However, the effect estimates of heart rate reduction on

atrial fibrillation are similar when using the GWAS results from Nielsen *et al.* which include both population-based and clinical cohorts [164]. The possibility that the effect is greater in healthy patients is also supported by the previously described association between low heart rate during physical activity and the increased incidence of atrial fibrillation [181].

2.6.2 Effect of *HCN4* on ischemic endpoints

We tested the association between the *HCN4* heart rate-reducing variant and various ischemic endpoints in the UK Biobank. The largest effect we observed was with unstable angina, which is coherent with the use of ivabradine to alleviate anginal symptoms. Nonetheless, the effect sizes of the association with CAD, angina and myocardial infarction were small and marginally significant in the UK Biobank and importantly they were not supported by results from larger GWAS consortia. This suggests that the effect of *HCN4* on CAD may be null or of a very small effect size so as to not be detectable in the context of a clinical trial such as in the SIGNIFY study [27]. We also investigated the possibility that the increased risk of atrial fibrillation offsets the beneficial effects on the SIGNIFY primary endpoint of myocardial infarction or cardiovascular death using a prospective competing risk analysis in individuals that did not develop atrial fibrillation and showed that accounting for atrial fibrillation had no impact on the risk for myocardial infarction or cardiovascular death. There was no detectable association of the *HCN4* heart rate-reducing variant with myocardial infarction or cardiovascular death in the cause-specific competing risk model. This was further supported by the bi-directional MR analysis that showed that CAD caused atrial fibrillation but not the opposite. Finally, the MR study did not show a causal link between heart rate and CAD suggesting that reducing heart rate is not sufficient to prevent the disease.

2.6.3 Relationship with clinical trials of ivabradine

The analysis of the subgroup of participants with angina class 2 or greater at baseline in SIGNIFY showed a nominal increase in the rate of the primary endpoint of cardiovascular death or myocardial infarction with ivabradine [25, 27]. Whether this observation represented a chance finding in the context of a neutral result in the overall SIGNIFY population or a potential signal of harm in this subset of patients was a matter of discussion. The results of the current analyses support neutral effects of I_f current inhibition on the composite endpoint of cardiovascular death and myocardial infarction, without evidence of harm.

In the SHIFT trial, ivabradine reduced the rate of the primary composite endpoint of

cardiovascular death or hospitalization for worsening heart failure in patients with heart failure with reduced ejection fraction and without atrial fibrillation. In the genetic model of ivabradine, the competing risk analysis accounting for atrial fibrillation showed that the *HCN4* heart rate-reducing variant protected against heart failure (HR= 0.90, 95% CI: 0.83-0.98, $P = 0.013$). The results from the marginal models and the competing risk analyses do suggest opposing effects of the heart rate-reducing *HCN4* variant on atrial fibrillation and heart failure. The importance of these effects is also highlighted by the bi-directional MR of atrial fibrillation and heart failure that confirmed that both diseases are mutually causal of one another.

2.6.4 Study limitations

As for any MR study, our analyses were subject to the assumptions of the underlying models and the possibility of unobserved horizontal pleiotropy. Additionally, our genetic model of ivabradine corresponds to a lifelong effect as opposed to an exposure after drug initiation. Generally, common variants also result in effects of smaller magnitude than ones resulting from pharmacological modulation and extrapolation is required to compare them. We also used data from individuals of predominantly European ancestry both in the UK Biobank and in summary statistics from large GWAS consortia which could limit the generalizability of our results to other populations both in terms of clinical profile and ancestry. Finally, we defined clinical variables based on combinations of hospitalization and death record codes in the UK Biobank which is likely to result in imperfect coding of disease status.

2.7 Conclusion

In conclusion, genetic modelling of ivabradine recapitulates its benefits in heart failure, promotion of atrial fibrillation, and neutral effect on myocardial infarction. This study supports the use of methods that leverage naturally occurring genetic variants to predict diverging results on different clinical outcomes and support the design of randomized clinical trials, even in a situation where more complex disease risks are at play.

2.8 Acknowledgements

This research has been conducted using the UK Biobank Resource under Application Number 20168.

2.9 Funding sources

This work was supported by the Health Collaboration Acceleration Fund from the Ministère de l'Économie et de l'Innovation du Gouvernement du Québec. M.-A.L. is supported by a Frederick Banting and Charles Best Canada Graduate Scholarship Doctoral Award from the Canadian Institutes of Health Research (CIHR). J.-C.T. holds the Canada Research Chair in personalized medicine and the Université de Montréal Pfizer-endowed research chair in atherosclerosis. S. d.D. holds the Université de Montréal Beaulieu-saucier Chair in Pharmacogenomics. SS is partly supported by a National Health and Medical Research Council (NHMRC) fellowship and NHMRC Program Grant 1113400. R.T.L. is supported by a UK Research and Innovation Rutherford Fellowship.

2.10 Disclosures

A patent pertaining to pharmacogenomics-guided CETP inhibition was granted and J.-C.Tardif and M.-P. Dubé are mentioned as authors. J.-C.Tardif and M.-P. Dubé have a minor equity interest in DalCor. J.-C.Tardif has received research support from Amarin, AstraZeneca, DalCor, Eli-Lilly, Ionis, Pfizer, RegenexBio, Sanofi and Servier, and honoraria from DalCor, Pfizer, Sanofi and Servier. M.-P. Dubé has received honoraria from Dalcor and research support (access to samples and data) from AstraZeneca, Pfizer, Servier, Sanofi and GlaxoSmithKline. B. Tyl is an employee of Laboratoires Servier. Simon de Denus has received grants from Pfizer, AstraZeneca, Roche Molecular Science, DalCor and Novartis. R. Thomas Lumbers has received research grants from Pfizer. The remaining authors have nothing to disclose.

CHAPITRE 3

Study of effect modifiers of genetically predicted CETP reduction

L'utilité de l'étude des variations génétiques situées dans les gènes encodant des cibles pharmacologiques dans le but de valider le potentiel thérapeutique étant bien établi, nous tournons maintenant notre intérêt vers l'utilisation de la génétique afin d'identifier des sous-groupes d'individus susceptibles de mieux répondre à un médicament. Cette application est particulièrement importante, car elle cible une limitation importante des essais randomisés, notamment leur pouvoir limité à estimer des effets de sous-groupe. À cette fin, nous utiliserons à titre d'exemple le cas de l'inhibition de la CETP. Les essais randomisés des inhibiteurs pharmacologiques de la CETP présentent tous une très faible proportion de femmes (entre 16 et 20%). Comme le nombre d'individus recrutés et la durée de suivi sont choisis afin de détecter l'effet escompté dans la population totale de l'étude, à moins d'être élaborés selon un devis stratifié prédéfini, les essais randomisés sur population globale n'ont pas une puissance statistique suffisante pour identifier une différence d'effet entre les sexes. Ce problème est prévalent dans le contexte des sous-groupes en général et n'est pas limité aux différences entre les sexes. La validité externe des essais cliniques dans des populations de patients présentant des caractéristiques différentes (et dans des proportions différentes) à celles testées dans les essais clinique, est un sujet d'importance en médecine basée sur l'évidence et en médecine de précision. La prémisse de notre étude est donc que l'utilisation de modèles génétiques puisse aider à prédire ou caractériser la validité externe des essais randomisés. En amont, il serait aussi possible d'utiliser de telles méthodes afin d'optimiser ces études pour cibler une population susceptible de bénéficier le plus d'un médicament dans un contexte de médecine de précision.

Dans cette étude, nous avons utilisé une variation génétique du gène CETP (rs1800775) connue pour altérer la liaison de facteurs de transcription, ainsi qu'un score génétique prédicteur de la concentration plasmatique de CETP, afin d'étudier l'effet d'une réduction de la CETP. Après avoir bien caractérisé l'impact d'une réduction génétiquement prédite de CETP sur différents biomarqueurs et maladies cardiovasculaires d'intérêt, nous avons testé comment ces effets étaient modifiés par le sexe et l'indice de masse corporelle (IMC). Ces variables ont été sélectionnées à cause d'évidence antérieure suggérant qu'elles modulaient possiblement l'effet de la CETP [182-187]. Dans une première itération de cette étude, nous avons aussi considéré la modification de l'effet de la CETP par le génotype à la variation rs1967309. Les analyses exploratoires ont cependant rapidement montré une grande complexité, car rs1967309 a des effets indépendants sur la capacité d'efflux de cholestérol et sur l'indice de masse corporelle, ce qui en complexifiait considérablement l'analyse. Nous avons donc décidé de reporter l'étude de cette variable à une étude future afin d'éviter de complexifier la présente contribution. Une présentation de l'interaction entre rs1967309, le sexe, et la variation de l'expression de *CETP*, modulée par une variation génétique située dans ce gène, est présentée dans l'article «An epistatic interaction between the *ADCY9* pharmacogene and the drug target *CETP*», sous évaluation pour publication et dont je suis coauteur.

Dans notre étude, les biomarqueurs considérés incluent les taux de cholestérol LDL et HDL, leurs apolipoprotéines constituantes et la lipoprotéine(a). Nous avons aussi testé l'effet sur la protéine C réactive, un biomarqueur d'inflammation systémique ainsi que sur la capacité d'efflux de cholestérol du plasma. Les maladies cardiovasculaires à l'étude étaient l'infarctus du myocarde, la maladie coronarienne ainsi que les procédures chirurgicales de revascularisation.

Nous avons observé de forts effets modulateurs du sexe et de l'IMC sur différents paramètres lipidiques. Une réduction génétique de la CETP était associée à une plus grande augmentation du cholestérol HDL et de l'efflux de cholestérol et une plus grande réduction du cholestérol LDL chez les femmes que chez les hommes. De façon similaire, les individus de plus faible IMC avaient une plus grande augmentation du cholestérol HDL et une plus grande réduction du cholestérol LDL que les individus de plus grand IMC. Aucune modification d'effet par l'IMC n'a cependant été observée sur l'efflux de cholestérol. Finalement, cette étude propose une caractérisation en profondeur de l'effet de la réduction génétique de la CETP et de la modulation de ces effets par deux variables d'intérêt clinique. La méthodologie utilisée dans cette étude est aussi pertinente au sens plus large de l'utilisation de la génétique des cibles pharmacologiques pour estimer des paramètres liés à la validité externe des essais randomisés et la médecine de précision.

Contributions : La concept de cette étude a été développé par Marc-André Legault, Marie-Pierre Sylvestre, Amina Barhdadi, Jean-Claude Tardif et Marie-Pierre Dubé. Isabel Gamache, Jean-Christophe Grenier et Julie G. Hussin ont contribué à l'évaluation et à la validation des résultats. Marc-André Legault, Jean-Claude Tardif et Marie-Pierre Dubé ont contribué au développement et à l'implémentation des algorithmes diagnostics utilisés pour définir les issues cardiovasculaires dans la UK Biobank. La méthodologie a été développée par Marc-André Legault, Amina Barhdadi, David Rhainds, Louis-Philippe Lemieux Perreault et Marie-Pierre Dubé. La production de graphiques, le développement de logiciels et les analyses statistiques ont été réalisées par Marc-André Legault et Amina Barhdadi. La supervision de ce projet a été faite par Marie-Pierre Dubé. Tous les auteurs ont lu et contribué à l'écriture du manuscrit et les contributions majeures ont été faites par Marc-André Legault, Julie G Hussin, Marie-Pierre Sylvestre et Marie-Pierre Dubé.

Study of effect modifiers of genetically predicted CETP reduction

To be submitted.

Marc-André Legault[1,2,3], Amina Barhdadi[1,2], Isabel Gamache[1,3], Audrey Lemaçon[1,2], Louis-Philippe Lemieux Perreault[1,2], Jean-Christophe Grenier[1], Marie-Pierre Sylvestre[4,5], Julie G. Hussin[1,6], David Rhainds[1], Jean-Claude Tardif[1,6], Marie-Pierre Dubé[1,2,6]

1. Montreal Heart Institute, Montreal, Canada
2. Université de Montréal Beaulieu-Saucier Pharmacogenomics Centre, Montreal, Canada
3. Université de Montréal, Department of biochemistry and molecular medicine, Montreal, Canada
4. Research Centre of the University of Montreal Hospital Centre, Montreal, Canada
5. Department of Social and Preventive Medicine, Université de Montréal, Montréal, Canada
6. Université de Montréal, Department of medicine, Montreal, Canada

3.1 Abstract

Genetic variants in drug targets can be used to predict the effect of drugs. Here, we extend this principle to assess how sex, body mass index and rs1967309 genotype may modify the effect of a genetically predicted lower CETP levels on biomarkers and cardiovascular outcomes. We confirm the genetically predicted results in users of dalcetrapib, a pharmacological CETP inhibitor.

We found sex and BMI to be modulators of the effect of genetically-predicted lower CETP on the lipid profile. Female sex and lower BMI are associated with higher HDL-c and lower LDL-c for the same genetically predicted reduction in CETP.

Our results provide insight on the clinical effects of CETP inhibitors. More broadly, we present a novel analytical approach to identify effect modification with implications for precision medicine and to evaluate external validity of clinical trials.

3.2 Keywords

CETP, drug target validation, precision medicine, effect modification

3.3 Introduction

Genetic variants in drug targets can be used to predict the effects of drugs [145, 188]. The identification of rare variants with strong effects on protein function led to the development of new drug classes, and there is a growing number of genetically supported drug targets along various phases of drug development [6, 9, 189]. However, few genetic studies of drug targets have focused on the identification of subgroups of individuals that could derive a greater benefit from the drug. This question is of paramount importance in our quest to improve precision medicine, which can be supported by genetic techniques. Randomized controlled trials (RCTs) are powered to detect benefit in the full study population and analyses in subgroups of individuals are typically reported as exploratory observations. When clinical or demographic subgroups are underrepresented or excluded from trials, external validity can be put into question [190]. In clinical trials of cardiovascular disease prevention, for example, women are frequently underrepresented, and only 31% of trials report sex-specific results [191]. This may be of importance as differences in body size and composition as well as hormonal and gender differences could all have an impact on drug response [192].

CETP inhibitors have a complex history of heterogeneous findings from RCTs and genetic studies. Three of the four trials of CETP inhibitors did not report benefit, except for the most recent and largest trial (Randomized EValuation of the Effects of Anacetrapib through Lipid-modification, REVEAL) that showed a small reduction in risk of cardiovascular outcomes in an at-risk population, with a rate ratio of 0.91 (0.85, 0.97) for the study primary efficacy endpoint [35].

In this paper we investigate how sex and body mass index (BMI) may modify the effect of genetically predicted CETP reduction on biomarkers and cardiovascular outcomes. We consider the effect on apolipoproteins and lipid fractions thought to be related to CETP inhibition, namely high density lipoprotein cholesterol (HDL-c), low density lipoprotein cholesterol (LDL-c), and their lipoprotein constituents apolipoproteinA (apoA) and apolipoproteinB (apoB), respectively. We also consider C-reactive protein levels, a measure of systemic inflammation and lipoprotein(a) an independent atherogenic lipoprotein which may be affected by CETP inhibition. The effect on the capacity of apoB-depleted plasma to efflux cholesterol which is related to HDL function is also assessed using an independent dataset.

This work is important both from a personalized medicine standpoint to identify subgroups of individuals that could derive a greater benefit from CETP inhibition, and from a fundamental research perspective to better understand the role of CETP in lipid homeostasis and disease pathology. Our approach is akin to a Mendelian Randomization study, and a justification for the causal interpretation is presented (Supplementary Methods, Section B.2.6).

3.4 Methods

3.4.1 Study populations

Data from the UK Biobank cohort were used for the analyses relating genetically predicted CETP with biomarkers and cardiovascular outcomes. The variable definitions and genetic quality control steps are described in the Supplementary Methods (Section B.2).

Data from the MHI Biobank cohort were used to conduct the analyses of genetically predicted CETP with cholesterol efflux (Supplementary Methods, Section B.2).

3.4.2 Genetic predictors of CETP activity

We evaluated various scores of CETP activity based on a GWAS of plasma CETP concentration or the MAGNETIC nuclear magnetic resonance GWAS [50, 193] constructed using the p-value thresholding and LD clumping method (Supplementary Methods, Section B.2.5). For this study, we selected the score based on the plasma CETP concentration as it directly relates to CETP activity. This choice is unlikely to affect our results because all the genetic scores were highly correlated with r^2 between 0.75 and 1.00 (Supplementary Table B.1).

Individual CETP variants have also been widely studied and the CETP variant -629C>A (rs1800775) in particular disrupts transcription factor binding at the *CETP* promoter and reduces CETP activity [194, 195]. Using such a variant as a proxy for CETP levels has the advantage of providing more directly interpretable results (*i.e.* in allelic units) and does not rely on weights which may be biased with respect to the population (both in terms of sex, clinical profile and ethnicity) in which they were estimated. For example, if genetic variants have different effects in men and women, scores based on these effects will be better predictors of the phenotype in individuals whose sex was most common in the study used to estimate the genetic effects. Given that genetic scores and variants have complementary advantages, we show results from both the CETP concentration score and the rs1800775 variant.

Because this study investigates the effect of genetically-predicted CETP levels, it is important to validate the statistical strength of our predictors. We estimated association strength using HDL-c and LDL-c levels as proxy variables because they are known to be affected by CETP modulation using univariable linear regression and we report F statistics and R^2 as is standard in Mendelian randomization studies [124]. When regressing on measured HDL-c in the UK Biobank, the F statistic was 10,400 ($F_{1;362,466}$) for the genetic score and 7,239 ($F_{1;362,466}$) for rs1800775, with corresponding R^2 of 0.028 and 0.020, respectively. Similarly, when regressing on LDL-c, the F statistics and R^2 were 201 ($F_{1;394,285}$) and $R^2 = 0.00051$ for the genetic score and 160 ($F_{1;394,285}$) and $R^2 = 0.00041$ for rs1800775. The genetic score has a stronger effect than rs1800775 on CETP and consequently on HDL-c and LDL-c. The rs1800775 variant remains a strong predictor of CETP and is not prone to bias due to weighting.

3.4.3 Statistical analyses

We assessed the effect of the genetic predictors of CETP on observed biomarkers and cardiovascular outcomes using linear and logistic regression models as appropriate. To estimate the effect modification by sex and BMI, we fit the models containing a product interaction term between the genetic predictor of CETP and the effect modifier of interest and fixed covariates including the component variables of the interaction and age, sex and ancestry principal components. We adjusted for these covariates to improve the precision of our estimates. We used hypothesis testing of a null product interaction coefficient as a test of effect modification and the p-value for this test is denoted as p_{itx}. In linear models, the hypothesis test is for an additive interaction and it can be interpreted as an additive deviation from the contribution of the interacting variables (Supplementary Methods, Section B.2.8). In logistic regression models, the test of the product interaction term assesses a multiplicative effect modification of the odds ratio (Supplementary Methods). In other words, in the latter case the test is of $\mathrm{OR}_{11}/(\mathrm{OR}_{01} \times \mathrm{OR}_{10}) = 1$ where OR_{ij} denotes the odds ratio when setting the first interacting variable to i and the 2^{nd} variable to j compared to the reference value for both covariables. Using the logistic regression model, we also computed interaction statistics on the additive scale namely the Relative Excess Risk due to Interaction (RERI) and the interaction contrast on the probability scale (Supplementary Methods). The additive effect statistics were developed for risk factors (and not preventive exposures) which prompted us to report interaction effects for the "male" sex, for a 1 s.d. increase in BMI and for a 1 s.d. increase in the CETP genetic score.

We also used interaction models to compute the marginal effects of the genetic CETP

predictors on the outcomes of interest at representative values of the tested modifiers. These marginal effects have the advantage of being directly interpretable. In models with more than a single interaction term, we used the R package "margins" to estimate the marginal effect at representative cases and corresponding 95% confidence interval, otherwise we computed the marginal effects directly by summing the relevant regression coefficients.

The effects of BMI and the CETP genetic score (and their interaction) could be nonlinear. To test this hypothesis and to account for nonlinear interactions, we fitted an interaction model with interacting restricted cubic splines with four knots for BMI and the CETP genetic score. We used ANOVA to test for interactions by simultaneously considering all interaction coefficients. For significant interactions in the nonlinear models, we then plotted the predicted effects at varying levels of BMI and of the CETP genetic score to visualize the nonlinear effects. We used the R "rms" package to conduct these analyses.

For analyses based on the CETP genetic score, all effects for biomarkers are reported in units of standard deviation (s.d.) of the biomarker per s.d. decrease of the score (representing a decrease in CETP concentration as for pharmacological CETP inhibition). For binary outcomes, the reported effects are odds ratios per s.d. decrease of the genetic score. For analyses based on the individual CETP variant, the results are per copy of the HDL-c increasing "A" allele at rs1800775 (CETP -629C>A). For causal interpretation of these analyses, see Supplementary Methods (Section B.2.6). No adjustment was made for multiple testing of phenotypes and effect modifiers. Estimates are reported with 95% confidence intervals.

3.4.4 Power analyses

We estimated the power of the different association models using simulations. In the simplest case, we simulated a normally distributed genetic score with a fixed effect on a standard normal outcome and computed the proportion of rejected null hypotheses across simulation replicates at $\alpha = 0.05$. We extended this model to account for interaction effects, and we used a latent variable logistic regression model when estimating the power for association with binary traits. The simulation model is described in Supplementary Methods (Section B.2.4).

3.5 Results

3.5.1 Study population

There were 413,138 unrelated participants from the UK Biobank included in the analyses (Supplementary Methods, Section B.2). The number of events under consideration for the cardiovascular outcomes as well as descriptive statistics for continuous measurements are presented in Table 3.1.

3.5.2 Effect of genetically-predicted reduction of CETP on biomarkers and cardiovascular outcomes

We first assessed the effect of the genetic CETP predictors on biomarkers and cardiovascular outcomes without any modifiers to contextualize future results (Table 3.2). As expected, the strongest association was with HDL-c, with an increase of 0.167 s.d. in HDL-c (corresponding to 0.064 mmol/l) per 1 s.d. decrease of the CETP genetic score, with concordant results for the rs1800775 (CETP -629C>A) SNP alone. To offer a comparison, pharmacological CETP inhibition with anacetrapib increased HDL-c levels by 1.11 mmol/l on average at trial midpoint, evacetrapib increased HDL-c levels by 1.52 mmol/l at 3 months whereas dalcetrapib increased HDL-c by about 0.34 mmol/l on average at 1 year [35, 196, 197]. In these examples, the pharmacological effect of CETP inhibition is about 17x stronger for anacetrapib, 24x for evacetrapib and 5x for dalcetrapib when compared to a 1 s.d. reduction of the CETP genetic score. This comparison is based only on the effect on HDL-c levels which may not represent the full spectrum of effects of CETP inhibitors.

The CETP genetic score was strongly associated with both basal and cAMP stimulated cholesterol efflux capacity as measured in plasma from 5,215 participants of the MHI Biobank (Table 3.2). A 1 s.d. decrease in the score increased basal cholesterol efflux by 0.105 s.d. (95% CI: 0.078, 0.130) and increased cAMP stimulated cholesterol efflux by 0.085 s.d. (95% CI: 0.059, 0.011).

The CETP genetic score was associated with LDL-c levels with a 0.023 s.d. decrease in LDL-c (corresponding to 0.020 mmol/l) per 1 s.d. decrease of the CETP genetic score ($p = 1.4 \times 10^{-45}$). Most RCTs of CETP inhibitors reported a decrease in LDL-c cholesterol, but not in the dal-OUTCOMES trial of dalcetrapib. Recently, anacetrapib was shown to decrease the production of lp(a) which could partly explain the benefit of this CETP inhibitor [198, 199]. A decrease in lp(a) levels was also observed with torcetrapib suggesting a possible class effect of CETP inhibitors [200]. We tested the association between the CETP genetic

Table 3.1 – **Descriptive statistics of the effect modifiers, biomarkers and cardiovascular outcomes in the UK Biobank study population.** If a transformation was needed to normalize the data, the statistics are given in both the original and transformed scale. The units represent the applied transformation.

	Women	Men	All
General characteristics			
n (%)	222,684 (54%)	190,454 (46%)	413,138
age – mean ± s.d.	56.6 ± 7.88	57.0 ± 8.06	56.8 ± 7.96
Effect modifiers			
Sex – n (%)	222,684 (100%)	190,454 (100%)	413,138 (100%)
Body mass index (original scale) (BMI, kg/m^2) - Mean ± s.d. (median ± IQR)	27.0 ± 5.15 (26.1 ± 6.21)	27.9 ± 4.24 (27.3 ± 5.07)	27.4 ± 4.77 (26.7 ± 5.74)
Body mass index (BMI, ln(kg/m^2)) - Mean ± s.d.	3.28 ± 0.18	3.32 ± 0.15	3.30 ± 0.17
Biomarkers as outcomes			
Lipoprotein(a) (lp(a), nmol/L) - Mean ± s.d.	44.6 ± 49.4	43.4 ± 49.3	44.0 ± 49.4
Apolipoprotein B (apoB, g/L) - Mean ± s.d.	1.04 ± 0.24	1.03 ± 0.24	1.03 ± 0.24
Low density lipoprotein cholesterol (LDL-c, mmol/L) - Mean ± s.d.	3.64 ± 0.87	3.48 ± 0.86	3.57 ± 0.87
Apolipoprotein A (apoA, g/L) - Mean ± s.d.	1.64 ± 0.27	1.43 ± 0.23	1.54 ± 0.27
High density lipoprotein cholesterol (HDL-c, mmol/L) - Mean ± s.d.	1.60 ± 0.38	1.29 ± 0.31	1.45 ± 0.38
C-reactive protein (original scale) (CRP, mg/L) - Mean ± s.d. (median ± IQR)	1.43 ± 2.97 (1.37 ± 2.30)	1.36 ± 2.77 (1.29 ± 1.88)	1.39 ± 2.89 (1.33 ± 2.09)
C-reactive protein - (CRP, ln(mg/L)) - Mean ± s.d.	0.36 ± 1.09	0.30 ± 1.02	0.33 ± 1.06
Cardiovascular outcomes			
Myocardial infarction (MI) - n (%)	4,747 (2%)	13,812 (7%)	18,559 (4%)
Percutaneous coronary intervention or coronary artery bypass graft (PCI/ CABG) - n (%)	3,395 (2%)	13,546 (7%)	16,941 (4%)
Coronary artery disease, soft (CAD) - n (%)	14,803 (7%)	29,910 (16%)	44,713 (11%)
Coronary artery disease, hard (CAD) - n (%)	6,825 (3%)	19,517 (11%)	26,342 (7%)

score and lp(a) levels measured in UK Biobank participants. A reduction of 1 s.d. in the CETP genetic score was associated with a decrease in lp(a) levels by 0.011 s.d. (95% CI: 0.008, 0.013); $p = 1.0 \times 10^{-14}$. In laboratory units of lp(a) levels, this corresponds to 0.524 nmol/l of lp(a) per s.d. of the CETP genetic score.

Table 3.2 – **Association of the CETP genetic score with biomarkers and cardiovascular events.**

Biomarkers - in s.d. units	**N**	**Standardized coefficient***	**p-value**
Lipoprotein(a)	315,214	-0.011 (-0.013, -0.008)	1.0×10^{-14}
C-reactive protein (CRP)	394,165	0.0026 (-0.0005, 0.0057)	0.098
HDL cholesterol	362,468	0.167 (0.164, 0.170)	$< 10^{-300}$
Apolipoprotein A	360,451	0.131 (0.129, 0.134)	$< 10^{-300}$
LDL cholesterol	394,287	-0.023 (-0.026, -0.019)	1.4×10^{-45}
Apolipoprotein B	393,089	-0.033 (-0.036, -0.030)	5.1×10^{-97}
Basal cholesterol efflux (MHI Biobank)	5,215	0.105 (0.079, 0.130)	1.9×10^{-15}
cAMP stimulated cholesterol efflux (MHI Biobank)	5,214	0.085 (0.059, 0.111)	2.7×10^{-10}
Cardiovascular outcomes	**N cases / N total**	**Odds ratio (95% CI)***	**p-value**
Coronary artery disease ("soft")	44,713 / 413,138	0.975 (0.965, 0.985)	8.4×10^{-7}
Coronary artery disease ("hard")	26,342 / 394,767	0.971 (0.958, 0.983)	6.0×10^{-6}
Myocardial infarction	18,559 / 413,138	0.980 (0.966, 0.995)	0.0096
Percutaneous coronary intervention or coronary artery bypass graft	16,941 / 413,138	0.967 (0.952, 0.982)	2.6×10^{-5}

* Coefficients for continuous variables (biomarkers) are from a linear regression model adjusted for age, sex and the first 10 principal components and are expressed in standard deviation units of the biomarkers outcome per standard deviation reduction in the CETP genetic score. Odds ratios for cardiovascular outcomes are estimated using a logistic regression model adjusted for the same covariates.

The association of the CETP genetic score with cardiovascular outcomes was concordant with the observed associations with the lipid profile. One s.d. reduction in the CETP genetic score was associated with CAD ("soft" definition, Supplementary Table B.2) with an OR of 0.97 (95% CI 0.96, 0.98) $p = 8.4 \times 10^{-7}$. Previous studies of genetic CETP reduction have also reported effects scaled by a 10 mg/dL reduction in apoB levels with an OR of 0.78 (95% CI 0.71, 0.86) [51]. After scaling for an effect of this magnitude, our estimate was comparable with an OR of 0.72 (95% CI 0.64, 0.82). We also repeated these analyses for the rs1800775 CETP promoter variant and obtained similar results (Supplementary Table B.3).

3.5.3 Female sex is associated with larger benefit of genetically lower CETP on the lipid profile

The large phase III RCTs of CETP inhibitors suffered from large sex imbalances, ranging between 16% of female participants (REVEAL) and 23% (ACCELERATE) and it is unlikely that these trials could have identified effect differences between men and women. In Figure 3.1, we show the reported drug effects from the major RCTs of CETP inhibitors stratified by sex. We conducted an inverse variance-weighted meta-analysis of the effect of the CETP inhibitor in the dal-OUTCOMES, REVEAL and ACCELERATE studies (Supplementary Methods, Section B.2). We did not include the ILLUMINATE trial with torcetrapib, as the drug had off-target deleterious effects. The calculated meta-analysis risk ratio is 0.96 (95% CI 0.91, 1.02) in men and 0.92 (95% CI 0.82, 1.04) in women. The test for heterogeneity between the male and female effects was not significant (p=0.50).

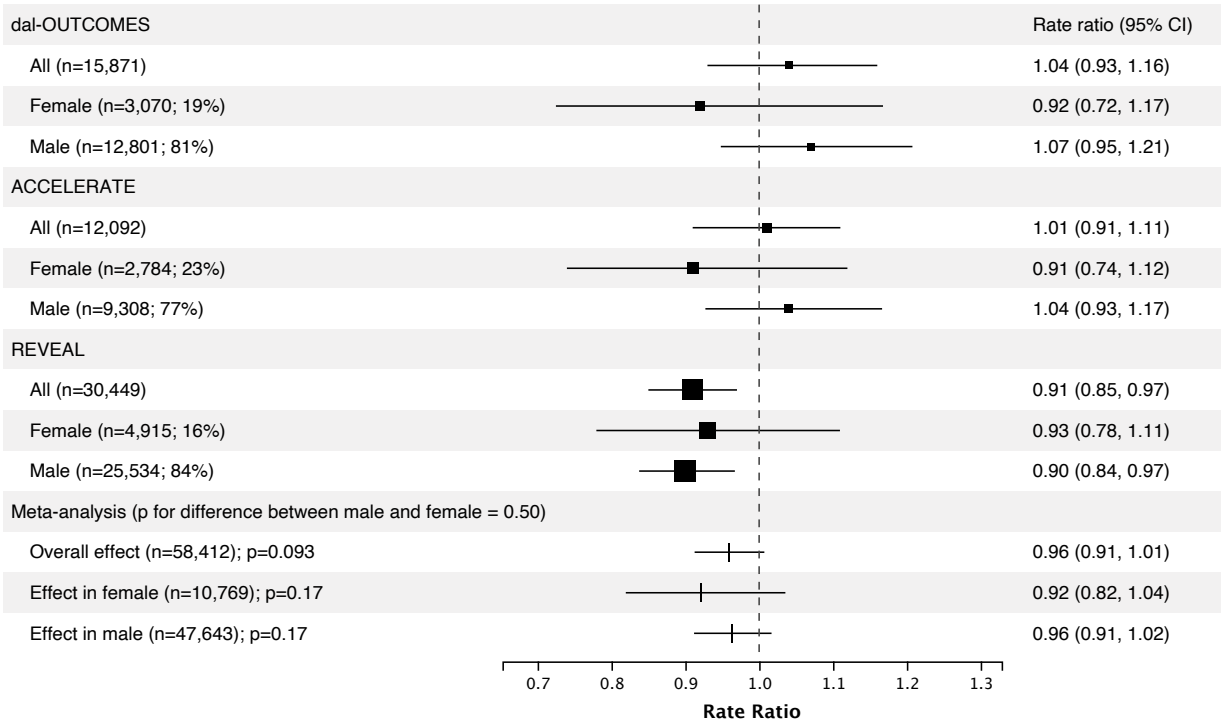


Figure 3.1 – **Effect of treatment from phase 3 trials of CETP inhibitors in the whole population and stratified by sex.** The Cochran Q statistics for heterogeneity between studies were 5.5 ($p = 0.063$) for the overall effect, 0.025 ($p = 0.99$) for the female-specific effect and 8.0 ($p = 0.018$) for the male-specific effect. Points are scaled with respect to the relative weight in the overall and sex-specific meta-analyses.

We used regression models to assess the interaction effect of sex with the CETP genetic score on biomarkers and cardiovascular outcomes in the UK Biobank (Figure 3.2). We observed statistically significant interactions for apoA and apoB as well as LDL-c and HDL-c levels. The strongest effect modification was with HDL-c and apoA. For instance, a one s.d. unit reduction in the CETP score increased HDL-c by 0.15 s.d. (95%CI 0.14, 0.15) in men ($p < 10^{-300}$) and by 0.18 s.d. (95% CI 0.18, 0.19) in women ($p < 10^{-300}$) and the interaction p-value was 5×10^{-32}. In general, genetically predicted lower CETP had a more beneficial effect on the lipid profile in women than men. Similar results were obtained with the rs1800775 variant alone (Supplementary Figure B.2).

We tested for the interaction of sex with the CETP genetic score on cholesterol efflux measured in a subgroup of participants of the MHI Biobank. For cAMP stimulated cholesterol efflux, a unit decrease in the score was associated with a 0.064 s.d. (95% CI 0.032, 0.095) increase in efflux for men ($p = 7 \times 10^{-5}$) and 0.13 (95% CI 0.086, 0.18) for women ($p = 5 \times 10^{-8}$) and the interaction p-value was 0.02. A similar effect was also observed for unstimulated efflux (Supplementary Figure B.3). Again, women had a more favourable cholesterol efflux profile than men, with higher cholesterol efflux with lower genetically-predicted CETP.

We considered whether the differences of the effect of CETP on biomarkers according to sex also led to differences in cardiovascular outcomes. On the multiplicative scale, the tested outcomes did not show statistically significant differences between men and women (interaction p-value of 0.56 for CAD). However, the RERI (0.082 95% CI [0.020, 0.146]) and interaction contrast (0.0017 95% CI [0.00024, 0.0032]) suggested a difference in the effect of the CETP score on CAD ("hard") in men compared to women (Supplementary Table B.5). Power analyses rule out a strong effect difference of genetically lower CETP concentration on cardiovascular outcomes between men and women (Supplementary Appendix, Section B.1). In sensitivity analyses, we show that results are robust to further adjustment for statin use (Supplementary Figure B.5).

3.5.4 Higher BMI reduces the benefit of genetically lower CETP on the lipid profile

Previous studies have reported that the effect of CETP on HDL-c is different across BMI classes [201, 202]. To assess how the effect of genetic CETP modulation is affected by BMI, we used interaction models from which we draw inferences and report marginal effects at fixed BMI values (Figure 3.3). We found BMI to be a significant modulator of the association between the CETP genetic score and HDL-c (interaction $p = 5.4 \times 10^{-73}$)

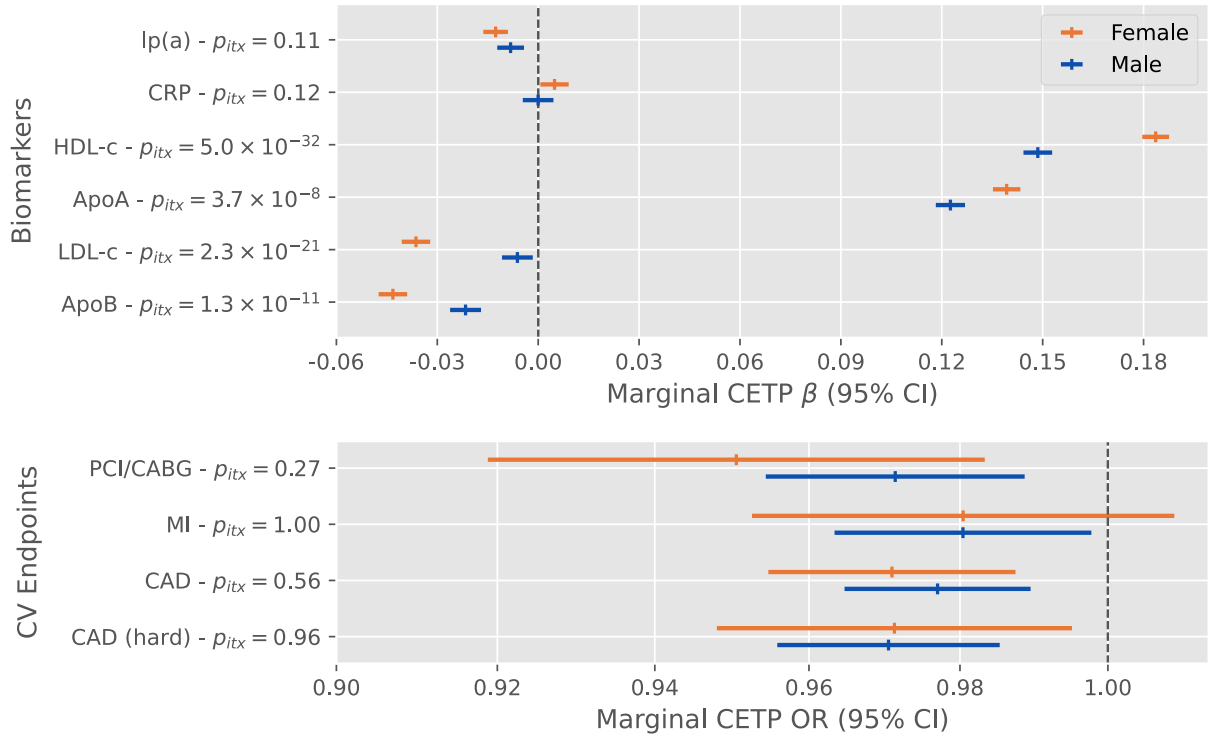


Figure 3.2 – **Effect modification of a 1 standard deviation decrease in the CETP concentration genetic score by sex on biomarkers and cardiovascular outcomes in the UK Biobank.** Displayed p-values (p_{itx}) are for the two-sided test of the product interaction term between the CETP score and a binary sex indicator variable.

and apoA levels (interaction $p = 5.2 \times 10^{-24}$) in the UK Biobank. Individuals with lower BMI had higher HDL-c and apoA levels per 1 s.d. decrease in the genetic CETP score. BMI was also a modulator of the association between the CETP genetic score and LDL-c (interaction $p = 0.00013$) and apoB levels (interaction $p = 0.0012$) with a lower BMI being associated with lower levels of LDL-c and apoB per s.d. decrease of the CETP genetic score (Figure 3.3). The interaction term was also significant for lp(a) supporting that BMI is also a modulator of the association between the CETP genetic score and lp(a) ($p = 0.0027$, Figure 3.3). In individuals of normal BMI, one s.d. reduction in the genetic CETP score was associated with 0.018 s.d. lower lp(a) (95% CI 0.014, 0.023); $p = 8 \times 10^{-15}$ and individuals of obese BMI had 0.0081 s.d. lower lp(a) (95% CI 0.0025, 0.014); $p = 0.004$ (Supplementary Figure B.6). The interaction between the genetic CETP score and BMI was not significant for C-reactive protein ($p = 0.24$). In the MHI Biobank, there was no interaction between BMI and the genetic CETP score on basal and cAMP-stimulated cholesterol efflux ($p = 0.31$ and $p = 0.84$, respectively). Results based on the rs1800775 variant were concordant with

those based on the genetic CETP score (Supplementary Figure B.7).

To allow for non-linear effects of BMI and the CETP genetic score, we used linear and logistic regression with interacting restricted cubic splines to model these two variables. There was evidence of nonlinear effects for BMI on all considered cardiovascular outcomes and biomarkers except for the cholesterol efflux (Supplementary Table B.6). The CETP genetic score exhibited possible nonlinear effects on LDL-c ($p = 0.0013$) and HDL-c ($p < 0.0001$) and their associated lipoproteins. When considering all linear and nonlinear interaction terms (9 degrees of freedom test), there was evidence for interaction between the CETP score and BMI for lipoprotein(a) levels ($p = 0.0233$), HDL-c ($p < 0.0001$), apolipoprotein A ($p < 0.0001$), LDL-c ($p = 0.0001$), and apolipoprotein B ($p = 0.0147$). There was no statistically significant nonlinear interaction with any of the cardiovascular outcomes. To facilitate the interpretation of the nonlinear interactions for biomarkers, we plotted the predicted value of the standardized outcome while varying the levels of the CETP genetic score and BMI (Supplementary Figure B.8).

We tested the modulatory effect of BMI on the relationship between the CETP genetic score and cardiovascular outcomes. None of the tested outcomes had statistically significant interactions between BMI and the CETP genetic score, but the trends were directionally consistent with the effects on the lipid profile on the multiplicative scale (Figure 3.3, Supplementary Table B.7). For instance, the interaction coefficient p-value for CAD was 0.060 and the marginal OR of the CETP score on CAD was 0.960 (95% CI 0.941, 0.980) when fixing BMI at 21.75 (normal) versus 0.986 (95% CI 0.971, 1.00) when fixing BMI at 33.75 (obese). Similar results are observed in subgroups of individuals based on their BMI class (Supplementary Figure B.6).

In sensitivity analyses, we show that the modulatory effects of BMI are robust to further adjustment for type II diabetes, ruling out the possibility of mediation of the BMI modulatory effect through diabetes (Supplementary Figure B.10).

Because both BMI and sex were important modifiers of the effect of CETP on biomarkers, we evaluated the possibility of a three-way interaction between sex, BMI and the CETP genetic score. There was sexual dimorphism in the interaction between BMI and genetic CETP levels on LDL-c (3-way interaction $p = 0.00041$) and apoB levels ($p = 0.001$). These results are further described in the Supplementary Appendix (Section B.1).

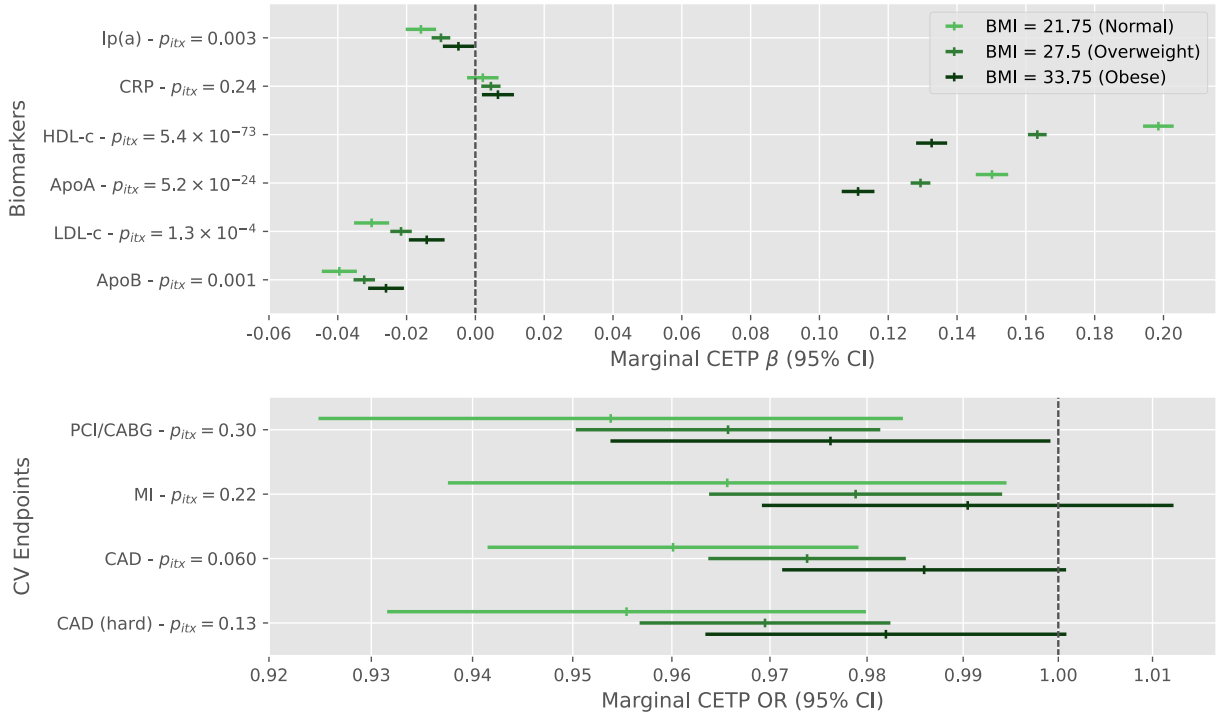


Figure 3.3 – **Effect modification by body mass index of a 1 standard deviation decrease in the CETP concentration genetic score on biomarkers and cardiovascular outcomes in the UK Biobank.** Displayed p-values (p_{itx}) are for the two-sided test of the product term between the CETP score and standardized body mass index.

3.6 Discussion

Using the large UK Biobank resource, supported by cholesterol efflux measurements in the MHI Biobank, we report on the effect of genetically lower CETP on lipid biomarkers, cholesterol efflux, CRP, and cardiovascular outcomes. We assessed how sex and BMI changed the effect of a genetically predicted decrease of CETP on those measurements and outcomes. We report significant modulatory effect of sex and BMI on the association of CETP with lipid biomarkers and cholesterol efflux, but we were unable to show that those benefits extended to cardiovascular outcomes.

In our analyses, we observed that a genetically predicted lower CETP concentration was strongly associated with higher HDL-c and apoA levels and, to a lesser extent, to lower LDL-c and apoB levels. These results are concordant with previous reports of the effect of CETP on lipids and lipoproteins [53, 54]. We also observed lp(a) levels to be slightly, but significantly, lower in individuals with a genetically predicted reduction in CETP. This observation is

interesting as lp(a) is an important risk factor for CAD that is largely independent of other lipoproteins. Previous studies have reported that both torcetrapib and anacetrapib could reduce lp(a) levels [198, 200, 203]. The added genetic support could be indicative of a class effect of CETP inhibitors. In a mendelian randomization study of lp(a) levels, a 10 mg/dl genetic reduction in lp(a) was associated with an OR of 0.942 for coronary heart disease supporting the causal role of lp(a) in coronary heart disease [151]. In our study, we show that a 1 s.d. decrease in the genetic CETP score was associated with a reduction in lp(a) of about 2 mg/dl which corresponds to an OR for coronary artery disease of 0.988 based on this previous MR study.

Genetically lower CETP was not associated with C-reactive protein. In the dal-OUTCOMES trial of dalcetrapib and the ACCELERATE trial of evacetrapib, CETP inhibition was associated with an increase in C-reactive protein, but there was no significant difference in the DEFINE trial of anacetrapib [196, 197, 204]. Given our high power to detect an association with biomarker measurements in the UK Biobank, it is unlikely that genetically predicted lifelong lower CETP levels has an effect on C-reactive protein levels.

We have also found that genetically lower CETP was associated with higher levels of cholesterol efflux. This observation supports a previous report of increased cholesterol efflux in patients treated with dalcetrapib [57]. Genetically lower CETP was also associated with lower rates of cardiovascular outcomes, with 2.5% fewer CAD events per s.d. decrease in the CETP genetic score. The observed protective effect was robust to adjustment for observed apoB levels at baseline in the UK Biobank, supporting that the protective effect of CETP may not be exclusively mediated by apoB levels. However, adjusting for observed apoB levels at a single time-point may not completely control for a lifetime reduction in apoB levels by CETP genetic variants. Nonetheless, we presented a unified portrait of the effect of genetically lower CETP which may help better understand class effects of CETP inhibitors and the diversity of pathways through which genetic variants in CETP may exert a protective effect on CAD.

The value of using human genetic variants to predict the effect of pharmacological modulation of drug targets is gaining recognition as more examples of drug target discoveries and predictions of ongoing randomized trials are reported [145, 205]. Human genetics can also inform on subgroup effects in the context of precision medicine, for clinical trial design, or to estimate external validity of drug effects to other patient populations. Genetic studies of CETP variants have highlighted possible effect modification by sex on HDL-c levels, the HDL-c / apoAI ratio and on the dynamics of postprandial triglyceride levels [183, 187, 206]. Sex differences were also observed in many traits thought to be involved in the athero-

protective effect of CETP such as the apoAI and apoA-II composition of HDL and CETP mediated cholesterol efflux [207, 208]. Whether this translates to cardiovascular outcomes remains unknown [184]. Considering that female sex is underrepresented in the majority of cardiovascular clinical trials, sex-differences may have important clinical implications as inferences drawn from the unbalanced trial populations are used to inform treatment.

In our analyses, we found sex to be a strong modulator of the effect of genetically lower CETP on lipid biomarkers and cholesterol efflux. Women with CETP lowering genetic variants had lower LDL-c and apoB levels and higher HDL-c, apoA and cholesterol efflux. In a substudy of the DEFINE trial, anacetrapib increased cholesterol efflux in men, but not women [209]. In that study, cholesterol efflux capacity was measured using fluorescently labeled cholesterol which mostly captures efflux through the ABCA1 pathway. Here, we used radiolabeled cholesterol and our measurements in cAMP stimulated conditions also include the contributions of ABCG1 and SR-B1. In healthy subjects, serum from women had more SR-B1 mediated efflux capacity whereas serum from men had more ABCA1-mediated efflux capacity [210]. We suggest that, on average, with genetically lower CETP, women have more overall cholesterol efflux capacity than men, but that specific pathways may show the inverse relationship. The modulatory effect of sex did not translate to strong changes in cardiovascular outcomes in our analysis even though the additive interaction model suggested that women had a stronger reduction in the relative risk of "hard" CAD than men for a same reduction in the CETP score. Because the UK Biobank contains relatively few cardiovascular events, especially in women, and the genetic variants in CETP have a limited effect size, a replication from a large, well-powered study is warranted to confirm these findings. A previous study has also reported sex differences in the association between CETP genetic variants and cardiovascular disease [211], but sex-interaction was not formally assessed and the sample size and power may have been limited in that study (n=866).

We conducted a meta-analysis of the sex-stratified results of three RCTs of CETP inhibitors and although we did observe a nominally greater protective effect in women (RR=0.92) than in men (RR=0.96), the difference in effect was not statistically significant (p=0.50). A total of 10,769 women pooled across all three studies were included in the meta-analysis for a total of 58,412 participants, which is still lower than the number of individuals included in the smallest study (n = 12,092 for ACCELERATE), suggesting limited power to identify a subgroup effect in women. We conclude that there is some indication of a stronger beneficial effect of CETP inhibition in women, but that confirmation using genetic datasets enriched for cardiovascular events or clinical trials with a greater representation of women would be needed.

Plasma CETP is mainly secreted by macrophages in the liver (Kupffer cells) and adipose tissue does not appear to be a clinically relevant contributor to circulating CETP levels despite statistically significant differences [212, 213]. However, higher BMI may affect hepatic and systemic inflammation and result in changes in lipid homeostasis that could alter the function of CETP without affecting plasma CETP concentration.

In our analyses, the atheroprotective profile of lipoproteins attributable to genetically lower CETP levels was stronger in individuals with lower BMI, as compared to individuals with higher BMI, with lower levels of LDL-c, apoB and lp(a) and higher levels of HDL-c and apoA with genetically lower CETP. Similar results were observed in models allowing for nonlinear effects on the biomarkers.

The modulatory effect of BMI on the relationship between CETP and biomarkers did not translate to cardiovascular outcomes, however, a larger dataset may be needed to assess the possible impact on outcomes with sufficient power.

We also found evidence of 3-way interactions of sex, BMI and genetically predicted CETP on LDL-c, apoB levels and cholesterol efflux. The attenuated effect of lower CETP on LDL-c reduction with increasing BMI was specific to men. In women, the increase in cholesterol efflux by genetic reduction in CETP was attenuated with increasing BMI, but this effect was sex-specific.

Our study had some limitations. We relied on common genetic variants to model pharmacological CETP inhibition, but these variants do not include rare mutations which can have much stronger effects on CETP function. CETP activity can also be modulated in more subtle ways than complete inhibition, with molecules such as dalcetrapib preserving pre-β-HDL formation, a function that is inhibited by anacetrapib [214]. This effect is due to differences in the CETP-mediated HDL remodeling through homotypic transfer of cholesteryl esters and may play an important role in atherosclerosis as pre-β-HDL are important acceptors for ABCA1-mediated cholesterol efflux. Our study could not distinguish between these two profiles of CETP modulation as measurements of ABCA1-mediated efflux or HDL subtypes were not available. CETP may also have an important intracellular role in storing triglycerides and cholesteryl esters in lipid droplets [215, 216]. Whether this activity was altered in our genetic models or played a role in the effect modification by BMI remains to be determined. Also, the estimated effects derived from the genetic variants relate to lifelong exposure to lower CETP concentrations, which may differ from the effects of short-term exposure to pharmacological inhibition of CETP. In addition, although the UK Biobank offers large numbers of study participants, the cohort has a limited number of cardiovas-

cular events. A case-control cohort with larger numbers of CAD events may help increase power to assess the translation of the detected modulatory effects of sex and BMI on lipid biomarkers and cholesterol efflux to cardiovascular outcomes. We did not adjust for multiple testing of phenotypes and effect modifiers in this study which evaluated several correlated phenotypes. We reported confidence intervals and provided power analyses to support the interpretation of results. The genetic variants at the CETP gene have concurrent effects on multiple biomarkers, making it difficult to disentangle the effect of the individual biomarkers on cardiovascular outcomes. The polygenic modeling of multivariable exposures may be an interesting approach to consider for this purpose [143].

In this study, we have evaluated the effect of a genetically predicted reduction in CETP concentration on lipoproteins, lipid fractions, cholesterol efflux and C-reactive protein. We have found results to be largely concordant with those obtained from clinical trials. Using statistical interaction models, we found that sex and BMI are modulators of the effect of CETP on lipid biomarkers and cholesterol efflux.

3.7 Acknowledgments

We thank the UK Biobank for providing the data under Application Number 20168. We thank the Montreal Heart Institute (MHI) Biobank for providing access to samples and data.

3.8 Funding Sources

The work was funded by the Canadian Institutes of Health Research (CIHR) and the Health Collaboration Acceleration Fund from the Government of Quebec. The Montreal Heart Institute (MHI) Biobank is funded by the MHI Foundation.

MAL holds a scholarship from Canadian Institutes of Health Research (CIHR); MPD holds the Canada Research Chair in Precision medicine data analysis. JCT holds the Canada Research Chair in Personalized Medicine and the Université de Montréal Pfizer-endowed research chair in atherosclerosis.

JGH is an IVADO Professor and a Fonds de la Recherche en Santé (FRQS) Junior 1 fellow. IG receives a PhD scholarship from the MHI Foundation.

3.9 Disclosures

JCT reports grants from the Government of Quebec, Amarin, Esperion, Ionis, Servier, RegenXBio; personal fees from AstraZeneca, Sanofi, Servier; and personal fees and minor equity interest from Dalcor. In addition, JCT is author on a submitted patent Methods of treating a coronavirus infection using Colchicine pending, and a patent Early administration of low-dose colchicine after myocardial infarction pending. MPD and JCT have a patent Methods for Treating or Preventing Cardiovascular Disorders and Lowering Risk of Cardiovascular Events issued to Dalcor, no royalties received, a patent Genetic Markers for Predicting Responsiveness to Therapy with HDL-Raising or HDL Mimicking Agent issued to Dalcor, no royalties received, and a patent Methods for using low dose colchicine after myocardial infarction, assigned to the Montreal Heart Institute. MPD reports personal fees and other from Dalcor and personal fees from GlaxoSmithKline, other from AstraZeneca, Pfizer, Servier, Sanofi.

JH has received speaker honoraria from Dalcor and District 3 Innovation Centre.

CHAPITRE 4

PheWAS analysis of human protein coding loci using a principal components approach

Les approches panphénomiques (PheWAS) sont bien adaptées à la caractérisation de cibles pharmacologiques, car elles permettent d'identifier rapidement des conséquences souhaitables ou indésirables associées à la modulation de gènes encodant ces cibles. Bien qu'il existe plusieurs bases de données recensant les associations génétiques basées sur les variations individuelles, il existe peu de bases de données utilisant une approche gène-centrique. Comme elles utilisent l'agrégation de plusieurs variations, de telles approches peuvent être plus puissantes si plus d'une variation causale existe et elles sont soumises à des corrections moins sévères pour les tests multiples. Finalement, elles peuvent être utilisées dans des analyses bio-informatiques basées sur des annotations géniques comme l'enrichissement de termes ontologiques, de voies biologiques ou de cibles pharmacologiques.

Face à ces avantages, nous avons mené un PheWAS gène-centrique. Pour tous les gènes autosomaux encodant des protéines, nous avons utilisé une approche d'analyse en composante principale pour tester l'association avec 1 210 phénotypes disponibles dans la UK Biobank. Ces phénotypes incluent des mesures anthropomorphiques et de laboratoire, des maladies autorapportées au recrutement à l'étude ainsi que des données sur les hospitalisations ou les registres de mortalité. Les résultats de cette étude sont présentés de façon interactive à l'aide d'un fureteur nommé *ExPheWas* que nous avons développé et qui est disponible publiquement (exphewas.statgen.org). Un accès informatique automatisé est aussi disponible à l'aide d'une interface de programmation d'application (*Application Programming Interface*, API). Afin de démontrer les capacités de l'approche, nous avons caractérisé les gènes associés à la fibrillation auriculaire en utilisant ExPheWas. Ces gènes étaient enrichis pour

plusieurs annotations pertinentes incluant la contraction du muscle cardiaque (terme GO de processus biologique) et les disques Z (terme GO de composante cellulaire). Les gènes associés à la fibrillation auriculaire étaient aussi enrichis parmi les cibles pharmacologiques de médicaments antiarythmiques. Finalement, nous avons identifié le gène de la myotiline comme contribuant à la fibrillation auriculaire.

Contributions : Marc-André Legault, Louis-Philippe Lemieux Perreault et Marie-Pierre Dubé ont participé à la conception de l'étude. Marc-André Legault a mené l'étude pheWAS. Marc-André Legault et Louis-Philippe Lemieux Perreault ont développé le fureteur de résultats. Marie-Pierre Dubé a supervisé le projet. Marc-André Legault et Louis-Philippe Lemieux Perreault ont participé à la configuration du serveur et au déploiement de l'application. Louis-Philippe Lemieux assurera le développement et le support futur de l'application. Les trois auteurs ont fait des contributions majeures au manuscrit.

***ExPheWas*: a browser for gene-based pheWAS associations**

To be submitted. Preprint available on *medRxiv*. doi:10.1101/2021.03.17.21253824v1

Marc-André Legault[1,2,3], Louis-Philippe Lemieux Perreault[1,2], Marie-Pierre Dubé[1,2,3]

1. Montreal Heart Institute, Montreal, Canada
2. Université de Montréal Beaulieu-Saucier Pharmacogenomics Centre, Montreal, Canada
3. Université de Montréal, Department of biochemistry and molecular medicine, Montreal, Canada

4.1 Structured Abstract

4.1.1 Motivation

The relationship between protein coding genes and phenotypes has the potential to inform on the underlying molecular function in disease etiology. We conducted a phenome-wide association study (pheWAS) of protein coding genes using a principal components analysis-based approach in the UK Biobank.

4.1.2 Results

We tested the association between 19,114 protein coding gene regions and 1,210 phenotypes including anthropometric measurements, laboratory biomarkers, cancer registry data, hospitalization and death record codes and algorithmically-defined cardiovascular outcomes. We report the pheWAS results in a user-friendly web-based browser. Taking atrial fibrillation, a common cardiac arrhythmia, as an example, ExPheWas identified genes that are known drug targets for the treatment of arrhythmias and genes involved in biological processes implicated in cardiac muscle function. We also identified MYOT as a possible atrial fibrillation gene.

4.1.3 Availability and implementation

The ExPheWas browser and API are available at https://exphewas.statgen.org/

4.2 Introduction

Uncovering the phenotypic roles of human protein coding genes is an important goal of genetics to help improve our understanding of molecular physiology and disease pathology. Genetic variation provides a tool for predicting the consequences of altering the functions of a protein and is an important first step in validating drug targets for the development of therapies. Rare loss-of-function variants can be used to assess the impact of disrupting protein function, but they are frequently untyped in large population cohorts that rely on genotyping platforms. In such cohorts, it may be possible to use many common variants with small effects to establish the association between a locus and phenotype. Gene-based pheWAS analysis can also support annotation efforts as follow-up to a GWAS discovery [72].

Many computational approaches to collapse common variants into gene-based association tests have been developed and an overview of methods is shown in Table C.1. In general, one of the major challenges is to account for the correlation between genetic variants due to linkage disequilibrium (LD) that induces correlation between the marginal association statistics. Methods have heterogeneous performance with respect to the underlying genetic architecture, the inclusion of common and rare variants and the presence of interactions [87, 91]. The required input may also be a concern as some methods require individual-level genotypes (*e.g.* SKAT or PLINK SNP set) whereas others rely on summary association statistics and a population-matched LD reference panel (*e.g.* VEGAS or GATES). Computational burden also varies with algorithms relying on permutation tests to derive empirical p-value typically being slower [217]. Among popular techniques are the use of a simulation-based approach [83], kernel-based association tests [92] and principal component analysis (PCA)-based methods [89]. More details and flexible software packages implementing different methods have been previously published [218]. Using a gene-based association testing approach reduces the multiple hypothesis testing burden and facilitates downstream bioinformatics analyses that require gene-level annotations, such as gene set enrichment analysis.

Here, we implemented an efficient PCA-based association testing method suitable for pheWAS analysis of protein coding genes in the UK Biobank. Our phenome-wide scan considered anthropomorphic measurements, laboratory measurements, hospitalization or death codes, algorithmically-defined cardiovascular outcomes as well as self-reported diseases. We

developed a web-based results browser available at http://exphewas.statgen.org/ and provide programmatic access through a publicly available API. We also characterize our association testing strategy in terms of power and hyperparameter sensitivity by using well-known drug target genes as examples. Finally, we used atrial fibrillation, a common cardiac arrhythmia to demonstrate how our resource may be used in practice.

4.3 Methods

4.3.1 UK Biobank and genetic quality control

The UK Biobank is a longitudinal population cohort of more than 500,000 individuals. All participants visited a recruitment center between 2006 and 2010 where urine and blood samples were collected allowing for the measurement of an extensive panel of biochemical markers. A touchscreen-based questionnaire was also filled and followed-up with a verbal interview with a nurse allowing participants to self-report a wide variety of diseases. Genetic data derived from a genome-wide genotyping array was also collected and imputed to about 96 millions genetic variants [106].

Because of the high throughput nature of our study, we conducted a strict genetic quality control to reduce the risk of bias due to poor genotyping or imputation quality or population stratification. We excluded all variants and individuals with more than 2% of data values missing. To avoid bias due to cryptic relatedness, we also randomly selected one individual from pairs predicted to be related using a kinship coefficient corresponding to a 3^{rd} degree relationship (0.0884) as a threshold. Our analysis also included only individuals from the largest genetically homogeneous population in the UK Biobank corresponding to individuals of European ancestry. We excluded individuals from different ancestry based on self-reported data or outliers from a manually defined cluster in the genome-wide PCA plot. After these steps, a total of 413,138 individuals remained for analysis.

4.3.2 Creation of gene-based PCs

To create a compact representation capturing genetic variability within autosomal gene regions, we conducted PCA with genetic variants of minor allele frequency (MAF) of 0.01 or greater at every locus. We used the Ensembl gene boundaries and added a padding of 3 kilobases (gene boundaries $\pm$ 1.5 kilobases), extracted all additively encoded genotype dosages and conducted the PCA using the implementation from scikit-learn in the Python programming language [219]. The projection on the space spanned by the PCs was saved

for all samples.

We excluded genes located on sexual chromosomes because systematic encoding differences between men and women risked creating spurious associations. For example, using the gene *ACE2* located on the X chromosome, we used a logistic regression model to test if the first 11 PCs, explaining 95% of the variance in the genotypes, could predict the individual's sex. The log-likelihood ratio test of the joint effect of the 11 included PCs, had a p-value of 0.032 suggesting that the gene-based PCs were different between men and women. The sex-stratified computation of gene-based PCs is likely to address this problem and will be considered for future releases.

4.3.3 Association testing

Different approaches have been developed for association testing of common variants at a gene locus [83, 89, 218]. Because our application, a pheWAS, required a large number of tests, we focused on an approach using dimensionality reduction as a first step to reduce the computational burden. We opted for a principal component analysis-based association model that was proposed in the MAGMA software [88, 89]. We also propose an extension of the method based on the analysis of deviance which can be used for non-continuous outcomes. This approach is valid for any model using maximum likelihood estimation including generalized linear models.

For continuous traits, an F test is used to compare two nested models as in MAGMA. First, a null model regresses the outcome on covariates such as age, sex and genome-wide principal components to adjust for confounding due to ethnicity. The association test is based on the gain in goodness of fit (or lack thereof) after adding gene-based PCs to the null model. More formally, given two nested models with P_1 and P_2 representing the set of parameters where $P_1 \subset P_2$, then one can compute the F statistic given by

$$F = \left(\frac{\mathrm{RSS}_1 - \mathrm{RSS}_2}{|P_2| - |P_1|} \right) \div \left(\frac{\mathrm{RSS}_2}{n - |P_2|} \right)$$

Where RSS_i and $|P_i|$ are the residual sum of squares and number of parameters of the i^{th} model, respectively, and n is the number of samples. This statistic follows an F distribution with $(|P_2| - |P_1|, n - |P_2|)$ degrees of freedom under the null hypothesis that the second model does not improve the residual sum of squares.

A similar approach for generalized linear models has been described based on the difference in deviance and was termed the analysis of deviance [220]. The difference between the

deviances of the nested models follows a χ^2 distribution with $|P_2| - |P_1|$ degrees of freedom under the null hypothesis. This extension can be used in the context of logistic regression as used in our analyses. In this setting, an equivalent approach based on the likelihood ratio test was also previously suggested [88].

Both the F-test and analysis of deviance based methods were implemented in R under the "anova" function that takes fitted models as parameters (*i.e.* output from the *lm* or *glm* functions). Because of the large number of tests in our analysis, we used the optimized fastglm (v0.0.1) package when fitting logistic regression models for binary outcomes (https://github.com/jaredhuling/fastglm). This package uses the Cholesky decomposition to optimize the iterative reweighted least squares algorithm which drastically reduced the time required for logistic regressions.

In the pheWAS study, we repeated the association tests about 25 million times. This certainly results in false positives and we used a false discovery rate (FDR) approach to control for multiple hypothesis testing. As we believe that considering all tests together is not suitable for most research questions, we integrated a FDR mechanism on a per gene or per phenotype basis. Specifically, we used the test p-values to compute corresponding q values designed so that if all q values ≤ 0.05 are considered significant, then 5% of the significant features will be false discoveries on average [221]. For example, when browsing results for a gene of interest, the displayed q-values will be based on the 1,127 tested phenotypes. Browsing the same results by "outcome" will result in different q-values because the latter would be based on all tested genes in association with the outcome of interest. In all cases, the uncorrected p-value and the Bonferroni corrected p-value are also provided. When browsing associations for a given gene, a quantile-quantile plot of association p-values is also displayed on the results browser along with the λ inflation factor corresponding to the ratio of the median observed association statistic to the median expected association statistic. The inflation factor is a quantitative measurement of deviation from the null and high λ (above one) values represent loci that contribute to many phenotypes (*i.e.* pleiotropic loci). For example, the *HLA-DQB1* gene is notoriously pleiotropic and has $\lambda = 2.48$.

In a sensitivity analysis exploring the optimal threshold for the variance explained for selected drug target genes, we repeated the association test while varying the number of included PCs. When the association p-values were numerically close to zero for many choices of PCs, we randomly selected a subset of individuals to avoid p-values numerically equal to zero hampering comparison between thresholds.

4.3.4 Power analyses

We simulated continuous phenotypes for every simulation replicate using two models representing a scenario with a single causal variant and a scenario where all variants contribute to the heritability in a random effects model. For the single causal variant model, we randomly sampled a causal variant and fixed its effect coefficient (β) as a simulation parameter. We then added gaussian noise so that the total variance in the continuous simulated phenotype was 1. In the second model, every variant had an effect sampled from a standard normal distribution and gaussian noise was added to achieve the heritability (h^2) set as a simulation parameter [222]. The simulations used real genotype data from the *PCSK9* locus in the UK Biobank containing 138 common variants as a typical example of phenotype-associated protein coding gene of median length. We randomly selected 20,000 individuals and repeated the simulation procedure 1,000 times. For simulated phenotypes, we tested the association between every genotype at the locus and the simulated phenotype using linear regression. We estimated the power for the conventional linear regression model as the fraction of simulation replicates where the minimum variant p-value was under a bonferroni corrected threshold of 0.05/138. For the gene-based association test, we used the fraction of F tests with a p-value under 0.05 across simulation replicates.

4.3.5 PheWAS analysis

For every gene, we tested the joint association of principal components explaining 95% of the genetic variance at the gene locus with 83 continuous phenotypes and 1,127 binary phenotypes. The continuous phenotypes include anthropometric measurements (such as height or body mass index), laboratory measurements (such as lipoprotein levels, glycemic traits or blood cell traits) and imaging measurements (such as the intima-media thickness). Due to the high throughput approach used, we manually transformed continuous variables so that they were approximately normally distributed. This step was taken to avoid violating the assumption of normally distributed residuals in linear regression as it was not feasible to run conventional model diagnostics for all gene-phenotype pairs individually. The description of the included continuous variables along with the transformation used, if needed, is presented in Supplementary Table C.2.

For the binary phenotypes, we used the self-reported diseases from the UK Biobank and the ICD10 codes reported in the hospitalization or death records. We also grouped the ICD10 codes according to two hierarchical levels to represent increasingly broad disease definitions. Specifically, we tested ICD10 blocks (*e.g.* I10-I15 Hypertensive diseases) as well

as more precise 3 character codes (*e.g.* I11 Hypertensive heart disease). For cancer data, the UK Biobank also provides linkage with a national cancer registry. For all ICD10 codes corresponding to neoplasms (C00-D49), we used the cancer registry data instead of the hospitalization data. Because of the shared etiology between cancer subtypes, we also excluded individuals with any cancer from being controls for analyses of other cancer subtypes. For hospitalization or death records, the initial analysis (data release v0.1) included hospitalization records up to 2016-03-13 and the latest recorded date of death was 2016-02-16. Codes from either the primary or secondary codes were used in the pheWAS to maximize the number of cases and statistical power.

We also included algorithmically-defined cardiovascular outcomes as phenotypes in the pheWAS. This includes coronary artery disease, angina, heart failure, myocardial infarction and coronary revascularization procedures (percutaneous coronary intervention and coronary artery bypass graft). The codes and procedure used to define these variables is presented in Supplementary Table C.3.

To efficiently conduct the PheWAS analysis of the previously described gene-based PCs and phenotypes, we used a custom R script (tested under R v3.6.0) that is available online (https://github.com/legaultmarc/UKBPheWAS).

4.3.6 Interface and API

To present the results of the PheWAS, we developed an API and a web-based results browser. The *Python* programming language (v3.8) and *Flask* web framework (v1.1.1) were used to develop both. We used the *SQLAlchemy* (v1.3.13) object relational mapper and SQL engine to build the results database and the publicly available instance uses *PostgreSQL* (v9.6.3). The code for the database models, API endpoints, web browser endpoints and the JavaScript frontend are publicly available online https://github.com/pgxcentre/ExPheWAS. Interactive visualizations integrated in the web application were developed using *D3.js* (https://d3js.org/). These include median tissue expression of genes in *GTEx* V8 used to contextualize the expression patterns of the tested protein coding loci as well as the drug target enrichment analyses.

To ensure long term availability, the service is hosted on the Compute Canada Cloud (https://www.computecanada.ca/home/).

4.3.7 Enrichment analyses

We integrated enrichment analysis utilities to the API and results browser. For a given phenotype, it is possible to test if the associated genes are enriched in drug targets. To achieve this, we used the ChEMBL database to map drug target genes to Anatomical Therapeutic Chemical Classification (ATC) codes representing drug classes [223]. The ATC codes are structured hierarchically in a 5 level system where the first level indicates the anatomical group (*e.g.* C represents drugs acting on the "cardiovascular system") and the fifth level represents individual drugs (*e.g.* C07AB07 represents bisoprolol, a specific beta-blocker molecule). For enrichment analyses, we used two complementary approaches. The first approach is based on a Fisher exact test of the 2×2 contingency table of the number of genes associated with the phenotype and drug class at a q-value ≤ 0.05 level. We also provided results from a *Fast Gene Set Enrichment Analysis* (FGSEA) implementation where ATC codes are treated as pathways and the association statistics are ranked to evaluate enrichment [224, 225]. One advantage of this approach over the Fisher exact test is that it does not categorize genes as associated or not based on a q-value threshold.

The results of enrichment analyses are displayed on the ExPheWas browser results page as an interactive tree with collapsible nodes. A color scale is used to represent enrichment p-value. For every ATC node, the fill color represents the node's enrichment p-value and the stroke color represents the minimum p-value in the subtree rooted at the current node.

For the atrial fibrillation example, we also tested enrichment of various ontology terms within the genes associated with atrial fibrillation with $q \leq 0.01$ (137 genes). The tested ontologies included the *Gene Ontology*, the *KEGG* pathways, and the *Human Phenotype Ontology*. This analysis was done using the *g:Profiler* web-based tool [226].

4.4 Results

4.4.1 Gene PCA

We used the Ensembl 87 database as a reference for the human protein coding genes. Out of the 20,356 genes, 926 were not on autosomes and another 316 genes either had no common variants in the UK biobank imputed dataset or did not converge because of unreasonable memory requirements during association testing. In total, 19,114 genes on autosomal chromosomes were included in the final analysis.

To efficiently represent the majority of the genetic variation at every protein coding locus,

we used PCA of the additively encoded genotypes and retained the number of principal components necessary to explain 95% of the total variance. As a sensitivity analysis, and for a subset of the genes, we also tested the inclusion of PCs explaining 99% of the total variance. Results for both analyses are presented in the ExPheWas browser. The first principal components generally assigned more weight to genetic variants that are highly correlated to many other variants (*i.e.* good "tagging" variants). Supplementary Figure C.1 shows a representative example of this pattern. We also note that the principal components (PCs) explaining a smaller portion of the variance (later components) tended to attribute larger weights to individual variants that are less correlated with other variants (smaller LD score). The number of principal components required to capture 95% of the genetic variance at a gene region varied greatly between genes, and was largely determined by the size of the region and LD structure.

4.4.2 Validation of the association testing approach

Marginal association of PCs

To gain insight into the effect of PCs on phenotypes, we used the well known example of the proprotein convertase subtilisin/kexin type 9 (PCSK9), a protein implicated in the recycling of low density lipoprotein (LDL) receptor and consequently of LDL cholesterol levels and coronary artery disease (CAD) [15]. When regressing individual *PCSK9* PCs in univariate models with LDL cholesterol and CAD, we observe that the first PCs have lower standard errors as they capture the overall effect of many common variants with small weights (Supplementary Figure C.2). Subsequent PCs that explain a smaller proportion of the variance and rely more on individual variants in low LD have larger standard errors, but may still have strong effects on the phenotypes. However, these effects may rely on individual variants for which traditional association tests may be more powerful. For example, PC14 in *PCSK9* is strongly associated with LDL-c levels ($\beta = 0.038$, 95% CI $[0.035, 0.041]$ mmol/l, $p = 2 \times 10^{-147}$) and CAD ($\beta = 0.026$, 95% CI $[0.015, 0.037]$ in log odds, $p = 2 \times 10^{-6}$) despite explaining 0.86% of the overall variance in genotypes. For this PC, the variant rs11591147 is an outlier of the component weight distribution and this missense *PCSK9* variant has a strong effect on LDL-c in the GLGC ($\beta = 0.50$, 95% CI $[0.46, 0.53]$, $p = 9 \times 10^{-143}$) and coronary artery disease in the CARDIoGRAMplusC4D consortium (allelic OR 1.29, 95% CI $[1.16, 1.45]$, $p = 7 \times 10^{-6}$) [62, 227].

It is also noticeable that some PCs are positively correlated with partial gain of function whereas others represent partial loss of function as can be seen by the positive and negative

effects in Supplementary Figure C.2. Finally, in the example of *PCSK9* the first PCs have small effect sizes on LDL cholesterol and CAD. This pattern is not systematically observed with other genes as it may depend on LD structure and selective pressures. For example, we observed the opposite trend when considering the *CETP* gene encoding the cholesteryl ester transfer protein a well known drug target associated with high density lipoprotein cholesterol levels. For this gene, the first PCs have among the strongest effects of all individual components (Supplementary Figure C.3).

Joint association testing of PC

The association model we used requires the selection of the number of PCs to include. This choice is important as including PCs with no phenotypic effect will result in decreased statistical power. If there is a single gene of interest, it is possible to tune this hyperparameter using cross-validation, but for a pheWAS approach and when the associated phenotype is not known a priori, using a fixed threshold is necessary. To assess power with respect to the number of included PCs, we used drug targets that are known to be associated with various continuous or discrete phenotypes (Supplementary Table C.4). We assessed the association p-value based on an increasing number of included PCs in association tests with selected phenotypes (Supplementary Figure C.4). There was no best choice in the number of PCs to include to maximize power that was shared for all genes. For example, a single PC explaining 36% of the variance maximizes power for the association between *HMGCR* and LDL cholesterol with decreasing benefit afterwards. On the contrary, power is maximized after including 29 PCs explaining 92% of the variance for *GLP1R* and glycated haemoglobin.

To increase our chance of detecting associations driven by less common variants and to avoid false negatives, we selected the threshold of 95% of the variance explained for the pheWAS analysis in the UK Biobank.

4.4.3 Power analyses

The first stop for many investigators in search of gene-phenotype associations are online catalogs such as the OpenTarget genetics platform, the GWAS catalog, the SAIGE-based PheWeb platform and PhenoScanner [94, 98, 102]. These portals mostly focus on reporting variant-based association statistics and so the strongest variant-phenotype associations within a gene are typically of interest. For this reason, gene-level association testing is most commonly based on using the minimum association p-value within the gene boundaries. As we aim to provide a complementary resource to these portals, we compared the behaviour of the PC-based association method to this approach in terms of power and false positive rate.

In a scenario where there is a single causal variant, the minimum p-value in the region approach is more powerful (Supplementary Figure C.5). However, under a random effects model where all variants make small contributions to the heritability, the PC-based approach is slightly more powerful and the estimated power was numerically higher for the PC-based approach for 97% of the simulated h^2 values (Supplementary Figure C.5). When the simulated heritability was zero (null model of no genetic effect), the fraction of simulation replicates where the null was rejected was 4.9% (95% CI 3.6%, 6.2%) and 1.8% (95% CI 1.0%, 2.6%) for the PC-based approach and minimum linear regression p-value approach, respectively. In both cases these false positive rates were close to the nominal levels ($\alpha = 0.05$). The lower false positive rate for the minimum linear regression p-value approach was likely due to the conservative Bonferroni correction which does not account for linkage disequilibrium.

4.4.4 PheWAS

To estimate the association of human protein coding genes with multiple phenotypes, we used a phenome-wide association study approach [72]. We tested the association between 19,114 protein coding gene regions and 1,210 phenotypes in the UK Biobank for a total of about 23 million tests (Figure 4.1). Some of these tests however are redundant as phenotypes may be correlated and genomic regions may overlap.

We also provide information to enable the assessment of the statistical significance of associations, including the raw p-value, the conservative Bonferroni p-value and the q-value that is related to the false discovery rate. When browsing results for a gene, the q-value will account for all tested phenotypes in association with that gene. When browsing results for a phenotype, then the q-value will control for all tested genes in association with that phenotype. This enables multiple testing adjustments for the specific questions: “what genes are associated with this phenotype?” or “what phenotypes are associated with this gene?”.

4.4.5 Application programming interface and web interface

The pheWAS results consist of a very large collection of association statistics between the tested gene loci and phenotypes. These results may be useful in a wide range of applications and to help answer different research questions, making the development of a convenient results browser important. We opted to offer both an application programming interface (API) to facilitate integration with other platforms or bioinformatics resources as well as a web-based interface to allow researchers to interactively browse the results. A summary of the data available in the ExPheWas browser along with representative vi-

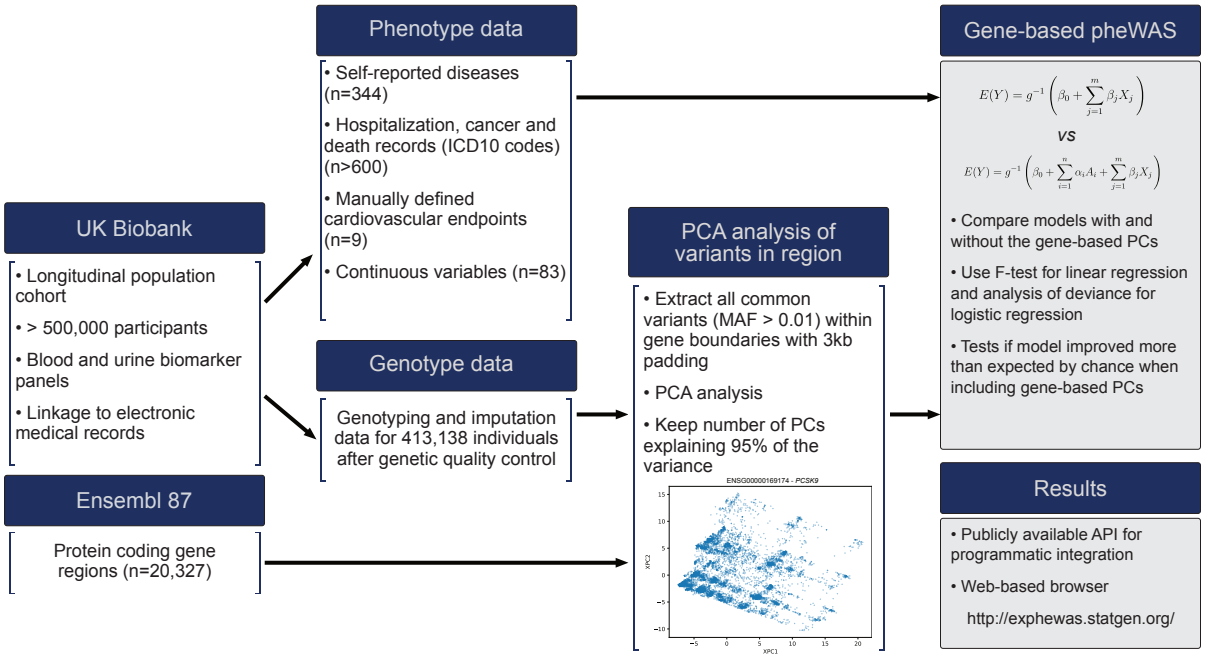


Figure 4.1 – **Schematic overview of the analysis from data sources to pheWAS results.** Contributions to this study: the results of the gene-based pheWAS and the web-browser are represented by shaded boxes. In the equations detailing the association model, β_j terms represent m included covariables X_j, and the α_i terms represent the fixed effects for n included gene-based PCs A_i. The g^{-1} function represents the generalized linear model link function (the identity for linear regression and the logit for logistic regression) and ϵ represents the error term. The association test measures the improvement of goodness of fit when including the n gene-based PCs to the model.

sualizations are presented in Figure 4.2. The main web-page to access these resources is http://exphewas.statgen.org. API documentation is provided along with examples.

4.4.6 Real data application

Atrial fibrillation is a common cardiac arrhythmia of the atria which can lead to severe complications including stroke, cardiomyopathy and heart failure [228, 229]. Recent GWAS including more than a million individuals have identified a large number of loci associated with atrial fibrillation [164, 170].

Our pheWAS study (data release v0.1) included 14,747 UK Biobank participants with evidence of atrial fibrillation in their hospitalization or death records (I48 ICD10 code, "Atrial fibrillation and flutter"). There were 137 genes associated with atrial fibrillation with a q-value ≤ 0.01. Out of these genes, 37 were also reported as the mapped gene (including

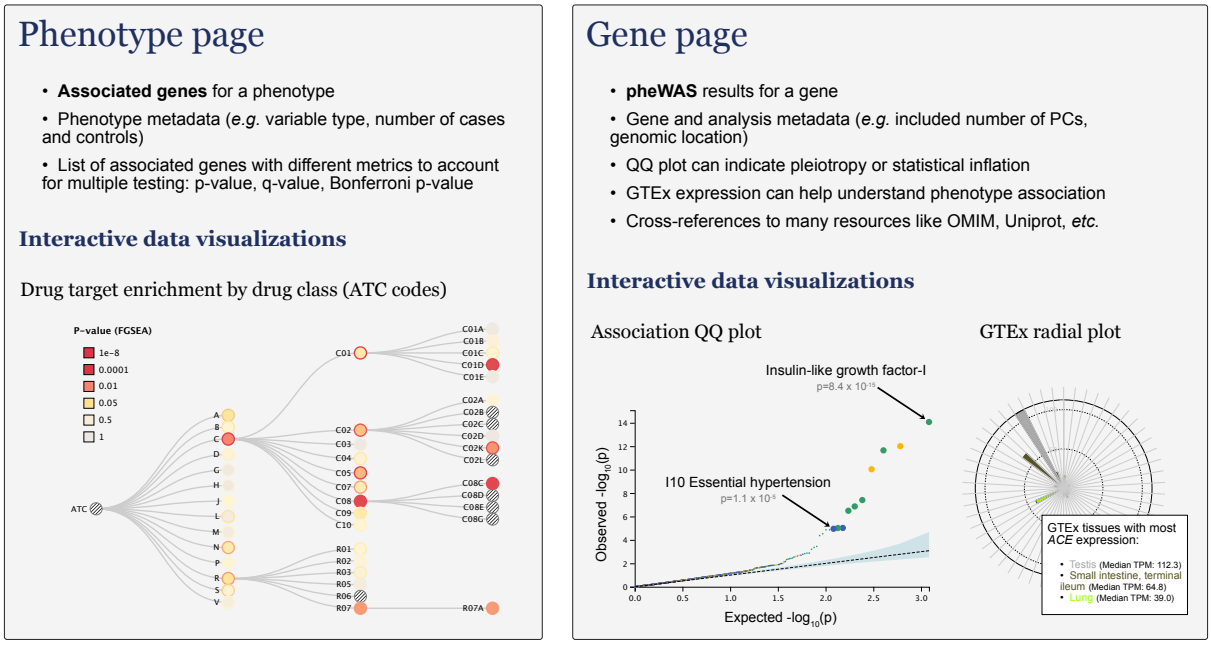


Figure 4.2 – **Summary of the data presented in the ExPheWas browser and adapted data visualization. Left.** Results presented on the phenotype results page. The example for the gene-set enrichment analysis is for the I10 Essential (primary) hypertension code. Enriched drug classes include C01D (vasodilators used in cardiac diseases, $p = 0.03$), C02K (other antihypertensives, $p = 0.016$), C08C (selective calcium channel blockers with mainly vascular effects, $p = 0.002$), R07 (other respiratory system products, $p = 0.019$). **Right.** Results presented on the gene results page. The example data visualisations are for the angiotensin I converting enzyme (*ACE* gene, ENSG00000159640). The leftmost plot is a QQ plot representing deviation from the expected distribution of p-values under the null hypothesis. The statistical inflation as measured by genomic control is $\lambda = 1.13$ (ratio of observed to expected median values on the QQ plot). The top 10 associated phenotypes were magnified and two were identified to emulate the interactivity available on the web version. The rightmost plot is a radial plot of median expression in GTEx v8 tissues which can be used to contextualize association findings.

upstream or downstream genes for intergenic *loci*) in GWAS Catalog associations [230]. We used *g:Profiler* to conduct an ontological term enrichment analysis of the 137 genes and found enrichment for supraventricular arrhythmia (Human Phenotype, HP:0005115; $p_{adj} = 3.7 \times 10^{-7}$), cardiac muscle contraction (Gene Ontology Biological Process, GO:0060048; $p_{adj} = 9.5 \times 10^{-6}$) and Z disk (Gene Ontology Cellular Component, GO:0030018; $p_{adj} = 0.001$) among other relevant terms (Supplementary Figure C.6, Supplementary Table C.5).

We also tested the enrichment of atrial fibrillation associated genes in drug targets based on the ChEMBL database [223] as implemented in the ExPheWas browser. We found en-

richment for class Ia and class Ib antiarrhythmics corresponding to ATC codes C01BA and C01BB, respectively. The Fisher's exact test enrichment p-values for these classes were 0.006 for both. This finding is concordant with pharmacological treatment of atrial fibrillation suggesting genetic support for these drug targets. However, there was also an enrichment with local anesthetics belonging to ATC code C05AD which may represent spurious associations driven by sodium channels genes including *SCN5A* and *SCN10A* both of which are robustly associated with atrial fibrillation [231].

We further explored gene associations that were not previously reported in the GWAS Catalog. Notably, myotilin encoded by the *MYOT* gene was associated with heart rate ($p = 8.6 \times 10^{-31}$), and atrial fibrillation (p=4.9×10^{-11}) in our pheWAS. This region is located in a long LD block spanning multiple genes (Supplementary Figure C.7a). Other credible genes that could drive the association signal in this region are *FAM13B* (atrial fibrillation $p = 1.9 \times 10^{-9}$), *PKD2L2* ($p = 3.1 \times 10^{-9}$), *WNT8A* ($p = 2.1 \times 10^{-11}$) and *NME5* ($p = 5.7 \times 10^{-6}$). The two top genes according to our analysis are *MYOT* and *WNT8A* which have p-values two orders of magnitude smaller than the others. However, *WNT8A* is not expressed in heart tissues in GTEx whereas *MYOT* is expressed in the heart. Myotilin is a component of the sarcomeric Z-disk, a structure implicated in muscle contraction and rare mutations in myotilin cause myofibrillar myopathy which often co-occurs with cardiomyopathy (OMIM:604103, [232, 233]). In single-cell RNA sequencing analysis, *MYOT* was found to be particularly expressed in left atrial and right ventricular cardiomyocytes in line with atrial fibrillation pathophysiology [234]. Even though it was located nearby, we did not consider *KLHL3* as it was revealed to represent an independent association signal through stepwise conditional analysis (Supplementary Figure C.7b). After two stages of the forward conditional stepwise analysis, no significant variants remained associated (Supplementary Figure C.7c).

4.5 Discussion

We used a PC-based association approach to test the association between all human protein coding loci and an exhaustive set of phenotypes in a gene-based pheWAS approach. This allowed us to build a web-based resource that provides a complementary repository to current databases providing gene to phenotype mappings. Using a gene-based approach favors analyses that rely on gene-level annotations such as drug target validation or ontological enrichment analyses. There was an important computational advantage in using a PCA-based approach for our pheWAS as it greatly reduced the number of models to fit when

compared to variant-based approaches. This association test captures the joint contribution of common genetic variants and complements models developed for the joint analysis of rare variants like the burden test or SKAT. We evaluated the power of PC-based approaches compared to conventional association testing, characterized the behaviour of individual PCs in marginal models and evaluated the effect of the choice of included PCs on the statistical association. These auxiliary results provide insight that may be useful for the development and improvement of PC-based methods in association testing and in other contexts such as Mendelian randomization [147]. Finally, we evaluated our gene-based pheWAS resource using a real world example by exploring genes associated with atrial fibrillation. We identified 137 atrial fibrillation-associated genes at a FDR of 0.01 and these genes were enriched for the targets of drugs used to treat atrial fibrillation and GO terms of heart development and physiology. We also focused on previously unreported genes that could be associated with atrial fibrillation and prioritized *MYOT* as an interesting candidate.

There are some limitations to our approach. First, we used gene boundaries and 3 kb of padding to define protein coding loci used for association analysis, but there is no guarantee that the association is driven by the gene product of interest and not by overlapping genes or other DNA elements. In regions of high linkage disequilibrium, it is also possible that an association is in fact due to correlation with a neighboring gene. These limitations are shared by most association testing approaches and can only be addressed by careful interrogation of candidate associations. Power was maximized when common variants in a gene made small contributions to the phenotype as opposed to when the effect is driven by a single causal variant. This is to be expected as the strength of our approach is in the collapsing of variants at a locus. Because of the multivariable nature of the association model, there is also no intuitive measure of effect size for which conventional methods based on single variants may be more appropriate.

To conclude we contribute an important atlas of gene to phenotype associations along with tools to interrogate, contextualize and interpret the results. ExPheWas enables gene-level discoveries at pheWAS scale based on common genetic variants which may otherwise be missed by traditional approaches. We believe that dimensionality reduction approaches such as PCA provide a natural way of addressing linkage disequilibrium in statistical genetics models that aggregate common variants. With the increasing number of large cohorts with available genotype data, it is essential to further develop gene-based methods that aggregate common variants with weak effects to complement other approaches such as the analysis of loss-of-function mutations.

4.6 Acknowledgements

We thank the UK Biobank for providing the data under Application Number 20168.

4.7 Data Availability

Access to the UK Biobank resource requires application through the Access Management System and instructions are available online: https://www.ukbiobank.ac.uk/enable-your-research/apply-for-access. The ExPheWas results browser is available online at https://exphewas.statgen.org/ and instructions for programmatic access through the API are provided in the documentation section. The code for the results browser including the database models, the web application and the data visualizations is open source and available at .

4.8 Funding

The work was supported by the Health Collaboration Acceleration Fund from the Government of Quebec. MAL holds a scholarship from Canadian Institutes of Health Research (CIHR); MPD holds the Canada Research Chair in Precision medicine data analysis.

CHAPITRE 5

Autres contributions scientifiques

5.1 «Pharmacogenomics of blood lipid regulation» Legault MA, Tardif JC and Dubé MP. *Pharmacogenomics* (2018)

Dans le cadre de ma thèse, nous avons rédigé un article de revue dont je suis le premier auteur et publié dans le journal *Pharmacogenomics* [235]. En premier lieu, cet article s'intéresse aux variations génétiques impliquées dans la réponse aux médicaments influençant les lipides plasmatiques correspondant à une approche plus classique en pharmacogénomique. Cette section inclut une discussion des bases génétique de la réponse aux médicaments déjà bien connus pour le traitement des différentes dyslipidémies : les statines, l'ezetimibe, les fibrates et la niacine. Par la suite, l'approche de validation génétique des cibles pharmacologiques est abordée et décrite à travers les exemples modernes des inhibiteurs de la CETP et de la PCSK9. Nous jetons ensuite un coup d'œil sur les médicaments en cours de développement et qui sont inspirés ou fortement supportés par la découverte de mutations génétiques de leurs cibles pharmacologiques. Un des exemples présentés est celui de l'*Angiopoietin Like 3*, une protéine encodée par le gène *ANGPTL3* et dont l'inhibition diminue les taux de triglycérides, de cholestérol LDL et de cholestérol HDL. L'intérêt pour cette cible comme médicament vient de la découverte d'une mutation LoF chez la souris qui influençait le métabolisme des lipides. Des études subséquentes ont montré que les mutations LoF du gène *ANGPTL3* protégeaient contre la maladie coronarienne chez l'humain [236, 237]. Récemment, l'essai clinique *Evinacumab Lipid Studies in Patients with Homozygous Familial Hypercholesterolemia* (ELIPSE HoFH) a ensuite été réalisé chez des patients atteints d'hypercholestérolémie familiale homozygote où le bénéfice de l'evinacumab, un anticorps monoclonal contre l'ANGPTL3, a été démontré. La FDA a approuvé le médicament dans ce contexte représentant la première

approbation de médicament de cette classe thérapeutique [238].

La section finale de cet article s'intitule «la transition polygénique en médecine de précision» (traduction libre) et discute des perspectives de futur dans le monde de la pharmacogénomique. L'utilisation des GRS afin d'identifier des groupes d'individus qui peuvent tirer un bénéfice supplémentaire d'un médicament est abordée et contrastée aux approches traditionnelles visant l'identification de variations génétiques qui affectent la réponse au médicament.

5.2 «*grstools* : A bioinformatics framework for the construction of genetic risk scores.» Legault MA, Lemieux Perreault LP, Lemieux S, Tardif JC and Dubé MP. (unpublished software)

Sur le plan technique, l'utilisation de GRS est derrière plusieurs de mes travaux de recherche. Au moment de mener les expériences, le choix d'outils bio-informatiques était limité et plusieurs nouvelles approches et marches à suivre ont été publiées depuis [114, 118, 239]. Ceci m'a mené à développer la suite de logiciels *grstools* qui facilite l'exécution de plusieurs tâches en lien avec la création et l'utilisation de GRS. Cette suite de logiciels est disponible publiquement à l'adresse github.com/legaultmarc/grstools et a fait l'objet d'une présentation par affiche dans le cadre de la rencontre des Instituts de Recherche en Santé du Canada en statistiques génétiques en 2017 (Figure Supplémentaire D.1). Étant donné la multitude d'outils maintenant disponibles, nous n'avons pas cru bon de tenter la publication d'un article scientifique. Cet outil a cependant été utilisé pour créer tous les scores génétiques décrits dans cette thèse et demeure utilisé au sein du laboratoire. Les différentes fonctionnalités des outils intégrés à *grstools* sont exhaustivement décrites dans la documentation de l'outil sur Github, et seront brièvement décrites ici.

L'utilitaire **grs-create** permet la création de GRS par l'approche C+T (décrite en plus de détails à la section 1.3). Brièvement, les variations sont triées en fonction de leurs valeurs-p, puis la construction du score se fait par l'ajout de variations approximativement indépendantes jusqu'à ce qu'un seuil de valeur-p prédéterminé soit atteint. L'utilitaire développé est flexible et permet des filtres par région génomique et par fréquence allélique ainsi que l'exclusion des variations génétiques dont le brin d'ADN peut porter à confusion (variation génétique A/T et G/C). Il est aussi possible de restreindre la sélection de variations en fonction d'une population de référence comme le *1000 Genomes Project* afin de faciliter l'arrimage des variations génétiques considérées [101]. Bien qu'il soit possible de réaliser ces opérations en utilisant des logiciels d'usage général comme *Plink*, l'utilisation de *grs-create*

est très conviviale et expressive [82].

Une fois la sélection des variations génétiques faite, soit par un tiers ou par l'utilisation de *grs-create*, il est possible d'utiliser l'outil **grs-compute** pour faire le calcul des scores individuels. Une fonctionnalité qui fait de *grs-create* un choix intéressant est qu'il permet la validation automatique du brin pour les variations ambiguës en utilisant les génotypes d'une population de référence séquencée. À cause de particularités des technologies de génotypage, il est difficile de distinguer le brin d'ADN lorsqu'une variation correspondant à un appariement Watson-Crick (A/T et G/C) est considérée sans étapes supplémentaires comme le phasage. Pour contourner ce problème, *grs-compute* compare la fréquence observée de l'allèle ambigu encodé à sa fréquence dans une référence issue de séquençage du génome et pour laquelle le brin est résolu. Cette validation de brin par comparaison des fréquences se fait à l'aide d'un intervalle de confiance de Clopper–Pearson permettant une certaine différence entre les fréquences alléliques due à l'échantillonnage. Cette fonctionnalité a été utile plusieurs fois dans notre laboratoire, car elle fournit une étape de contrôle de qualité supplémentaire et qu'elle facilite l'inclusion de variations de brin ambigu qui sont parfois incluses dans les GRS publiés [240].

L'utilitaire **grs-utils** intègre plusieurs sous-commandes de gestion de données. Sa fonctionnalité la plus intéressante a été en grande partie développée avec l'aide de Camille Rochefort-Boulanger, une stagiaire sous la supervision de Marie-Pierre Dubé qui s'est jointe à notre laboratoire durant l'été 2017 et dont j'ai contribué à l'encadrement pour la durée de son stage. Comme les GRS sont habituellement calculés à partir de statistiques sommaires d'associations génétiques, il est pertinent de comparer ces statistiques externes à celles qui sont observées dans le jeu de données d'intérêt. Par exemple, une différence dépassant l'erreur aléatoire d'échantillonnage peut être due à des erreurs d'encodage des allèles ou à une différence de structure génétique. La sous-commande «*beta-plot*» permet donc de comparer les effets des variations génétiques entre les statistiques sommaires et les données individuelles disponibles avant la construction d'un GRS. Cette étape représente un contrôle de qualité important qui peut permettre l'identification d'erreurs ponctuelles comme l'inversion de l'encodage d'allèles ou de biais systématiques. Un exemple du diagramme généré par l'utilisation de cette commande est présenté à la Figure 5.1.

L'utilitaire **grs-mr** permet l'utilisation de GRS pour estimer l'effet causal d'une exposition sur une issue par la méthode du ratio. La méthode du ratio est valide sous les suppositions conventionnelles des variables instrumentales décrites à la section 1.4.1. L'estimation ponctuelle est donnée par le ratio des coefficients de régression de l'issue sur le GRS et du coefficient de régression de l'exposition sur le GRS. L'utilitaire *grs-mr* calcule ensuite

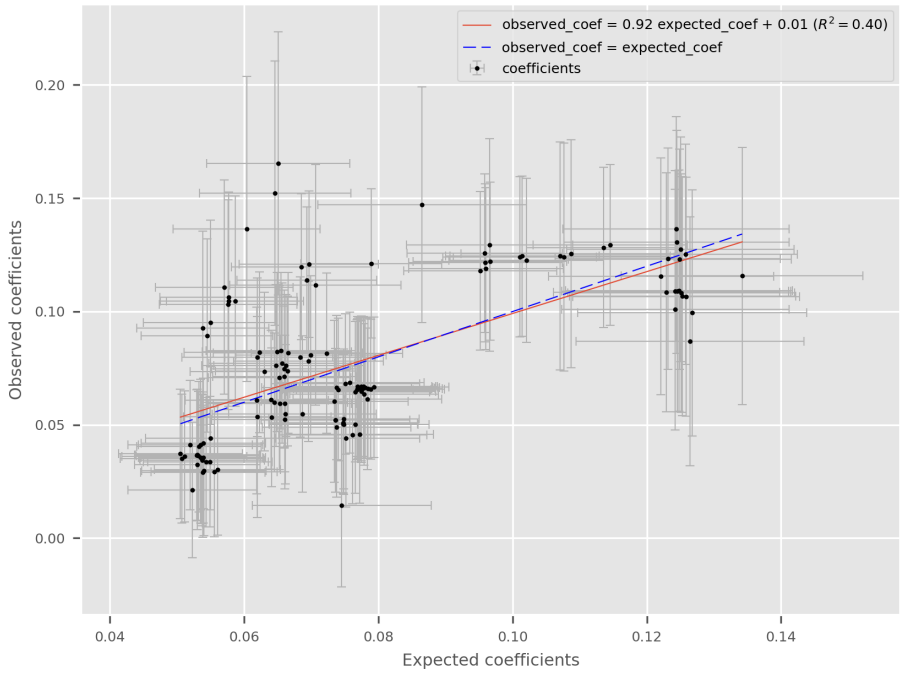


FIGURE 5.1 – **Exemple de graphique «beta-plot» illustrant la concordance entre l'effet des variations génétiques estimé à même les données individuelles et l'effet issu de statistiques sommaires de GWAS.** Des données aberrantes identifiées sur ce graphique pourraient représenter des erreurs de codage d'allèles à vérifier en amont de la construction de GRS. L'absence de corrélation pourrait aussi suggérer une différence grave dans la structure génétique des populations.

l'erreur type et l'intervalle de confiance à 95% en utilisant le rééchantillonage *bootstrap*.

Finalement, les fonctionnalités majeures de la suite d'outils *grstools* ont été brièvement décrites. Il s'agit d'un logiciel facile d'utilisation qui propose une interface unifiée pour toutes les tâches liées à la création, construction, évaluation et utilisation de GRS. Bien qu'il ne représente pas une contribution nouvelle sur le plan algorithmique, il s'agit d'un outil utile et flexible qui a bien servi les objectifs de cette thèse.

5.3 Publications en tant que co-auteur

Durant mes études doctorales, j'ai aussi contribué à différents projets de recherche qui ont mené à des publications, énumérées ici-bas, du plus récent au plus vieux.

5.3.1 Articles publiés dans des journaux revus par les pairs

1. **Genetics of symptom remission in outpatients with COVID-19**. Dube, M.-P., Lemacon, A., Barhdadi, A., Lemieux Perreault, L.-P., Oussaid, E., Asselin, G., Provost, S., Sun, M., Sandoval, J., Legault, M.-A., Mongrain, I., Dubois, A., Valois, D., Dedelis, E., Lousky, J., Choi, J., Goulet, E., Savard, C., Chicoine, L.-M., Cossette, M., Chabot-Blanchet, M., Guertin, M.-C., de Denus, S., Bouabdallaoui, N., Marchand, R., Bassevitch, Z., Nozza, A., Gaudet, D., L'Allier, PL., Hussin, J., Boivin, G., Busseuil, D., Tardif, J.-C. (2021). *Scientific Reports*, 11 (1), 1-10. https://doi.org/10.1038/s41598-021-90365-6
   Ce manuscrit présente la sous-étude pharmacogénomique de l'essai clinique COLCORONA qui évaluait l'utilisation de la colchicine chez des patients atteints de la COVID-19. Une approche d'analyse de survie portant sur le temps avant la fin des symptômes a été utilisée dans le GWAS. J'ai contribué à cette étude en révisant le manuscrit et en effectuant des analyses de corrélation génétique qui n'ont finalement pas été incluses dans l'article.
2. **Genetic meta-analysis of cancer diagnosis following statin use identifies new associations and implicates human leukocyte antigen (HLA) in women**. Sun, M., Lemaçon, A., Legault, M.-A., Asselin, G., Provost, S., Aschard, H., Barhdadi, A., Zada, Y. F., Valois, D., Mongrain, I., Tardif, J.-C., & Dubé, M.-P. (2021). *The Pharmacogenomics Journal*, 1–12. https://doi.org/10.1038/s41397-021-00221-z
   Ce projet a révélé une association génétique avec l'incidence de cancer chez les utilisatrices de statines. J'y ai contribué sur le plan technique en participant à la préparation des données de la UK Biobank pour permettre les analyses de survie avec le modèle

de Cox.

3. **Pharmacogenomics of the Efficacy and Safety of Colchicine in COLCOT**. Dubé, M.-P., Legault, M.-A., Lemaçon, A., Lemieux Perreault, L.-P., Fouodjio, R., Waters, D.D., Kouz, S., Pinto, F.J., Maggioni, A.P., Diaz, R., Berry, C., Koenig, W., Lopez-Sendon, J., Gamra, H., Kiwan, G.S., Asselin, G., Provost, S., Barhdadi, A., Sun, M., Cossette, M., Blondeau, L., Mongrain, I., Dubois, A., Rhainds, D., Bouabdallaoui, N., Samuel, M., de Denus, S., L'Allier, P.L., Guertin, M.-C., Roubille, F., Tardif, J.-C. (2021). *Circulation : Genomic and Precision Medicine* https://doi.org/10.1161/CIRCGEN.120.003183
   Cet article décrit la sous-étude pharmacogénomique de l'essai clinique *COLCOT* évaluant un traitement à la colchicine pour la prévention d'évènements coronariens. J'ai participé aux analyses de suivi bio-informatique post-GWAS. Les analyses de colocalisation que j'ai menées ont révélé que le *locus* sur le chromosome 6 associé aux effets indésirables gastro-intestinaux en réponse à la colchicine est aussi associé à la maladie de Crohn. La région du chromosome 9 qui est possiblement associée avec l'efficacité du traitement chez les hommes était aussi associée avec l'expression du gène *HAUS6* qui est impliqué dans la régulation des microtubules.

4. **genipe : an automated genome-wide imputation pipeline with automatic reporting and statistical tools**. Lemieux Perreault, L.-P., Legault, M.-A., Asselin, G., & Dubé, M.-P. (2016). Bioinformatics , 32(23), 3661–3663.
   https://doi.org/10.1093/bioinformatics/btw487
   Cet outil bio-informatique facilite l'imputation des variations génétiques à partir d'un jeu de données issu du génotypage par puce à ADN. J'ai contribué au développement logiciel de l'outil en intégrant des tests d'association génétique.

5. **Machine learning and data mining in complex genomic data – a review on the lessons learned in Genetic Analysis Workshop 19**. König, I. R., Auerbach, J., Gola, D., Held, E., Holzinger, E. R., Legault, M.-A., Sun, R., Tintle, N., & Yang, H.-C. (2016). *BMC Genetics*, 17 Suppl 2(S2), 1. http://www.biomedcentral.com/1471-2156/17/S2/1
   J'ai participé au congrès *Genetic Analysis Workshop 19* au sein du groupe d'apprentissage machine. Durant le congrès, les contributions individuelles étaient partagées en groupe, puis un résumé des discussions était présenté devant tous les participants du congrès. J'ai été volontaire pour présenter les résultats de mon groupe et j'ai ensuite contribué à l'écriture du manuscrit sommaire.

5.3.2 Manuscrits en révision

1. **An epistatic interaction between the *ADCY9* pharmacogene and the drug target *CETP***. Gamache, I., Legault, M.-A., Grenier, J.-C., Sanchez, R., Rhéaume É., Asgari, S., Barhdadi, A., Feroz Zada, Y., Trochet H., Luo Y., Lecca L., Murray, M., Raychaudhuri, S., Tardif, J.-C., Dubé, M.-P., Hussin, J. *Manuscrit soumis en révision.*
   Ce projet démontre une relation évolutive entre le gène *ADCY9* et le gène *CETP*. J'ai participé à ce projet en caractérisant les effets phénotypiques de la variation génétique rs158477 du gène *CETP*. L'accent était sur la modification de l'effet par le sexe et la variation rs1967309 du gène *ADCY9*. L'interaction d'ordre 3 entre rs158477, rs1967309 et le sexe était significative ce qui appuyait les résultats de l'étude qui démontre une corrélation entre ces deux variations, mais seulement chez les hommes.

5.3.3 Manuscrits en pré-publication

1. **Deep interpretability for GWAS**. Sharma, D., Durand, A., Legault, M.-A., Perreault, L.-P. L., Lemaçon, A., Dubé, M.-P., & Pineau, J. (2020). In arXiv [cs.LG]. Preprint posted to *arXiv*. http://arxiv.org/abs/2007.01516
   Ce manuscrit décrit l'utilisation d'outils pour permettre l'interprétation des réseaux de neurones dans le contexte de la prédiction du diabète à partir de variations génétiques. J'ai contribué à cette étude en épaulant le travail du premier auteur tout au long du projet.
2. **Diet networks : thin parameters for fat genomics**. Romero, A., Carrier, P. L., Erraqabi, A., Sylvain, T., Auvolat A., Dejoie E., Legault, M.-A., Dubé, M.-P., Hussin J.G., Bengio Y. (2016). Preprint posted to *arXiv*. https://arxiv.org/abs/1611.09340
   Cette étude démontre l'utilisation d'un réseau auxiliaire pour aider à l'apprentissage dans un régime où le nombre de paramètres dépasse largement le nombre d'exemples d'entraînement. J'ai contribué à ce projet en guidant l'utilisation des données génétiques.

CHAPITRE 6

Conclusion et perspectives

6.1 Conclusion

D'abord, nous avons utilisé une approche de validation des cibles pharmacologiques afin d'étudier le cas de l'ivabradine, un médicament avec un profil d'efficacité complexe. Notre étude a démontré la grande similarité entre l'effet des variations génétiques du gène *HCN4* et de l'ivabradine. Nous avons observé une réduction du risque d'insuffisance cardiaque et une augmentation du risque de fibrillation auriculaire en lien avec une réduction génétiquement prédite de la fréquence cardiaque par le HCN4. Nos observations étaient largement concordantes avec les résultats des essais randomisés. La méthodologie utilisée pour arriver à cette conclusion est importante, car elle illustre le besoin de tenir compte de l'effet des comorbidités dans une approche génétique de cibles pharmacologiques. La variation génétique étudiée, rs8038766, a des effets de directions opposées sur deux maladies qui ont une relation de causalité mutuelle. Une approche naïve aurait prédit un effet nul sur l'insuffisance cardiaque comme c'est le cas dans les données les plus récentes issues des grands consortiums GWAS [165]. L'ajustement statistique n'aurait pas non plus été adéquat à cause du risque d'introduire un biais du collisionneur (*collider bias*). La méthodologie complémentaire proposée a donc permis de bien modéliser la situation malgré le nombre comparativement limité de cas d'insuffisance cardiaque rapportés dans les données à l'étude. En plus d'étudier la cible pharmacologique de l'ivabradine, nous avons aussi évalué l'hypothèse au sens plus large du bénéfice de la réduction de fréquence cardiaque. Notre étude de MR suggère que la fréquence cardiaque n'est pas un biomarqueur causal pour la maladie coronarienne et donc que le développement d'une thérapie agissant uniquement sur ce paramètre est susceptible de ne pas être cliniquement efficace. L'étude de MR montre aussi que la réduction génétique

de la fréquence cardiaque augmente le risque fibrillation auriculaire, mais n'a pas d'effet sur le risque d'insuffisance cardiaque. Le rôle de la fibrillation auriculaire comme médiateur n'a pas été exploré dans ces analyses. Une telle étude nécessite l'utilisation de modèles de MR en réseau ou de médiation dépassant l'objectif de notre étude mais qui seraient des approches intéressantes à considérer dans de futures études [241, 242].

Cette étude comportait une certaine part de difficulté. Par exemple, la création d'une variable instrumentale génétique était difficile, car le gène *HCN4* est fortement intolérant aux mutations génétiques. Peu de variations génétiques avaient donc un effet robuste sur la fréquence cardiaque à ce locus et nous avons dû axer notre étude sur l'effet d'une seule variation génétique. Certaines questions soulevées dans les essais randomisés, par exemple l'interaction significative entre la fréquence cardiaque à l'entrée de l'étude et la réponse à l'ivabradine dans l'étude SHIFT, ne peuvent cependant pas être étudiées par une approche génétique [25]. Cette limite des études génétiques est due à la différence entre l'effet des variations génétiques qui débute à la naissance en contraste à la prise d'un médicament initiée durant un essai clinique. De plus, l'augmentation du risque de fibrillation auriculaire associée aux variations génétiques de *HCN4* était largement supérieure à l'effet observé dans les essais randomisés. Il est improbable que la totalité de l'effet observé soit due à l'effet sur la fréquence cardiaque qui n'était que modestement affectée. Comme certaines mutations rares de *HCN4* sont associées à des cardiomyopathies et que le rôle de ce gène dans la cardiogénèse d'organismes modèles a déjà été établi, il est plausible que les variations génétiques de *HCN4* aient d'autres effets sur la structure du myocarde [177, 179].

Les essais randomisés où le critère de jugement est l'incidence d'évènements cardiovasculaires sévères ont souvent recours à des populations à haut risque afin de réduire le nombre d'individus à recruter et le temps de suivi. De plus, l'effet estimé représente l'effet moyen du traitement alors que la réponse au traitement peut être hétérogène dans des sous-groupes de patients présentant des caractéristiques particulières. L'intérêt d'utiliser des devis d'études complémentaires aux essais randomisés a été évoqué dans l'optique de pallier ces lacunes [243]. Dans le Chapitre 3, nous avons sélectionné deux variables susceptibles de modifier l'effet de la CETP en fonction d'études antérieures : le sexe et l'IMC. Nous avons étudié comment ces variables pouvaient moduler l'effet de la réduction de la CETP sur différents biomarqueurs et maladies ischémiques. Les biomarqueurs étudiés incluaient les fractions lipidiques comme le cholestérol LDL et le cholestérol HDL, mais aussi la lipoprotéine(a), la protéine C-réactive et la capacité d'efflux de cholestérol. Une réduction génétiquement prédite de la concentration de CETP avait des effets variables entre les hommes et les femmes. En comparaison, les femmes avaient de plus hauts taux de cholestérol HDL et une plus

grande capacité d'efflux de cholestérol, tandis que leur taux de cholestérol LDL était plus petit en moyenne. En comparant les individus avec un plus faible IMC aux individus ayant un plus grand IMC, la réduction génétique de la CETP était associée à de plus hauts taux de cholestérol HDL. Quant à eux, les taux de cholestérol LDL et de lipoprotéine(a) étaient plus petits. Le IMC ne modulait pas l'effet entre la CETP et l'efflux de cholestérol. Chez les utilisateurs de dalcetrapib (un inhibiteur de la CETP), nous avons confirmé la plus forte augmentation du cholestérol HDL et la plus forte diminution du cholestérol LDL en réponse au médicament tel que prédit par le modèle génétique. Cette observation était d'autant plus intéressante que l'effet du dalcetrapib sur le cholestérol LDL était nul dans l'étude originale. Dans cette étude, nous n'avons pas été en mesure de prouver une modification de l'effet de l'inhibition de la CETP sur la maladie coronarienne. Il est possible que cette observation négative soit due à un manque de puissance statistique, car le nombre d'évènements cardiovasculaires est limité dans la UK Biobank dont les participants sont en général en bonne santé [180].

Les résultats de notre étude aident à mieux caractériser l'effet de la réduction pharmacologique de la CETP prédite dans différentes sous-populations. Une meilleure connaissance de ces effets est importante à la fois d'un point de vue clinique et pour notre compréhension fondamentale des voies biologiques impliquées dans les effets de la CETP. Dans un cadre plus large, notre étude soulève la possibilité d'utiliser une approche par validation génétique des cibles pharmacologiques pour évaluer la validité externe des essais randomisés. D'un point de vue médical, l'étude de la modulation de l'effet pharmacologique entre les hommes et les femmes est particulièrement intéressante, car ces dernières sont fréquemment sous-représentées dans les grands essais cliniques de phase 3 en cardiologie, même si elles se voient administrer les mêmes traitements que les hommes. Le sexe est aussi moins susceptible à des biais qui pourraient complexifier l'interprétation des résultats, car il demeure habituellement inchangé au cours de la vie. L'idée de tester la modification d'effet est nouvelle dans la littérature de la MR et peu d'estimateurs ont été proposés [244]. Le développement méthodologique est aussi freiné par le nombre limité de jeux de données de statistiques sommaires de GWAS incluant des statistiques adaptées au calcul d'interactions (*p. ex.* coefficients de régression chez les hommes et les femmes séparément ou présentation du coefficient de l'interaction). Certains consortiums, comme *DIAbetes Genetics Replication And Meta-analysis Consortium* (DIAGRAM), ont présenté des statistiques sommaires stratifiées par le sexe, ce qui a permis des études MR stratifiées où la modification d'effet par le sexe peut être qualitativement décrite [245, 246].

Au Chapitre 4, les résultats d'une étude d'association gène-centrique sont présentés à

l'aide du fureteur *ExPheWas*. Dans cette étude, nous avons utilisé une approche PheWAS pour caractériser l'effet de variations génétiques de 19 114 gènes portés par des autosomes et encodant des protéines. Les phénotypes considérés dans l'étude étaient diversifiés et incluaient les maladies cardiovasculaires, des mesures de laboratoire, des mesures anthropométriques, des maladies autorapportées et des codes d'hospitalisation ou de décès. Une telle étude PheWAS est utile dans le contexte de la validation de cibles pharmacologiques, car elle peut être utilisée pour prédire les effets indésirables de la modulation de cibles ou pour identifier des opportunités de repositionnement de médicaments. En comparaison aux approches basées sur les variations génétique individuelles, l'approche gène-centrique peut avoir une plus grande puissance statistique sous certaines architectures génétiques. Elle facilite aussi les analyses bio-informatiques comme l'enrichissement de termes ontologiques ou de cibles pharmacologiques qui réfèrent à des concepts définis au niveau des protéines et non des variations génétiques. Par exemple, sur la page de présentation des résultats pour un phénotype, l'enrichissement des gènes partagés entre des cibles des différentes classes médicamenteuses est présenté. Cette approche établit la relation entre l'étiologie génétique d'un trait et des classes de médicaments et pourrait révéler de nouvelles indications possibles pour des médicaments déjà connus. Nous avons aussi utilisé la plate-forme ExPheWas afin d'identifier des gènes associés à la fibrillation auriculaire dans une démonstration de faisabilité. Plusieurs annotations pertinentes étaient surreprésentées parmi les 137 gènes associés à la fibrillation auriculaire. Notamment, ils étaient plus fréquemment impliqués dans la contraction cardiaque et associés aux disques Z ou aux arythmies supraventriculaires. Ces gènes étaient aussi enrichis parmi les cibles d'antiarythmiques utilisés pour le traitement de la fibrillation auriculaire. Nous avons identifié le gène de la myotiline comme acteur possible dans le développement de la fibrillation auriculaire.

Cette étude a présenté plusieurs défis techniques en raison de sa portée. D'abord, au niveau bio-informatique, nous avons développé un outil permettant l'utilisation de tests gène-centriques dans un contexte PheWAS. L'optimisation de la gestion des données et du calcul du test statistique a été nécessaire afin de réduire le fardeau computationnel. Néanmoins, l'étude présentée inclut plus de 23 millions de tests impliquant plus de 400 000 individus, et nous avons utilisé des ressources de calcul à haute performance de *Calcul Canada* pour y parvenir. Nous avons développé une base de données pour organiser tous ces résultats et un portail web pour en faciliter la consultation. Beaucoup d'outils d'analyse et de visualisation des données sont inclus dans ce portail, notamment les graphiques quantile-quantile interactifs ou l'arbre d'enrichissement des classes médicamenteuses. Bien que l'utilisation de la bio-informatique ait été centrale aux études présentées dans ma thèse, la plate-forme *ExPheWas* est certainement un exemple où les compétences en bio-informatique et en dé-

veloppement logiciel ont été primordiales. Une limitation importante de cette étude est que les variations génétiques retrouvées à un gène n'agissent pas forcément sur le produit de ce gène. Il est possible que les variations génétiques aient des effets en *trans* ou qu'elles agissent sur un élément génétique inconnu, par exemple. Nous avons validé que les résultats sont tel qu'attendu pour plusieurs cibles pharmacologiques et l'étude de la fibrillation auriculaire démontre un enrichissement d'associations biologiquement plausibles. Néanmoins, les associations doivent être interprétées avec prudence et la plate-forme doit être utilisée en conjonction avec des approches complémentaires comme la colocalisation avec des données d'expression ou de pQTL par exemple.

L'idée de prédire l'effet de médicaments en exploitant les registres de données et de résultats d'études génétiques et phénotypiques est exaltante. Dans cette thèse, nous avons utilisé cette approche pour mieux comprendre les effets de l'ivabradine et des inhibiteurs de la CETP, deux exemples complexes de médicaments qui ont soulevé beaucoup de questions dans la communauté médicale suite aux résultats des essais randomisés avec ces molécules. Sur un plan plus technique, nous avons fait usage d'approches statistiques parfois non conventionnelles et dont le contexte d'application dépasse les cibles étudiées. Par exemple, l'utilisation de modèles de risques en compétition pourrait avoir des applications en MR multivariée pour l'étude de l'effet d'expositions sur des maladies comorbides. La présentation des effets marginaux de la CETP dans différentes sous-populations est un autre exemple qui bénéficie des avantages des modèles d'interaction, comme de permettre l'inférence sur le terme d'interaction, tout en conservant l'interprétabilité des modèles de stratification. Finalement, pour avoir utilisé la plate-forme *ExPheWas* dans le contexte de mes travaux de recherche depuis maintenant plusieurs mois, je suis excité par son potentiel et j'espère que d'autres utilisateurs tireront profit de cette plate-forme pour caractériser rapidement les effets phénotypiques de la modulation d'un gène par des variations génétiques communes. Nous avons mis en place un plan de pérennité pour maintenir cette plateforme active dans les années qui suivent qui sera géré par les membres de l'équipe de recherche de Marie-Pierre Dubé. Les approches gène-centriques pourraient servir à l'étude des cibles pharmacologiques en favorisant l'intégration des connaissances sur la fonction des gènes, comme les ressources ontologiques ou des bases de données de voies biologiques.

6.2 Perspectives

En 2015, l'année du début de mon doctorat, 225 articles portant sur la MR ont été publiés et ce nombre était 4 fois plus élevé en 2020 [1]. Cet engouement coïncide avec l'émergence et la démocratisation de méthodes qui permettent l'inférence de relations causales à partir d'observations. Ce concept est la prémisse longtemps contestée derrière la discipline de l'inférence causale qui gagne de plus en plus d'adeptes en économétrie et en épidémiologie. Cet essor aura des répercussions sur les méthodes utilisées en MR qui risquent de connaître un développement accéléré dans les années à venir. Par exemple, des approches de MR non linéaire ont été développées pour estimer les effets dose-dépendants de l'exposition sur une issue [247]. Silverwood *et coll.* ont évalué l'effet de la consommation d'alcool sur, entre autres, la pression artérielle et le taux de cholestérol non HDL démontrant que ces relations étaient non linéaires [248]. À l'aide d'une approche similaire, Sun *et coll.* ont établi que la relation entre l'IMC et la mortalité, inférée par MR, était non-linéaire et en forme de «J» [249]. Il s'agit probablement d'un phénomène épidémiologique fréquent, où le même changement d'une exposition peut être presque neutre près de la moyenne, mais avoir un impact sévère près des extrêmes. On peut également envisager des modèles qui intègrent simultanément plusieurs expositions et plusieurs issues tout en modélisant leurs interactions et en tenant compte de la non-linéarité. De tels modèles relèvent de l'intelligence artificielle et sont à l'heure actuelle hors de portée, mais on peut prédire que des développements simultanés en génétique, statistique et apprentissage machine pourraient les rendre accessibles. Cela nous permettrait de comprendre l'effet de multiples variables et de leurs interactions, améliorant ainsi notre compréhension de la physiopathologie des maladies et la robustesse des modèles prédictifs. Un algorithme s'appuyant sur un modèle causal de la maladie n'aura probablement pas un meilleur pouvoir prédictif, mais il est susceptible d'avoir une plus grande validité externe, un facteur clé de la pertinence clinique de ces modèles. Au niveau de la génétique des cibles pharmacologiques, ce type d'approches permettrait de beaucoup plus facilement utiliser les approches de MR pour guider la conception d'essais randomisés en permettant de cibler les populations de patients les plus sujets à bénéficier d'un médicament et d'optimiser les objectifs thérapeutiques en fonction de biomarqueurs, *p. ex.* adapter la posologie du médicament pour viser les niveaux de biomarqueurs prédits comme était optimaux en MR. Les progrès de la recherche en apprentissage automatique qui promettent d'estimer la réponse individuelle au traitement méritent également d'être surveillés et pourraient avoir des implications importantes en pharmacogénomique [250, 251].

1. Selon une recherche sur le Clarivate Web of Science (http://www.webofknowledge.com/) menée en Mars 2021

Bibliographie

1. Stancu, C. & Sima, A. Statins : mechanism of action and effects. en. *J. Cell. Mol. Med.* **5,** 378-387 (oct. 2001).
2. Knouff, C. W. *et al.* Pharmacological effects of lipid-lowering drugs recapitulate with a larger amplitude the phenotypic effects of common variants within their target genes. en. *Pharmacogenet. Genomics* **18,** 1051-1057 (déc. 2008).
3. Ference, B. A., Majeed, F., Penumetcha, R., Flack, J. M. & Brook, R. D. Effect of naturally random allocation to lower low-density lipoprotein cholesterol on the risk of coronary heart disease mediated by polymorphisms in NPC1L1, HMGCR, or both : a 2 × 2 factorial Mendelian randomization study. en. *J. Am. Coll. Cardiol.* **65,** 1552-1561 (avr. 2015).
4. Cook, D. *et al.* Lessons learned from the fate of AstraZeneca's drug pipeline : a five-dimensional framework. en. *Nat. Rev. Drug Discov.* **13,** 419-431 (juin 2014).
5. Holmes, M. V., Richardson, T. G., Ference, B. A., Davies, N. M. & Davey Smith, G. Integrating genomics with biomarkers and therapeutic targets to invigorate cardiovascular drug development. *Nat. Rev. Cardiol.* (mar. 2021).
6. Nelson, M. R. *et al.* The support of human genetic evidence for approved drug indications. en. *Nat. Genet.* **47,** 856-860 (août 2015).
7. Plenge, R. M., Scolnick, E. M. & Altshuler, D. Validating therapeutic targets through human genetics. *Nat. Rev. Drug Discov.* **12,** 581-594 (août 2013).
8. Kathiresan, S. Developing medicines that mimic the natural successes of the human genome : lessons from NPC1L1, HMGCR, PCSK9, APOC3, and CETP. en. *J. Am. Coll. Cardiol.* **65,** 1562-1566 (avr. 2015).
9. King, E. A., Davis, J. W. & Degner, J. F. Are drug targets with genetic support twice as likely to be approved ? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. en. *PLoS Genet.* **15,** e1008489 (déc. 2019).

10. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits : From Polygenic to Omnigenic. en. *Cell* **169,** 1177-1186 (juin 2017).
11. Seidah, N. G. *et al.* The secretory proprotein convertase neural apoptosis-regulated convertase 1 (NARC-1) : liver regeneration and neuronal differentiation. en. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 928-933 (fév. 2003).
12. Abifadel, M. *et al.* Mutations in PCSK9 cause autosomal dominant hypercholesterolemia. en. *Nat. Genet.* **34,** 154-156 (juin 2003).
13. Maxwell, K. N. & Breslow, J. L. Adenoviral-mediated expression of Pcsk9 in mice results in a low-density lipoprotein receptor knockout phenotype. en. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 7100-7105 (mai 2004).
14. Cohen, J. *et al.* Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. en. *Nat. Genet.* **37,** 161-165 (fév. 2005).
15. Cohen, J. C., Boerwinkle, E., Mosley, T. H. & Hobbs, H. H. Sequence Variations in PCSK9, Low LDL, and Protection against Coronary Heart Disease. *N. Engl. J. Med.* **354,** 1264-1272 (mar. 2006).
16. Benn, M., Nordestgaard, B. G., Grande, P., Schnohr, P. & Tybjaerg-Hansen, A. PCSK9 R46L, low-density lipoprotein cholesterol levels, and risk of ischemic heart disease : 3 independent studies and meta-analyses. en. *J. Am. Coll. Cardiol.* **55,** 2833-2842 (juin 2010).
17. Schmidt, A. F. *et al.* PCSK9 genetic variants and risk of type 2 diabetes : a mendelian randomisation study. en. *Lancet Diabetes Endocrinol* **5,** 97-105 (fév. 2017).
18. Ference, B. A. *et al.* Variation in PCSK9 and HMGCR and Risk of Cardiovascular Disease and Diabetes. en. *N. Engl. J. Med.* **375,** 2144-2153 (déc. 2016).
19. Sabatine, M. S. *et al.* Evolocumab and Clinical Outcomes in Patients with Cardiovascular Disease. en. *N. Engl. J. Med.* **376,** 1713-1722 (mai 2017).
20. Shapiro, M. D., Tavori, H. & Fazio, S. PCSK9 : From Basic Science Discoveries to Clinical Trials. en. *Circ. Res.* **122,** 1420-1438 (mai 2018).
21. DiFrancesco, D. The role of the funny current in pacemaker activity. en. *Circ. Res.* **106,** 434-446 (fév. 2010).
22. Custodis, F. *et al.* Vascular pathophysiology in response to increased heart rate. en. *J. Am. Coll. Cardiol.* **56,** 1973-1983 (déc. 2010).
23. Fox, K. *et al.* Ivabradine for patients with stable coronary artery disease and left-ventricular systolic dysfunction (BEAUTIFUL) : a randomised, double-blind, placebo-controlled trial. en. *Lancet* **372,** 807-816 (sept. 2008).

24. Borer Jeffrey S., Fox Kim, Jaillon Patrice & Lerebours Guy. Antianginal and Antiischemic Effects of Ivabradine, an If Inhibitor, in Stable Angina. *Circulation* **107,** 817-823 (fév. 2003).
25. Swedberg, K. *et al.* Ivabradine and outcomes in chronic heart failure (SHIFT) : a randomised placebo-controlled study. en. *Lancet* **376,** 875-885 (sept. 2010).
26. Fox, K. *et al.* Rationale, design, and baseline characteristics of the Study assessInG the morbidity-mortality beNefits of the If inhibitor ivabradine in patients with coronarY artery disease (SIGNIFY trial) : a randomized, double-blind, placebo-controlled trial of ivabradine in patients with stable coronary artery disease without clinical heart failure. en. *Am. Heart J.* **166,** 654-661.e6 (oct. 2013).
27. Fox, K. *et al.* Ivabradine in stable coronary artery disease without clinical heart failure. en. *N. Engl. J. Med.* **371,** 1091-1099 (sept. 2014).
28. Martin, R. I. R. *et al.* Atrial fibrillation associated with ivabradine treatment : meta-analysis of randomised controlled trials. en. *Heart* **100,** 1506-1510 (oct. 2014).
29. Anter Elad, Jessup Mariell & Callans David J. Atrial Fibrillation and Heart Failure. *Circulation* **119,** 2516-2525 (mai 2009).
30. Qiu, X. *et al.* Crystal structure of cholesteryl ester transfer protein reveals a long tunnel and four bound lipid molecules. en. *Nat. Struct. Mol. Biol.* **14,** 106-113 (fév. 2007).
31. Gordon D J *et al.* High-density lipoprotein cholesterol and cardiovascular disease. Four prospective American studies. *Circulation* **79,** 8-15 (jan. 1989).
32. Inazu, A. *et al.* Increased high-density lipoprotein levels caused by a common cholesteryl-ester transfer protein gene mutation. en. *N. Engl. J. Med.* **323,** 1234-1238 (nov. 1990).
33. Tall, A. R. & Rader, D. J. Trials and Tribulations of CETP Inhibitors. en. *Circ. Res.* **122,** 106-112 (jan. 2018).
34. Barter, P. J. *et al.* Effects of torcetrapib in patients at high risk for coronary events. en. *N. Engl. J. Med.* **357,** 2109-2122 (nov. 2007).
35. HPS3/TIMI55–REVEAL Collaborative Group *et al.* Effects of Anacetrapib in Patients with Atherosclerotic Vascular Disease. en. *N. Engl. J. Med.* **377,** 1217-1227 (sept. 2017).
36. Ouimet, M., Barrett, T. J. & Fisher, E. A. HDL and Reverse Cholesterol Transport. en. *Circ. Res.* **124,** 1505-1518 (mai 2019).
37. Rohatgi, A. *et al.* HDL cholesterol efflux capacity and incident cardiovascular events. en. *N. Engl. J. Med.* **371,** 2383-2393 (déc. 2014).

38. Khera, A. V. *et al.* Cholesterol efflux capacity, high-density lipoprotein function, and atherosclerosis. en. *N. Engl. J. Med.* **364,** 127-135 (jan. 2011).
39. Rhainds, D. & Tardif, J.-C. From HDL-cholesterol to HDL-function : cholesterol efflux capacity determinants. en. *Curr. Opin. Lipidol.* **30,** 101-107 (avr. 2019).
40. Assanasen, C. *et al.* Cholesterol binding, efflux, and a PDZ-interacting domain of scavenger receptor-BI mediate HDL-initiated signaling. en. *J. Clin. Invest.* **115,** 969-977 (avr. 2005).
41. Thomas, T. *et al.* CETP (Cholesteryl Ester Transfer Protein) Inhibition With Anacetrapib Decreases Production of Lipoprotein(a) in Mildly Hypercholesterolemic Subjects. en. *Arterioscler. Thromb. Vasc. Biol.* **37,** 1770-1775 (sept. 2017).
42. Calkin, A. C. *et al.* Reconstituted high-density lipoprotein attenuates platelet function in individuals with type 2 diabetes mellitus by promoting cholesterol efflux. en. *Circulation* **120,** 2095-2104 (nov. 2009).
43. Ronsein, G. E. & Vaisar, T. Inflammation, remodeling, and other factors affecting HDL cholesterol efflux. en. *Curr. Opin. Lipidol.* **28,** 52-59 (fév. 2017).
44. Brodeur, M. R. *et al.* Dalcetrapib and anacetrapib differently impact HDL structure and function in rabbits and monkeys. en. *J. Lipid Res.* **58,** 1282-1291 (juil. 2017).
45. Kuivenhoven, J. A. *et al.* The role of a common variant of the cholesteryl ester transfer protein gene in the progression of coronary atherosclerosis. The Regression Growth Evaluation Statin Study Group. en. *N. Engl. J. Med.* **338,** 86-93 (jan. 1998).
46. Boekholdt, S. M. *et al.* Cholesteryl ester transfer protein TaqIB variant, high-density lipoprotein cholesterol levels, cardiovascular risk, and efficacy of pravastatin treatment : individual patient meta-analysis of 13,677 subjects. en. *Circulation* **111,** 278-287 (jan. 2005).
47. Thompson, A. *et al.* Association of cholesteryl ester transfer protein genotypes with CETP mass and activity, lipid levels, and coronary risk. en. *JAMA* **299,** 2777-2788 (juin 2008).
48. Johannsen, T. H., Frikke-Schmidt, R., Schou, J., Nordestgaard, B. G. & Tybjærg-Hansen, A. Genetic inhibition of CETP, ischemic vascular disease and mortality, and possible adverse effects. en. *J. Am. Coll. Cardiol.* **60,** 2041-2048 (nov. 2012).
49. Niu Wenquan & Qi Yue. Circulating Cholesteryl Ester Transfer Protein and Coronary Heart Disease. *Circ. Cardiovasc. Genet.* **8,** 114-121 (fév. 2015).
50. Blauw Lisanne L. *et al.* CETP (Cholesteryl Ester Transfer Protein) Concentration. *Circulation : Genomic and Precision Medicine* **11,** e002034 (mai 2018).

51. Ference, B. A. *et al.* Association of Genetic Variants Related to CETP Inhibitors and Statins With Lipoprotein Levels and Cardiovascular Risk. en. *JAMA* **318,** 947-956 (sept. 2017).
52. Nomura, A. *et al.* Protein-Truncating Variants at the Cholesteryl Ester Transfer Protein Gene and Risk for Coronary Heart Disease. en. *Circ. Res.* **121,** 81-88 (juin 2017).
53. Kettunen, J. *et al.* Lipoprotein signatures of cholesteryl ester transfer protein and HMG-CoA reductase inhibition. en. *PLoS Biol.* **17,** e3000572 (déc. 2019).
54. Blauw, L. L. *et al.* Mendelian randomization reveals unexpected effects of CETP on the lipoprotein profile. en. *Eur. J. Hum. Genet.* **27,** 422-431 (mar. 2019).
55. Tardif, J.-C. *et al.* Pharmacogenomic determinants of the cardiovascular effects of dalcetrapib. en. *Circ. Cardiovasc. Genet.* **8,** 372-382 (avr. 2015).
56. Rautureau, Y. *et al.* ADCY9 (Adenylate Cyclase Type 9) Inactivation Protects From Atherosclerosis Only in the Absence of CETP (Cholesteryl Ester Transfer Protein). en. *Circulation* **138,** 1677-1692 (oct. 2018).
57. Tardif, J.-C. *et al.* Genotype-Dependent Effects of Dalcetrapib on Cholesterol Efflux and Inflammation : Concordance With Clinical Outcomes. en. *Circ. Cardiovasc. Genet.* **9,** 340-348 (août 2016).
58. Nissen, S. E. *et al.* ADCY9 Genetic Variants and Cardiovascular Outcomes With Evacetrapib in Patients With High-Risk Vascular Disease : A Nested Case-Control Study. en. *JAMA Cardiol* **3,** 401-408 (mai 2018).
59. Hopewell, J. C. *et al.* Impact of ADCY9 Genotype on Response to Anacetrapib. en. *Circulation* (juil. 2019).
60. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. en. *Nucleic Acids Res.* **47,** D1005-D1012 (jan. 2019).
61. Wood, A. R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. en. *Nat. Genet.* **46,** 1173-1186 (nov. 2014).
62. Nikpay, M. *et al.* A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. en. *Nat. Genet.* **47,** 1121-1130 (oct. 2015).
63. Leusink, M., Onland-Moret, N. C., de Bakker, P. I. W., de Boer, A. & Maitland-van der Zee, A. H. Seventeen years of statin pharmacogenetics : a systematic review. en. *Pharmacogenomics* **17,** 163-180 (2016).
64. Das, S. *et al.* Next-generation genotype imputation service and methods. en. *Nat. Genet.* **48,** 1284-1287 (oct. 2016).

65. Taliun, D. *et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program* en. Mar. 2019.
66. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. en. *Nature* **447,** 661-678 (juin 2007).
67. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. en. *Nat. Genet.* **47,** 291-295 (mar. 2015).
68. Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. en. *Nat. Rev. Genet.* **19,** 581-590 (sept. 2018).
69. Inouye, M. *et al.* Genomic Risk Prediction of Coronary Artery Disease in 480,000 Adults : Implications for Primary Prevention. en. *J. Am. Coll. Cardiol.* **72,** 1883-1893 (oct. 2018).
70. Zhu, Z. *et al.* Causal associations between risk factors and common diseases inferred from GWAS summary data. en. *Nat. Commun.* **9,** 224 (jan. 2018).
71. Roden, D. M. *et al.* Development of a large-scale de-identified DNA biobank to enable personalized medicine. en. *Clin. Pharmacol. Ther.* **84,** 362-369 (sept. 2008).
72. Denny, J. C. *et al.* PheWAS : demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* **26,** 1205-1210 (mai 2010).
73. Rastegar-Mojarad, M., Ye, Z., Kolesar, J. M., Hebbring, S. J. & Lin, S. M. Opportunities for drug repositioning from phenome-wide association studies. *Nat. Biotechnol.* **33,** 342-345 (avr. 2015).
74. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506,** 376-381 (fév. 2014).
75. Kingsmore, K. M., Grammer, A. C. & Lipsky, P. E. Drug repurposing to improve treatment of rheumatic autoimmune inflammatory diseases. en. *Nat. Rev. Rheumatol.* **16,** 32-52 (jan. 2020).
76. Fang, H. *et al.* A genetics-led approach defines the drug target landscape of 30 immune-related traits. en. *Nat. Genet.* **51,** 1082-1091 (juil. 2019).
77. Rao Abhiram S. *et al.* Large-Scale Phenome-Wide Association Study of PCSK9 Variants Demonstrates Protection Against Ischemic Stroke. *Circulation : Genomic and Precision Medicine* **11,** e002162 (juil. 2018).
78. Sabatine, M. S. PCSK9 inhibitors : clinical evidence and implementation. en. *Nat. Rev. Cardiol.* **16,** 155-165 (mar. 2019).
79. Giugliano, R. P. *et al.* Cognitive Function in a Randomized Trial of Evolocumab. en. *N. Engl. J. Med.* **377,** 633-643 (août 2017).

80. Panagiotou, O. A., Ioannidis, J. P. A. & Genome-Wide Significance Project. What should the genome-wide significance threshold be ? Empirical replication of borderline genetic associations. en. *Int. J. Epidemiol.* **41,** 273-286 (fév. 2012).
81. Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-variant association analysis : study designs and statistical tests. *Am. J. Hum. Genet.* **95,** 5-23 (juil. 2014).
82. Purcell, S. *et al.* PLINK : A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **81,** 559-575 (sept. 2007).
83. Liu, J. Z. *et al.* A versatile gene-based test for genome-wide association studies. en. *Am. J. Hum. Genet.* **87,** 139-145 (juil. 2010).
84. Li, M.-X., Gui, H.-S., Kwan, J. S. H. & Sham, P. C. GATES : a rapid and powerful gene-based association test using extended Simes procedure. en. *Am. J. Hum. Genet.* **88,** 283-293 (mar. 2011).
85. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **57,** 289-300 (1995).
86. Lamparter, D., Marbach, D., Rueedi, R., Kutalik, Z. & Bergmann, S. Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics. en. *PLoS Comput. Biol.* **12,** e1004714 (jan. 2016).
87. Wang, M. *et al.* COMBAT : A Combined Association Test for Genes Using Summary Statistics. en. *Genetics* **207,** 883-891 (nov. 2017).
88. Gauderman, W. J., Murcray, C., Gilliland, F. & Conti, D. V. Testing association between disease and multiple SNPs in a candidate gene. en. *Genet. Epidemiol.* **31,** 383-395 (juil. 2007).
89. De Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA : generalized gene-set analysis of GWAS data. en. *PLoS Comput. Biol.* **11,** e1004219 (avr. 2015).
90. Kwee, L. C., Liu, D., Lin, X., Ghosh, D. & Epstein, M. P. A powerful and flexible multilocus association test for quantitative traits. en. *Am. J. Hum. Genet.* **82,** 386-397 (fév. 2008).
91. Wu, M. C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. en. *Am. J. Hum. Genet.* **89,** 82-93 (juil. 2011).
92. Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J. D. & Lin, X. Sequence kernel association tests for the combined effect of rare and common variants. en. *Am. J. Hum. Genet.* **92,** 841-853 (juin 2013).
93. Lee, S. *et al.* Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. en. *Am. J. Hum. Genet.* **91,** 224-237 (août 2012).

94. Gagliano Taliun, S. A. *et al.* Exploring and visualizing large-scale genetic associations by using PheWeb. en. *Nat. Genet.* **52,** 550-552 (juin 2020).
95. Ghoussaini, M. *et al.* Open Targets Genetics : systematic identification of trait-associated genes using large-scale genetics and functional genomics. en. *Nucleic Acids Res.* **49,** D1311-D1320 (jan. 2021).
96. Lambert, S. A. *et al. The Polygenic Score Catalog : an open database for reproducibility and systematic evaluation* mai 2020.
97. Staley, J. R. *et al.* PhenoScanner : a database of human genotype-phenotype associations. en. *Bioinformatics* **32,** 3207-3209 (oct. 2016).
98. Kamat, M. A. *et al.* PhenoScanner V2 : an expanded tool for searching human genotype–phenotype associations. en. *Bioinformatics* **35,** 4851-4853 (juin 2019).
99. McKusick, V. A. Mendelian Inheritance in Man and its online version, OMIM. en. *Am. J. Hum. Genet.* **80,** 588-604 (avr. 2007).
100. Kamat, M. A. *et al.* PhenoScanner V2 : an expanded tool for searching human genotype-phenotype associations. en. *Bioinformatics* **35,** 4851-4853 (nov. 2019).
101. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. en. *Nature* **526,** 68-74 (oct. 2015).
102. Ochoa, D. *et al.* Open Targets Platform : supporting systematic drug-target identification and prioritisation. en. *Nucleic Acids Res.* **49,** D1302-D1310 (jan. 2021).
103. Pruim, R. J. *et al.* LocusZoom : regional visualization of genome-wide association scan results. en. *Bioinformatics* **26,** 2336-2337 (sept. 2010).
104. Dube, M.-P. *et al. Genetics of symptom remission in outpatients with COVID-19* mar. 2021.
105. Dubé Marie-Pierre *et al.* Pharmacogenomics of the Efficacy and Safety of Colchicine in COLCOT. *Circulation : Genomic and Precision Medicine* **0.**
106. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. en. *Nature* **562,** 203-209 (oct. 2018).
107. Goldstein, J. A. *et al.* LabWAS : Novel findings and study design recommendations from a meta-analysis of clinical labs in two independent biobanks. en. *PLoS Genet.* **16,** e1009077 (nov. 2020).
108. Hemani, G. *et al.* The MR-Base platform supports systematic causal inference across the human phenome. en. *Elife* **7** (mai 2018).
109. Lambert, S. A. *et al.* The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation. *Nat. Genet.* (mar. 2021).
110. Dudbridge, F. Power and predictive accuracy of polygenic risk scores. *PLoS Genet.* **9,** e1003348 (mar. 2013).

111. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. en. *Nat. Genet.* **50,** 1219-1224 (sept. 2018).
112. Khera, A. V. *et al.* Genetic Risk, Adherence to a Healthy Lifestyle, and Coronary Disease. *N. Engl. J. Med.,* NEJMoa1605086 (nov. 2016).
113. Mega, J. L. *et al.* Genetic risk, coronary heart disease events, and the clinical benefit of statin therapy : an analysis of primary and secondary prevention trials. en. *Lancet* **385,** 2264-2271 (juin 2015).
114. Choi, S. W., Mak, T. S.-H. & O'Reilly, P. F. Tutorial : a guide to performing polygenic risk score analyses. en. *Nat. Protoc.* **15,** 2759-2772 (sept. 2020).
115. Vilhjálmsson, B. J. *et al.* Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. en. *Am. J. Hum. Genet.* **97,** 576-592 (oct. 2015).
116. Privé, F., Arbel, J. & Vilhjálmsson, B. J. LDpred2 : better, faster, stronger. en. *Bioinformatics* (déc. 2020).
117. Mak, T. S. H., Porsch, R. M., Choi, S. W., Zhou, X. & Sham, P. C. Polygenic scores via penalized regression on summary statistics. en. *Genet. Epidemiol.* **41,** 469-480 (sept. 2017).
118. Choi, S. W. & O'Reilly, P. F. PRSice-2 : Polygenic Risk Score software for biobank-scale data. en. *Gigascience* **8** (juil. 2019).
119. Euesden, J., Lewis, C. M. & O'Reilly, P. F. PRSice : Polygenic Risk Score software. en. *Bioinformatics* **31,** 1466-1468 (mai 2015).
120. Katan, M. APOLIPOPROTEIN E ISOFORMS, SERUM CHOLESTEROL, AND CANCER. *Lancet* **327,** 507-508 (mar. 1986).
121. Thomas, D. C. & Conti, D. V. Commentary : the concept of 'Mendelian Randomization'. en. *Int. J. Epidemiol.* **33,** 21-25 (fév. 2004).
122. Lousdal, M. L. An introduction to instrumental variable assumptions, validation and estimation. en. *Emerg. Themes Epidemiol.* **15,** 1 (jan. 2018).
123. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments : effect estimation and bias detection through Egger regression. en. *Int. J. Epidemiol.* **44,** 512-525 (avr. 2015).
124. Davies, N. M., Holmes, M. V. & Davey Smith, G. Reading Mendelian randomisation studies : a guide, glossary, and checklist for clinicians. en. *BMJ* **362,** k601 (juil. 2018).
125. Wallace, C. *A more accurate method for colocalisation analysis allowing for multiple causal variants* en. Fév. 2021.

126. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. en. *PLoS Genet.* **10,** e1004383 (mai 2014).

127. Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M. & Smoller, J. W. Pleiotropy in complex traits : challenges and strategies. *Nat. Rev. Genet.* **14,** 483-495 (juil. 2013).

128. Burgess, S., Small, D. S. & Thompson, S. G. A review of instrumental variable estimators for Mendelian randomization. en. *Stat. Methods Med. Res.* **26,** 2333-2355 (oct. 2017).

129. Efron, B. & Tibshirani, R. J. *An Introduction to the Bootstrap* en (CRC Press, mai 1994).

130. Pierce, B. L. & Burgess, S. Efficient design for Mendelian randomization studies : subsample and 2-sample instrumental variable estimators. en. *Am. J. Epidemiol.* **178,** 1177-1184 (oct. 2013).

131. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. en. *Nat. Genet.* **48,** 481-487 (mai 2016).

132. Bowden, J. Misconceptions on the use of MR-Egger regression and the evaluation of the InSIDE assumption. en. *Int. J. Epidemiol.* **46,** 2097-2099 (déc. 2017).

133. Minelli, C. *et al. The use of two-sample methods for Mendelian randomization analyses on single large datasets* en. Mai 2020.

134. Bowden, J., Davey Smith, G., Haycock, P. C. & Burgess, S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. en. *Genet. Epidemiol.* **40,** 304-314 (mai 2016).

135. Verbanck, M., Chen, C.-Y., Neale, B. & Do, R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. en. *Nat. Genet.* **50,** 693-698 (mai 2018).

136. Burgess, S., Foley, C. N., Allara, E., Staley, J. R. & Howson, J. M. M. A robust and efficient method for Mendelian randomization with hundreds of genetic variants. en. *Nat. Commun.* **11,** 376 (jan. 2020).

137. O'Connor, L. J. & Price, A. L. Distinguishing genetic correlation from causation across 52 diseases and complex traits. en. *Nat. Genet.* **50,** 1728-1734 (déc. 2018).

138. Brower, M. A. *et al.* Bidirectional Mendelian randomization to explore the causal relationships between body mass index and polycystic ovary syndrome. en. *Hum. Reprod.* **34,** 127-136 (jan. 2019).

139. Randomised trial of cholesterol lowering in 4444 patients with coronary heart disease : the Scandinavian Simvastatin Survival Study (4S). en. *Lancet* **344,** 1383-1389 (nov. 1994).

140. White, J. *et al.* Association of Lipid Fractions With Risks for Coronary Artery Disease and Diabetes. *JAMA Cardiology* **1,** 692-699 (sept. 2016).

141. Voight, B. F. *et al.* Plasma HDL cholesterol and risk of myocardial infarction : a mendelian randomisation study. en. *Lancet* **380,** 572-580 (août 2012).

142. Burgess, S. & Thompson, S. G. Multivariable Mendelian randomization : the use of pleiotropic genetic variants to estimate causal effects. en. *Am. J. Epidemiol.* **181,** 251-260 (fév. 2015).

143. Burgess, S., Freitag, D. F., Khan, H., Gorman, D. N. & Thompson, S. G. Using multivariable Mendelian randomization to disentangle the causal effects of lipid fractions. en. *PLoS One* **9,** e108891 (oct. 2014).

144. Richardson, T. G. *et al.* Evaluating the relationship between circulating lipoprotein lipids and apolipoproteins with risk of coronary heart disease : A multivariable Mendelian randomisation analysis. en. *PLoS Med.* **17,** e1003062 (mar. 2020).

145. Schmidt, A. F. *et al.* Genetic drug target validation using Mendelian randomisation. en. *Nat. Commun.* **11,** 3255 (juin 2020).

146. Gkatzionis, A., Burgess, S. & Newcombe, P. J. Statistical Methods for cis-Mendelian Randomization. arXiv : 2101.04081 [q-bio.QM] (jan. 2021).

147. Burgess, S., Zuber, V., Valdes-Marquez, E., Sun, B. B. & Hopewell, J. C. Mendelian randomization with fine-mapped genetic data : Choosing from large numbers of correlated instrumental variables. en. *Genet. Epidemiol.* **41,** 714-725 (déc. 2017).

148. Burgess, S. & Thompson, S. G. Use of allele scores as instrumental variables for Mendelian randomization. en. *Int. J. Epidemiol.* **42,** 1134-1144 (août 2013).

149. Rees, J. M. B., Foley, C. N. & Burgess, S. Factorial Mendelian randomization : using genetic variants to assess interactions. en. *Int. J. Epidemiol.* **49,** 1147-1158 (août 2019).

150. Cannon, C. P. *et al.* Ezetimibe Added to Statin Therapy after Acute Coronary Syndromes. *N. Engl. J. Med.* **372,** 2387-2397 (juin 2015).

151. Burgess, S. *et al.* Association of LPA Variants With Risk of Coronary Disease and the Implications for Lipoprotein(a)-Lowering Therapies : A Mendelian Randomization Analysis. en. *JAMA Cardiol* **3,** 619-627 (juil. 2018).

152. Ference, B. A. *et al.* Variation in PCSK9 and HMGCR and Risk of Cardiovascular Disease and Diabetes. *N. Engl. J. Med.* **375,** 2144-2153 (déc. 2016).

153. Ference, B. A. *et al.* Mendelian Randomization Study of ACLY and Cardiovascular Disease. en. *N. Engl. J. Med.* **380,** 1033-1042 (mar. 2019).
154. Athey, S., Chetty, R. & Imbens, G. Combining Experimental and Observational Data to Estimate Treatment Effects on Long Term Outcomes. arXiv : 2006.09676 [stat.ME] (juin 2020).
155. Sun, B. B. *et al.* Genomic atlas of the human plasma proteome. en. *Nature* **558,** 73-79 (juin 2018).
156. Zheng, J. *et al.* Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases. en. *Nat. Genet.* **52,** 1122-1131 (oct. 2020).
157. Burgess, S., Thompson, S. G. & CRP CHD Genetics Collaboration. Avoiding bias from weak instruments in Mendelian randomization studies. en. *Int. J. Epidemiol.* **40,** 755-764 (juin 2011).
158. Smith, C. Drug target validation : Hitting the target. en. *Nature* **422,** 341, 343, 345 passim. issn : 0028-0836 (mar. 2003).
159. Kamstrup, P. R., Tybjaerg-Hansen, A., Steffensen, R. & Nordestgaard, B. G. Genetically elevated lipoprotein(a) and increased risk of myocardial infarction. en. *JAMA* **301,** 2331-2339 (juin 2009).
160. Hemani, G., Bowden, J. & Davey Smith, G. Evaluating the potential role of pleiotropy in Mendelian randomization studies. en. *Hum. Mol. Genet.* **27,** R195-R208 (août 2018).
161. DiFRANCESCO, D. The contribution of the 'pacemaker'current (if) to generation of spontaneous activity in rabbit sino-atrial node myocytes. *J. Physiol.* **434,** 23-40 (1991).
162. Moosmang, S. *et al.* Cellular expression and functional characterization of four hyperpolarization-activated pacemaker channels in cardiac and neuronal tissues. en. *Eur. J. Biochem.* **268,** 1646-1652 (mar. 2001).
163. Nelson, C. P. *et al.* Association analyses based on false discovery rate implicate new loci for coronary artery disease. en. *Nat. Genet.* **49,** 1385-1391 (sept. 2017).
164. Nielsen, J. B. *et al.* Biobank-driven genomic discovery yields new insight into atrial fibrillation biology. en. *Nat. Genet.* **50,** 1234-1239 (sept. 2018).
165. Shah, S. *et al.* Genome-wide association and Mendelian randomisation analysis provide insights into the pathogenesis of heart failure. en. *Nat. Commun.* **11,** 163 (jan. 2020).
166. Austin, P. C., Lee, D. S. & Fine, J. P. Introduction to the Analysis of Survival Data in the Presence of Competing Risks. en. *Circulation* **133,** 601-609 (fév. 2016).

167. Eppinga, R. N. *et al.* Identification of genomic loci associated with resting heart rate and shared genetic predictors with all-cause mortality. en. *Nat. Genet.* **48,** 1557-1563 (déc. 2016).

168. Burgess, S., Butterworth, A. & Thompson, S. G. Mendelian randomization analysis with multiple genetic variants using summarized data. en. *Genet. Epidemiol.* **37,** 658-665 (nov. 2013).

169. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. en. *Nature* **581,** 434-443 (mai 2020).

170. Roselli, C. *et al.* Multi-ethnic genome-wide association study for atrial fibrillation. en. *Nat. Genet.* **50,** 1225-1233 (juin 2018).

171. Wolf, P. A., Dawber, T. R., Thomas Jr, H. E. & Kannel, W. B. Epidemiologic assessment of chronic atrial fibrillation and risk of stroke : the Framingham study. en. *Neurology* **28,** 973-977 (oct. 1978).

172. Malik, R. *et al.* Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. en. *Nat. Genet.* **50,** 524-537 (avr. 2018).

173. Wang, T. J. *et al.* Temporal relations of atrial fibrillation and congestive heart failure and their joint influence on mortality : the Framingham Heart Study. en. *Circulation* **107,** 2920-2925 (juin 2003).

174. Larsson, S. C., Drca, N., Mason, A. M. & Burgess, S. Resting Heart Rate and Cardiovascular Disease. en. *Circ Genom Precis Med* **12,** e002459 (mar. 2019).

175. Wolf, P. A., Dawber, T. R., Thomas, H. E. & Kannel, W. B. Epidemiologic assessment of chronic atrial fibrillation and risk of stroke : The fiamingham Study. *Neurology* (1978).

176. Fox, K. *et al.* Bradycardia and atrial fibrillation in patients with stable coronary artery disease treated with ivabradine : an analysis from the SIGNIFY study. en. *Eur. Heart J.* **36,** 3291-3296 (déc. 2015).

177. Milano, A. *et al.* HCN4 mutations in multiple families with bradycardia and left ventricular noncompaction cardiomyopathy. en. *J. Am. Coll. Cardiol.* **64,** 745-756 (août 2014).

178. Milanesi, R., Baruscotti, M., Gnecchi-Ruscone, T. & DiFrancesco, D. Familial sinus bradycardia associated with a mutation in the cardiac pacemaker channel. en. *N. Engl. J. Med.* **354,** 151-157 (jan. 2006).

179. Pitcairn, E. *et al.* Coordinating heart morphogenesis : A novel role for hyperpolarization-activated cyclic nucleotide-gated (HCN) channels during cardiogenesis in Xenopus laevis. en. *Commun. Integr. Biol.* **10,** e1309488 (mai 2017).

180. Fry, A. *et al.* Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. en. *Am. J. Epidemiol.* **186,** 1026-1034 (nov. 2017).

181. Grundvold, I. *et al.* Low heart rates predict incident atrial fibrillation in healthy middle-aged men. en. *Circ. Arrhythm. Electrophysiol.* **6,** 726-731 (août 2013).

182. Papp, A. C. *et al.* Cholesteryl Ester Transfer Protein (CETP) polymorphisms affect mRNA splicing, HDL levels, and sex-dependent cardiovascular risk. en. *PLoS One* **7,** e31930 (mar. 2012).

183. Anagnostopoulou, K. K. *et al.* Sex-associated effect of CETP and LPL polymorphisms on postprandial lipids in familial hypercholesterolaemia. en. *Lipids Health Dis.* **8,** 24 (juin 2009).

184. Cai, G., Shi, G. & Huang, Z. Gender specific effect of CETP rs708272 polymorphism on lipid and atherogenic index of plasma levels but not on the risk of coronary artery disease : A case-control study. en. *Medicine* **97,** e13514 (déc. 2018).

185. Dedoussis, G. V. *et al.* Cholesteryl ester-transfer protein (CETP) polymorphism and the association of acute coronary syndromes by obesity status in Greek subjects : the CARDIO2000-GENE study. en. *Hum. Hered.* **63,** 155-161 (fév. 2007).

186. Christen, T. *et al.* Mendelian randomization analysis of cholesteryl ester transfer protein and subclinical atherosclerosis : A population-based study. en. *J. Clin. Lipidol.* **12,** 137-144.e1 (jan. 2018).

187. Kark, J. D. *et al.* Taq1B CETP polymorphism, plasma CETP, lipoproteins, apolipoproteins and sex differences in a Jewish population sample characterized by low HDL-cholesterol. en. *Atherosclerosis* **151,** 509-518 (août 2000).

188. Szustakowski, J. D. *et al. Advancing Human Genetics Research and Drug Discovery through Exome Sequencing of the UK Biobank* nov. 2020.

189. Cohen, J. C., Boerwinkle, E., Mosley, T. H. & Hobbs, H. H. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *N. Engl. J. Med.* **354,** 1264-1272 (mar. 2006).

190. Rothwell, P. M. Factors that can affect the external validity of randomised controlled trials. en. *PLoS Clin. Trials* **1,** e9 (mai 2006).

191. Melloni Chiara *et al.* Representation of Women in Randomized Clinical Trials of Cardiovascular Disease Prevention. *Circ. Cardiovasc. Qual. Outcomes* **3,** 135-142 (mar. 2010).

192. Yakerson, A. Women in clinical trials : a review of policy development and health equity in the Canadian context. en. *Int. J. Equity Health* **18,** 56 (avr. 2019).

193. Kettunen, J. *et al.* Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. en. *Nat. Commun.* **7,** 11122 (mar. 2016).
194. Merriman Tony R. Application of Genetic Epidemiology to CETP (Cholesteryl Ester Transfer Protein) Concentration and Risk of Cardiovascular Disease. *Circulation : Genomic and Precision Medicine* **11,** e002138 (mai 2018).
195. Dachet, C., Poirier, O., Cambien, F., Chapman, J. & Rouis, M. New functional promoter polymorphism, CETP/-629, in cholesteryl ester transfer protein (CETP) gene related to CETP mass and high density lipoprotein cholesterol levels : role of Sp1/Sp3 in transcriptional regulation. en. *Arterioscler. Thromb. Vasc. Biol.* **20,** 507-515 (fév. 2000).
196. Schwartz, G. G. *et al.* Effects of dalcetrapib in patients with a recent acute coronary syndrome. *N. Engl. J. Med.* **367,** 2089-2099 (nov. 2012).
197. Lincoff, A. M. *et al.* Evacetrapib and Cardiovascular Outcomes in High-Risk Vascular Disease. en. *N. Engl. J. Med.* **376,** 1933-1942 (mai 2017).
198. Thomas, T. *et al.* CETP (Cholesteryl Ester Transfer Protein) Inhibition With Anacetrapib Decreases Production of Lipoprotein(a) in Mildly Hypercholesterolemic Subjects. en. *Arterioscler. Thromb. Vasc. Biol.* **37,** 1770-1775 (sept. 2017).
199. Arai, H. *et al.* Efficacy and safety of the cholesteryl ester transfer protein inhibitor anacetrapib in Japanese patients with heterozygous familial hypercholesterolemia. en. *Atherosclerosis* **249,** 215-223 (juin 2016).
200. Arsenault, B. J. *et al.* Effect of atorvastatin, cholesterol ester transfer protein inhibition, and diabetes mellitus on circulating proprotein subtilisin kexin type 9 and lipoprotein(a) levels in patients at high cardiovascular risk. en. *J. Clin. Lipidol.* **12,** 130-136 (jan. 2018).
201. Cole, C. B. *et al.* Adiposity significantly modifies genetic risk for dyslipidemia. en. *J. Lipid Res.* **55,** 2416-2422 (nov. 2014).
202. Sull, J. W., Kim, S. & Jee, S. H. Effects of Obesity and Family History of Diabetes on the Association of CETP rs6499861 with HDL-C Level in Korean Populations. en. *J Lipid Atheroscler* **8,** 252-257 (sept. 2019).
203. Bloomfield, D. *et al.* Efficacy and safety of the cholesteryl ester transfer protein inhibitor anacetrapib as monotherapy and coadministered with atorvastatin in dyslipidemic patients. en. *Am. Heart J.* **157,** 352-360.e2 (fév. 2009).
204. Cannon, C. P. *et al.* Safety of anacetrapib in patients with or at high risk for coronary heart disease. en. *N. Engl. J. Med.* **363,** 2406-2415 (déc. 2010).
205. Minikel, E. V. *et al.* Evaluating drug targets through human loss-of-function genetic variation. en. *Nature* **581,** 459-464 (mai 2020).

206. Klingel, S. L. *et al.* Sex Differences in Blood HDL-c, the Total Cholesterol/HDL-c Ratio, and Palmitoleic Acid are Not Associated with Variants in Common Candidate Genes. en. *Lipids* **52,** 969-980 (déc. 2017).

207. Villard, E. F. *et al.* Genetic determination of plasma cholesterol efflux capacity is gender-specific and independent of HDL-cholesterol levels. en. *Arterioscler. Thromb. Vasc. Biol.* **33,** 822-828 (avr. 2013).

208. Schaefer, E. J. *et al.* Human apolipoprotein A-I and A-II metabolism. en. *J. Lipid Res.* **23,** 850-862 (août 1982).

209. Metzinger, M. P. *et al.* Effect of Anacetrapib on Cholesterol Efflux Capacity : A Substudy of the DEFINE Trial. en. *J. Am. Heart Assoc.* **9,** e018136 (déc. 2020).

210. Catalano, G. *et al.* Cellular SR-BI and ABCA1-mediated cholesterol efflux are gender-specific in healthy subjects. en. *J. Lipid Res.* **49,** 635-643. issn : 0022-2275 (mar. 2008).

211. Papp, A. C. *et al.* Cholesteryl Ester Transfer Protein (CETP) polymorphisms affect mRNA splicing, HDL levels, and sex-dependent cardiovascular risk. en. *PLoS One* **7,** e31930 (mar. 2012).

212. Blauw, L. L. *et al.* Serum CETP concentration is not associated with measures of body fat : The NEO study. en. *Atherosclerosis* **246,** 267-273 (mar. 2016).

213. Wang, Y. *et al.* Plasma cholesteryl ester transfer protein is predominantly derived from Kupffer cells. en. *Hepatology* **62,** 1710-1722 (déc. 2015).

214. Niesor, E. J. *et al.* Modulating cholesteryl ester transfer protein activity maintains efficient pre-β-HDL formation and increases reverse cholesterol transport. en. *J. Lipid Res.* **51,** 3443-3454. issn : 0022-2275, 1539-7262 (déc. 2010).

215. Izem, L. & Morton, R. E. Possible role for intracellular cholesteryl ester transfer protein in adipocyte lipid metabolism and storage. en. *J. Biol. Chem.* **282,** 21856-21865 (juil. 2007).

216. Izem, L., Liu, Y. & Morton, R. E. Exon 9-deleted CETP inhibits full length-CETP synthesis and promotes cellular triglyceride storage. en. *J. Lipid Res.* **61,** 422-431. issn : 0022-2275, 1539-7262 (mar. 2020).

217. Li, M.-X., Gui, H.-S., Kwan, J. S. H. & Sham, P. C. GATES : a rapid and powerful gene-based association test using extended Simes procedure. en. *Am. J. Hum. Genet.* **88,** 283-293 (mar. 2011).

218. Svishcheva, G. R., Belonogova, N. M., Zorkoltseva, I. V., Kirichenko, A. V. & Axenovich, T. I. Gene-based association tests using GWAS summary statistics. en. *Bioinformatics* **35,** 3701-3708 (oct. 2019).

219. Pedregosa, F. *et al.* Scikit-learn : Machine Learning in Python. *J. Mach. Learn. Res.* **12,** 2825-2830 (2011).
220. McCullagh, P. & Nelder, J. A. *Generalized Linear Models* en (Chapman et Hall, 1983).
221. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. en. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 9440-9445 (août 2003).
222. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. en. *Nat. Genet.* **42,** 565-569 (juil. 2010).
223. Bento, A. P. *et al.* The ChEMBL bioactivity database : an update. *Nucleic Acids Res.* **42,** D1083-90 (jan. 2014).
224. Korotkevich, G., Sukhov, V. & Sergushichev, A. *Fast gene set enrichment analysis* en. Oct. 2019.
225. Subramanian, A. *et al.* Gene set enrichment analysis : a knowledge-based approach for interpreting genome-wide expression profiles. en. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 15545-15550 (oct. 2005).
226. Raudvere, U. *et al.* g :Profiler : a web server for functional enrichment analysis and conversions of gene lists (2019 update). en. *Nucleic Acids Res.* **47,** W191-W198 (juil. 2019).
227. Global Lipids Genetics Consortium *et al.* Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45,** 1274-1283 (nov. 2013).
228. Rahman, F., Kwan, G. F. & Benjamin, E. J. Global epidemiology of atrial fibrillation. en. *Nat. Rev. Cardiol.* **11,** 639-654 (nov. 2014).
229. Qin, D., Mansour, M. C., Ruskin, J. N. & Heist, E. K. Atrial Fibrillation-Mediated Cardiomyopathy. en. *Circ. Arrhythm. Electrophysiol.* **12,** e007809 (déc. 2019).
230. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42,** D1001-6 (jan. 2014).
231. McNair, W. P. *et al.* SCN5A mutation associated with dilated cardiomyopathy, conduction disorder, and arrhythmia. en. *Circulation* **110,** 2163-2167 (oct. 2004).
232. Selcen, D. & Engel, A. G. Mutations in myotilin cause myofibrillar myopathy. en. *Neurology* **62,** 1363-1371 (avr. 2004).
233. McKusick, V. A. *Mendelian inheritance in man : a catalog of human genes and genetic disorders* (JHU Press, 1998).
234. Tucker, N. R. *et al.* Transcriptional and Cellular Diversity of the Human Heart. en. *Circulation* **142,** 466-482 (août 2020).
235. Legault, M.-A., Tardif, J.-C. & Dubé, M.-P. Pharmacogenomics of blood lipid regulation. en. *Pharmacogenomics* **19,** 651-665 (mai 2018).

236. Dewey, F. E. *et al.* Genetic and Pharmacologic Inactivation of ANGPTL3 and Cardiovascular Disease. en. *N. Engl. J. Med.* **377,** 211-221 (juil. 2017).

237. Koishi, R. *et al.* Angptl3 regulates lipid metabolism in mice. en. *Nat. Genet.* **30,** 151-157 (fév. 2002).

238. Regeneron Pharmaceuticals, Inc. *FDA Approves First-in-class Evkeeza™ (evinacumab-dgnb) for Patients with Ultra-rare Inherited Form of High Cholesterol* https://www.prnewswire.com/news-releases/fda-approves-first-in-class-evkeeza-evinacumab-dgnb-for-patients-with-ultra-rare-inherited-form-of-high-cholesterol-301227183.html. Accessed : 2021-3-13. Fév. 2021.

239. Wand, H. *et al.* Improving reporting standards for polygenic scores in risk prediction studies. *Nature* **591,** 211-219 (mar. 2021).

240. Van der Harst, P. & Verweij, N. Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease. en. *Circ. Res.* **122,** 433-443 (fév. 2018).

241. Burgess, S. *et al.* Dissecting Causal Pathways Using Mendelian Randomization with Summarized Genetic Data : Application to Age at Menarche and Risk of Breast Cancer. en. *Genetics* **207,** 481-487. issn : 0016-6731, 1943-2631 (oct. 2017).

242. Burgess, S., Daniel, R. M., Butterworth, A. S., Thompson, S. G. & EPIC-InterAct Consortium. Network Mendelian randomization : using genetic variants as instrumental variables to investigate mediation in causal pathways. en. *Int. J. Epidemiol.* **44,** 484-495. issn : 0300-5771, 1464-3685 (avr. 2015).

243. Frieden, T. R. Evidence for Health Decision Making - Beyond Randomized, Controlled Trials. en. *N. Engl. J. Med.* **377,** 465-475 (août 2017).

244. North, T.-L. *et al.* Using Genetic Instruments to Estimate Interactions in Mendelian Randomization Studies. en. *Epidemiology* **30,** e33-e35 (nov. 2019).

245. Morris, A. P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. en. *Nat. Genet.* **44,** 981-990 (sept. 2012).

246. Peters, T. M. *et al.* Sex Differences in the Risk of Coronary Heart Disease Associated With Type 2 Diabetes : A Mendelian Randomization Analysis. en. *Diabetes Care* **44,** 556-562 (fév. 2021).

247. Burgess, S., Davies, N. M., Thompson, S. G. & EPIC-InterAct Consortium. Instrumental variable analysis with a nonlinear exposure-outcome relationship. en. *Epidemiology* **25,** 877-885 (nov. 2014).

248. Silverwood, R. J. *et al.* Testing for non-linear causal effects using a binary genotype in a Mendelian randomization study : application to alcohol and cardiovascular traits. en. *Int. J. Epidemiol.* **43,** 1781-1790 (déc. 2014).
249. Sun, Y.-Q. *et al.* Body mass index and all cause mortality in HUNT and UK Biobank studies : linear and non-linear mendelian randomisation analyses. en. *BMJ* **364,** l1042 (mar. 2019).
250. Shalit, U., Johansson, F. D. & Sontag, D. Estimating individual treatment effect : generalization bounds and algorithms. arXiv : 1606.03976 [stat.ML] (juin 2016).
251. Alaa, A. M., Weisz, M. & van der Schaar, M. Deep Counterfactual Networks with Propensity-Dropout. arXiv : 1706.05966 [cs.LG] (juin 2017).
252. Lemieux Perreault, L.-P., Provost, S., Legault, M.-A., Barhdadi, A. & Dubé, M.-P. pyGenClean : efficient tool for genetic data clean up before association testing. en. *Bioinformatics* **29,** 1704-1705 (juil. 2013).
253. Burgess, S. & Thompson, S. G. Interpreting findings from Mendelian randomization using the MR-Egger method. en. *Eur. J. Epidemiol.* **32,** 377-389 (mai 2017).
254. Verweij, N., Eppinga, R. N., Hagemeijer, Y. & van der Harst, P. Identification of 15 novel risk loci for coronary artery disease and genetic risk of recurrent events, atrial fibrillation and heart failure. en. *Sci. Rep.* **7,** 2761 (juin 2017).
255. Legault, M.-A. *et al.* A genetic model of ivabradine recapitulates results from randomized clinical trials. en. *PLoS One* **15,** e0236193 (juil. 2020).
256. Gao, X., Starmer, J. & Martin, E. R. A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. en. *Genet. Epidemiol.* **32,** 361-369 (mai 2008).
257. Pearl, J., Glymour, M. & Jewell, N. P. *Causal Inference in Statistics : A Primer* en (John Wiley & Sons, mar. 2016).
258. Hernán, M. A. & Robins, J. M. *Causal Inference : What If* (Chapman & Hall/CRC, Boca Raton, 2020).
259. Lincoff, A. M. *et al.* Evacetrapib and Cardiovascular Outcomes in High-Risk Vascular Disease. en. *N. Engl. J. Med.* **376,** 1933-1942 (mai 2017).
260. Magi, R., Lindgren, C. M. & Morris, A. P. Meta-analysis of sex-specific genome-wide association studies. en. *Genet. Epidemiol.* **34,** 846-853 (déc. 2010).
261. Knol, M. J., van der Tweel, I., Grobbee, D. E., Numans, M. E. & Geerlings, M. I. Estimating interaction on an additive scale between continuous determinants in a logistic regression model. en. *Int. J. Epidemiol.* **36,** 1111-1118 (oct. 2007).
262. VanderWeele, T. J. & Knol, M. J. A Tutorial on Interaction. *Epidemiol. Method.* **3.** issn : 2194-9263, 2161-962X (jan. 2014).

263. Zou, G. Y. On the estimation of additive interaction by use of the four-by-two table and beyond. en. *Am. J. Epidemiol.* **168,** 212-224. issn : 0002-9262, 1476-6256 (juil. 2008).

264. Zhan, X., Hu, Y., Li, B., Abecasis, G. R. & Liu, D. J. RVTESTS : an efficient and comprehensive tool for rare variant association analysis using sequence data. en. *Bioinformatics* **32,** 1423-1426 (mai 2016).

ANNEXE A

Matériel supplémentaire pour le Chapitre 2

A.1 Supplementary Material

A.1.1 Supplementary Methods

Additional information on UK Biobank variable selection

Phenotype data based on a touchscreen-based questionnaire followed by a verbal interview with a trained nurse was gathered. Hospitalization records are also available through linkage to the Health Episode Statistics (HES). For this project we used data from hospitalization episodes between the beginning of the HES linkage (April 1st 1997) and the last available date for the current data release (March 1st 2016). The date and cause of death were also available from death records made available through linkage to the National Health Services records for England, Scotland and Wales. We defined clinically relevant variables based on combinations of self-reported diseases, operation codes and hospitalization or death record ICD9/ICD10 codes. For the definition of most variables, self-reported diseases were included. However, for myocardial infarction we noticed that many self-reported events were unsupported by HES data even though they occurred within the time period of the HES linkage. We used the baseline resting heart rate measurement (variable #102), prioritizing the manual reading (variable #95) if available and taking the average value if many readings were available. We used age at recruitment defined in variable #21022 and sex in variable #31. Many of the self-reported myocardial infarction events co-occurred with ICD10 codes for related but distinct disorders such as I25.1 (atherosclerotic heart disease), I20.0 (unstable angina) or R07.4 (chest pain) without diagnostic codes for myocardial infarction, suggesting ischemic disease without a myocardial infarction event.

For this reason, we ignored self-reported events for myocardial infarction as well as for angina, unstable angina and coronary artery disease as participants may incorrectly report them. For prospective analyses we used time from first baseline assessment centre visit (extracted from variable #53) in years. The censure date was defined as the date of death or date of end of follow up. The end of follow-up date was set to 2016-03-01 for England and Wales and to 2015-11-30 for Scotland as defined in the UK Biobank documentation (https://biobank.ctsu.ox.ac.uk/b̃bdatan/death_cancer_report_Sept16.pdf). Individuals were assigned to countries based on the location of the UK Biobank assessment centre visited for the baseline visit.

UK Biobank additional genetic quality controls

All UK Biobank participants were previously genotyped using two similar arrays, the UK BiLEVE Axiom Array and the UK Biobank Axiom Array and genome-wide imputation was conducted using the Haplotype Reference Consortium as the main reference panel. Additional genetic quality control was done using pyGenClean version 1.8.3 [252]. Variants or individuals with more than 2% missing genotypes (per sample and per variant, respectively) were filtered out. The self-reported sex and the genetic sex based on sexual chromosome was compared and individuals with discrepancies or with aneuploidies were removed from the analysis. We only considered individuals of European descent for this study as they represent the major population in the UK Biobank. We used the computed principal components from the UK Biobank and defined a region in principal components space using individuals identified as "white British ancestry" as a reference population [106]. To avoid including related individuals, we used the kinship estimates from the UK Biobank and randomly selected an individual for pairs with a kinship coefficient > 0.0884 corresponding to individuals with 2nd degree relationships or less [106]. The resulting post QC dataset included 413,083 individuals. There were 1,165 available imputed variants located at the *HCN4* gene region (chr15:73,612,200-73,661,605 $\pm$ 200kb padding) with a MAF above 1% and that were bi-allelic.

External summary statistics

The CARDIoGRAMplusC4D consortium published a 1000 Genomes based meta-analysis of myocardial infarction and CAD of 60,810 CAD cases [62] and a more recent meta-analysis that adds the UK Biobank and the MIGen / CARDIOGRAM exome chip study [163]. The main data release was based on the UK Biobank "soft" CAD definition that includes self-reported chronic ischemic heart disease and angina patients as well as the cases for the "hard"

CAD definition of previous myocardial infarction or revascularization.

Two recent GWAS of atrial fibrillation were used. The first dataset was published by Nielsen *et al.* [164] and was based on 60,620 atrial fibrillation patients of European ancestry from six studies. The second dataset was published by Roselli *et al.* [170] and included 65,446 atrial fibrillation patients predominantly of European ancestry. This study conducted a trans-ethnic association analysis that also included participants of Japanese (12.5%), African American (1.3%) and Brazilian and Hispanic (1.3%) populations.

For heart failure, we summary association statistics from the HERMES case-control consortium including 47,309 cases and 910,014 controls [165].

Finally, for stroke, we obtained summary associated statistics from the MEGASTROKE consortium who conducted GWAS for many stroke subtypes including ischemic stroke, large artery stroke, cardioembolic stroke and small vessel strokes. There were two available datasets, one based on European individuals including 40,585 stroke cases and a trans-ethnic GWAS dataset including 67,162 stroke cases [172].

Mendelian randomization

Mendelian randomization is a technique to infer the causal effect of an exposure such as heart rate on an outcome such as atrial fibrillation. Because genetic variants are randomly assigned at birth and are generally not influenced by the environment, they represent an unconfounded way of modulating the exposure that may be suitable for causal inference. Unfortunately, MR is susceptible to other methodological issues some of which can be accounted for in more sophisticated models. A commonly used model that assumes the classical instrument variable conditions is the inverse variance weighted (IVW) approach [168]. In this approach, the causal estimates for multiple variants are averaged with weights corresponding to the precision of the causal estimates without accounting for the possibility of invalid genetic instruments. The MR-Egger test is a similar approach that extends the IVW by allowing an intercept term that corresponds to directional pleiotropy [123]. Directional pleiotropy is detectable if the mean direct effect of genetic variants on the outcome is different from zero. The MR-Egger approach relaxes the exclusion-restriction (or IV3) assumption that requires genetic variants to be conditionally independent of the outcome given the exposure and covariates. Instead, MR-Egger requires the instrument strength independent of direct effect (InSIDE) assumption which stipulates that the effect of genetic variants on the exposure should be independent from the direct effects. MR-Egger is useful method, but the InSIDE assumption is hard to verify and is likely to fail in many biologically

plausible scenarios leading to possibly biased estimates [253]. Moreover, the IVW and MR-Egger are greatly influenced by outlier variants as they are based on a linear regression of individual variant effects. A more recent set of MR methods further relax these assumptions and rely on the hypothesis that the largest set of variants with homogeneous effects is likely to represent the set of valid instrument variables. The contamination mixture method uses a mixture model to assign variants to a distribution of valid causal effects and a distribution of noisy variants with a null expected causal effect and a large variance [136]. Similarly, the Mendelian Randomization Pleiotropy RESidual Sum and Outlier (MR-PRESSO) iteratively eliminates variants whose effects are outliers when compared to the others until a homogeneous signal remains corresponding to the estimated causal effect [135]. These methods are interesting because they are less dependent on hard to verify assumptions, they are robust even when some invalid instruments are included, and they allow variants to be individually tested for their heterogeneity and further investigated.

For MR with the heart rate GRS, we used the two-stage method which is akin to the previously described IVW method, but uses individual level data [128]. Heart rate expressed in units of 10 bpm reduction was predicted based on a fitted model including the GRS as a continuous variable and covariates (age, sex and PCs). The predicted heart rates were then used in the second stage to estimate the causal effect on CAD, heart failure and atrial fibrillation using logistic regression adjusted for the same covariates. The standard errors for the causal effect were estimated using the percentile method based on 5,000 bootstrap resamples which is an empirical way of estimating standard errors without assuming their distribution. This approach does not account for violations of the instrument variable assumptions, but the estimates rely on a strong genetic instrument whose effect is closer to pharmacological effects than what is observed using individual genetic variants making extrapolations less problematic.

Analyses were performed with the "MendelianRandomization" R package (https://cran.r-project.org/web/packages/MendelianRandomization/) and MR-PRESSO (https://github.com/rondolab/MR-PRESSO).

Bi-directional MR

Bi-directional MR is a method used to infer the direction of the causality between two traits [138]. Genetic variants are ascertained for their association with the first trait and used as an instrument variable to estimate its effect on the second trait. The procedure is then repeated with instruments for the second trait (with the first trait as the outcome). The estimated causal effects can then be compared, and the direction of effect can be eluci-

dated. Here, we were interested in the causality between atrial fibrillation and heart failure, CAD and myocardial infarction. We used 11 genome-wide significant independent variants from the HERMES consortium and estimated the causal effect of heart failure on atrial fibrillation using these variants. The previously described MR methods were used for the bi-directional analysis as well. To test the effect of atrial fibrillation on heart failure, we selected uncorrelated variants that reached genome-wide significance in the Nielsen et al. study 5. The selection was done using grstools and the 1000 Genomes Phase III Europeans as a reference panel for linkage disequilibrium. Variants with a linkage disequilibrium r^2 above 0.15 were clumped together, keeping the most significant variant. The selection was stopped when no genome-wide significant variant remained leaving 152 variants to be used as genetic instruments.

The variants for the myocardial infarction genetic instrument were selected as for the atrial fibrillation instrument but using the CARDIoGRAMplusC4D summary statistics. A total of 31 variants were selected all of which were available in the atrial fibrillation summary statistics. For the CAD instrument, we used the 71 variants comprising the genetic risk score described in Verweij *et al.* [254].

Genetic risk score for heart rate

For the construction of the heart rate Genetic Risk Score (GRS), we used the 64 genome-wide significant heart rate associated SNPs from Eppinga *et al.* [167] To ensure that there was no strand confusion due to ambiguous alleles (*i.e.* A/T and G/C SNPs), we compared the observed allele frequencies to the expected distribution using the 1000 Genomes Phase III Europeans as a reference panel. If the observed allele frequency fell in a 95% Clopper-Pearson confidence interval around the reference panel frequency estimate, the strand was considered to be validated and the SNP was used as-is. When variants had a minor allele frequency (MAF) above 40%, an unambiguous SNP in LD was automatically selected to avoid strand confusion as the frequencies were close to 50%. A total of 8 of the 10 ambiguous variants were validated based on allele frequency. The other 2 variants had a MAF above 40% and were replaced by tag SNPs in LD. The variant rs13165531 was replaced by rs6887889 ($r^2 = 1$) and rs3951016 was replaced by rs9401060 ($r^2 = 0.91$). GRS weights were adjusted by multiplying the r^2 value for both SNPs. The final set of variants and their corresponding weights are shown in Supplementary Table A.5. The software used to compute the genetic risk scores is publicly available at https://github.com/legaultmarc/grstools.

A.1.2 Supplementary Tables

Supplementary Table A.1 – Summary of ivabradine cardiovascular outcomes trials.

Study / Intervention	**Patient population**	**Main cardiovascular exclusions**	**Primary efficacy endpoint**	**Results**
SHIFT 2.5-7.5 mg bid versus placebo	• Resting heart rate >=70 bpm • Symptomatic chronic HF for at least 4 weeks • LVEF <= 35% • Recent hospitalization for worsening HF	• HF caused by congenital heart disease or primary severe valvular disease • Recent MI, atrial fibrillation or flutter	CV death or hospitalization for worsening HF	HR 0.82 (0.75, 0.90) p<0.0001. Effect driven by hospitalization for worsening HF. Adverse events: Atrial fibrillation 9% in ivabradine vs 8% in placebo (p=0.012).
BEAUTIFUL 5-7.5 mg bid versus placebo	• Stable CAD (previous MI, PCI/CABG or angiographic evidence of obstruction of at least 50%) • LVEF <= 40% • Resting heart rate >= 65 bpm	• Recent revascularization or MI • Recent stroke or TIA • NYHA class IV HF • Implanted pacemaker, cardioverter or defibrillator • Valvular disease, SSS, sinoatrial block, severe hypertension	CV death or MI or hospitalization for worsening HF	Primary endpoint was not significant (p=0.94). In a prespecified subgroup with baseline heart rate >= 70bpm, there was a reduction for ischemic endpoints, which led to conduct SIGNIFY.
SIGNIFY 5-10 mg bid (treat to target of 55-60 bpm) versus placebo	• Stable CAD • No heart failure • Resting heart rate >= 70 bpm	• Patients with LVEF <= 40%	CV death or MI	Primary endpoint and its individual components were not significant In patients with severe angina (>= CCS II) there was an increase of the primary endpoint with ivabradine HR 1.18 (1.03, 1.35) p=0.02. Individuals in this group also found better anginal symptom improvement on ivabradine Atrial fibrillation was 5.3% in ivabradine vs 3.8% in placebo

CABG, coronary artery bypass graft; CAD, coronary artery disease; CV, cardiovascular; HF, heart failure; HR, hazard ratio; LVEF, left ventricular ejection fraction; MI, myocardial infarction; PCI, percutaneous coronary intervention; TIA, transient ischemic attack

Supplementary Table A.2 – Self-reported, hospitalization (ICD10) and operation (OPCS) codes used to define clinical variables based on the UK Biobank available data.

	Included codes		
Variable	Self-reported disease (variable #20002)	ICD9/10 for HES primary or secondary hospitalization codes or primary cause of death	Operations (OPCS)
Angina	-	ICD9: 413 ICD10: I20	
Unstable angina	-	ICD10: I20.0 (as the primary hospitalization code only to ensure acute event)	
Myocardial infarction	-	ICD9: 410, 412, 411.0, 429.79 ICD10: I21, I22, I23, I25.2	
Coronary artery disease	-	ICD9: 410-414 (except 414.1) ICD10: I20-I25	K40, K41, K42, K43, K44, K45, K46, K49, K50, K75
Stroke (any)	1583, 1081, 1086, 1491	ICD9: 430, 431, 434, 436 ICD10: I60, I61, I63, I64	
Stroke - Ischemic	1583	ICD9: 434, 436 ICD10: I63, I64	
Atrial fibrillation	1471	ICD9: 427.3 ICD10: I48	
Heart failure	1076	ICD9: 428, 425 ICD10: I50, I42	

Supplementary Table A.3 – Results from the NHGRI-EBI GWAS catalog mapped to the HCN4 gene. LD with the lead independent heart-rate associated variants in the UK Biobank identified through forward stepwise conditional analysis are reported.

Variant and risk allele	**Beta**	**P value**	**Trait**	**LD with rs8038766** [a]	**LD with rs3743496** [a]
rs7173389-T	0.539	1.00E-32	Resting heart rate	1	0.022
rs7173389-A	0.528	2.00E-09	Resting heart rate	1	0.022
rs8040516-T	0.219	3.00E-06	Nickel levels	0.001	0.532
rs74022964-T	0.113	4.00E-36	Atrial fibrillation	0.956	0.022
rs478438-G	0.019	2.00E-09	Heel bone mineral density	0.237	0.050
rs142859932-G	0.487	3.00E-06	Post bronchodilator FEV1	-	-
rs16957893-C	1.72	2.00E-08	Cold medicine-related Stevens-Johnson syndrome/toxic epidermal necrolysis (SJS/TEN) with severe ocular complications	0.025	0.004
rs7164883-G	0.17	3.00E-17	Atrial fibrillation	0.978	0.022
rs7183206-A	0.12	8.00E-12	Atrial fibrillation	0.956	0.018
rs2680344-A	0.024	5.00E-11	Heart rate variability traits (SDNN)	0.575	0
rs2680344-A	0.024	3.00E-11	Heart rate variability traits (SDNN)	0.575	0
rs2680344-A	0.032	1.00E-10	Heart rate variability traits (RMSSD)	0.575	0
rs2680344-A	0.046	3.00E-06	Heart rate variability traits (pvRSA/HF)	0.575	0
rs4489968-T	0.513	4.00E-20	Heart rate	1	0.022
rs7172038-G	0.10	2.00E-27	Atrial fibrillation	0.993	0.022
rs74022964-T	0.10	1.00E-27	Atrial fibrillation	0.956	0.022
rs11072405-A	0.021	1.00E-08	Waist-to-hip ratio adjusted for BMI	0.007	0.621
rs11072405-A		5.00E-07	Waist-to-hip ratio adjusted for BMI×sex×age interaction (4df test)	0.007	0.621

[a] LD measurements (r^2) are for individuals of European (EUR) descent from the 1000 Genomes Project (phase 3) and were obtained using LDlink

Supplementary Table A.4 – Variants and weights used for the computation of the heart rate GRS.

Variant	Chr.	Position	Reference allele	Risk allele	P value	Effect
rs145358377	1	6272136	G	GA	1.94E-11	-0.259
rs272564	1	45012273	A	C	4.51E-21	0.351
rs2152735	1	87893132	G	A	7.23E-18	-0.306
rs41317993	1	207961732	G	A	5.42E-31	0.630
rs11454451	1	217722890	C	CT	1.29E-11	0.256
rs1260326	2	27730940	T	C	4.29E-16	-0.275
rs12713404	2	60006705	G	T	9.33E-09	-0.199
rs564190295	2	175547672	G	GCCGCCGCCCCC	4.95E-10	-0.355
rs151041685	2	179725237	G	T	7.86E-75	1.061
rs62172372	2	188242369	A	G	5.99E-16	0.337
rs907683	2	220299541	G	T	1.02E-20	-0.334
rs4608502	2	228134155	T	C	1.85E-12	0.249
rs13002735	2	232268884	A	C	1.29E-17	-0.331
rs41312411	3	38621237	C	G	1.34E-11	-0.320
rs3749237	3	49770032	G	A	3.09E-13	0.258
rs2358740	3	53455569	G	T	3.58E-09	-0.208
rs1483890	3	69410725	A	G	2.54E-15	0.284
rs11920570	3	122090102	G	A	5.18E-13	0.268
rs7612445	3	179172979	G	T	2.41E-24	-0.428
rs12501032	4	23951018	C	G	1.83E-15	0.288
rs6845865	4	148974602	T	C	2.25E-14	-0.342
rs6887889 (tag for rs13165531)	5	30893205	T	G	3.57E-09	-0.221
rs1468333	5	137552970	T	C	9.53E-14	-0.255
rs4868243	5	172643118	G	A	4.08E-16	-0.361
rs236349	6	36820565	A	G	1.01E-15	0.281
rs9401060 (tag for rs3951016)	6	118561348	A	G	4.04E-33	0.473
rs1320761	6	122168138	C	T	1.22E-64	0.902
rs58437978	7	35258277	T	C	2.61E-12	-0.240
rs180239	7	93550415	G	C	4.54E-21	-0.326
rs17881696	7	100493359	G	A	1.18E-41	0.578
rs41748	7	116446573	T	G	7.14E-09	-0.193
rs11563648	7	126970046	G	C	4.42E-10	-0.231
rs138186803	7	130965408	AT	A	1.27E-16	-0.333
rs73158705	7	136576100	A	G	2.81E-18	0.393
rs56233017	8	144981488	G	A	1.09E-15	-0.666
rs10739663	9	128278739	A	G	9.62E-16	-0.266

Variant	Chr.	Position	Reference allele	Risk allele	P value	Effect
rs12576326	11	44980383	A	G	1.20E-12	0.253
rs174536	11	61551927	A	C	1.65E-30	0.399
rs75190942	11	128764571	C	A	1.19E-16	-0.496
rs2283274	12	2184466	G	C	7.21E-20	-0.405
rs10841486	12	20472202	T	C	2.98E-09	-0.238
rs4963772	12	24758480	G	A	3.23E-53	-0.714
rs1050288	12	27955296	C	T	2.74E-09	-0.213
rs1994135	12	33682405	T	C	7.19E-34	0.400
rs10880689	12	37930102	A	G	8.10E-10	0.208
rs867400	12	64976850	T	C	4.58E-19	0.298
rs12579753	12	82219376	C	T	4.81E-10	-0.246
rs12889267	14	21542766	A	G	3.61E-20	0.416
rs422068	14	23864804	T	C	1.52E-100	0.731
rs17180489	14	72885471	G	C	9.15E-19	-0.490
rs1549118	14	78379684	C	T	4.67E-08	0.200
rs17201923	14	85796564	A	G	6.55E-29	-0.41
rs4900069	14	91583373	A	C	5.38E-09	0.200
rs7173389	15	73663903	A	T	1.31E-32	-0.539
rs3915499	16	15910743	G	A	1.24E-17	0.303
rs7194801	16	65286870	T	C	3.58E-18	-0.291
rs79121763	17	15195279	C	T	7.17E-14	-0.471
rs11083258	18	25766218	A	C	5.51E-10	-0.276
rs61735998	18	34289285	G	T	2.06E-14	-0.834
rs16974196	19	40833470	G	A	1.11E-11	0.244
rs12721051	19	45422160	C	G	5.23E-11	-0.287
rs6123471	20	36840156	T	C	6.63E-72	-0.595
rs17265513	20	39832628	T	C	1.12E-08	0.240
rs2076028	22	39150450	G	A	5.45E-16	-0.295

Variants that were substituted by a tag SNP are identified in parenthesis and the p-value from the original GWAS is reported. For these variants, the weight is computed as the effect of the original variant weighted by the LD in Europeans ($r^2 \times \beta$). Chromosomal positions for GRCh37. GRS, genetic risk score.

Supplementary Table A.5 – MR estimates based on 64 heart-rate associated variants and their effect on outcomes in the UK Biobank. Reported effects are per genetically predicted s.d. decrease in heart rate (1 s.d. is 11.1 bpm in the UK Biobank).

			Intercept* (MR-Egger only)		Causal Estimate	
Exposure	Outcome	Method	Estimate (95% CI)	p-value	OR (95% CI)	p-value
Heart rate reduction (1 s.d. or 11.1 bpm)	Atrial fibrillation	IVW			1.25 (0.97, 1.62)	0.083
		MR-Egger	0.015 (-0.004, 0.034)	0.13	0.85 (0.48, 1.49)	0.57
		Contamination mixture			1.54 (1.22, 1.79)	-
		MR-PRESSO			1.54 (1.34, 1.78)	1.3E-07
	Heart failure	IVW			1.02 (0.87, 1.21)	0.77
		MR-Egger	0.007 (-0.005, 0.020)	0.60	0.84 (0.58, 1.21)	0.35
		Contamination mixture			1.16 (0.92, 1.43)	-
		MR-PRESSO †			-	-
	Coronary artery disease	IVW			1.01 (0.89, 1.15)	0.86
		MR-Egger	0.004 (-0.006, 0.014)	0.53	0.91 (0.68, 1.23)	0.44
		Contamination mixture			0.97 (0.87, 1.09)	-
		MR-PRESSO			1.01 (0.91, 1.12)	0.86

* The MR-Egger estimate intercepts represent directional pleiotropy and are not converted to the OR scale because they do not have an intuitive interpretation on this scale.

† The MR-PRESSO did not provide adjusted estimates as the global test did not detect significant pleiotropy (p=0.31).

Supplementary Table A.6 – MR estimates based on the effect of 64 heart-rate associated variants in external summary statistics from large GWAS consortia. Reported effects are per genetically predicted s.d. decrease in heart rate (1 s.d. is 11.1 bpm).

				Intercept* (MR-Egger only)		Causal estimate	
Exposure	Dataset (for the outcome)	Outcome	Method	Estimate (95% CI)	P value	OR (95% CI)	p-value
Heart rate reduction (1 s.d. or 11.1 bpm)	Nielsen et al.	Atrial fibrillation	IVW			1.24 (1.00, 1.53)	0.047
			MR-Egger	0.167 (-0.056, 0.378)	0.14	0.85 (0.49, 1.46)	0.55
			Contamination mixture			1.56 (1.40, 1.56)	-
			MR-PRESSO			1.36 (1.24, 1.49)	6.3E-08
	HERMES case/control	Heart failure	IVW			1.03 (0.97, 1.12)	0.34
			MR-Egger	-0.011 (-0.089, 0.067)	0.79	1.06 (0.88, 1.29)	0.55
			Contamination mixture			1.12 (1.00, 1.12)	-
			MR-PRESSO			-	-
	CARDIoGRAMplusC4D + UKB SOFT + MiGen	Coronary artery disease	IVW			1.07 (0.96, 1.19)	0.27
			MR-Egger	0.044 (-0.078, 0.167)	0.45	0.96 (0.71, 1.29)	0.77
			Contamination mixture			1.00 (1.00, 1.12)	-
			MR-PRESSO			1.04 (0.95, 1.14)	0.36

* The MR-Egger estimate intercepts represent directional pleiotropy and are not converted to the OR scale because they do not have an intuitive interpretation on this scale.

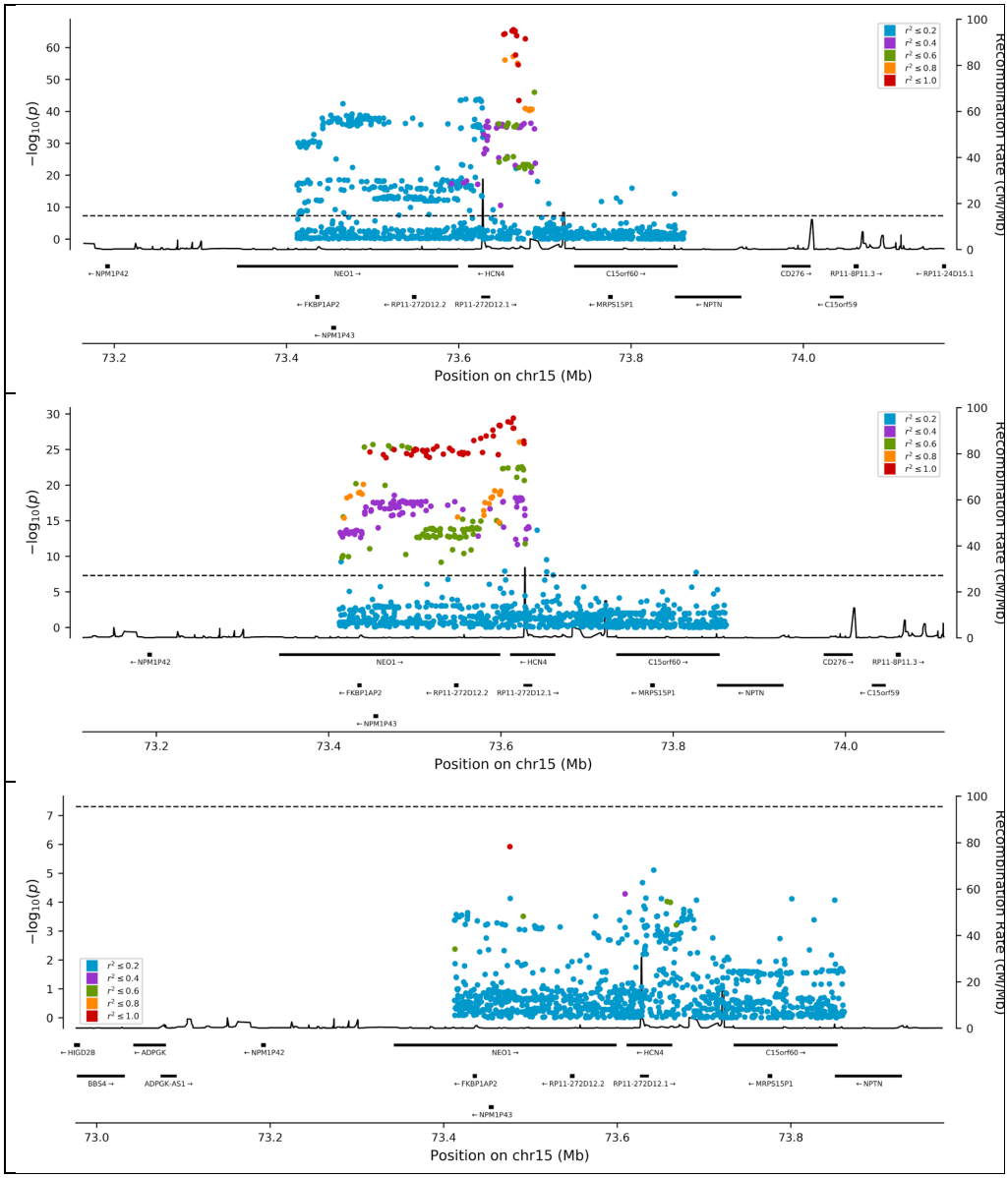
Supplementary Table A.7 – Bi-directional MR estimates using summary GWAS results for IVW and MR models more robust to invalid instruments.

			Intercept (MR-Egger only)		Causal estimate	
Exposure	Outcome	Method	Estimate (95% CI)	P-value	OR (95% CI)	P-value
Atrial fibrillation (152 variants)	Heart failure	IVW			1.23 (1.20, 1.27)	3.7E-52
		MR-Egger	0.001 (-0.004, 0.006)	0.77	1.22 (1.15, 1.30)	8.5E-10
		Contamination mixture			1.25 (1.22, 1.27)	-
		MR-PRESSO			1.23 (1.20, 1.26)	3.0E-33
Atrial fibrillation (152 variants)	Coronary artery disease	IVW			1.00 (0.98, 1.03)	0.76
		MR-Egger	0.004 (-0.001, 0.009)	0.09	0.96 (0.90, 1.02)	0.17
		Contamination mixture			1.01 (0.99, 1.03)	-
		MR-PRESSO[a]			-	-
Atrial fibrillation (152 variants)	Myocardial infarction	IVW			0.98 (0.95, 1.02)	0.30
		MR-Egger	0.003 (-0.003, 0.009)	0.32	0.95 (0.89, 1.02)	0.18
		Contamination mixture			0.98 (0.95, 1.01)	-
		MR-PRESSO			0.98 (0.95, 1.01)	0.23
Heart failure (11 variants)	Atrial Fibrillation	IVW			1.45 (1.11, 1.90)	0.0067
		MR-Egger	0.038 (0.014, 0.063)	0.002	1.21 (0.97, 1.52)	0.094
		Contamination mixture			6.82 (4.62, 9.20)	-
		MR-PRESSO			1.94 (1.66, 2.25)	0.07
Coronary artery disease (68 variants)	Atrial Fibrillation	IVW			1.15 (1.11, 1.21)	1.7E-10
		MR-Egger	0.003 (-0.005, 0.011)	0.47	1.12 (1.01, 1.23)	0.03
		Contamination mixture			1.17 (1.13, 1.21)	-
		MR-PRESSO			1.15 (1.11, 1.19)	2.5E-10
Myocardial infarction (31 variants)	Atrial Fibrillation	IVW			1.11 (1.06, 1.16)	1.3E-05
		MR-Egger	-0.010 (-0.021, 0.001)	0.07	1.22 (1.09, 1.36)	6.6E-04
		Contamination mixture			1.15 (1.12, 1.19)	-
		MR-PRESSO			1.11 (1.07, 1.16)	1.4E-05

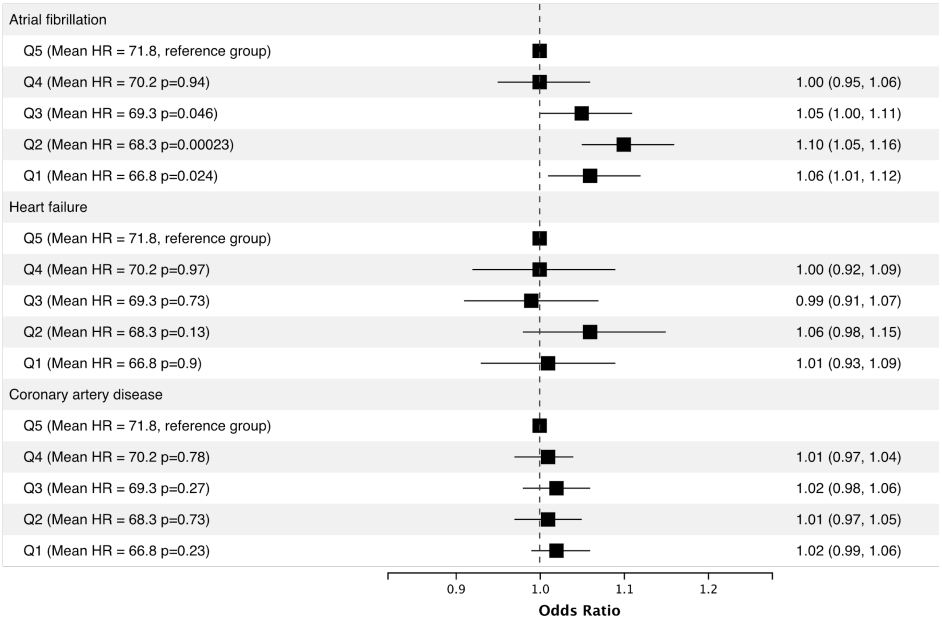
Summary statistics for atrial fibrillation taken from *Nielsen et al.* [164] for myocardial infarction and CAD from CARDIoGRAMplusC4D and CARDIoGRAMplusC4D + UKB SOFT + MiGen for CAD [62, 163], and for heart failure from the HERMES consortium [165]. The contamination mixture model does not provide P-values. IVW, inverse-variance weighted; MR, Mendelian randomization; OR, odds ratio.

[a] MR-PRESSO did not provide adjusted estimates as it detected no outliers.

A.1.3 Supplementary Figures



Supplementary Figure A.1 – **Results from the stepwise forward regression using 1,165 variants in the HCN4 region tested for association with heart rate in the UK Biobank, and adjusted for age, sex and the first 10 principal components.** The lead variant identified was rs8038766, $\beta = -0.574$ (95% CI -0.640, -0.509), $P = 2.76 \times 10^{-66}$ (results shown in the top panel). For the second stage, we conditioned on the lead variant from Stage 1 (rs8038766) and repeated the analysis. The lead variant identified in Stage 2 was rs3743496, $\beta = -0.297$ (95% CI -0.349, -0.246), $P = 3.96 \times 10^{-30}$ (results shown in the second panel). We repeated the analysis again conditioning on the lead variants from both previous stages (rs8038766 and rs3743496), but no additional variant crossed the genome-wide significance threshold (third panel). The first y-axis shows the negative $\log_{10}$ of P-values, the second y-axis shows the recombination rate from HapMap reference samples (black line). Genes are displayed below the x-axis from Ensembl (build37), the degree of linkage disequilibrium (r^2) of each genetic variant with the lead variant.



Supplementary Figure A.2 – **Effect of heart rate genetic risk score quintiles on atrial fibrillation, heart failure and coronary artery disease in the UK biobank dataset.** For every outcome, the highest heart rate group (5^{th} quintile) is used as the reference group and the reported odds ratios are adjusted for age, sex and the first 10 principal components.

ANNEXE B

Matériel supplémentaire pour le Chapitre 3

B.1 Supplementary Appendix

B.1.1 Effect modification by sex and BMI

In our analyses, both sex and BMI modified how a genetically predicted CETP reduction influenced lipid and lipoprotein levels. There is also evidence that the interaction between adiposity and lipids exhibits sexual dimorphism hinting at a possible 3-way interaction.

The three way interaction term including sex, BMI and genetic CETP reduction was significant for LDL-c levels ($p = 4.1 \times 10^{-4}$) and had concordant effects on apoB levels ($p = 0.001$). As seen on the marginal effect plots (Supplementary Figure B.11), BMI influences the effect of CETP on LDL-c and apoB in men with higher BMI values associated with a smaller decrease in LDL-c. This pattern is not seen in women and BMI does not influence the reduction of LDL-c with genetically lower CETP concentration. We did not detect a significant 3-way interaction (sex, BMI, genetic CETP reduction) for cardiovascular outcomes, but statistical power was likely limited. In the analysis based on rs1800775, results for biomarkers were similar to those obtained with the score (Supplementary Figure B.12).

We also estimated how sex and BMI may modifiy the effect of a genetically lower CETP on cholesterol efflux in the MHI Biobank. There was evidence for a 3-way interaction between sex, BMI and the CETP score on basal efflux ($p = 0.037$) and stimulated efflux ($p = 0.007$). The marginal effects in men and women with fixed BMI are presented in Supplementary Figure B.13. In men, increasing BMI may increase the CETP associated cAMP stimulated

cholesterol efflux (subgroup CETP by BMI interaction $p = 0.11$) but not its effect on basal efflux ($p = 0.53$), whereas in women increasing BMI reduced the effect of the CETP score on both stimulated ($p = 0.045$) and unstimulated efflux ($p = 0.050$).

For cardiovascular events, the 3-way interaction p-values with the genetic score were 0.43 for revascularization procedures, 0.062 for MI, 0.058 for CAD ("soft" definition) and 0.15 for CAD ("hard" definition). In results based on rs1800775, the direction of the interaction in men was inconsistent with the observed effects on biomarkers (increasing BMI conferred a stronger protective effect for all cardiovascular endpoints, Supplementary Figure B.12). The slope of the effect modification of CETP by BMI was inverted in women compared to men for cardiovascular endpoints.

B.1.2 Results from power analyses

Effect modification by sex

We conducted power analyses to assess our limit of detection based on the sample size and the prevalence of CAD in men and women in the UK Biobank dataset. For the association of one s.d. reduction in the CETP genetic score with CAD, we calculated that for an OR of 0.98 in men an OR of 0.95 or less in women was sufficient to reach 80% power to detect a significant interaction effect between the CETP genetic score and sex (Supplementary Figure B.4).

Effect modification by BMI

Using simulations, we estimated the smallest detectable effect modification by BMI of the association between genetically-predicted CETP levels and cardiovascular outcomes (Supplementary Figure B.9) at 80% power to be $\beta_{itx} = 0.015$. This represents an OR per s.d. decrease in the CETP score of 0.96 for the normal BMI range versus 0.99 for obese individuals. In our analyses, the product term between the CETP score and BMI was 0.010 (95% CI -0.00042, 0.020) and we estimate our power to detect a statistically significant effect of this magnitude at 48%.

B.2 Supplementary Methods

B.2.1 UK Biobank

To assess the effect of genetic CETP modulation on biomarkers and cardiovascular diseases by sex and BMI, we used the UK Biobank resource. The UK Biobank is a population-based longitudinal cohort of more than 500,000 individuals [106]. At recruitment, a touchscreen-based questionnaire followed-up by a verbal interview with a nurse was used to assess self-reported diseases and medications. Linkage with hospitalization and death registries also allows acute events such as cardiovascular events to be well captured. Anthropomorphic measurements were also taken and an extensive panel of blood and urine biomarkers captures many laboratory measurements including major lipoprotein fractions. Finally, genetic data based on a genotyping array and imputation is available on all participants.

Biochemical markers were measured from samples of urine, packed red blood cells and serum collected at baseline for all participants. For logistics reasons, the UK Biobank opted for centralized processing of samples in a high throughput facility. Participants were not required to fast before the collection of biological samples.

For the analysis of blood biomarkers including lipoproteins and cholesterol levels, we transformed the variables as needed to obtain an approximately normal distribution and then standardized the values (units are reported in the main text Table 3.1). Because we included different biomarkers in our analyses, standardization allows unified reporting of effects in units of standard deviation. For all continuous measurements, we used values from the baseline visit and we used the mean if multiple measurements were available.

To define type 2 diabetes, we opted to rely on the self-reported diseases from the verbal interview with a nurse. We extracted all individuals who self-reported “diabetes” (coded as 1220 in variable #20002) or its children variables (“gestational diabetes” coded 1221, “type 1 diabetes” coded 1222, “type 2 diabetes” coded 1223 or “diabetes insipidus” coded 1521). Then, we excluded from the analyses individuals reporting “type 1 diabetes” or reporting “gestational diabetes” or “diabetes insipidus” with no record of “type 2 diabetes”. The remaining individuals were assumed to have type 2 diabetes and individuals that did not report any of the preceding codes were used as controls.

For cardiovascular events, preliminary analyses revealed that self-reported data could lack precision and so we relied on ICD10 codes from hospitalization or death records. Revascularization procedures (percutaneous coronary interventions [PCI] and coronary artery bypass

graft [CABG]) were defined using OPCS procedure codes in the linkage with hospitalization data. The specific codes used to define the events (based on hospitalizations, deaths or procedures) are described in Supplementary Table B.2.

For the statin use status, we extracted all the self-reported medications (coded 20003). All drugs classified under the ATC code C10B were used to define statin users. We used the UK Biobank Self Reported Medication Data parsing and matching software (available https://github.com/PhilAppleby/ukbb-srmed/) to achieve this coding. This procedure resulted in the identification of 83,385 statin users based on 10 drugs. The specific codes used to define this composite are described in Supplementary Table B.4.

B.2.2 Genetic quality control

The genetic quality control steps used to avoid risk of confounding due to ethnicity, relatedness or incorrect genotyping has been previously described [255].

Briefly, we excluded variants or individuals with a missing rate above 2%. We compared the genetic and self-reported sex variables to validate sample matching between the genetic and phenotypic data. Individuals with discrepencies or aneuploidies were removed from the analysis dataset. To avoid bias due to population stratification, we only selected individuals from the majority population in the UK Biobank. We excluded individuals not self-reporting as of European descent or falling outside of a manually defined region based on the principal component analysis projections. Related individuals were also excluded from analysis using a kinship coefficient of 0.0884 as the cutoff (corresponding to a 3rd degree relationship). After genetic quality control, a total of 413,138 individuals remained.

B.2.3 Montreal Heart Institute Biobank

Participants to the Montreal Heart Institute (MHI) Biobank and hospital cohort were recruited from different MHI departments and its affiliated prevention centre (EPIC) between 2006 and 2016. Blood, DNA, and plasma are collected at baseline and stored at the Pharmacogenomics Centre at MHI. All MHI Biobank participants provided informed consent and the study was approved by the MHI scientific and ethics review committees.

Cholesterol efflux measurements

Basal and cAMP stimulated cholesterol efflux were measured using plasma from a subset of 5,215 participants of the MHI Biobank. The procedure for the efflux measurements was previously published elsewhere [57]. Briefly, plasma was obtained from venous blood

samples collected on potassium-EDTA coated tubes (BD Vacutainers), centrifuged as per the manufacturer's protocol and frozen at -80°C until analysis. Samples were then thawed at 4°C and the cholesterol efflux capacity was measured in vitro with J774 macrophages. Cells were grown for 24 hours in presence of tritiated (^3H) cholesterol (2 $\mu Ci/ml$). After an 18 hours equilibration period, patient plasma depleted of apoB-containing lipoproteins with PEG6000 or control serum was added in triplicate wells for 4 hours. The concentration measurement was selected from dose-response curves obtained from pooled human plasma to avoid saturation of the efflux signal and to allow detection of changes in efflux capacity in any direction. Aliquots of cell-free culture medium and J774 cell homogenates were used to measure ^3H cholesterol counts with a beta counter (Tricarb, Perkin-Elmer). Cholesterol efflux capacity was defined as the ratio of ^3H-cholesterol found in the medium to the total cholesterol label in each well. Sample batches were tested in parallel with the same pool of normolipidemic human serum to calculate the sample/control ratio. We rejected cholesterol efflux capacity measurements that were outside of the historical mean ± 2 standard deviations.

To avoid batch effects in the statistical analysis, we used the residuals of the multivariable regression of efflux measurements on a factor variable representing the day of analysis. The residuals were subsequently standardized and used as the dependant variable in our analyses.

B.2.4 Simulation-based power analyses

Simulation for an interaction with body mass index

We are interested in assessing whether BMI modifies the effect of genetically-predicted reduction in CETP concentration on coronary artery disease (CAD). To estimate statistical power to detect such an effect, we simulate interaction terms while fixing the overall disease prevalence, the allele frequencies and marginal effect of the CETP genetic score to observed values from the data. We repeat the simulation and estimate the fraction of simulated datasets where the null hypothesis of $\beta_{itx} = 0$ is rejected at $\alpha = 0.05$ using a conventional logistic regression. The simulation parameters are described below.

Parameter	**Description**	**Value**
n	Number of individuals	413,138
$P(Y=1)$	Prevalence of coronary artery disease in the overall population	10.8%
Y	Coronary artery disease (CAD), the outcome.	Simulated as described in the "Simulation" section
X	The CETP genetic score expressed so that 1 unit of X corresponds to a 1 standard deviation decrease in the CETP concentration score (negative of the standardized score of CETP concentration).	Simulated $\mathcal{N}(0,1)$
BMI	The body mass index	Simulated $\mathcal{N}(0,1)$
β_0	Intercept for the logistic regression model. Corresponds to the logistic of the prevalence when other covariables are 0.	See section "Simulation model"
β_x	Marginal effect of a 1 unit increase in X (CETP) on Y (coronary artery disease). In our analyses, a 1 unit increase in X represents a 1 standard deviation reduction in the genetic score of CETP concentration. This is the β_x as estimated in a model with no interaction terms.	-0.0255
β_b	Coefficient of BMI (direct effect on Y). This value is the observational effect of BMI on Y as estimated in the UK Biobank. We used the estimate from a multivariable logistic regression model including age, sex and principal components as covariates.	0.366
β_{itx}	The additive effect modification of β_x per unit increase in BMI (*i.e.* product interaction term between BMI and X).	Simulated values between -0.03 and 0.03

The association model is the following logistic regression model:

$$\ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_x \cdot X + \beta_b \cdot \text{BMI} + \beta_{itx} \cdot X \cdot \text{BMI} \tag{B.1}$$

Where the parameters are as previously defined.

Simulation When simulating the outcome (CAD), we want to control for the prevalence ($P(CAD = 1)$) and obtain the desired regression coefficients. To achieve this, we use a latent variable model of the logistic regression.

$$Y^* = \beta_0 + \beta_x \cdot X + \beta_b \cdot \text{BMI} + \beta_{itx} \cdot X \cdot \text{BMI} + \epsilon \tag{B.2}$$

With

$$\begin{aligned} \epsilon &\sim \mathcal{L}(0, 1) \\ X &\sim \mathcal{N}(0, 1) \\ \text{BMI} &\sim \mathcal{N}(0, 1) \end{aligned}$$

and the coefficients set as previously described except for β_0 which we will discuss later. The resulting continuous latent variable (Y^*) is used to simulate the outcome as follows:

$$Y = \begin{cases} 1 & \text{if } Y^* > 0 \\ 0 & \text{otherwise} \end{cases}$$

We then use β_0 to control the prevalence (*i.e.* $P(Y = 1)$) as Y^* is positive if and only if

$$-\beta_0 < \beta_x \cdot X + \beta_b \cdot \text{BMI} + \beta_{itx} \cdot X \cdot \text{BMI} + \epsilon$$

Using the normal approximation for the right hand side of the former, we have:

$$
\begin{aligned}
&Z = \beta_x \cdot X + \beta_b \cdot \text{BMI} + \beta_{itx} \cdot X \cdot \text{BMI} + \epsilon \\
&Z \sim \mathcal{N}\left(0, Var\left(\beta_x X + \beta_b \text{BMI} + \beta_{itx} \cdot X \cdot \text{BMI} + \epsilon\right)\right) && \text{Normal approx. and } X \text{ standard normal} \\
&Z \sim \mathcal{N}\left(0, \beta_x^2 + \beta_b^2 + Var(\beta_{itx} \cdot X \cdot \text{BMI}) + \frac{\pi^2}{3})\right) && \text{Variance of standard normal and logistic} \\
&Z \sim \mathcal{N}\left(0, \beta_x^2 + \beta_b^2 + \beta_{itx}^2 + \frac{\pi^2}{3}\right) && \text{Assuming X and BMI independant}
\end{aligned}
$$

And we can use the normal quantile function to determine the value of β_0 resulting in the observed prevalence.

$$
\begin{aligned}
P(Y = 1) = p &= P(-\beta_0 < Z) \\
&= 1 - \Phi_Z(-\beta_0) \iff \\
\beta_0 &= -\Phi_Z^{-1}(1 - p)
\end{aligned}
$$

Implementation (R) We implemented the simulation in R. Power computations are done by simulating an outcome, fitting a logistic regression model and empirically estimating the power as the fraction of all simulation iterations where the null was rejected at $\alpha = 0.05$.

```r
# Fixed parameters
n <- 413138
prevalence <- 44713 / n

# Estimated from CAD ~ bmi + age + sex + PCs model
b_bmi <- 0.3659062

# Marginal effect of CETP at mean BMI
b_cetp <- -0.025514967819954

bmi <- rnorm(n)
cetp <- rnorm(n)

b0 <- -qnorm(
    1 - prevalence,
    mean = 0,
    sd = sqrt(b_cetp^2 + b_bmi^2 + b_itx^2 + pi^2 / 3)
```

```
)

y_star <- b0 + b_cetp * cetp + b_bmi * bmi + b_itx * cetp * bmi + rlogis(n)
y <- as.numeric(y_star > 0)
```

Power estimation We estimated power by repeatadly simulating datasets with fixed parameters and estimating the proportion of all simulations where the null hypothesis is rejected using the conventional hypothesis testing framework. Specifically, we test the hypothesis that the interaction coefficient is different from zero at an $\alpha = 0.05$ threshold and using the Wald statistic.

The standard error of the estimated power is $\sqrt{\hat{p}(1-\hat{p})/n}$ where $\hat{p}$ is the proportion of simulated datasets where the null is rejected and n is the number of simulated datasets. When plotting power estimates, the standard error is used to construct the 95% confidence interval.

Simulation for an interaction with sex

The simulation model for the effect modification by sex was a little bit different. Because the levels in the simulation were binary, we simulated men and women separately and concatenated the results. The proportion of men and women from our dataset and the prevalence of CAD were fixed at observed values. The sex-specific effect of CETP was then used as the simulation parameter using the same strategy as before to control for the prevalence of CAD (the latent logistic regression model).

B.2.5 Construction of CETP activity scores

To construct a weighted allele score of CETP activity, we considered various approaches. We constructed a total of 5 scores based on various summary statistics from previous GWAS of plasma CETP concentration [50] or lipid phenotypes measured by nuclear magnetic resonance by the MAGNETIC consortium [193]. In the end, we only selected one representative score for the analyses to avoid redundancy as all the scores we generated were strongly correlated (Supplementary Table B.1). All scores were constructed using the LD clumping and p-value thresholding approach based on variants at the CETP locus, except from the score based on the independently associated variants as presented in the original publication of the CETP concentration GWAS. To determine an appropriate p-value threshold for the genetic scores, we estimated the effective number of independent pairs of lipid and variant associations akin to the simpleM correction [256].

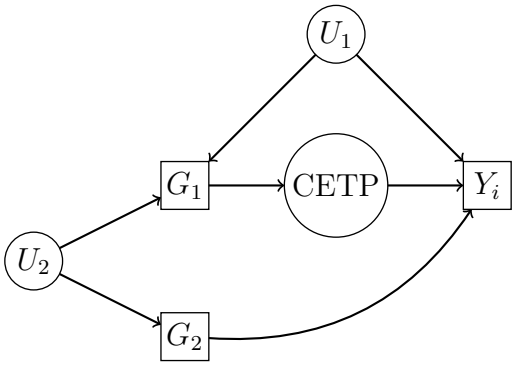
Specifically, we estimated the number of independent genetic variants at the CETP locus using a Principal Component Analysis (PCA) of the genotype matrix. Including 50 principal components explained 99% of the variance in CETP genotypes and was used to estimate the effective number of tested variants. Similarly, to estimate the effective number of lipid phenotypes to account for, we used hierarchical clustering of the correlation between variant association coefficients in the MAGNETIC GWAS. We used this dataset to estimate the number of independent phenotypes because it includes an exhaustive number of lipid traits. Manual inspection of the hierarchical clustering dendrogram revealed that there were three major classes of lipids independently affected by CETP. We picked HDL cholesteryl ester, small VLDL triglycerides and large VLDL triglycerides as cluster representatives. Association coefficients for these lipids were thus used when generating candidate scores of CETP activity. Based on the estimates of the effective number of variants and lipids, the p-value threshold was set to $0.05/(3 \times 50) = 3.3 \times 10^{-4}$ for all scores. We fixed the LD threshold at $r^2 = 0.1$ as to not exclude variants correlated by chance and a minimum allele frequency of 1% was used.

B.2.6 Causal interpretation

Genetic variants have an important property that makes them a good tool for causal inference: they have very few causes. Because they are determined at birth and fixed throughout life, they are not susceptible to reverse causation or environmental confounders, two major sources of bias in observational studies. Mendelian randomization (MR) leverages this advantage of genetic variants to draw causal inferences of the effect of heritable exposures [122, 128].

The current study is similar in spirit to an MR study and we will formalize how the presented associations can be causally interpreted based on the same assumptions as any instrumental variable study. We represented the causal structure of the current experiment as a causal directed acyclic graph (DAG) below: [257]

In this DAG, we can see features that are common in instrumental variable studies such as the classical MR setup. The G_1 node represents a genetic predictor of the unobserved CETP activity. In our study, this corresponds to either the CETP concentration score or the CETP -629C>A (rs1800775) variant. The unobserved CETP node represents all aspects of a genetic disruption of CETP in terms of plasma concentration, isoform prevalence, enzymatic activity and substrate preference. In our study this is an unobserved variable that encompasses different aspects of CETP function. The Y_i node represents the outcomes of interest. In our study, it is used to represent alternatively C-reactive protein, lipid fractions, lipoprotein



Supplementary Figure B.1 – Directed acyclic graph representing the causal structure of the experiment. Squares are used to denote observed variables and circles are used to denote unobserved variables. Arrows represent causal effects such that changes in a parent variable will result in changes in its child.

levels, plasma cholesterol efflux capacity or cardiovascular outcomes, captured by the index on the Y_i variable.

Two latent variables (U_1 and U_2) corresponding to potential confounders are also included. The first variable (U_1) is used to represent an effect of population stratification. Specifically, the distribution of G_1 is expected to vary between populations simply due to differences in population histories. The distribution of Y_i may also vary across populations because of lifestyle or environmental differences. For example a diet common in one area of the world may influence risk of heart attack in a way that is independent from genetics. If unaccounted for, this common cause of G_1 and Y_i (ancestry) will bias the estimate of the effect of G_1 on Y_i. In our study, we ensured that this unobserved variable was controlled for by only using the largest genetically homogeneous subset of the UK Biobank consisting of individuals of European descent. Additionally, we included principal components capturing population structure to our regression models, which is the conventional way to account for population stratification in genetic association studies.

The second possible confounder is depicted on the graph by the U_2 node. This node represents unobserved linkage disequilibrium (LD) inducing a correlation between the G_1 variable and a second locus (G_2) with a direct effect on Y_i. As is conventional in MR, we will assume no direct effect of G_1 on Y which includes the effect of the $G_1 \rightarrow U_2 \rightarrow G_2 \rightarrow Y$ path. This assumption also justifies that no arrow from G_1 to Y was included in the DAG. In instrumental variable analysis and MR this assumption is called the exclusion restriction. As in any MR study, the causal interpretation of our results requires this unverifiable, but plausible, assumption to hold. Another reason to include a direct effect of G_1 on Y is horizontal pleiotropy, a term used to describe a genetic variant with simultaneous effects

on distinct pathways. Because the score only includes CETP locus variants associated with plasma concentration, and the rs1800775 "C" allele is known to repress CETP promoter activity and Sp3 transcription factor binding, it is unlikely that the observed effects are through another pathway. In general, studies based on well known variants at a single locus are less likely to suffer from bias due to pleiotropy.

To summarize, we make the following assumptions:

1. **There are no confounders of the $G_1 - Y_i$ relationship (independance)**. U_1 is accounted for by using a genetically homogeneous subset of the UK Biobank and including principal components in our regression models.
2. $G_1 \rightarrow$ **CETP is non-null (relevance)**. This is given by external data validating the effect of rs1800775 on CETP levels and by construction for the score of CETP concentration. We also argue that the effect observed effects on HDL-c levels, a well-known consequence of CETP disruption, supports this assumption.
3. G_1 **only affects** Y_1 **through CETP (exclusion restriction)**. More formally, $G_1 \perp\!\!\!\perp Y_1 \mid \text{CETP}$. We minimized the risk of violating this assumption by only relying on genetic variables known to affect CETP function and located at the CETP locus.

These assumptions are the same as for instrumental variable or MR analyses and allow us to interpret the multivariable regression coefficients described in the main text as the total effect of G_1 on Y_i. Because this total effect is only mediated by CETP, under our assumptions, it represents the average causal effect of genetic CETP disruption.

In our manuscript, the focus is on effect modifiers of the genetically predicted reduction in CETP on biomarkers and cardiovascular outcomes. The question of effect modification in causal effects is interesting and currently scarcely discussed in the literature, especially in the parametric and multivariable context. Stratification has been suggested as the natural way to identify effect modification and so we reported marginal subgroup effects for all of the considered effect modifiers in the main text [258].

Because the CETP node is unobserved, we are not able to distinguish effect modification in the $G_1 \rightarrow CETP$ from the $CETP \rightarrow Y_i$ effects. This distinction is important and represents a limitation of our study if generalization to pharmacological CETP inhibition is sought. Concretely, if the effect modification only influences the effect of the genetic variant or genetic score on CETP activity, then the observed effects are strictly genetic and do not represent a broadly applicable feature of CETP inhibition.

B.2.7 Meta-analysis of randomized controlled trials of CETP inhibitors

We conducted a sex-stratified fixed-effects meta-analysis of three RCTs of CETP inhibitors. In the REVEAL trial, rate ratios (RRs) were provided for all analyses. For ACCELERATE, the hazard ratio was provided for the main study, but RRs were reported in subgroup analyses. We calculated the RR for the main study and it was numerically identical to the hazard ratio at two decimal places. Similarly, for dal-OUTCOMES, the RR and hazard ratio were numerically identical in the main study. In sex-stratified analyses only the hazard ratios were available for this study and we could not calculate the RRs because the sex-stratified event counts were not provided. We used the hazard ratio in place of the RR for these effects. The effect estimates are summarized below along with the relevant references:

Study (drug)	**Group (n; %)**	**Rate ratio (95% CI)**	**Ref.**
dal-OUTCOMES (dalcetrapib)	All (n=15,871)	1.04 (0.93, 1.16)	[196]
	Male (n=12,801; 81%)	Hazard ratio: 1.07 (0.95, 1.21)	
	Female (n=3,070; 19%)	Hazard ratio: 0.92 (0.72, 1.16)	
ACCELERATE (evacetrapib)	All (n=12,092)	1.01 (0.91, 1.11)	[259]
	Male (n=9,308; 77%)	1.04 (0.93, 1.17)	
	Female (n=2,784; 23%)	0.91 (0.74, 1.12)	
REVEAL (anacetrapib)	All (n=30,449)	0.91 (0.85, 0.97)	[35]
	Male (n=25,534; 84%)	0.90 (0.84, 0.97)	
	Female (n=4,915; 16%)	0.93 (0.78, 1.11)	

For the dal-OUTCOMES (main trial) and ACCELERATE (main trial and all subgroups), we estimated the standard error of the rate ratio on the natural log scale as $\sqrt{y_0^{-1} + y_1^{-1}}$ where y_0 and y_1 denote the number of events in the treatment arm and in the placebo arm, respectively. The 95% confidence intervals were then calculated on the natural log scale and are reported on the rate ratio scale.

The variances for the meta-analysis were calculated from the 95% confidence interval for all studies as:

$$V_i = \left(\frac{\ln(\mathrm{UCL}_i) - \ln(\mathrm{LCL}_i)}{2 \times \Phi^{-1}(0.05/2)} \right)^2$$

Where i indexes studies, UCL is the upper confidence interval limit, LCL is the lower confidence interval limit and Φ^{-1} is the normal quantile function.

The meta-analysis weights were the inverse of the variances ($W_i = V_i^{-1}$). The meta analysis effect on the log scale is then calculated as:

$$\beta = \frac{\sum_i \ln(\mathrm{RR}_i) W_i}{\sum_i W_i}$$
$$Var(\beta) = \left(\sum_i W_i \right)^{-1}$$

For every group or subgroup.

Now, we denote the sex-specific meta-analysis estimates as β_M for the male-only estimate and β_F as the female-only estimate. We can conduct hypothesis testing of the null hypothesis of no effect using the following Wald statistics (following a 1 d.f. χ^2 under the null):

$$\chi^2_C = \beta^2 / Var(\beta)$$
$$\chi^2_M = \beta^2_M / Var(\beta_M)$$
$$\chi^2_F = \beta^2_F / Var(\beta_F)$$

It is also possible to test for heterogeneity in the sex-specific effects using the $\chi^2_H = \chi^2_M + \chi^2_F - \chi^2_C$ statistic having a χ^2 distribution with 1 degree of freedom as described in Magi *et al.* (2010) [260]. This test is equivalent to a sex-interaction test.

B.2.8 Note on interaction scales

Interpretation of the product interaction term in linear regression

In a linear regression model, the product interaction term represents an interaction contrast.

For example, assume the interaction model with one continous outcome Y and two interacting variables X_1 and X_2.

$$\mathbb{E}(Y|X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{itx} X_1 X_2$$

We define the interaction contrast as the effect difference when both $X_1 = X_2 = 1$ compared to the sum of their individual effects ($X_1 = 1$ or $X_2 = 1$). All effects are taken using the absence of both covariables as the reference.

Concretely, the interaction contrast (IC) is:

$$\begin{aligned} IC &= Y_{11} - Y_{00} - [(Y_{10} - Y_{00}) + (Y_{01} - Y_{00})] \\ &= Y_{11} - Y_{10} - Y_{01} + Y_{00} \\ &= (\beta_0 + \beta_1 + \beta_2 + \beta_{itx}) - (\beta_0 + \beta_1) - (\beta_0 + \beta_2) + \beta_0 \\ &= \beta_{itx} \end{aligned}$$

Denoting $Y_{ij} = \mathbb{E}(Y|X_1 = i, X_2 = j)$.

Hence, the product interaction coefficient in a linear regression is the added effect attributable to the co-occurence of both risk factors in addition to the sum of their individual effects.

Interpretation of the product interaction term in logistic regression

Using the same interaction of two covariables (X_1 and X_2) as in the previous section, we express the following logistic regression model on a binary outcome variable (Y).

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{itx} X_1 X_2$$

Where $p = P(Y = 1|X_1, X_2)$.

In this model, the individual effects of X_1 and X_2 expressed as odds ratios are represented by $e^{\beta_1} = \mathrm{OR}_{10}$ and $e^{\beta_2} = \mathrm{OR}_{01}$, respectively. Their combined effect when compared to no effect of X_1 and X_2 can be derived as follows:

$$
\begin{aligned}
\ln\left(\mathrm{OR}_{11}\right) &= \ln\left(\frac{\mathrm{odds}(Y|X_1=1,X_2=1)}{\mathrm{odds}(Y|X_1=0,X_2=0)}\right) \\
&= \ln(\mathrm{odds}(Y|X_1=1,X_2=1) - \ln(\mathrm{odds}(Y|X_1=0,X_2=0)) \\
&= (\beta_0+\beta_1+\beta_2+\beta_{itx}) - \beta_0 \\
&= \beta_1+\beta_2+\beta_{itx} \\
&\rightarrow OR_{11} = e^{\beta_1+\beta_2+\beta_{itx}} = e^{\beta_1}\cdot e^{\beta_2}\cdot e^{\beta_{itx}} = \mathrm{OR}_{10}\cdot\mathrm{OR}_{01}\cdot e^{\beta_{itx}}
\end{aligned}
$$

This reveals how the interaction coefficient (β_{itx}) represents a multiplicative change from the combined individual effects of the covariables on the odds ratio scale [261]:

$$
e^{\beta_{itx}} = \frac{\mathrm{OR}_{11}}{\mathrm{OR}_{10}\ \mathrm{OR}_{01}}
$$

Estimating additive interactions from a logistic regression model

It is often argued that additive interactions are most relevant to the study of disease etiology or public health [262]. For this reason, we calculated additive interactions from the logistic regression results. We used two complementary strategies. First, we calculated the Relative Excess Risk due to Interaction (RERI) on the odds ratio scale as defined in [261]:

$$
\begin{aligned}
\mathrm{RERI} &= (\mathrm{RR}_{11}-1) - (\mathrm{RR}_{10}-1) - (\mathrm{RR}_{01}-1) \\
&= \mathrm{RR}_{11} - \mathrm{RR}_{10} - \mathrm{RR}_{01} + 1 \\
&\approx e^{\hat{\beta}_1+\hat{\beta}_2+\hat{\beta}_{itx}} - e^{\hat{\beta}_1} - e^{\hat{\beta}_2} + 1
\end{aligned}
$$

Which holds if the odds ratio approximates the relative risk as is the case with rare outcomes. This statistic can be seen as the additive deviation on the relative risk scale due to the interaction [261]. In our analyses, the RERI were computed from the logistic regression fit using the "interactionR" R package and using the "mover" method to compute the confidence interval [263].

Because this measure is a deviation on the relative risk scale, we also considered interaction contrasts on the probability scale which may be more interpretable. For this analysis, we

computed marginal predicted probabilities from the logistic regression fit by using a weighted average across observed covariable levels while fixing the interaction variables. We then used the interaction contrast (IC) as defined in Section B.2.8:

$$IC = P_{11} - P_{10} - P_{01} + P_{00}$$

With $P_{ij} = P(Y = 1|X_1 = i, X_2 = j)$

We used the "boot" R package to construct 95% confidence intervals for this statistic using the percentile method and 2,000 bootstrap replicates.

B.3 Supplementary Tables and Figures

B.3.1 Supplementary Tables

Supplementary Table B.1 – **Correlation coefficient between the different CETP genetic scores or the rs1800775 variant.** The upper triangular matrix shows the Pearson correlation coefficient and the lower triangular shows the squared coefficients. All scores were calculated using the p-value thresholding and LD clumping approach, except for the "Blauw *et al.* condi. indep." score which is based on the conditionally independent variants reported in the original study.

	MAGNETIC HDL diameter	MAGNETIC L-HDL-CE	MAGNETIC S-VLDL-TG	Blauw *et al.* GRS	Blauw *et al.* condi. Indep.	SNP rs1800775
MAGNETIC HDL diameter	1	1.00	-1.00	-0.87	-0.94	0.72
MAGNETIC L-HDL-CE	1.00	1	-1.00	-0.87	-0.94	0.72
MAGNETIC S-VLDL-TG	0.99	0.99	1	0.87	0.93	-0.71
Blauw *et al.* GRS	0.76	0.76	0.76	1	0.87	-0.71
Blauw et al. condi. indep.	0.89	0.89	0.86	0.75	1	-0.74
SNP rs1800775	0.51	0.51	0.51	0.51	0.55	1

Supplementary Table B.2 – **Codes used to define cardiovascular endpoint events in the UK Biobank.** Events from either the hospitalization or death records were used and the procedure codes (OPCS codes) are from the hospitalization records.

Variable	**Definition**
Myocardial infarction (MI)	ICD10 codes: I21, I22, I23, I25.2 or ICD9 codes: 410, 412, 411.0, 429.79
Percutaneous coronary intervention or coronary artery bypass graft (PCI/CABG)	OPCS codes: K40, K41, K42, K43, K44, K45, K46, K49, K50, K75
Coronary artery disease (CAD) – soft	PCI/CABG or ICD9 codes 410-414 (except for aneurysms: 414.1) or ICD10 codes I20-I25
Coronary artery disease (CAD) – hard	Combination of MI, PCI/CABG or hospitalization or death for unstable angina coded a I20.0 as the primary cause of death or hospitalization. For this variable, cases for the "soft" CAD definition that would be controls for the "hard" CAD definition are set as missing so that they are excluded from the analyses.

Supplementary Table B.3 – **Effect per alternative allele of rs1800775 (CETP -629C>A) on biomarkers and cardiovascular events.** Coefficients for continuous variables are from a linear regression model adjusted for age, sex and the first 10 principal components and are expressed in standard deviation of the outcome per "A" allele of the rs1800775 SNP. Odds ratios for cardiovascular outcomes are estimated using a logistic regression model adjusted for the same covariates.

Outcome	n **or** n **cases**	**Coefficient or odds ratio (95% CI)**	**p-value**
Biomarkers			
lipoprotein(a)	315,214	-0.0143 (-0.0182, -0.0106)	1.4×10^{-13}
C-reactive protein	394,165	0.0041 (-0.0003, 0.0084)	0.070
HDL cholesterol	362,468	0.198 (0.194, 0.203)	$< 10^{-300}$
Apolipoprotein A	360,451	0.156 (0.151, 0.160)	$< 10^{-300}$
LDL cholesterol	394,287	-0.028 (-0.033, -0.024)	3.2×10^{-36}
Apolipoprotein B	393,089	-0.041 (-0.045, -0.037)	1.4×10^{-73}
Cardiovascular outcomes			
Coronary artery disease ("soft")	44,713	0.973 (0.959, 0.987)	0.00015
Coronary artery disease ("hard")	26,342	0.968 (0.950, 0.986)	0.00047
Myocardial infarction	18,559	0.982 (0.962, 1.000)	0.10
Percutaneous coronary intervention or coronary artery bypass graft	16,941	0.970 (0.948, 0.992)	0.0068

Supplementary Table B.4 – **Drug codings to define statin users in the UK Biobank.**

UK biobank Data-Coding 4	**Drug**
1140861958	simvastatin
1141146234	atorvastatin
1141146138	lipitor 10mg tablet
1141192410	rosuvastatin
1141192736	ezetimibe
1140888648	pravastatin
1141188146	simvador 10mg tablet
1141192414	crestor 10mg tablet
1140881748	zocor 10mg tablet
1141200040	zocor heart-pro 10mg tablet

Supplementary Table B.5 – **Interaction measure between the CETP genetic score and sex on the additive scale (RERI and interaction contrast) based on the logistic regression model in the UK Biobank.**

Outcome	Odds ratio RERI (95% CI)	Interaction contrast (bootstrap 95% CI)
Coronary artery disease ("soft")	0.031 (-0.006, 0.070)	0.0011 (-0.00074, 0.0029)
Coronary artery disease ("hard")	0.082 (0.020, 0.146)	0.0017 (0.00024, 0.0032)
Myocardial infarction	0.050 (-0.018, 0.121)	0.00078 (-0.00042, 0.0019)
Percutaneous coronary intervention or coronary artery bypass graft	0.091 (-0.003, 0.188)	0.00099 (-0.00012, 0.0021)

Supplementary Table B.6 – **ANOVA results for the nonlinear interaction models of the CETP genetic score and BMI on biomarkers and cardiovascular outcomes.** Linear regression was used for biomarkers and logistic regression was used for cardiovascular outcomes. The presented association statistics are F statistics for the former and χ^2 for the latter. Both the CETP score and BMI are modeled using restricted cubic splines with 4 knots.

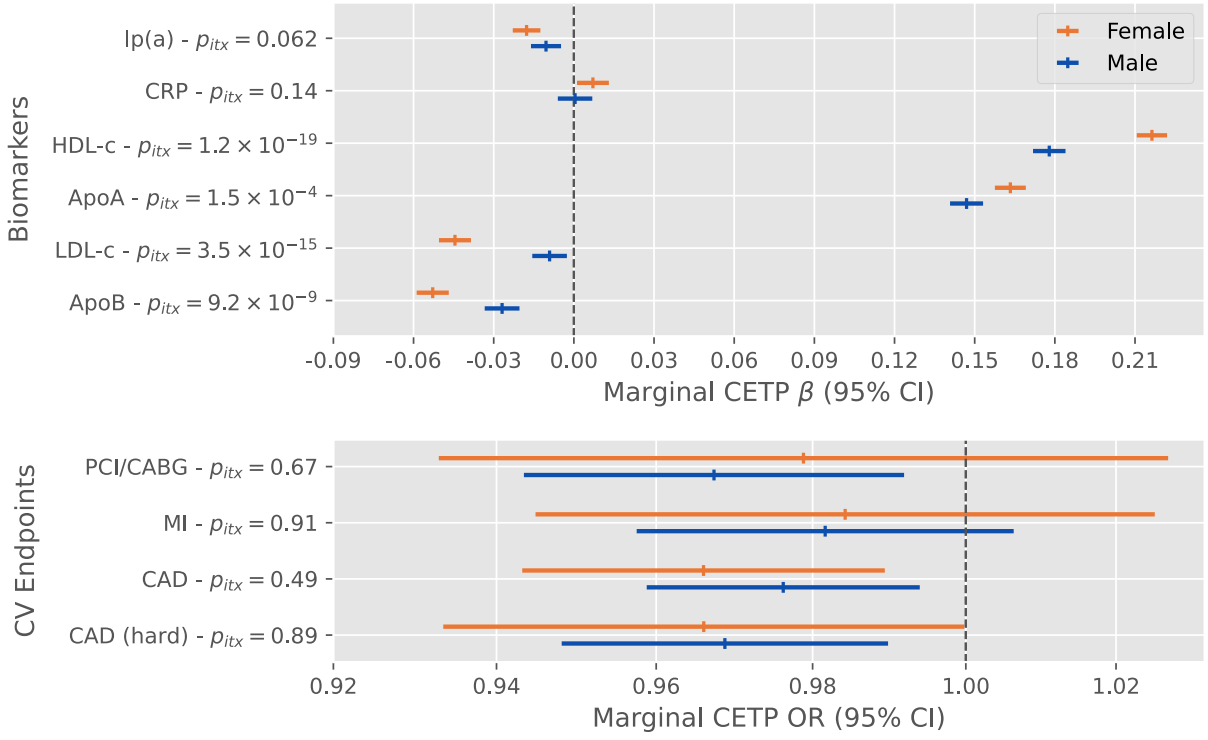
Outcome	Factor	Statistic (F or χ^2)	d.f.	p-value
Biomarkers (modeled using linear regression)				
Lipoprotein(a)*	All interactions (BMI and CETP)	2.14	9	0.0233
	BMI	15.55	12	< 0.0001
	Nonlinear effects	22.09	8	< 0.0001
	CETP genetic score	6.91	12	< 0.0001
	Nonlinear effects	0.94	8	0.4793
C-reactive protein	All interactions (BMI and CETP)	0.89	9	0.5299
	BMI	7662.84	12	< 0.0001
	Nonlinear effects	19.81	8	< 0.0001
	CETP genetic score	1.47	12	0.1265
	Nonlinear effects	0.66	8	0.7247
HDL cholesterol	All interactions (BMI and CETP)	37.41	9	< 0.0001
	BMI	4975.76	12	< 0.0001
	Nonlinear effects	189.61	8	< 0.0001
	CETP genetic score	1242.72	12	< 0.0001
	Nonlinear effects	8.51	8	< 0.0001
Apolipoprotein A	All interactions (BMI and CETP)	11.72	9	< 0.0001
	BMI	2632.38	12	< 0.0001
	Nonlinear effects	77.75	8	< 0.0001
	CETP genetic score	683.50	12	< 0.0001
	Nonlinear effects	2.05	8	0.0365
LDL cholesterol	All interactions (BMI and CETP)	3.68	9	0.0001
	BMI	521.46	12	< 0.0001
	Nonlinear effects	704.12	8	< 0.0001
	CETP genetic score	20.57	12	< 0.0001
	Nonlinear effects	3.17	8	0.0013
Apolipoprotein B	All interactions (BMI and CETP)	2.29	9	0.0147
	BMI	733.80	12	< 0.0001
	Nonlinear effects	638.04	8	< 0.0001

Outcome	Factor	Statistic (F or χ^2)	d.f.	p-value
	CETP genetic score	38.91	12	< 0.0001
	Nonlinear effects	2.70	8	0.0058
Basal cholesterol efflux (MHI Biobank)	All interactions (BMI and CETP)	0.66	9	0.7473
	BMI	17.55	12	< 0.0001
	Nonlinear effects	1.91	8	0.0537
	CETP genetic score	5.84	12	< 0.0001
	Nonlinear effects	0.64	8	0.7437
cAMP stimulated cholesterol efflux (MHI Biobank)	All interactions (BMI and CETP)	0.49	9	0.8815
	BMI	4.27	12	< 0.0001
	Nonlinear effects	0.50	8	0.8537
	CETP genetic score	3.67	12	< 0.0001
	Nonlinear effects	0.64	8	0.7439
Cardiovascular outcomes (modeled using logistic regression)				
Coronary artery disease ("soft")	All interactions (BMI and CETP)	8.10	9	0.5244
	BMI	4892.66	12	< 0.0001
	Nonlinear effects	128.51	8	< 0.0001
	CETP genetic score	30.67	12	0.0022
	Nonlinear effects	4.84	8	0.7742
Coronary artery disease ("hard")	All interactions (BMI and CETP)	6.16	9	0.72
	BMI	2798.71	12	< 0.0001
	Nonlinear effects	68.01	8	< 0.0001
	CETP genetic score	25.36	12	0.0132
	Nonlinear effects	3.19	8	0.9219
Myocardial infarction	All interactions (BMI and CETP)	8.64	9	0.4716
	BMI	1857.02	12	< 0.0001
	Nonlinear effects	69.72	8	< 0.0001
	CETP genetic score	15.49	12	0.2157
	Nonlinear effects	5.71	8	0.6798
Percutaneous coronary intervention or coronary artery bypass graft	All interactions (BMI and CETP)	5.49	9	0.7893
	BMI	1254.46	12	< 0.0001
	Nonlinear effects	122.46	8	< 0.0001
	CETP genetic score	22.25	12	0.0349
	Nonlinear effects	3.37	8	0.9091

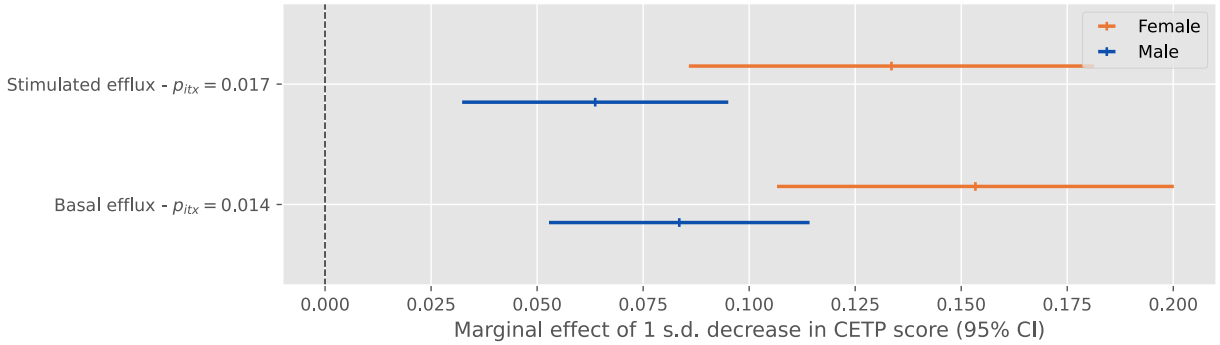
Supplementary Table B.7 – **Interaction measure between the CETP genetic score and body mass index on the additive scale (RERI and interaction contrast) based on the logistic regression model in the UK Biobank.**

Outcome	Odds ratio RERI (95% CI)	Interaction contrast (bootstrap 95% CI)
Coronary artery disease ("soft")	-0.002 (-0.017, 0.012)	-0.00030 (-0.0013, 0.00076)
Coronary artery disease ("hard")	-0.001 (-0.020, 0.018)	-9.5×10^{-5} (-0.00090, 0.00068)
Myocardial infarction	-0.005 (-0.025, 0.016)	-0.00014 (-0.00073, 0.00043)
Percutaneous coronary intervention or coronary artery bypass graft	0.00039 (-0.021, 0.022)	-5.8×10^{-6} (-0.00051, 0.00051)

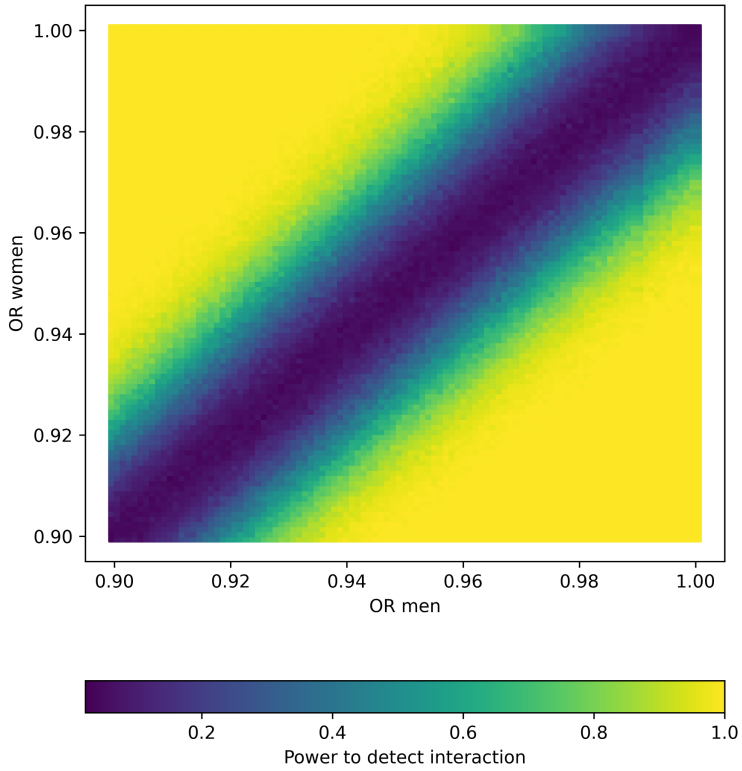
B.3.2 Supplementary Figures



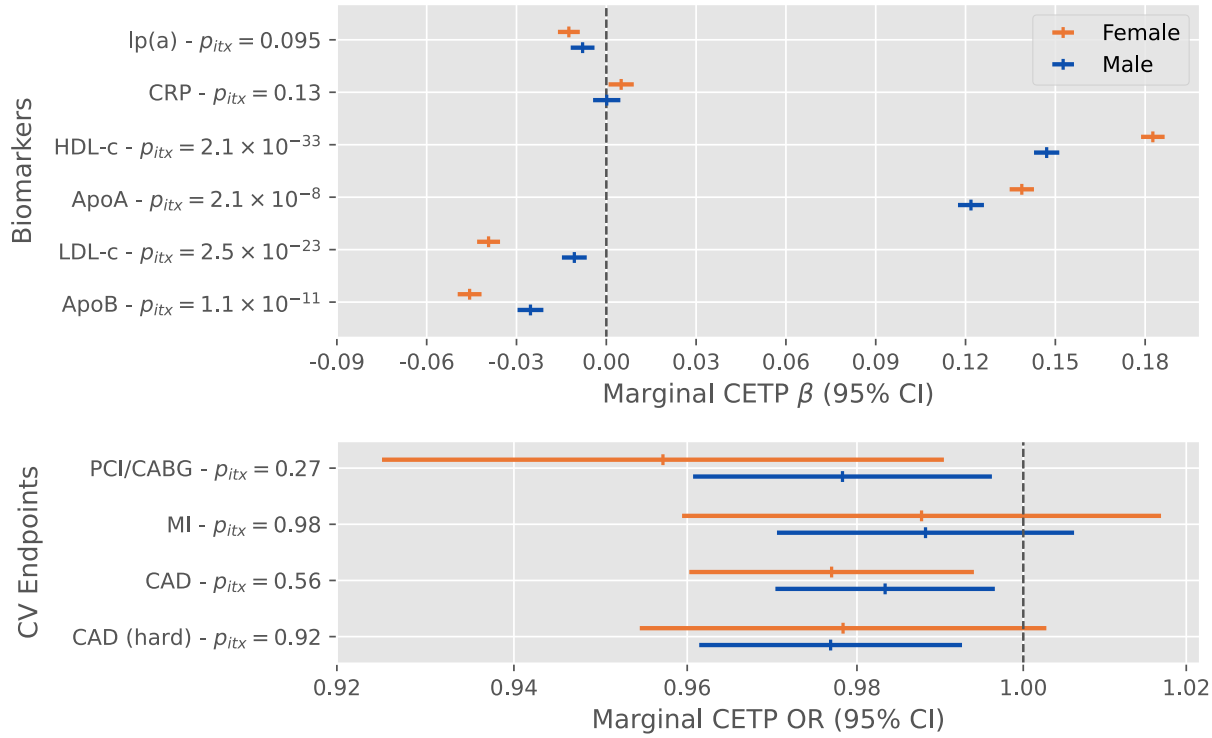
Supplementary Figure B.2 – **Effect modification of rs1800775 (CETP -629C>A) alternative alleles by sex on biomarkers and cardiovascular endpoints in the UK Biobank.** Displayed p-values (p_{itx}) are for the two sided test of the product term between the SNP and a binary sex indicator variable.



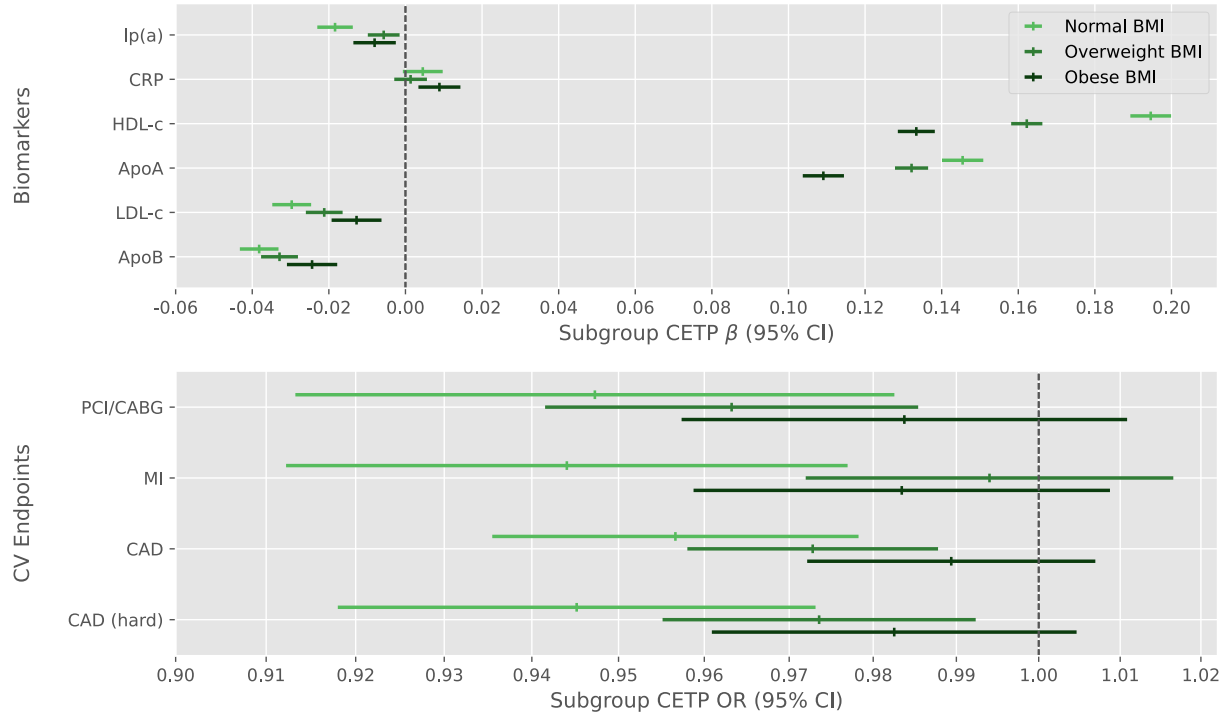
Supplementary Figure B.3 – **Effect modification of a 1 standard deviation decrease in the CETP concentration genetic score by sex on cholesterol efflux in the MHI Biobank.** Displayed p-values (p_{itx}) are for the two sided test of the product term between the CETP score and a binary sex indicator variable.



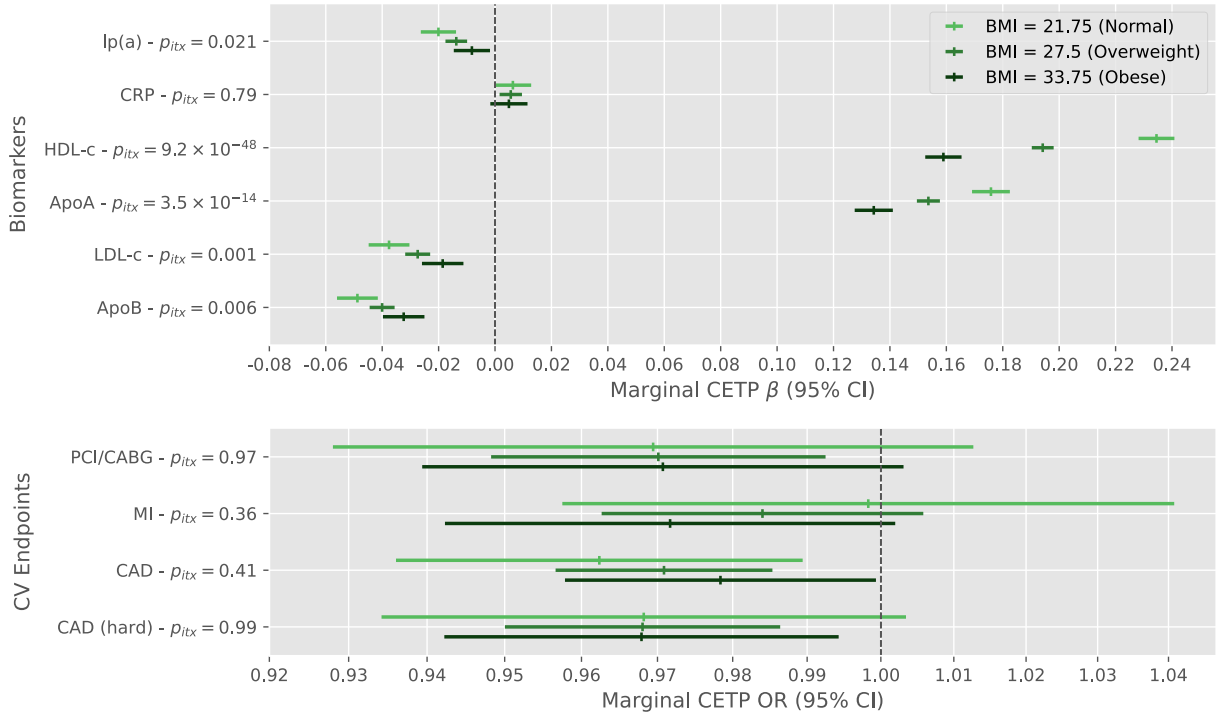
Supplementary Figure B.4 – **Power to detect a difference between men and women in the association between a standardized genetic score and coronary artery disease.** The results of this simulation are for 500 simulated datasets for different combinations of OR in men and women and given the sample size ($n = 413,138$) and prevalence of coronary artery disease ("soft" definition, 10.8%) in our dataset.



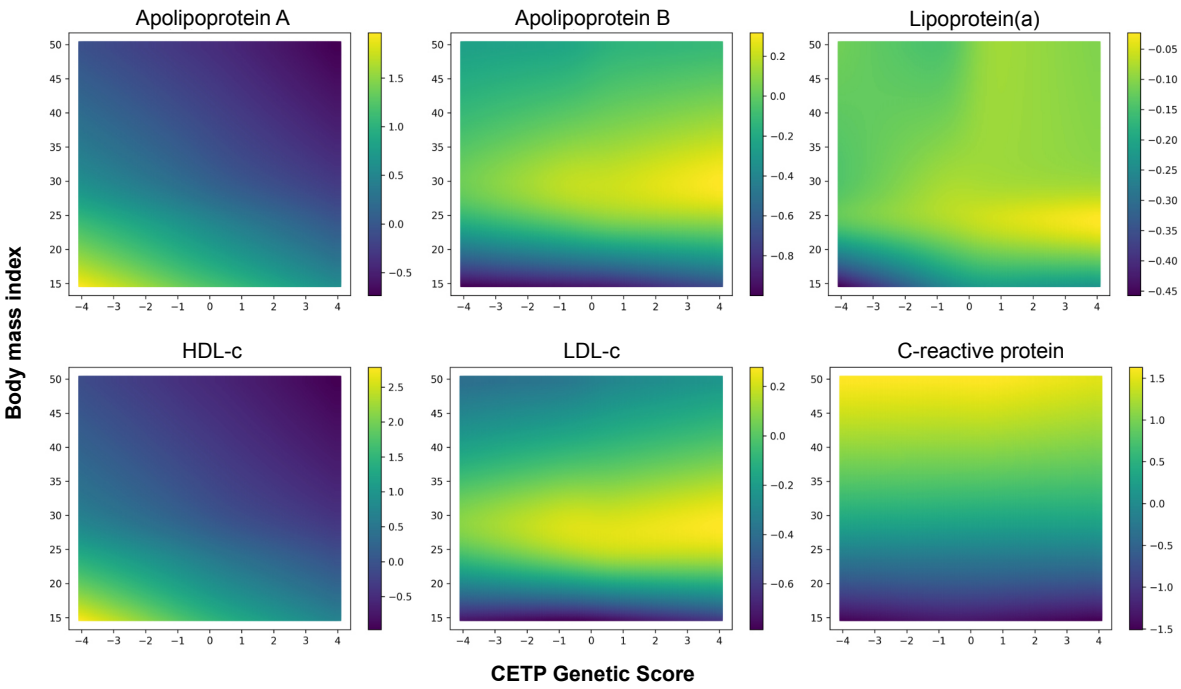
Supplementary Figure B.5 – **Effect modification of a 1 standard deviation decrease in the CETP concentration genetic score by sex on biomarkers and cardiovascular endpoints adjusting for self-reported statin use in the UK Biobank.** Displayed p-values (p_{itx}) are for the two sided test of the product term between the CETP score and a binary sex indicator variable.



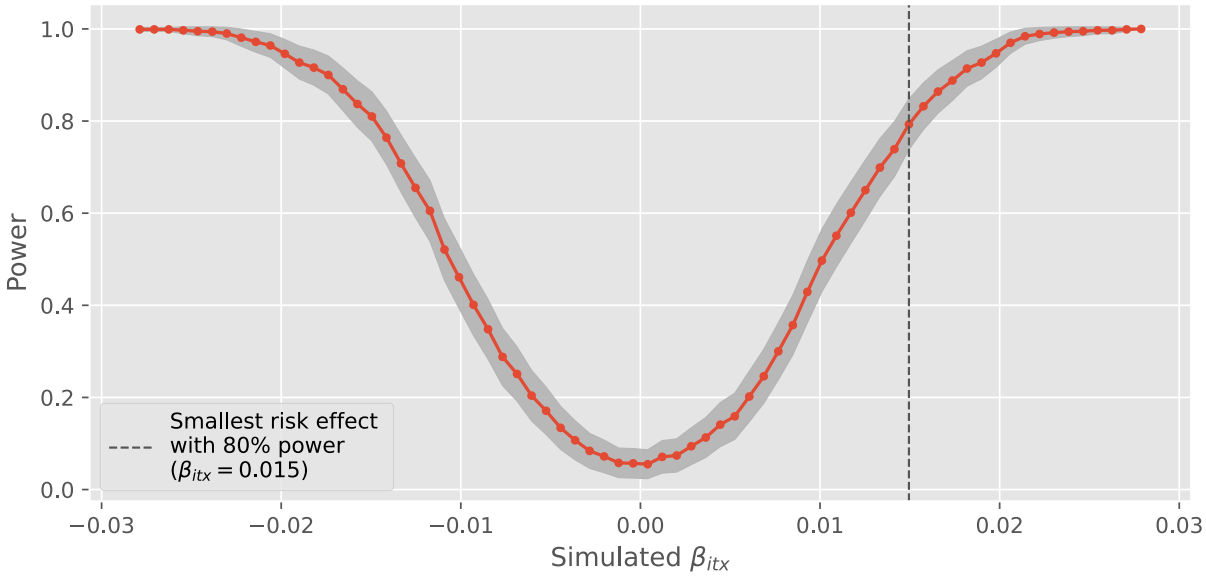
Supplementary Figure B.6 – **Subgroup effect of a 1 standard deviation decrease in the CETP concentration genetic score on biomarkers and cardiovascular endpoints by BMI class in the UK Biobank.**



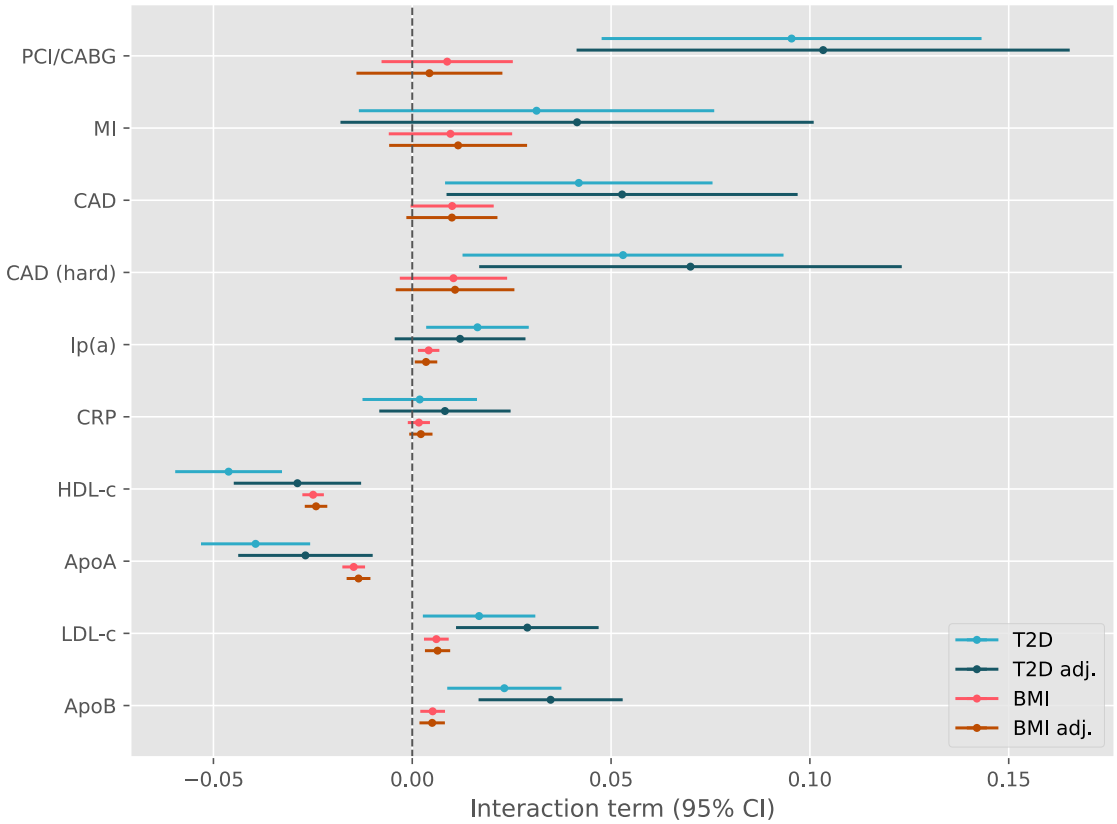
Supplementary Figure B.7 – **Effect modification of rs1800775 (CETP -629C>A) alternative alleles by BMI on biomarkers and cardiovascular endpoints in the UK Biobank.** Displayed p-values (p_{itx}) are for the two sided test of the product term between the SNP and standardized body mass index.



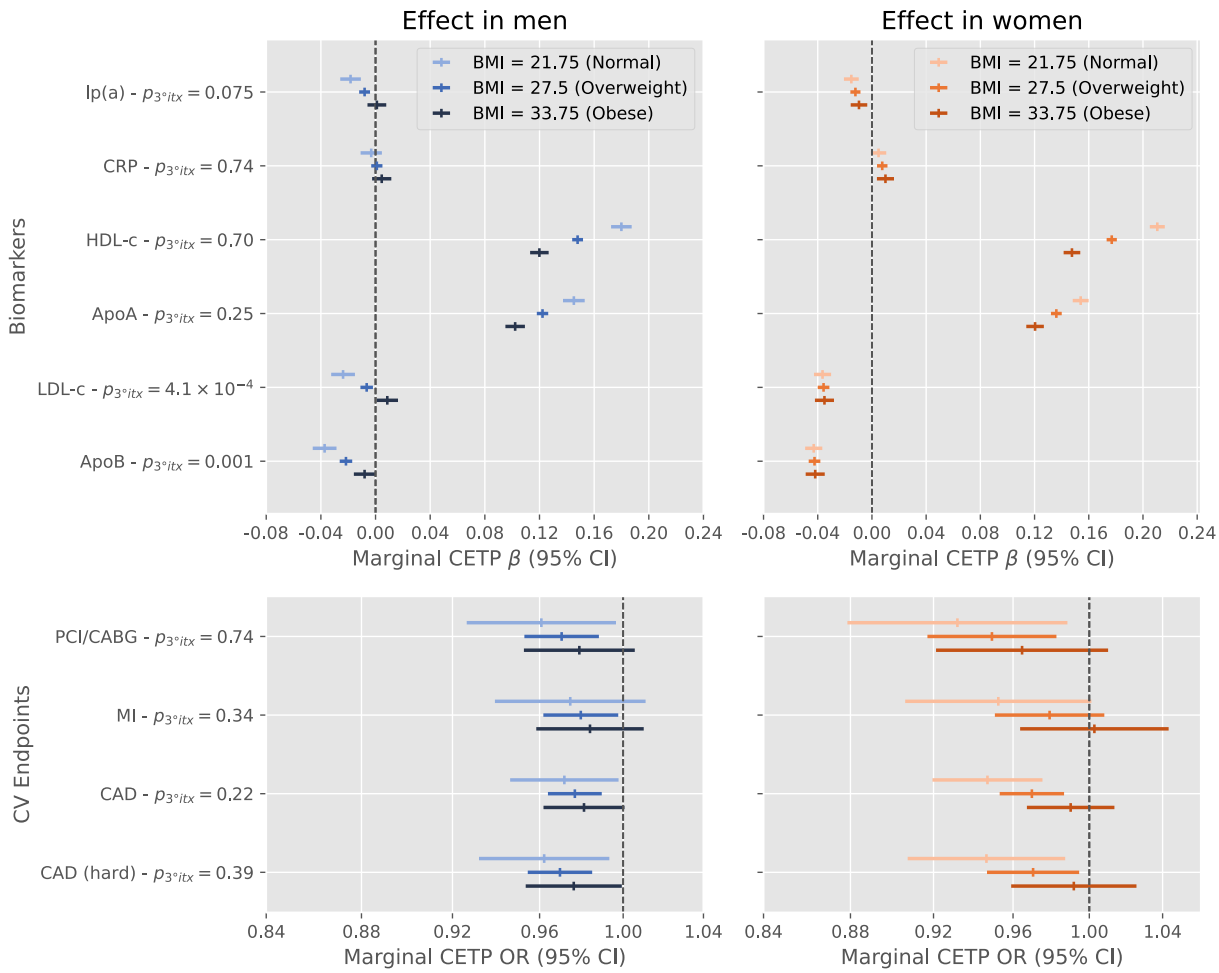
Supplementary Figure B.8 – **Expected value of the standardized biomarkers predicted by linear regression models including interacting splines for the BMI and the genetic score.** The CETP genetic score was set to values between -4 and 4 (x axis) and the BMI was set to values between 15 and 50 (y axis).



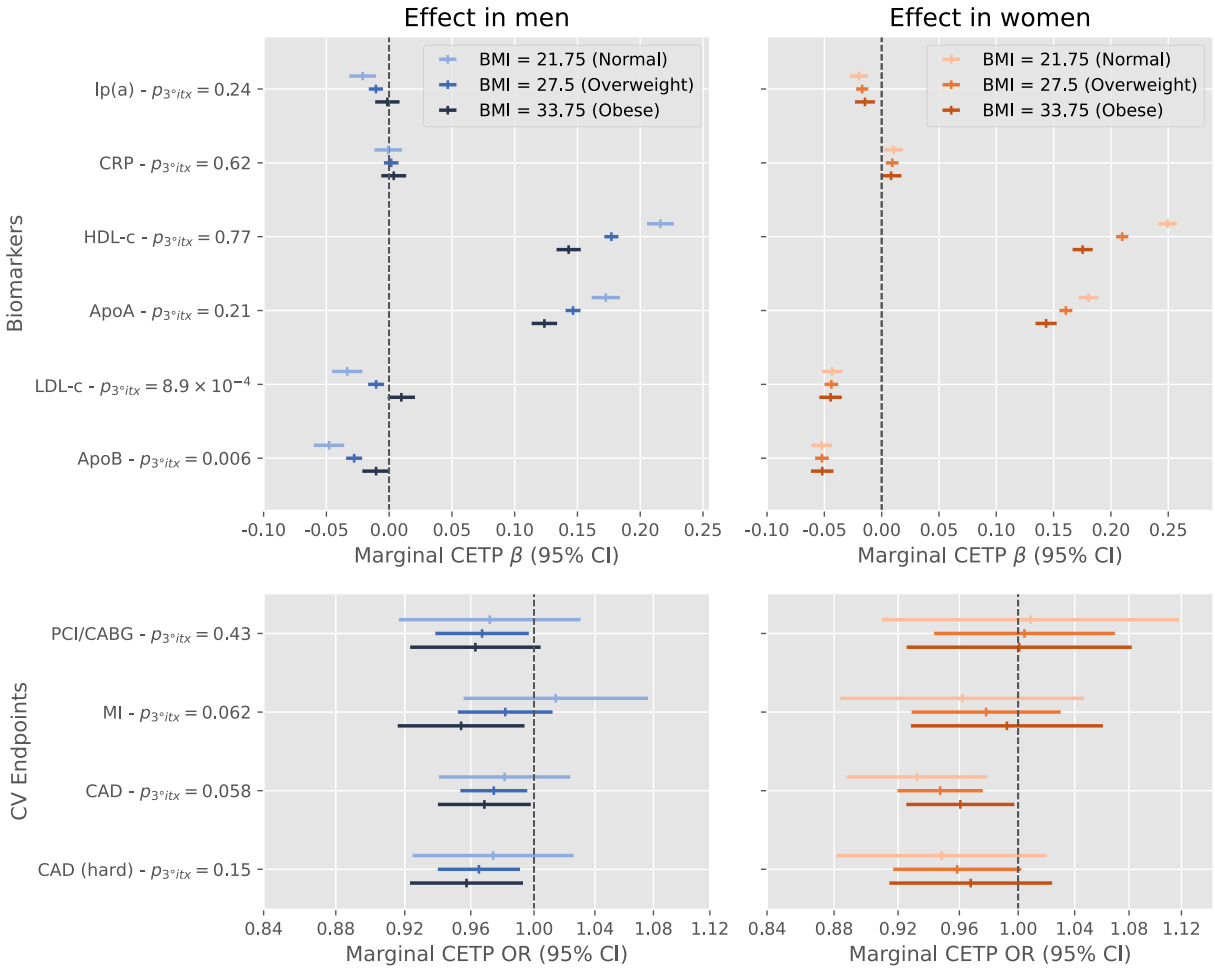
Supplementary Figure B.9 – **Simulation-based power analysis to detect an additive interaction effect between a continuous CETP score and standardized BMI.** This simulation sets as static parameters the prevalence of CAD and the mean effect of a reduction in the CETP score. Results are based on 200 simulation replicates of 413,138 individuals. A 5 data points moving average was used to soften the curve.



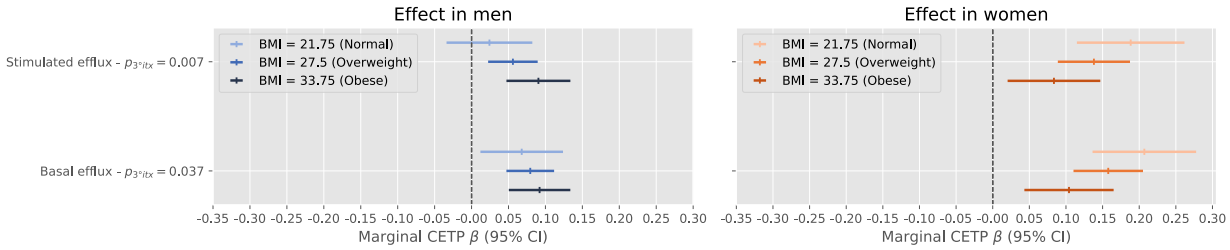
Supplementary Figure B.10 – **Interaction coefficients between the CETP genetic score and type II diabetes and BMI with and without adjustment for the other variable in the UK Biobank.** The unadjusted model includes the product interaction term between CETP and type II diabetes or BMI whereas the adjusted model includes the three-way product interaction.



Supplementary Figure B.11 – **Effect modification of a 1 standard deviation decrease in the CETP concentration genetic score by BMI and sex on biomarkers and cardiovascular endpoints in the UK Biobank.** Displayed p-values ($p_{3°itx}$) are for the two-sided test of the three way product term between the CETP score, standardized body mass index and a binary variable for sex.



Supplementary Figure B.12 – **Effect modification of rs1800775 (CETP -629C>A) alternative alleles by BMI and sex on biomarkers and cardiovascular endpoints in the UK Biobank.** Displayed p-values ($p_{3°itx}$) are for the two-sided test of the 3-way product term between the SNP, a binary variable for sex and standardized body mass index.



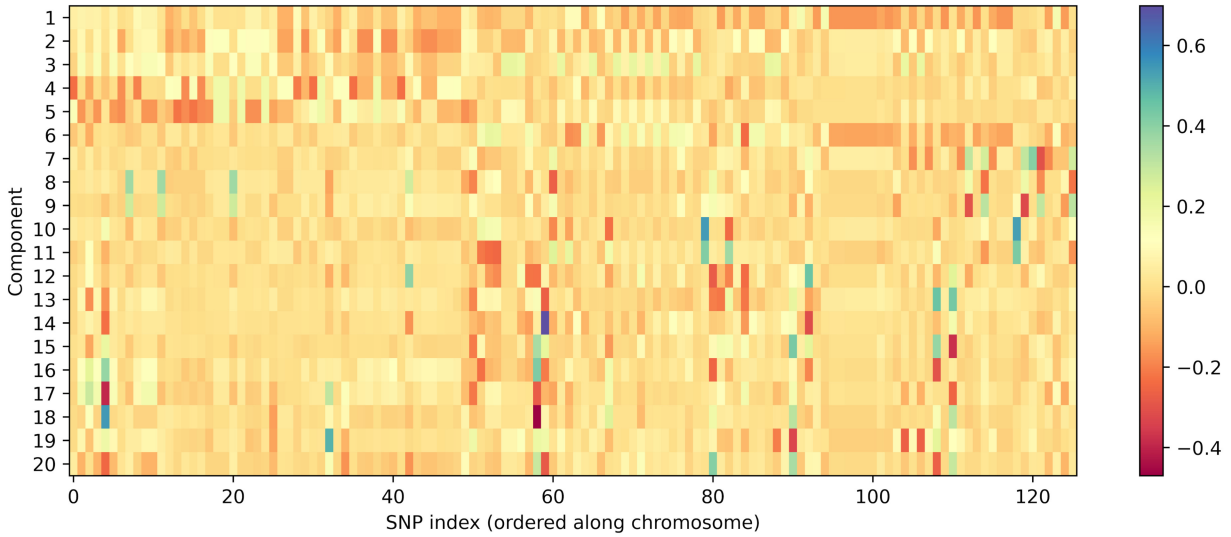
Supplementary Figure B.13 – **Effect modification of the CETP score on cholesterol efflux in the MHI Biobank by sex and BMI.**

ANNEXE C

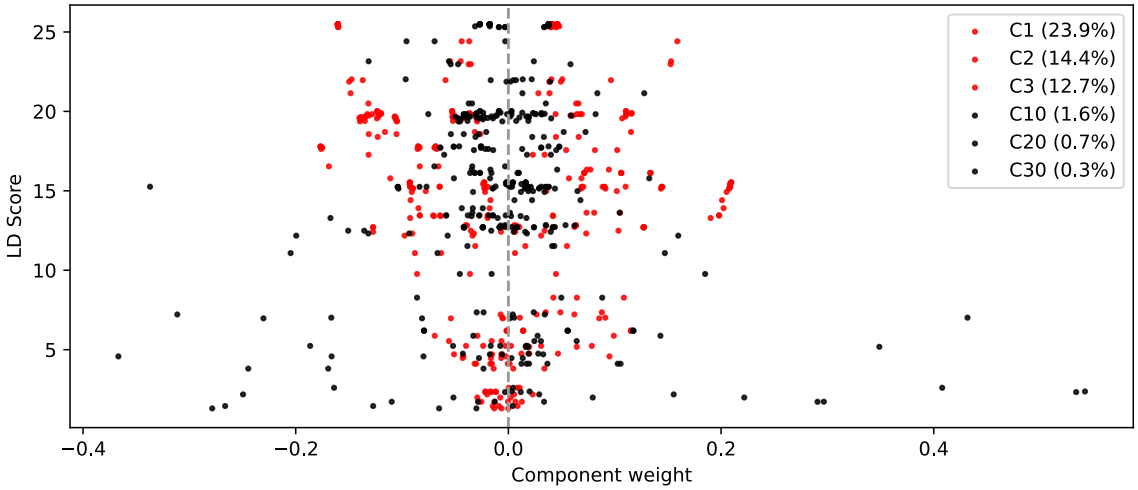
Matériel supplémentaire pour le Chapitre 4

C.1 Supplementary Material

C.1.1 Supplementary Figures

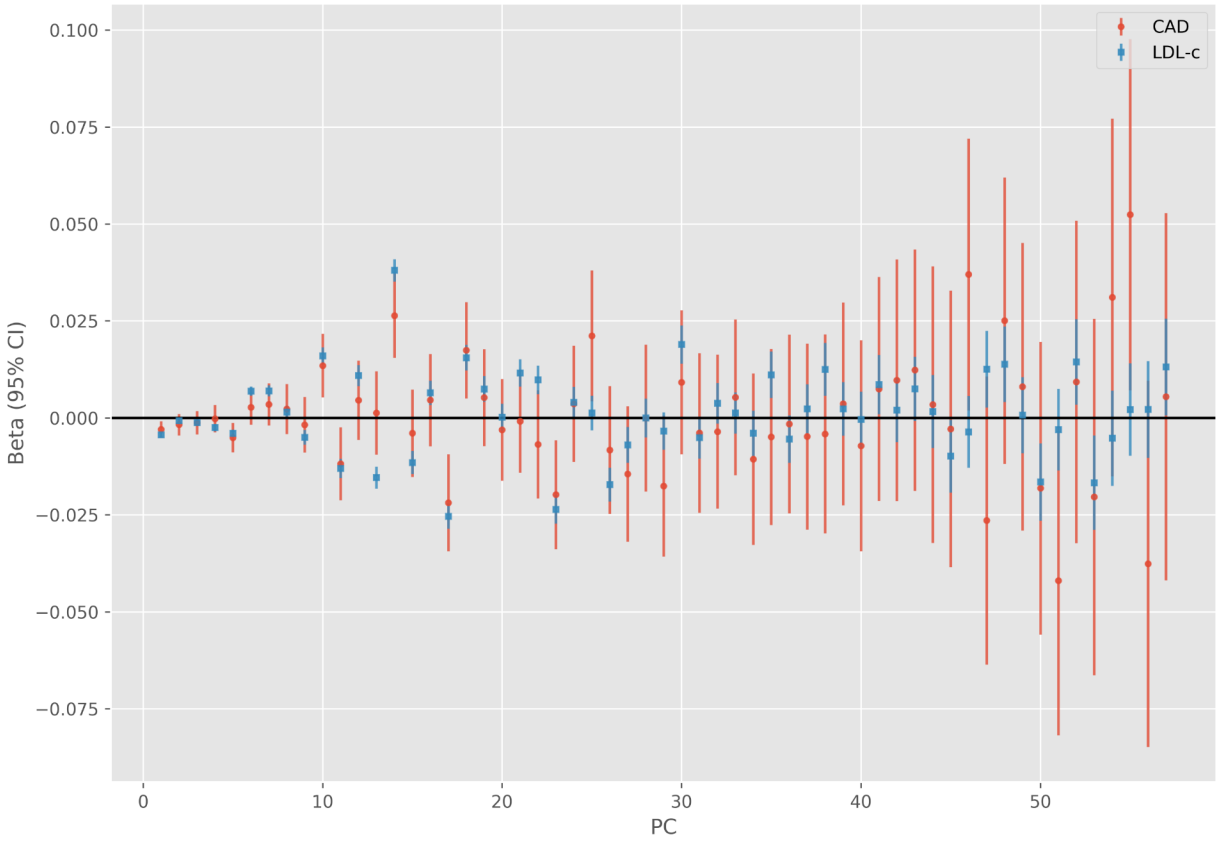


(a) PCA weights for the first 20 principal components computed in 503 unrelated European participants of the 1000 Genomes Project (phase 3) using genotypes for 126 common bi-allelic variants at the *PCSK9* gene locus.

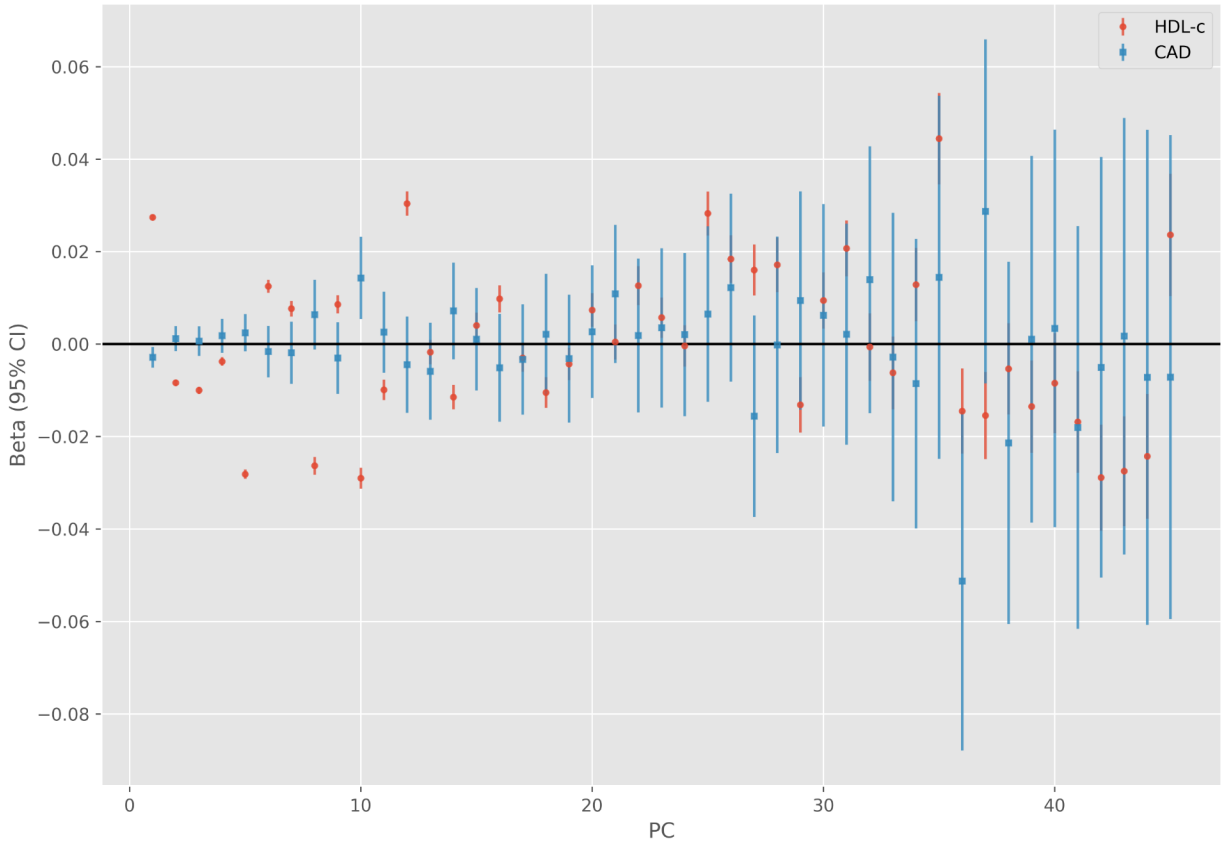


(b) LD Score for the variants with respect to component weights. We selected representative components explaining a large portion of the variance in genotypes (components 1 to 3, in red) and components with smaller eigenvalues explaining a smaller proportion of the variance (components 10, 20 and 30, in black). We notice that the first components attribute larger weights to variants with higher LD scores whereas the later components attribute higher weights to variants with smaller LD scores.

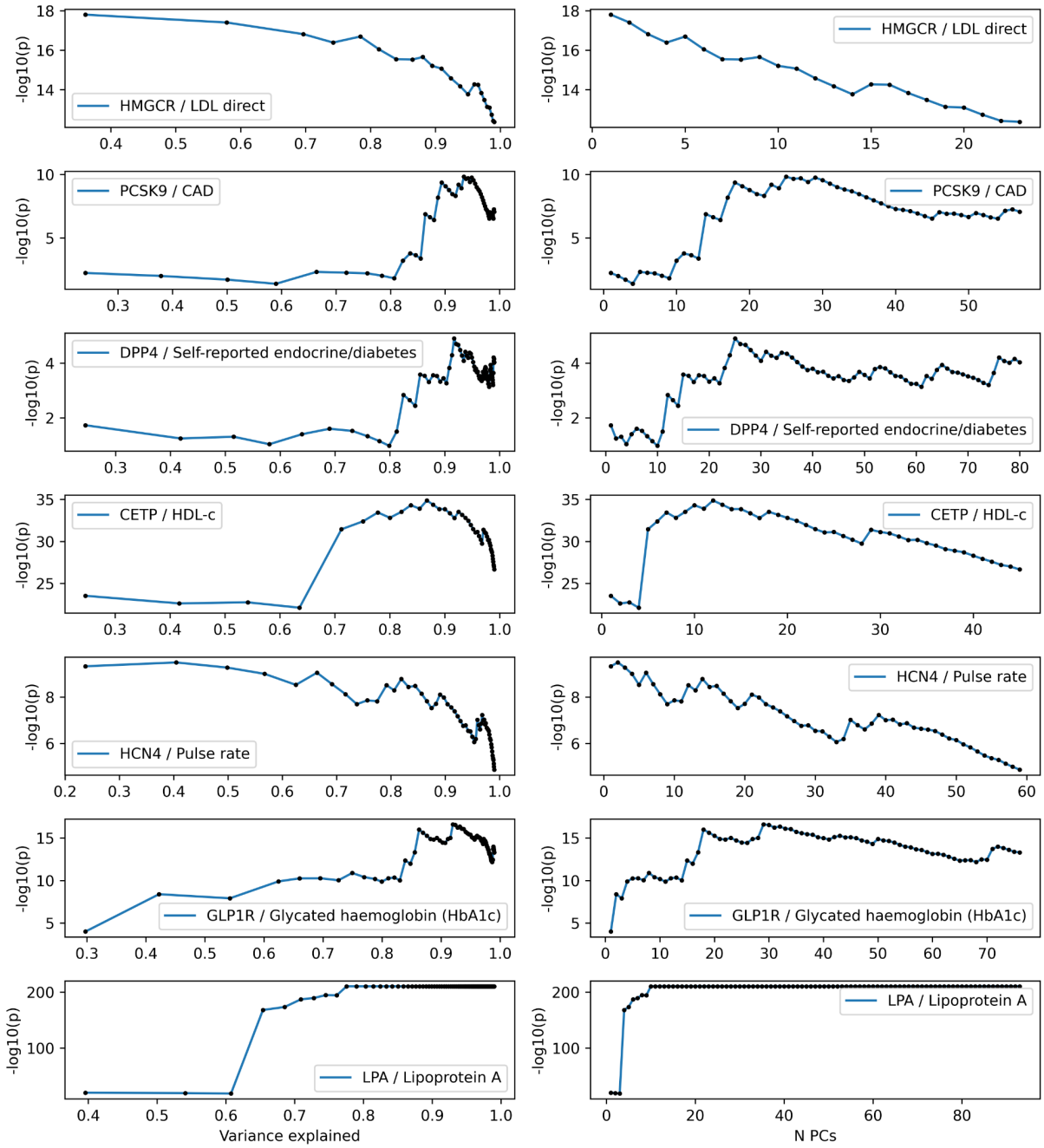
Supplementary Figure C.1 – **Visualisation of the PCA components along with their position on the chromosome and LD score.**



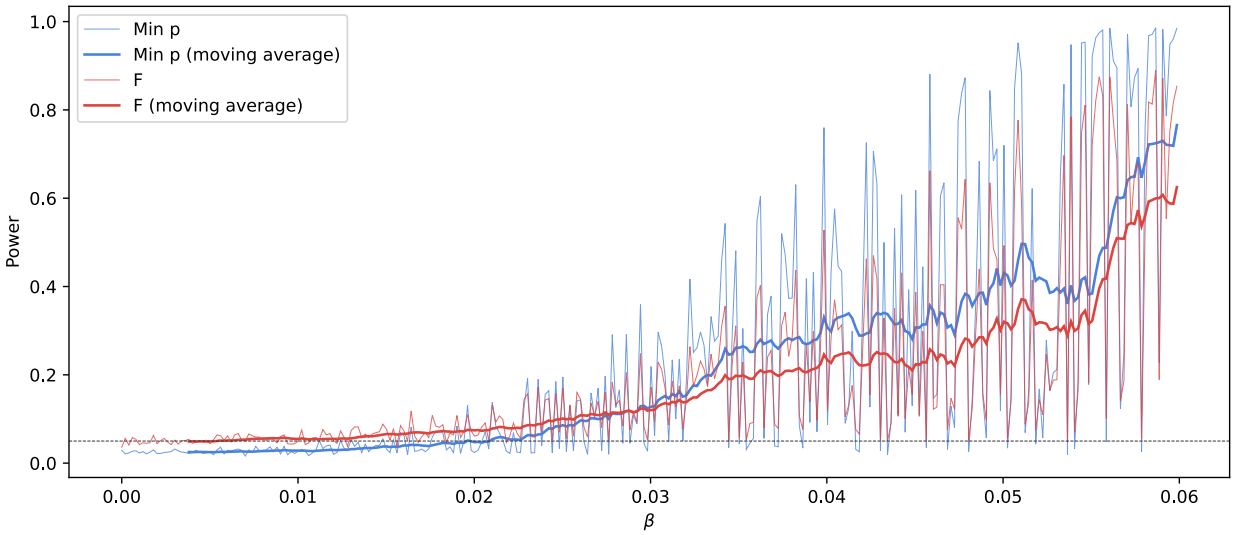
Supplementary Figure C.2 – **Marginal association between genetic PCs based on genotypes at the *PCSK9* locus and low density lipoprotein cholesterol (blue) and coronary artery disease (red).**



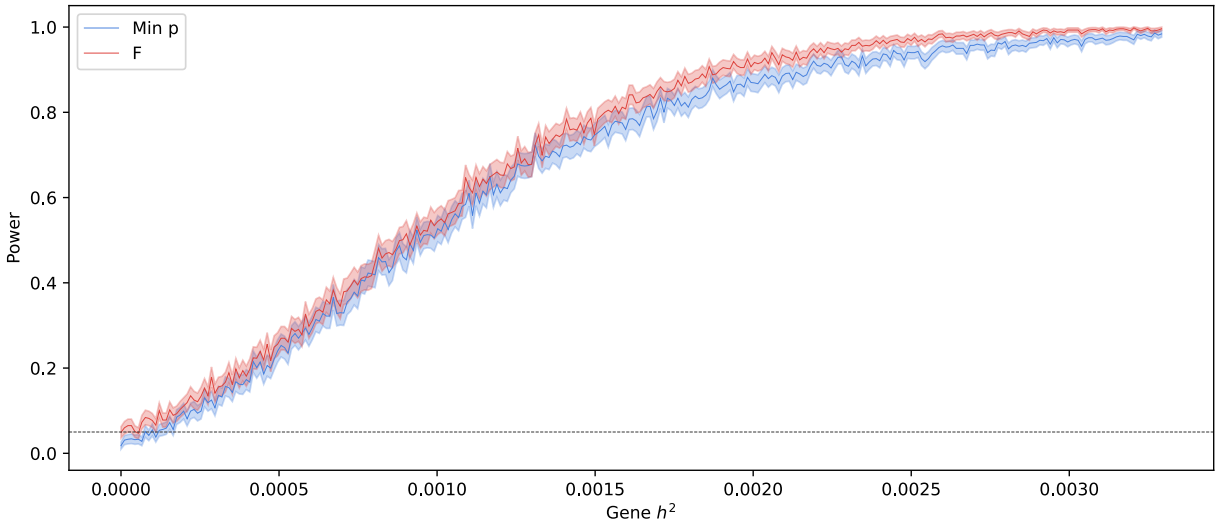
Supplementary Figure C.3 – **Marginal association between genetic PCs based on genotypes at the *CETP* locus and low density lipoprotein cholesterol (blue) and coronary artery disease (red).**



Supplementary Figure C.4 – **Association p-value for selected drug target genes and phenotypes for different choices of included PCs.** The choice of included PCs is expressed as a proportion of cumulative variance explained (left) and as the actual number of included PCs (right). To avoid saturation of the association p-values at 0, some phenotypes were subsampled by randomly selecting individuals.

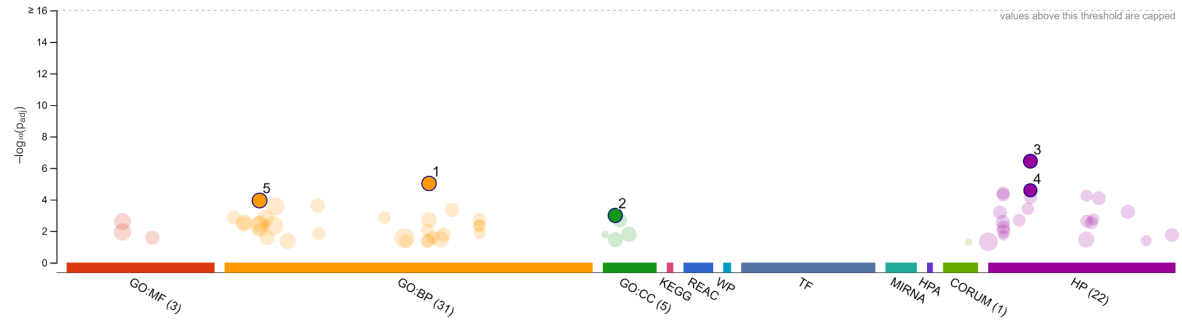


(a) We used a simulation model with a single randomly selected causal variant of increasing effect size (β). The wider lines represent a 20 points moving average.



(b) We used a random effects model where every variant at the locus has an effect drawn from a normal distribution collectively explaining the simulated heritability (h^2). The shaded region represents the 95% confidence interval for the power estimate. When taking the minimum p-value in the region we rejected the null hypothesis using a bonferroni adjusted p-value threshold based on the number of tested variants and a $\alpha = 0.05$ level. Both simulation models include the null model (β or $h^2 = 0$) and the dashed line represents the nominal type 1 error rate of 0.05.

Supplementary Figure C.5 – **Results from 1,000 simulation replicates of 20,000 randomly sampled individuals assessing the power of the PC-based association model and the minimum linear regression p-value within gene boundaries approach.**

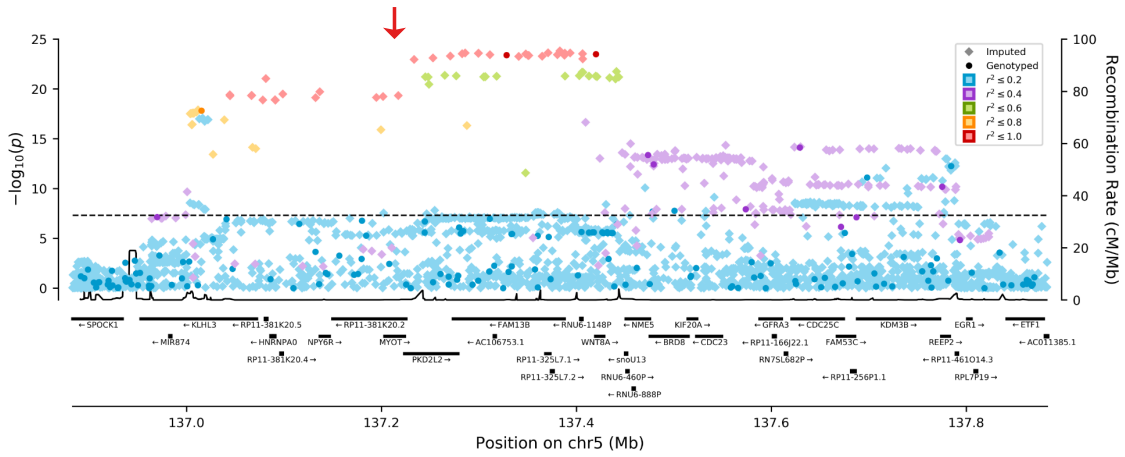


ID	Source	Term ID	Term Name	p_{adj} (query_1)
1	GO:BP	GO:0060048	cardiac muscle contraction	9.515×10^{-6}
2	GO:CC	GO:0030018	Z disc	1.007×10^{-3}
3	HP	HP:0005115	Supraventricular arrhythmia	3.684×10^{-7}
4	HP	HP:0005110	Atrial fibrillation	2.580×10^{-5}
5	GO:BP	GO:0006941	striated muscle contraction	1.151×10^{-4}

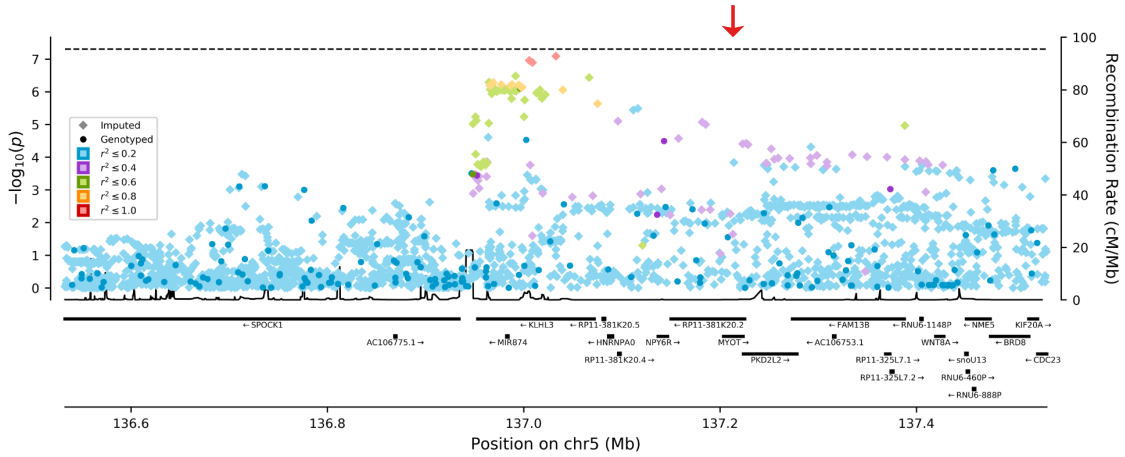
version	e102_eg49_p15_7a9b4d6
date	2/10/2021, 11:29:21 AM
organism	hsapiens

g:Profiler

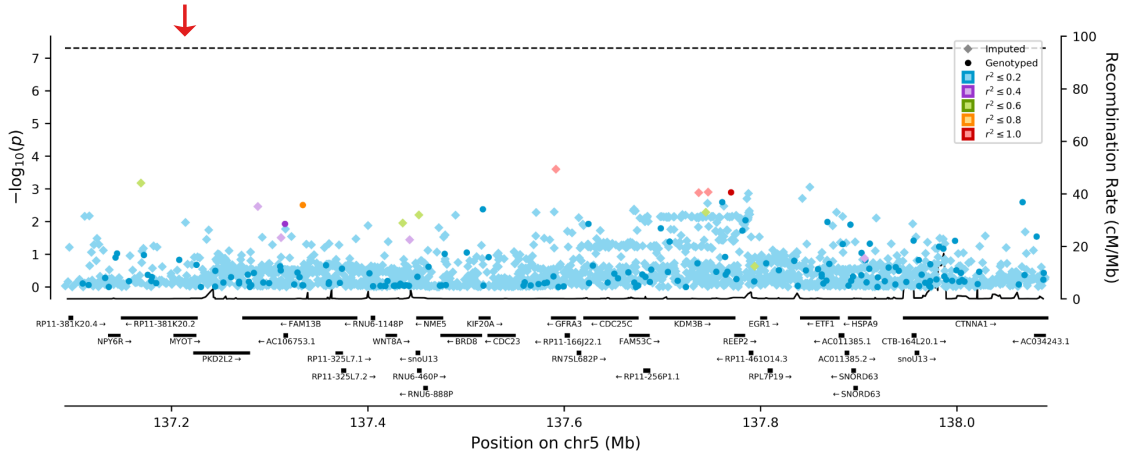
Supplementary Figure C.6 – **Results from a gene set enrichment analysis using g:Profiler for the 137 atrial fibrillation-associated genes with $q \leq 0.01$.** Some representative top enrichments are identified on the plot. The full results are available from the g:Profiler website at https://biit.cs.ut.ee/gplink/l/hMthhT44Rl



(a) Initial scan conditional only on age, sex and the first 10 PCs.



(b) Association scan conditional on the lead variant from stage 1, rs148378888.



(c) Association scan conditional on the lead variants from previous stages: rs148378888 and rs12653760.

Supplementary Figure C.7 – **Stepwise forward conditional analysis of atrial fibrillation associated variant at the chr5:136,883,078-137,883,078 (GRCh37) locus in the UK Biobank.** The red arrows indicate the position of the *MYOT* gene.

C.1.2 Supplementary Tables

Supplementary Table C.1 – **Summary of gene or locus-based association tests. A brief description of the methods is provided along with major characteristics.** The computational cost is given as a qualitative estimate and accounts for the computation of single variant association statistics for tests that rely on them.

Name	**Description**	**Allows weighting**	**Used with summary statistics**	**Genetic variant frequency**	**Comp. cost**	**Method to account for LD**	**Ref.**
PLINK Set-based test	Mean of χ^2 statistics amongst strongest associations	No	No	Common	+++	Empirical P-value from phenotype permutations	[82]
Versatile gene-based association study (VEGAS)	Sum of χ^2 statistics from single-variant models	No	Yes	Common	++	Simulates association statistics under the null using the LD matrix Empirical P-value computation	[83]
GATES	Adaptation of the Simes test for multiple testing correction Estimates the effective number of P-values (tests) from their correlation matrix	Yes	Yes	Common	+	Estimation of the effective number of tests in the P-value correction	[217]
Burden tests	Test based on the count or distribution of alternative alleles	Yes	No*	Rare	-	None (no strong impact for rare variants)	Reviewed in ([81]. Popular implementation described in [264]
Kernel-based tests (*e.g.* SKAT)	The kernel function can be seen as a scalar genotypic similarity measure for pairs of individuals allowing for complex relationships Semi-parametric models are use for statistical testing	Yes	No*	Common and rare	+	No formal modeling of LD Depends on the choice of kernel	[90, 91]

Combined Association Test (COMBAT)	Ensemble learning approach to combine results from different methods	-	Yes	Common	++	-	[87]
Principal Component Analysis (PCA)-based methods	Uses PCA to construct linear combinations of genetic variants that form an orthogonal basis Inference is based on testing for the gain in goodness of fit from including principal components in conventional regression models	No**	No*	Common	-	LD is accounted for because the principal components are orthogonal	[88, 89]
PASCAL	Optimized implementation of tests based on the mean and maximum χ^2 statistics P-value is computed using Monte Carlo simulation or from an optimized numerical integration algorithm	No	Yes	Common	+	LD is modeled through the correlation matrix of summary statistics	[86]

* Even though it was not possible in their original description, the tool SUMmary statistics Fast REGional Association Tests provides implementations for these algorithms that allow for use with summary statistics [218].
** Weighting may be possible as is done in PCA-based Mendelian randomization methods [147].

Supplementary Table C.2 – Summary of the included continuous variables and transformations used to obtain approximately normally distributed variables.
[Excel file available online as Supplementary Material]
https://www.medrxiv.org/content/10.1101/2021.03.17.21253824v1

Supplementary Table C.3 – **Definition of the algorithmically-defined outcomes.**

Outcome	**Definition***
Death outcomes	
Any death	Any entry in the death records variable #40001
Cardiovascular death	Any "I" code as the primary cause of death
Coronary artery disease death	I20–I25 as the primary cause of death
Cardiovascular composite outcomes	
Myocardial infarction	ICD9 codes: 410, 412, 411.0, 429.79 ICD10 codes: I21, I22, I23, I25.2 in the hospitalization or death records
Percutaneous coronary intervention / Coronary artery bypass graft (PCI/CABG)	OPCS procedure codes: K40, K41, K42, K43, K44, K45, K46, K49, K50, K75
Unstable angina	I20.0 code as the primary reason for hospitalization or cause of death
Any angina	ICD9 code 413 or ICD10 code I20 in the hospitalization or death records
Coronary artery disease	ICD9 codes: 410–414 except for aneurysms (414.1) ICD10 codes: I20–I25 in the hospitalization or death records or operation codes for PCI/CABG as previously fined
Heart failure	ICD9 codes: 425, 428 ICD10 codes: I42, I50 in the hospitalization or death records

* Unless otherwise specified, codes were taken in both the primary and secondary reasons for hospitalization, but only the primary cause of death was used.

Supplementary Table C.4 – **Selected drug target genes and phenotypes and the optimal choice in the number of PCs to include to maximise the association strength.**

Gene (symbol)	**Tested phenotype**	**Associated drug class**	**Gene length (quantile)**	**Optimal number of PCs (cumulative variance explained)**	**n PCs explaining 95% of the variance**
3-hydroxy-3-methylglutaryl-CoA reductase (*HMGCR*)	LDL cholesterol	Statins (*e.g.* atorvastatin, simvastatin)	26 kb (0.50)	1 (36%)	14
Proprotein convertase subtilisin/kexin type 9 (*PCSK9*)	Coronary artery disease	PCSK9 inhibitors (*e.g.* alirocumab, evolocumab)	25 kb (0.50)	25 (93%)	30
Dipeptidyl-peptidase 4 (*DPP4*)	Self-reported endocrine/diabetes	DPP4 inhibitors (*e.g.* linagliptin)	82 kb (0.79)	25 (92%)	37
Cholesteryl ester transfer protein (*CETP*)	High density lipoprotein cholesterol	CETP inhibitors (*e.g.* anacetrapib, evacetrapib, dalcetrapib)	22 kb (0.46)	12 (88%)	23
Hyperpolarization activated cyclic nucleotide-gated potassium channel 4 (*HCN4*)	Pulse rate	Ivabradine	49 kb (0.67)	6 (66%)	33
Glucagon-like peptide 1 receptor (*GLP1R*)	Glycated haemoglobin (HbA1c)	GLP1R agonists (*e.g.* liraglutide, albiglutide)	39 kb (0.61)	29 (92%)	39
lipoprotein, Lp(a) (*LPA*)	Lipoprotein(a)	Reduced by PCSK9 inhibitors, antisense-based therapies in development	135 kb (0.88)	25 (90%)	45

Supplementary Table C.5 – Full g:Profiler ontological enrichment results for the genes associated with atrial fibrillation.

[Excel file available online as Supplementary Material]
https://www.medrxiv.org/content/10.1101/2021.03.17.21253824v1

ANNEXE D

Matériel supplémentaire pour le Chapitre 5

A bioinformatics framework for the construction of genetic risk scores

Legault MA., Lemieux Perreault LP., Lemieux S., Tardif JC. and Dubé MP.

Background

- Large GWAS and meta-analysis summary statistics datasets provide precise estimates of the effect of single variants on human traits and diseases
- Such datasets can be used to construct genetic risk scores (GRS) which aggregate the effect of multiple individual variants into a weighted score
- GRS have been shown to be useful as genetic instruments in Mendelian Randomization studies as well as in a risk prediction setting
- Bioinformatics tools can help address challenges faced when constructing and evaluating the performance of GRS

Methods and Results

Choosing SNPs (*grs-create*)

Genetic variants are usually selected from large GWAS or meta-analysis summary statistics datasets.

A widely-used algorithm is the p-value sorting and thresholding method consisting of the following steps:

1. Rank variants by increasing p-value
2. Select the top variant
3. Remove all variants in LD with the selected variant
4. Repeat steps 1. to 3. until the desired p-value threshold is reached

Here, we used summary association statistics from the GLGC consortium to build a GRS of LDL levels.

Computing the GRS (*grs-compute*)

Given a file containing the SNPs to include in the score, "grs-compute" can then compute the GRS using individual-level genotype data.

The weighted GRS are computed as:

$$GRS_i = \sum_{j=0}^{m} \beta_j \cdot X_{ij}$$

After selecting SNPs from the GLGC, we computed the GRS using the imputed genotypes from the MHI Biobank.

Manipulating constructed GRS (*grs-utils*)

Multiple utilities have been implemented to manipulate constructed GRS. The "grs-utils" makes it easy to:

- Generate a histogram of GRS values
- Standardize a GRS
- Measure the correlation between two computed GRS to evaluate the effect of the choice of parameters

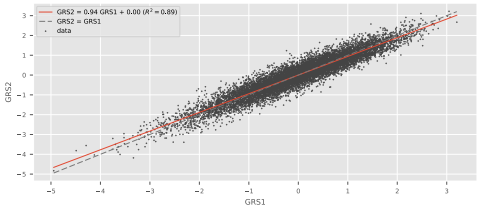


Figure 1. Correlation between two GRS built using different MAF thresholds (plot generated by grs-utils).

Evaluate the predictive performance (grs-evaluate)

Using "grs-evaluate", linear and logistic regressions can be used to evaluate the predictive performance of the GRS.

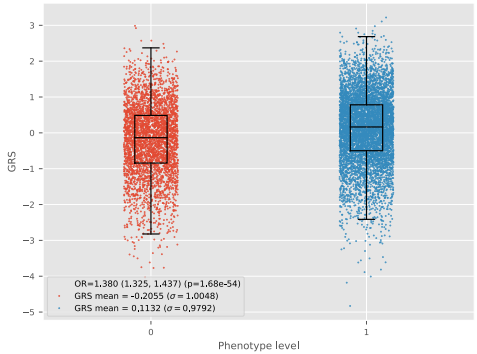


Figure 2. Distribution of the LDL genetic risk score in individuals with and without dyslipidemia as recorded through the MHI Biobank medical questionnaire. On average, there is a 1.38 increase in odds of dyslipidemia per s.d. increase in the GRS (plot generated by grs-evaluate).

GRS are often stratified into groups with low or high genetic risk. This dichotomization can be achieved by comparing extreme quantiles of the GRS, but individuals falling in the intermediate group are often excluded from analyses. Hence, there is a tradeoff between the effect size and the number of individuals in the extreme groups.

To help guide the selection of thresholds, the "dichotomize-plot" sub-command can be used.

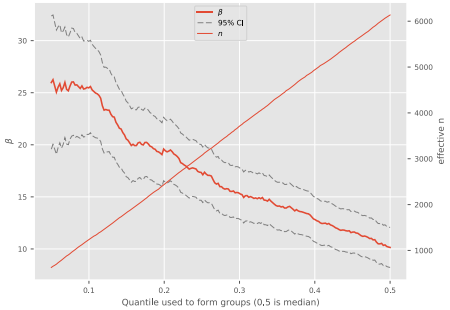


Figure 3. Dichotomization plot showing the estimated increase in LDL in the "high" GRS group when compared to the "low" group. The number of individuals retained by the classification is shown as the diagonal line. In this case, a dichotomization based on the 15 percentile can be used to achieve a mean LDL difference between groups of 20 mg/dl, retaining approximatly 3,200 individuals (plot generated by grs-evaluate).

MHI Biobank

- Longitudinal study of 22,000 individuals
- More than 10,600 genotyped individuals after quality control
- Dense phenotypic data (especially for cardiovascular diseases) using a questionnaire and medical records
- Pharmacotherapy data available as free text

Software availability

This software was written in Python and the source code and documentation are freely available online:

http://github.com/legaultmarc/grstools

The command-line utilities can be installed using a simple command:

```
pip install grstools
```

Discussion

Using *grstools*, it is possible to quickly construct genetic risk scores using GWAS summary statistics or any other weighting scheme. It also allows easy manipulation and assessment of the performance of the scores.

To demonstrate our tool, we have computed a GRS that can be used as a genetic instrument to predict the effect of varying LDL levels.

email: marc-andre.legault.1@umontreal.ca

FIGURE D.1 – Affiche scientifique décrivant la suite d'outils *grstools* pour la création, le calcul et l'évaluation de scores de risque génétique.