University of Helsinki
Dissertationes Universitatis Helsingiensis 234/2026

Improving Cross-Biobank Genetics Through Phenotype Refinement and Machine Learning

Lisa Eick

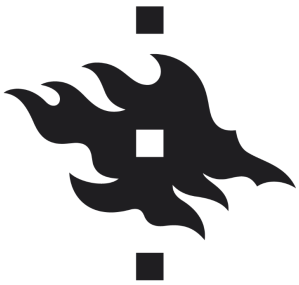
Institute for Molecular Medicine Finland (FIMM)
Helsinki Institute of Life Science (HiLIFE)
University of Helsinki, Finland

Doctoral Program in Population Health
University of Helsinki Doctoral School
Faculty of Medicine
University of Helsinki, Finland

ACADEMIC DISSERTATION

To be presented, with the permission of the Faculty of Medicine of the University of Helsinki,
for public examination in Lecture Hall 3, Biomedicum Helsinki,
on May 29th 2026, at 13.00.

Helsinki 2026



HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

FIMM
Institute for Molecular Medicine Finland
Nordic EMBL Partnership for Molecular Medicine

HiLIFE UNIT

Publisher: University of Helsinki
Series: Dissertationes Universitatis Helsingiensis 234/2026
Cover Image: Lisa Eick

ISBN 978-952-84-2342-3 (paperback)
ISBN 978-952-84-2341-6 (PDF)
ISSN 2954-2898 (print)
ISSN 2954-2952 (online)

PunaMusta, Joensuu 2026

Supervisors Associate Professor **Andrea Ganna**, Ph.D.
Institute for Molecular Medicine Finland (FIMM)
HiLIFE, University of Helsinki, Finland, and
The Broad Institute of MIT and Harvard,
Cambridge, Massachusetts, USA, and
Massachusetts General Hospital,
Boston, Massachusetts, USA

Dr. **Zhiyu Yang**, Ph.D.
Institute for Molecular Medicine Finland (FIMM),
HiLIFE, University of Helsinki, Finland

Reviewers Associate Professor **Sandra Sanchez-Roige**, Ph.D.
Department of Psychiatry, School of Medicine,
University of California San Diego, USA and
Department of Medicine, Division of Genetic Medicine,
Vanderbilt University Medical Center, USA

Associate Professor **Mikko Arvas**, Ph.D.
Finnish Red Cross Blood Service.
Helsinki, Finland.

Opponent Associate Professor **Sandra Sanchez-Roige**, Ph.D.
Department of Psychiatry, School of Medicine,
University of California San Diego, USA and
Department of Medicine, Division of Genetic Medicine,
Vanderbilt University Medical Center, USA

"To [her] that will, ways are not wanting."

– George Herbert, Jacula Prudentum (1640)

To my friends, family and partner, who walked with me every step of the way.

Abstract

Large-scale population biobanks linking genomic data to electronic health records, prescription registries, and longitudinal clinical information have transformed human genetics. Resources such as UK Biobank, FinnGen, and the Estonian Biobank enable genome-wide association studies of highly polygenic traits and the construction of polygenic scores. While increasing sample size has improved statistical power, it has also highlighted persistent limitations related to cohort heterogeneity, phenotype definitions, and inconsistencies observed in real-world clinical data, particularly in cross-biobank analyses. Differences in healthcare systems, coding practices, and population structure can strongly affect genetic discovery and predictive performance. This thesis investigates sources of heterogeneity in cross-biobank genetic analyses and evaluates phenotype refinement and machine learning as strategies to improve robustness and transferability.

The thesis comprises three interconnected studies. Study I evaluates polygenic prediction across multiple biobanks and polygenic scores construction methods. No single method consistently outperforms others across traits or cohorts. Instead, ensemble approaches achieve the most robust performance. Variance decomposition shows that differences between biobanks explain more predictive variability than differences between methods, identifying cohort heterogeneity as a primary limitation.

Study II focuses on phenotype quality in psychiatric genetics, where diagnostic uncertainty and comorbidity lead to substantial misclassification in electronic health records-derived phenotypes. Using simulations and analyses of ten psychiatric traits in FinnGen, we show that stricter control definitions excluding individuals with psychiatric comorbidity or related medication use (SuperControl) increase genetic signal and locus discovery. This improvement occurs despite smaller sample sizes and with comparable replication performance. A machine-learning framework (PRISMA) integrates predicted liability for excluded individuals, further increasing discovery but also decreasing disease specificity and reducing replication in some settings, highlighting trade-offs between statistical power, specificity, and noise.

Study III extends phenotype refinement to prescription-derived traits using opioid intake as a proxy for pain burden. Leveraging longitudinal prescription records from FinnGen, UK Biobank, and the Estonian Biobank, harmonized opioid exposure and dose phenotypes are constructed. Genome-wide analyses reveal extensive polygenicity and pleiotropy, with genetic overlap between opioid intake, pain-related traits, and mental health outcomes. Machine-learning-derived early-onset and overuse phenotypes distinguish general pain liability from opioid-specific mechanisms, with the overuse phenotype implicating impaired opioid metabolism at the *CYP2D6/CYP2D7* locus.

Together, these studies show that increasing sample size alone does not resolve limitations driven by heterogeneity and phenotype definition. Phenotype construction emerges as a central design choice shaping statistical power, biological specificity, and cross-biobank transferability, underscoring the need to explicitly manage trade-offs in biobank-based genetic research.

Tiivistelmä

Laajat väestökohortit, jotka yhdistävät genomitiedon pitkittäisiin kliinisiin rekisteritietoihin, ovat mahdollistaneet valtavan murroksen ihmisen genetiikan tutkimuksessa. Resurssit, kuten UK biopankki, FinnGen ja Estonian biopankki, mahdollistavat genominlaajuisten assosiaatiotutkimusten toteuttamisen erittäin polygeenisistä piirteistä ja polygeenisten riskiarvojen rakentamisen. Tutkittavien kohorttien kasvu on parantanut analyysien tilastollista merkitsevyyttä samalla kuitenkin korostaen analyysien rajoituksia, jotka liittyvät kohorttien heterogeenisyyteen, ilmiasujen määritelmiin ja kliinisen datan epäjohdonmukaisuuksiin, erityisesti biopankkien välisissä analyyseissä. Erot terveydenhuoltojärjestelmissä, koodauskäytännöissä ja väestörakenteessa voivat vaikuttaa voimakkaasti geneettisiin löytöihin ja ennustettavuuteen. Tässä väitöskirjassa tutkitaan biopankkien välisten geneettisten analyysien heterogeenisyyden lähteitä, arvioidaan ilmiasujen määrittelyä ja koneoppimisen mallien hyödyntämistä, joilla parannetaan geneettisten analyysien luotettavuutta ja siirtokelpoisuutta.

Väitöskirja koostuu kolmesta tutkimuksesta. Tutkimuksessa I arvioidaan polygeenistä ennustamista eri biopankeissa ja polygeenisten riskiarvojen rakennusmenetelmissä. Mikään yksittäinen menetelmä ei johdonmukaisesti toimi muita paremmin tutkituissa piirteissä tai kohorteissa. Menetelmät yhdistettyinä saavuttavat sen sijaan parhaan suorituskyvyn. Varianssien jakauma osoittaa, että biopankkien väliset erot selittävät ennustettavuuden vaihtelua enemmän kuin menetelmien väliset erot osoittaen kohorttien heterogeenisyyden olevan ensisijainen heikkous käytettävien menetelmien sijaan.

Tutkimus II keskittyy ilmiasujen laatuun psykiatrisessa genetiikassa, jossa diagnoosien epävarmuus ja vahva komorbiditeetti ilmiasujen välillä johtavat usein luokitteluvirheisiin suoraan sähköisistä potilastietojärjestelmistä määritellyissä ilmiasuissa. Käyttämällä simulaatioita ja genominlaajuisia analyysejä kymmenestä psykiatrisesta piirteestä FinnGen-tietokannassa osoitimme, että tarkennetut verrokkimääritelmät, jotka sulkevat pois psykiatrisen komorbiditeetin ja lääkkeiden käytön (SuperControl), lisäävät geneettistä signaalia ja todennäköisyyttä geneettisten lokuksien löytämiselle pienemmästä otoskoosta huolimatta. Näistä saadut tulokset ovat myös vertailukelpoisia. Koneoppimisen kehys (PRISMA) integroi ennustetun alttiuden pois alun perin analyysistä poistetuille yksilöille, mikä lisää löydöksiä samalla lisäten myös geneettistä korrelaatiota ja vähentäen tulosten toistettavuutta. Nämä asiat korostavat muodostuneita kompromisseja tilastollisen voiman, spesifisyyden ja aineiston kohinan välillä.

Tutkimus III laajentaa ilmiasujen määrittelyn reseptilääkkeistä johdettuihin piirteisiin hyödyntämällä opioidien käyttöä kivun voimakkuuden mittarina. FinnGenin, UK biopankin ja Estonian biopankin pitkittäistutkimusten reseptitietojen avulla muodostettiin yhdenmukaistetut ilmiasut sisältäen opioideille altistumisen sekä annostusten määrät. Koko genomin kattavat analyysit paljastavat laajaa polygeenisyyttä ja pleiotropiaa, joissa opioidien käyttö, kipuun liittyvät ominaisuudet ja mielenterveys jakavat geneettisiä tekijöitä. Koneoppimisella johdetut opioidien varhaisen käytön ja liikakäytön ilmiasut erottavat yleisen kipualttiuden opioideille ominaisista mekanismeista. Erityisesti opioidien liikakäyttö nostaa esille heikentyneeseen opioidiaineenvaihduntaan viitatun *CYP2D6*/*CYP2D7*-lokuksen.

Yhdessä nämä tutkimukset osoittavat, että ainoastaan otoskoon kasvattaminen ei ratkaise heterogeenisyydestä ja ilmiasujen määritelmistä johtuvia heikkouksia. Ilmiasujen rakentaminen nousee keskeiseksi osaksi genomin laajuisissa analyyseissa muokaten tilastollista voimaa, biologista spesifisyyttä ja biopankkien välistä siirtokelpoisuutta. Tämä korostaa tarvetta huomioida tarkemmin kompromisseja, jotka biopankkiaineistoihin perustuvissa genomin laajuisissa tutkimuksissa usein hyväksytään.

Table of Contents

Abstract ..... **I**
Tiivistelmä ..... **II**
Table of Contents ..... **III**
List of Original Publications ..... **V**
Abbreviations ..... **VI**
1 Introduction **1**
2 Literature Review **3**
2.1 Human Genetics and Large-Scale Biobanks 3
2.1.1 Molecular Basis of Genetics 3
2.1.2 The Genetic Architecture and the Biobank Revolution 5
2.1.3 Genetic Analysis Approaches 7
2.1.4 Bias and Limitations in Large Biobank Data 10
2.2 Genetics of Psychiatric Traits 13
2.2.1 Diagnostic Challenges in Psychiatry 14
2.2.2 Genetic Architecture of Psychiatric Traits 14
2.2.3 Study Design in Psychiatric Genetics 16
2.3 The Genetic Link between Psychiatric Disorders and Pain 17
2.4 Applications of Machine Learning in Human Genetics 18
2.4.1 Machine Learning Models for Polygenic Scores 18
2.4.2 Machine Learning for Phenotype Refinement 19
3 Aims of the Study **24**
4 Methods **25**
4.1 Study Cohorts 25
4.1.1 FinnGen 25
4.1.2 UK Biobank (UKBB) 26
4.1.3 Estonian Biobank (EstBB) 26
4.1.4 Other Cohorts 27
4.2 Phenotype Definitions and Machine Learning Refinement 28
4.2.1 Phenotype Definitions for Polygenic Score Analyses (Study I) 28
4.2.2 Psychiatric Traits and SuperControl Framework (Study II) 29
4.2.3 Opioid Prescription Phenotypes and Opioid Intake Harmonization (Study III) 31
4.2.4 FinnGen Harmonization for Machine Learning Application 32
4.2.5 Psychiatric Risk Prediction and PRISMA Phenotype Construction 32
4.2.6 Opioid Intake Prediction Model 33
4.3 Genetic Analysis 35
4.3.1 Simulation of SuperControl Design and Misclassification 35
4.3.2 Genome-wide Association Analyses (GWAS) 35
4.3.3 Post-GWAS Downstream Analysis 36
4.4 Polygenic Score (PGS) Analyses 38
4.4.1 Target Genotype Processing and Ancestry Matching 38
4.4.2 Polygenic Score Construction and Ensemble Modeling 38

4.4.3 Performance Evaluation of Polygenic Score Methods 40
4.4.4 Meta-Analysis of Polygenic Score Methods 41
4.4.5 Cross-Cohort Prediction and Regional Polygenic Score Analyses 42
4.6 Ethics 43
5 Results 44
5.1 Variability of Polygenic Prediction Across Methods and Biobanks (Study I) 44
5.1.1 Performance of Polygenic Score Methods within Biobanks 44
5.1.2 Variance in Polygenic Score Performance 46
5.2 Refining Psychiatric Phenotypes to Reduce Misclassification (Study II) 47
5.2.1 Simulation of Misclassification and SuperControl Bias 47
5.2.2 Empirical Application of SuperControl Phenotypes in EHR 48
5.2.3 Recovering Excluded Samples using Machine Learning 51
5.2.4 Replication and Cross-Biobank Validation 52
5.3 Genetic Architecture of Prescription Opioid Intake (Study III) 53
5.3.1 Genome-Wide Associations of Opioid Exposure and Cumulative Dose 53
5.3.2 Machine Learning-Derived Overuse and Early-Onset Phenotypes 56
6 Discussion 58
6.1 Sources of Heterogeneity in Cross-Biobank Analysis 58
6.2 Phenotype Definition in Electronic Health Records: Limits and Opportunities 60
6.3 Machine Learning as an Opportunity for Signal and a Potential Source of Noise 64
7 Conclusion and Future Directions 67
Acknowledgements 70
Appendix A 74
References 76
Original Publication 97

List of Original Publications

The work presented in this thesis is based on the following publications:

Study I: Monti R.*, **Eick L.***, Hudjashov G., Läll K., Kanoni S., Wolford B. N., Wingfield B., Pain O., Wharrie S., Jermy B., McMahon A., Hartonen T., Heyne H., Mars N., Lambert S., Genes and Health Research Team, Hveem K., Inouye M., van Heel D. A., Mägi R., Marttinen P., Ripatti S., Ganna A., Lippert C., “Evaluation of polygenic scoring methods in five biobanks shows larger variation between biobanks than methods and finds benefits of ensemble learning”, *The American Journal of Human Genetics*, Volume 111, Issue 7, Pages 1431-1447, (2024)
Doi: https://doi.org/10.1016/j.ajhg.2024.06.003
*These authors contributed equally

Study II: Eick L., FinnGen, Ganna A., Yang Z., “Reducing misclassification bias in electronic health record-based GWAS of psychiatric traits using the SuperControl framework”, medRxiv, submitted to *PLOS Genetics*
Doi: https://doi.org/10.64898/2025.12.14.25342213

Study III: Eick L., Luitva L. B., Krebs K., Jukarainen S., Kulju S., FinnGen, Estonian Biobank research team, Marttinen M., Rivas M. A., Milani L., Ganna A., Yang Z., Kiiskinen T., “Machine learning augmented genome-wide meta-analysis of prescription opioid use in 860,000 individuals”, medRxiv, submitted to *Genome Medicine*
Doi: https://doi.org/10.64898/2025.12.07.25341785

Abbreviations

ADHD	Attention-Deficit/Hyperactivity Disorder
AFR	African (1000 Genomes superpopulation)
AMR	Admixed American (1000 Genomes superpopulation)
ANX	Anxiety Disorder
ATC	Anatomical Therapeutic Chemical (classification system)
AUROC	Area Under the Receiver Operating Characteristic Curve
BD	Bipolar Disorder
BMI	Body Mass Index
BNF	British National Formulary (medication coding system in UKBB)
CKD	Chronic Kidney Disease
CNS	Central Nervous System
CV	Cross-validation
DF12	FinnGen Data Freeze 12
DNA	Deoxyribonucleic Acid
EAS	East Asian (1000 Genomes superpopulation)
eGFR	Estimated Glomerular Filtration Rate
EHR	Electronic Health Record
EstBB/EBB	Estonian Biobank
EUR	European (1000 Genomes superpopulation)
GBR	Gene-based replication
GNH	Genes & Health cohort
GWAS	Genome-Wide Association Study
h^2	SNP-based heritability
HbA1c	Glycated hemoglobin
HDL	High-density lipoprotein cholesterol
HLA	Human Leukocyte Antigen (genomic region)
HRC	Haplotype Reference Consortium
HUNT	Trøndelag Health Study
HWE	Hardy-Weinberg equilibrium
I^2	I-squared heterogeneity metric (meta-analysis heterogeneity)
ICD	International Classification of Diseases
LBR	Locus-based replication
LD	Linkage Disequilibrium
LDSC	Linkage Disequilibrium Score Regression

MAF	Minor Allele Frequency
MDD	Major Depressive Disorder
METAL	METAL meta-analysis software
MHD	Mental Health Disorder (composite trait in Study II)
ML	Machine Learning
MPE	Moderate prior evidence (replication category)
MRI	Magnetic resonance imaging
mRNA	Messenger Ribonucleic Acid
MTAG	Multi-Trait Analysis of GWAS
NSE	No supporting evidence (replication category)
OCD	Obsessive-Compulsive Disorder
OF	Opioid Factor
OME	Oral Morphine Equivalent
PC	Principal Component
PGC	Psychiatric Genomics Consortium
PGS	Polygenic Score
PRISMA	PRedicted excluded Individuals and SuperControls via MTAG Analysis
PTSD	Post-Traumatic Stress Disorder
r^2	proportion of variance explained by the model
r_g	Genetic correlation
RxDosePop	Population-wide cumulative opioid dose phenotype
RxDoseUsers	Cumulative opioid dose among exposed individuals phenotype
RxExpPop	Ever-exposed-to-opioids phenotype
SAS	South Asian (1000 Genomes superpopulation / ancestry group used)
SCZ	Schizophrenia
SLD	Sleep Disorders
SNOMED	Systematized Nomenclature of Medicine
SNP	Single-Nucleotide Polymorphism
SPE	Strong prior evidence (replication category)
SUD	Substance Use Disorder
UKBB	UK Biobank
WES	Whole-Exome Sequencing
WGS	Whole-Genome Sequencing
XGBoost	Extreme Gradient-Boosted Decision Trees

1 Introduction

Large-scale biobanks that link genomic data with electronic health records (EHR), registry information, and prescription data have reshaped human genetics. Initiatives such as UK Biobank (UKBB), FinnGen, and the Estonian Biobank (EstBB) combine genetic, clinical, and lifestyle information for hundreds of thousands of participants, enabling large-scale analyses of how genetic variation contributes to human traits and disease.

Genome-wide association studies (GWAS) leverage these data to identify variants associated with traits, including those influenced by many variants of small effect, so-called polygenic traits. The aggregate influence of these variants can be summarized through polygenic scores (PGS), which estimate an individual's genetic liability for a trait. PGS are now widely applied in research and increasingly considered as a tool to improve existing disease risk models[1,2], in disease prevention, risk stratification, and personalized medicine[3,4]. However, PGS performance can vary substantially across biobanks, particularly in the presence of subtle ancestry differences[5].

Other sources of variability are methodological, as different PGS models can yield different estimates[6,7]. Yet an even greater contributor is how phenotypes are defined[8]. EHR-based traits reflect the structure of national healthcare systems, coding conventions, and data availability, leading to substantial differences between cohorts. Diagnostic codes often collapse complex, gradual, or heterogeneous clinical presentations into binary case-control labels that may overlook subthreshold symptoms, fail to capture clinical nuance, or inadvertently misclassify individuals due to incomplete histories, diagnostic biases, or stigma[9–11]. These issues are particularly severe for psychiatric disorders, which lack objective biomarkers and are frequently underdiagnosed[12,13]. As a result, psychiatric phenotypes in biobanks often contain large sample sizes but limited specificity, and even modest misclassification has been shown to reduce genetic signal and bias downstream analyses[14]. Attempts to overcome this problem through supplemental data, such as clinical interviews, questionnaires[15,16], or family history[17], improve specificity but are only available for subsets of participants.

Despite these limitations, the richness of EHR data also provides opportunities for more refined phenotyping[18]. Many biobanks contain detailed longitudinal prescription records, making it possible to reconstruct medication-use patterns across the population and enabling research on drug response, including pharmacogenetics[19–21]. Drug response research is particularly valuable for drugs with major clinical and societal implications, such as opioids. The sharp increase in opioid prescribing in the late 1990s, and the subsequent rise in misuse, opioid use disorder, and overdose deaths, highlighted the need to disentangle the factors driving prescription-derived opioid exposure[22,23]. Although opioids remain essential for managing severe pain, they also pose substantial risks, ranging from sedation and respiratory depression to dependence[24,25]. Individuals differ greatly in their cumulative opioid intake, and understanding what contributes to this variability, including genetic contributions, has become an important scientific and public-health concern.

Recent work has increasingly turned to machine learning (ML) to address the limitations of noisy EHR-derived phenotypes and to improve pharmacogenetic analysis[26]. ML models can integrate thousands of heterogeneous predictors, capture nonlinear structure, and exploit correlations across traits. These capabilities make them well-suited to refine or reconstruct phenotypes[16], turning incomplete or noisy case-control labels into more accurate and often continuous estimates of disease liability[27,28]. ML-based imputation approaches restore missing information at scale[27,29], effectively expanding sample sizes and uncovering associations that would otherwise remain undetected. ML has also been used to detect underdiagnosed conditions, extract subtle physiological signatures from raw clinical measurements, and generate refined trait values that enhance genetic discovery[16,30]. Models that integrate comorbidities, symptoms, biomarkers, or family history further reduce misclassification and improve locus discovery[30,31]. Finally, ML-derived risk models, which are trained directly on EHR data, often surpass traditional clinical predictors and standalone PGS, and combining ML-refined phenotypes with genetic information yields more accurate and robust predictions[2,32].

Building on recent advances in ML, this thesis develops a framework for improving genetic prediction through cross-biobank evaluation, phenotype refinement, and ML. Study I systematically assesses how predictive performance varies across biobanks and PGS methods, quantifying the sources of variability and identifying modeling strategies that generalize best across data sets. Study II focuses on phenotype quality, showing that refining psychiatric traits through stringent control definitions and ML-based imputation enhances genetic signal, biological specificity, and predictive transferability. Study III extends these concepts to prescription data, constructing harmonized opioid intake phenotypes across biobanks and using ML to derive refined overuse traits that capture latent behavioral and pharmacogenetic patterns.

Together, these studies first identify limits in genetic prediction, then examine how phenotype refinement reshapes those limits, and finally apply ML models to gain further insights. They illustrate that advancing prediction in human genetics requires both understanding sources of heterogeneity and building more accurate phenotypes within clinical data. Through this integrated framework, the thesis outlines an approach to genetic prediction that is more robust, interpretable, and transferable across diverse biobank datasets.

2 Literature Review

2.1 Human Genetics and Large-Scale Biobanks

Genetics forms the language through which life encodes, transmits, and reshapes information across generations. It provides the foundation for all analyses discussed in this thesis, from the identification of genetic associations to the prediction of disease risk. Therefore, I provide a brief introduction to the molecular basis of heredity and how genetic variation contributes to diversity in human traits.

2.1.1 Molecular Basis of Genetics

DNA and the Molecular Basis of Inheritance

At the core of all life stands deoxyribonucleic acid (DNA), the molecule that stores the genetic instructions of every organism. DNA consists of four chemical bases: adenine (A), thymine (T), cytosine (C), and guanine (G)**,** which pair specifically (A with T, C with G) to form the rungs of a double-helical ladder. Each repeating unit of a base attached to a sugar and a phosphate group forms a nucleotide. During cell division, the DNA double helix unwinds and each strand serves as a template for the synthesis of a complementary strand, ensuring that genetic information is copied according to the original sequence. Nevertheless, occasional replication errors, for instance, introduce permanent changes in the DNA sequence, providing the raw material for genetic variation. Inside the cell nucleus, DNA is compacted by winding around small structural proteins called histones. Together, DNA and histones form nucleosomes, which coil into chromatin and ultimately into the familiar thread-like chromosomes. Humans possess 23 pairs of chromosomes, 22 pairs of autosomes and one pair of sex chromosomes (XX or XY), which together contain the complete genome, as the full set of genetic instructions required to build and maintain the human body[33] (Figure 1).

Genes and Gene Expression

A gene is a specific segment of DNA that contains the information needed to produce a protein or a functional ribonucleic acid (RNA) molecule. RNA is a single-stranded nucleic acid transcribed from DNA that serves as an intermediate or regulatory molecule in gene expression. The DNA sequence of a gene is read in triplets of bases called codons, each corresponding to one amino acid, the building blocks of proteins. Gene expression occurs in two main steps: transcription and translation. During transcription, a complementary messenger RNA (mRNA) copy of the gene is made. During translation, the mRNA sequence is read by ribosomes, which assemble the corresponding chain of amino acids into a protein that folds into its functional shape (Figure 1). Although only ~1-2 % of the genome directly encodes proteins[34,35], the remaining non-coding DNA has essential roles in regulating gene activity. Regulatory elements such as promoters, enhancers, and silencers control when and where a gene is active, while various non-coding RNAs fine-tune expression after the transcription[36]. Additional control occurs through epigenetic modifications, chemical changes

to DNA or histones, such as methylation or acetylation, that influence gene accessibility without altering the DNA sequence itself. These mechanisms ensure that cells with identical DNA can express distinct sets of genes, allowing for the specialized tissues and functions that characterize multicellular organisms[37,38].

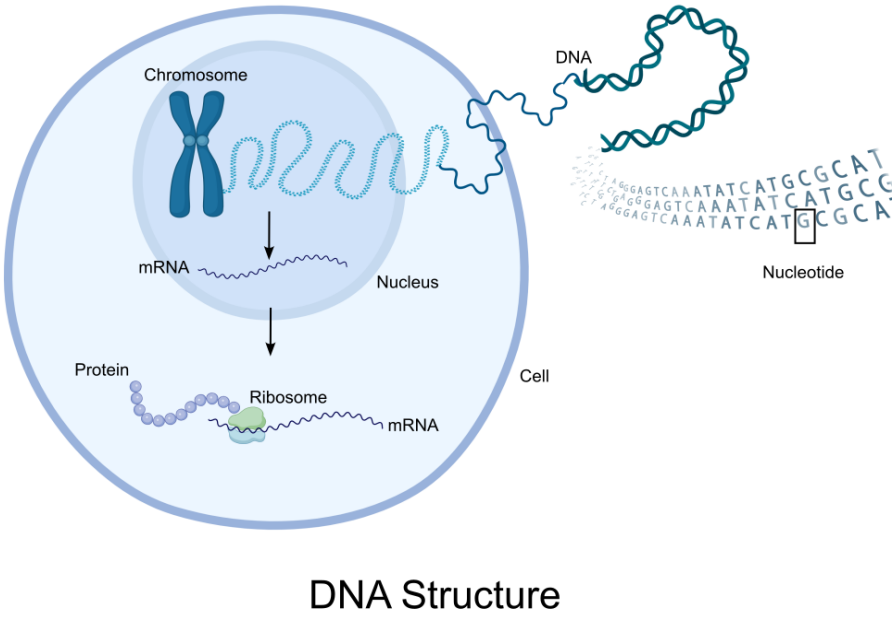


Figure 1 | DNA structure and gene expression.
DNA is organized into chromosomes within the cell nucleus. Specific DNA sequences are transcribed into mRNA, which is exported from the nucleus into the cytoplasm. There, ribosomes translate the mRNA sequence into a protein. The DNA double helix is composed of nucleotides, each consisting of a base, sugar, and phosphate group, and the linear nucleotide sequence encodes genetic information (Created with Biorender.com).

Genetic Variation and Mutation

While the human genome is 99.9 % identical among individuals, the remaining 0.1 % accounts for nearly all biological diversity[39,40]. Variation arises through mutations, which may occur spontaneously during DNA replication or be triggered by external factors such as radiation or toxins. The simplest and most common form of variation is the single-nucleotide polymorphism (SNP), a change in one base at a specific genomic position. Each position in the genome is referred to as a locus, and alternative versions of the sequence at that position are called alleles. An individual's combination of alleles across loci constitutes their genotype. Other types of variation include small insertions or deletions (indels) of nucleotides, and larger structural variants such as copy-number variations, inversions, or translocations of DNA segments[40]. Mutations can arise in different cellular contexts, most commonly in germline cells or in somatic cells. Germline mutations happen in reproductive cells (egg or sperm) and can be passed to offspring, making them heritable. Somatic

mutations arise in non-reproductive cells during an individual's lifetime and are not inherited, although they can cause local effects such as cancer. Over generations, the accumulation of germline variation contributes to genetic diversity within and between populations, forming the basis of evolution and population genetics[41,42].

2.1.2 The Genetic Architecture and the Biobank Revolution

The Genetic Architecture of Complex Traits

The phenotype of an organism refers to its observable characteristics, from physical traits, like height and eye color, to psychological and behavioral features. Phenotypes arise from interactions between an individual's genotype and environmental influences[43]. The Mendel's pea plant experiments first demonstrated that some traits follow simple, predictable inheritance patterns controlled by single genetic variants[44,45]. If a trait is caused by a mutation in a single gene, it is referred to as a monogenic trait. However, most human traits are complex or polygenic, shaped by the combined influence of thousands of variants of small effects together with environmental and lifestyle factors[46,47]. To quantify the genetic contribution to such traits, researchers use heritability, which measures the proportion of phenotypic variation attributable to inherited differences within a population[48].
Gene-environment interactions further illustrate how exposures such as diet, stress, or medication can modify genetic risk, helping explain why individuals with similar genotypes may experience different health outcomes[49]. A distinctive feature of complex traits is pleiotropy, where one genetic variant affects multiple traits. In vertical pleiotropy, one variant affects a cascade of traits through causal relationships, while in horizontal pleiotropy, it independently influences distinct traits. Some variants even show antagonistic pleiotropy, where the same allele can be beneficial in one context but harmful in another[50]. Recognizing these overlapping effects is essential for interpreting shared genetic signals across diseases. Complex traits are also shaped by interactions between genetic variants. This phenomenon, known as epistasis, occurs when the effect of one variant depends on another, making their combined impact non-additive. Such interactions can amplify, diminish, or otherwise alter genetic effects and add an additional layer of complexity to the architecture of human traits[51].

Population Genetics Characteristics

Genetic variation takes many forms, but what matters most for association studies is how variants are distributed across populations and structured along the genome. Variant frequency reflects evolutionary history: common variants (minor allele frequency ≥ 1% (MAF)) tend to be older and shared across populations, whereas rare variants are often younger and restricted to specific ancestries. These patterns arise from demographic and evolutionary forces such as mutation, genetic drift, and natural selection[52]. Under a set of idealized conditions, common variants are expected to remain common in a population, and the relationship between allele frequencies and genotype frequencies follows a predictable pattern. If a population is large, randomly mating, and unaffected by mutation, migration, selection, or genetic drift, allele frequencies remain constant across generations and genotype frequencies remain stable. This state is known as Hardy-Weinberg equilibrium (HWE)[33]. Population history further shapes allele frequencies. Bottlenecks, migration, and long-term isolation can enrich particular alleles or haplotypes (blocks of SNPs inherited together), producing distinctive genetic signatures. Classic founder effects, as seen in

Finland, have increased the frequency of variants that are extremely rare elsewhere, contributing to the “Finnish disease heritage”[53]. Variation is also organized along the genome[54]. Because recombination is uneven, nearby variants are inherited together, generating blocks of correlated alleles. This linkage disequilibrium (LD) reflects the non-random association of alleles at neighboring loci and leads to haplotypes, which differ across populations due to demographic history and recombination landscapes[33].

Human Genetics Analysis in Large-Scale Biobank Studies

Advances in the understanding of genetic variation and inheritance enabled the transition toward large-scale, population-based biobank studies. In contrast, earlier human genetics research relied on hypothesis-driven study designs. These studies focused on narrowly defined case cohorts, family-based analyses, and small clinical samples. They emphasized deep phenotyping through detailed clinical assessments tailored to specific biological hypotheses and typically targeted individual genes or variants[55]. However, this approach was limited by small sample size, low statistical power, and reduced generalizability. A turning point came with the completion of the Human Genome Project in 2003[35,56], which produced the first reference genome after more than a decade of international collaboration and continues to be refined by the Genome Reference Consortium[57] and the 1000 Genomes Project[58]. In parallel, the International HapMap Project cataloged common haplotypes enabling efficient interrogation of genetic variation across the genome[59,60]. Together, these foundational resources made it possible to move beyond targeted genetic testing toward GWAS. These developments enabled hypothesis-free approaches: rather than testing a limited number of biologically motivated variants, hypothesis-free studies systematically scan the entire genome without prior assumptions about which genes are involved[55].

Lower genotyping and sequencing costs have enabled national and population-level initiatives such as biobanks to collect DNA at scale and link it to routine healthcare data. These resources are typically organized as cohorts, groups of individuals sharing key characteristics such as ancestry or recruitment setting. Because whole-genome and whole-exome sequencing (WGS/WES) are still limited by cost, most biobanks instead rely on high-density SNP arrays that genotype participants at up to one million variants. These arrays are designed to capture GWAS backbones through SNPs selected based on LD, common variants with high MAF, population-specific variants, known disease-associated loci, and functional sites, within technical assay constraints[61]. To ensure data quality, biobanks apply extensive genotype quality-control procedures. These procedures include filtering variants that deviate from Hardy-Weinberg equilibrium, removing very rare variants based on MAF, and excluding variants with low call rate. In addition, batch or plate effects are identified, which occur when genotype distributions differ systematically across processing workflows[62]. Beyond directly genotyped variants, biobanks expand genomic coverage through imputation, using population-matched reference panels to infer millions of untyped variants based on local LD structure. Accurate imputation depends critically on matching allele frequencies, LD patterns, and population structure between the study cohort and the reference panel[63,64]. Genotyped individuals are then linked to their corresponding EHR, which provide the phenotypic information used in genetic analyses. EHR contain, for example, longitudinal clinical data including diagnoses, medications, laboratory results, procedures, and healthcare encounters.

The systematic collection of genomic data alongside standardized longitudinal health records has enabled the emergence of large population biobanks as a central infrastructure for human genetic research. Cohort studies such as FinnGen[54,65], EstBB[66,67], UKBB[68,69], the Trøndelag Health Study (HUNT)[70], and Genes & Health (GNH)[71]. Whereas early biobanks mainly connected genotypes to EHR data, modern biobanks increasingly expand and improve their data resources, with some providing deep phenotyping through anthropometric, biomarker, questionnaire, and imaging data[72,73]. Although biobanks now share similar core data, each cohort retains unique features that preserve its scientific value. FinnGen is built around nationwide Finnish health registries and reflects Finland's unique population history, including population bottlenecks and founder effects[54]. These features lead to elevated frequencies of certain genetic variants, increasing power to detect their phenotypic effects. The EstBB represents a high-coverage national cohort, with participation approaching 20% of the adult population, supporting analyses that benefit from broad population representativeness[66]. UKBB combines large sample size with deep phenotyping, including detailed baseline assessments, longitudinal follow-up, lifestyle and environmental data, large-scale imaging, and comprehensive primary care and hospital records[68]. Ongoing WGS further expands its value as a reference resource for genetic research[69]. The HUNT Study adds a longitudinal and multigenerational perspective, with repeated survey waves, family-linked participants, and long-term follow-up enabling the study of disease trajectories, aging, and biomarker dynamics[70]. In contrast, GNH focuses on British Pakistani and British Bangladeshi populations, increasing ancestry diversity and enabling genetic analyses in populations shaped by distinct demographic histories[71].

2.1.3 Genetic Analysis Approaches

With access to hundreds of thousands of genotyped individuals linked to routine clinical data, human genetics has transitioned from small, hypothesis-driven studies focused on individual genes toward large-scale, population-based biobank research. This shift enabled hypothesis-free approaches, in which the entire genome is systematically scanned for associations using broad, shallow phenotypes derived directly from biobanks.

Genome-Wide Association Study (GWAS) and Meta Analysis

A central tool in this setting is the GWAS, which scans the genome for variants associated with a trait. GWAS are typically performed either in a binary case-control framework, for example comparing individuals diagnosed with and without depression[74–76], or by associating genetic variants with continuous quantitative traits such as height[77]. Linear regression is used for continuous outcomes and logistic regression for binary outcomes. To reduce confounding bias, GWAS models are adjusted for common covariates including age, sex, population structure, and technical factors such as genotyping batch. Adjustment refers to the inclusion of such additional variables in the regression model to account for sources of variation that could otherwise distort the relationship between a genetic variant and the trait. By controlling for these covariates, the analysis isolates the specific contribution of each SNP and reduces false associations arising from systematic differences between individuals. Each tested variant yields an effect size and a p-value, which quantifies the statistical evidence against the null hypothesis of no association. Because millions of variants are examined simultaneously, GWAS are highly susceptible to multiple testing bias and therefore require stringent correction. The conventional threshold of $p < 5 \times 10^{-8}$ approximates a

Bonferroni correction for the effective number of independent genomic tests[78]. Results are typically visualized using Manhattan plots, which display genomic association peaks, and Quantile-Quantile plots, which assess deviations from the null distribution. To date, this framework has led to thousands of published GWAS and forms the foundation of modern statistical human genetics[79,80] (Figure 2).

Meta-analysis is a central component of genome-wide discovery, combining GWAS summary statistics from multiple cohorts to increase power, stabilize effect-size estimates, and enable analyses that would be underpowered in individual studies[81]. Because it operates on harmonized summary statistics rather than individual-level data, meta-analysis facilitates large-scale collaboration across cohorts and biobanks with differing privacy constraints and study designs. Fixed-effect models assume a shared underlying genetic effect across studies, whereas random-effects models allow effects to vary, capturing heterogeneity related to ancestry, phenotype definition, or cohort characteristics[79,81]. Tools such as METAL implement GWAS meta-analysis by aligning alleles and combining effect estimates using inverse-variance or sample-size weighting, while also quantifying between-study heterogeneity using statistics such as Cochran's Q[82].

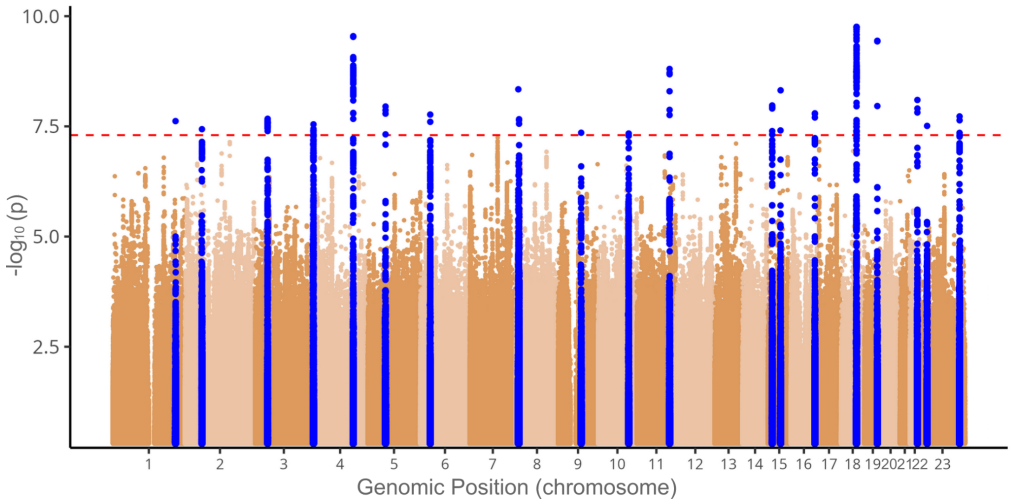


Figure 2 | Example Manhattan plot illustrating GWAS results.
The x-axis shows genomic position across chromosomes, and the y-axis shows the negative logarithm of the association p-value. Peaks indicate genomic loci with strong evidence of association; the horizontal line denotes the genome-wide significance threshold.

Heritability and Genetic Correlation

GWAS can provide insight into the overall genetic architecture of a trait. At the population level, the combined influence of all measured common genetic variants is summarized by SNP-based heritability (h^2), which reflects the proportion of phenotypic variation attributable to common SNPs. The remaining variation may arise from environmental factors, rare variants, measurement error, or other influences. Methods such as LD score regression (LDSC) estimate SNP heritability directly from GWAS summary statistics by exploiting the

relationship between LD score and test statistic inflation to distinguish polygenic signal from confounding bias[48].

Partitioned heritability quantifies how total SNP-based heritability is distributed across different functional regions of the genome, including coding regions, regulatory elements, and cell type-specific epigenomic annotations. Rather than providing a single genome-wide estimate, this approach identifies which biological layers of the genome contribute disproportionately to genetic risk. By testing whether specific annotations explain more heritability than expected given their genomic size, partitioned heritability highlights relevant tissues, cell types, and regulatory mechanisms underlying complex traits[83,84].

The same genome-wide framework can be extended to compare traits. Genetic correlation (r_g) measures the extent, to which two traits share underlying genetic influences: positive correlations indicate shared risk-increasing alleles, whereas negative correlations reflect opposing genetic effects. LDSC provides a robust estimator of genetic correlation using only GWAS summary statistics, and large-scale analyses have revealed extensive shared genetic architecture across many complex phenotypes, particularly among psychiatric traits, highlighting the pervasive role of pleiotropy[48,85].

Fine-Mapping

Due to LD, SNPs that are physically close within a locus tend to be correlated with one another and therefore often show association signals together in GWAS. As a result, SNPs that reach genome-wide significance are not necessarily causal themselves but may instead tag the true causal variant through LD. To identify which variants are most likely to be truly causal, fine-mapping methods have been developed that leverage detailed LD structure to disentangle correlated association signals. Most fine-mapping approaches compute posterior inclusion probabilities for each variant and derive credible sets that contain the true causal variant with high probability[86,87]. SuSiE, the method used in this thesis, extends traditional fine-mapping by explicitly modeling multiple independent causal effects within a region, offering improved robustness in the presence of strong LD and producing separate credible sets for each effect. This approach makes it particularly well suited for complex loci with multiple causal signals[88,89].

Polygenic Score

Finally, findings from GWAS can be aggregated into PGS, which combine thousands of effect-weighted variants into an individual-level measure of genetic liability. PGS quantifies relative, not deterministic, risk and is widely used to evaluate how well genetic information can predict complex traits[90,91]. PGS methods differ primarily in how they select variants, model LD, and shrink effect sizes. Simple clumping-and-thresholding approaches, such as pt + clump, retain approximately independent SNPs based on LD and association significance and are computationally efficient, but may discard many variants with small polygenic effects. Penalized regression methods, including lassosum[92], apply regularization to jointly estimate SNP weights, improving stability in the presence of correlated variants but requiring careful parameter tuning. Specific tools such as LDpred2[93], PRS-CS[94], DBSLMM[95], SBayesR[96] and MegaPRS[97], explicitly model LD and assume prior distributions on effect sizes, often yielding higher predictive accuracy at the cost of increased computational complexity. More recently, ML-based approaches aim to capture non-linear effects and

interactions between variants, although their interpretability and generalizability remain active areas of investigation[6,7,98].

2.1.4 Bias and Limitations in Large Biobank Data

Human populations are diverse, diagnoses can be imperfect, and EHR were primarily created for clinical rather than research purposes. These factors make case-control studies vulnerable to systematic distortions, known as biases, which can obscure true biological signals. In particular, four sources of bias and limitations are especially relevant: methodological limitations, information bias, selection bias, and confounding bias.

Methodological Limitations of GWAS

Methodological limitations arise from constraints imposed by both data generation and analytical modeling. In biobank-based studies, such limitations already emerge at the level of genetic data generation, before statistical analysis is performed.

Many modern biobanks rely still primarily on SNP-based genotyping arrays rather than WGS. While this approach is cost-effective and scalable, it captures only a fraction of genomic variation, mainly common SNPs and selected coding regions, leaving most regulatory and structural variation unobserved[61]. Because SNPs included on these arrays are selected based on prior discovery in limited populations, this design introduces ascertainment bias and preferentially captures common variants while underrepresenting rare variation[99,100]. As genotyping arrays directly measure only a limited number of variants, most remaining genetic variation is inferred through imputation using reference panels, with accuracy depending strongly on ancestry matching between the study cohort and the reference population[101]. In contrast, WES and WGS directly sequence individual genomes, reducing reliance on preselected markers and enabling detection of rare and structural variants. Although increasingly used, WES and WGS remain constrained by cost and computational demand[69]. As a result, many rare and structural variants, including copy number variations, inversions, and translocations, remain effectively invisible in standard GWAS.

Beyond data generation, GWAS are subject to additional methodological constraints arising from their analytical framework. GWAS tests one genetic variant at a time under an additive model, assuming additive and independent effects across the genome[78,102]. While this framework has enabled the discovery of thousands of robust associations, it does not capture dominance effects, rare variants, structural variation, gene-gene interactions, gene-environment interactions, or other non-linear relationships, collectively referred to here as non-additive genetic effects, particularly for highly polygenic traits[78,103]. Moreover, many associated variants lie in non-coding regions, and statistical fine-mapping often depends strongly on population-specific LD patterns, introducing uncertainty in the identification of truly causal variants[104,105]. Consequently, GWAS is highly effective at detecting statistical associations but remains limited in resolving biological causality.

These methodological constraints are reflected in heritability estimation. Heritability describes the proportion of phenotypic variation in a population attributable to genetic differences between individuals. GWAS-based methods estimate narrow-sense heritability,

capturing only the additive genetic variance tagged by common SNPs, while non-additive genetic effects fall largely outside their scope[106]. In contrast, twin and family studies estimate broad-sense heritability, which encompasses the full spectrum of genetic contributions but relies on assumptions about shared environments. The systematic gap between SNP-based and family-based estimates is referred to as missing heritability and is particularly pronounced for highly polygenic complex traits[107].

For the same reasons that GWAS and SNP-based heritability capture only a restricted subset of genetic variation, PGS explain only a modest fraction of total heritable risk[108,109]. In addition, PGS are subject to substantial methodological variability. Different construction methods vary in how they select variants, model LD, and shrink effect-size estimates, such that the same individual can receive meaningfully different risk estimates depending on the applied method[6,90,110]. A second major limitation concerns limited cross-population portability. Most GWAS underlying current PGS are conducted in cohorts of predominantly European ancestry. Differences in allele frequencies, LD structure, environmental exposures, and trait architecture across populations lead to marked reductions in predictive performance when PGS are transferred across cohorts, and especially across ancestries[111].

Information bias

Information bias arises from systematic errors in how exposures, diagnoses, or traits are measured and recorded. In biobank-based genetic studies, this bias is pervasive because phenotypic data are derived from routine clinical care rather than generated in controlled research settings. As a result, EHR-derived phenotypes reflect clinical workflows designed for diagnosis, treatment, and billing, not for standardized phenotypic ascertainment.

This mismatch introduces several interconnected sources of error. Early-stage, undiagnosed, or misdiagnosed individuals may initially be classified as controls, while incomplete medical histories, inconsistent coding practices, and delayed diagnoses blur the boundary between cases and controls[9,10,18,112]. Such misclassification is not random but systematically related to healthcare access, disease severity, and help-seeking behavior, thereby distorting case-control contrasts and attenuating genetic signal.

Information bias is further amplified in cross-biobank and international analyses by differences in healthcare systems and coding conventions. Although structured classification systems such as the International Classification of Diseases (ICD) for diagnoses and the Anatomical Therapeutic Chemical (ATC) system for medications provide a shared framework, their implementation varies substantially across countries. Differences in reimbursement rules, diagnostic thresholds, and clinical practice mean that the same ICD code can correspond to different disease severities, stages, or diagnostic certainty across biobanks. Beyond these coding systems, different countries can additionally have their own diagnostic practices and nomenclature. For example, UK healthcare records frequently rely on Systematized Nomenclature of Medicine (SNOMED) or British National Formulary (BNF)-based medication classifications, which do not map cleanly onto ICD or ATC systems. As a consequence, each biobank effectively operates within its own EHR ecosystem, where nominally identical codes may represent distinct underlying clinical realities[9,113,114].

Selection Bias

Selection bias arises when the individuals included in a study do not adequately represent the population to which the results are intended to generalize. In biobank-based genetic studies, this bias often emerges because participation and inclusion are not random: study design, ancestry composition, population structure, and healthcare access influence who enters the cohort and how frequently conditions are recorded. As a result, disease prevalence, allele frequencies, and environmental exposures can differ systematically from those in the broader population[9].

In practice, selection bias in biobanks primarily reflects non-random patterns of recruitment and participation that shape who enters a cohort and how diseases are observed. The UKBB is enriched for healthier, more health-conscious and more socioeconomically advantaged participants[115–117]. FinnGen over-represents women and hospital-treated individuals, partly due to recruitment through maternity and hospital registries[65,118]. Participant composition in the EstBB also differs from the national population, with overrepresentation of women and younger individuals and underrepresentation of the Russian minority[67]. These differences influence disease prevalence, environmental exposures, and comorbidity profiles, shaping the associations detected in each cohort. A special form of selection bias is called survivorship bias, whereby individuals with early-onset or severe disease are less likely to be recruited into biobanks. This bias leads to an underrepresentation of the most affected cases and can result in underestimation of genetic risk, for example among individuals with early-onset severe psychiatric disorders[119–121].

Beyond recruitment mechanisms, selection bias is further compounded by the strong overrepresentation of individuals of European ancestry in most large-scale genetic biobanks. This overrepresentation reflects historical investment patterns in genomic research and the high cost of large-scale sequencing, which still limits representation from other global populations[115,122,123]. As a result, FinnGen, UKBB, EstBB, and HUNT all capture primarily European genetic backgrounds, while GNH provides an important, but comparatively smaller, contribution from British Pakistani and British Bangladeshi communities. This imbalance means that most genetic discoveries are optimized for European allele frequency spectra and LD structures. Consequently, associations that replicate well within these cohorts may not hold in ancestries that remain underrepresented, and PGS trained in these biobanks often show steep drops in predictive accuracy when applied to other populations[111].

Confounding Bias

Confounding Bias refers to a situation in which an observed association between two traits is driven by a third factor that is related to both. A simple illustration is the apparent association between carrying a lighter and lung cancer: without accounting for smoking, one might falsely infer a causal relationship. In genetic biobank studies, confounding can distort effect estimates when additional risk factors are unevenly distributed across the study population and are not adequately measured or adjusted for. Environmental and lifestyle factors such as socioeconomic status, education, smoking behavior, or healthcare access may vary systematically across participants and influence recorded disease risk. These factors are often incompletely captured in the EHR, allowing residual confounding despite statistical adjustment. Confounding can further arise from clinical pathways. Individuals with certain

traits may have more frequent healthcare contact, increasing the likelihood of receiving diagnoses recorded in the EHR. This can create apparent associations between genetic markers and disease that reflect healthcare utilization rather than biology[124]. To mitigate confounding, analyses typically adjust for covariates such as age, sex, and ancestry components, although unobserved confounders and complex interactions can leave residual bias.

In Study I of this thesis, I directly interrogate two key sources of uncertainty in polygenic prediction: methodological variability across PGS construction methods and limited portability across biobanks and ancestries. I systematically compare multiple state-of-the-art PGS approaches across five independent cohorts spanning two ancestries, allowing us to disentangle method-dependent performance from population-driven effects, identifying which PGS methods remain robust under these real-world cross-biobank challenges.

2.2 Genetics of Psychiatric Traits

EHRs have enabled the large sample sizes required for modern psychiatric GWAS, but they also shift research from deep to shallow phenotyping. Deep phenotypes rely on structured diagnostic interviews, validated assessments, and screened controls, individuals actively evaluated to rule out the disorder of interest. This approach yields highly specific case-control definitions but is typically feasible only in relatively small cohorts. In contrast, shallow phenotypes rely on routine clinical codes and classify controls as individuals without a recorded diagnosis (Figure 3).

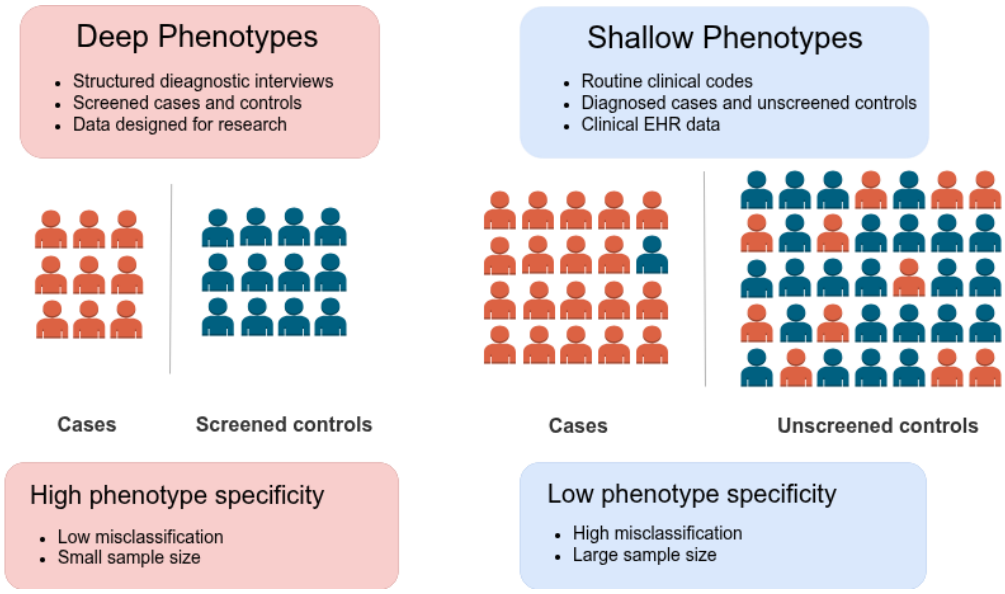


Figure 3 | Deep versus shallow phenotype definition
Schematic comparison of structured, research-based phenotyping with screened controls (left) and biobank- or EHR-based phenotyping using routine clinical codes and unscreened controls (right).

For psychiatric traits, this assumption is particularly weak. Many symptomatic individuals remain undiagnosed, are coded inconsistently, enter clinical care only later in the disease course, or encounter other sources of misclassification that further contribute to this problem. As a result, shallow psychiatric phenotypes are especially susceptible to diluted case-control contrasts and reduced genetic signal[125,126].

2.2.1 Diagnostic Challenges in Psychiatry

Psychiatric disorders possess several intrinsic clinical features that complicate their use in genetic studies, independent of the biases introduced by EHR data. A foundational challenge is the absence of objective biomarkers. Diagnosis relies on clinician judgment, patient self-report, and behavioral observation, processes that vary across clinicians, cultures, and health systems[127]. Without biological anchors, psychiatric nosology has evolved repeatedly over time, and revisions to diagnostic criteria (e.g., Diagnostic and Statistical Manual of Mental Disorders updates) mean that the same condition may be defined differently across decades[128,129]. These issues are particularly relevant for long-running health registries. For example, FinnGen includes diagnostic codes dating back to the 1970s[65], a period during which diagnostic criteria, coding standards, and clinical norms have changed substantially.

These subjective diagnostic processes contribute directly to symptom heterogeneity. Individuals with the same diagnosis often markedly differ in symptom profile, severity, and course. Many conditions span a continuum of liability rather than forming discrete categories, making binary case-control definitions an imperfect reflection of underlying biology[130]. Diagnostic boundaries are similarly porous. High rates of comorbidity, for example, between depression and anxiety, or between substance use and other psychiatric disorders, reflect overlapping symptom domains and shared etiological pathways. This intrinsic overlap complicates attempts to isolate disorder-specific diagnosis[131].

Psychiatric diagnoses are also characterized by diagnostic instability. Symptoms often evolve over time, individuals may transition between diagnostic categories, and age of onset and chronicity vary widely. Such temporal variability reflects genuine clinical trajectories rather than measurement error, yet it produces inconsistencies in the mapping between clinical labels and the underlying disease processes relevant for genetic analysis[128,129].

Finally, stigma and help-seeking behavior introduce additional layers of complexity. Individuals may delay or avoid clinical evaluation because of societal stigma or limited access to mental health care. This behavioral component means that diagnostic absence does not reliably indicate true absence of illness, which is an inherent challenge for any phenotype definition based on clinical diagnosis[126].

2.2.2 Genetic Architecture of Psychiatric Traits

These challenges motivated the establishment of large-scale collaborative efforts that aim to reduce diagnostic noise through harmonized methods, standardized analyses, and very large sample sizes. The Psychiatric Genomics Consortium (PGC)[132], established in 2007, has been central to this effort. The PGC coordinates data sharing, unified quality control, and

meta-analysis across hundreds of cohorts worldwide, enabling the largest GWAS for schizophrenia[133], anxiety[134], bipolar disorder[135], major depression[75], Attention-Deficit/Hyperactivity Disorder (ADHD)[136], Obsessive-Compulsive Disorder (OCD)[137], Post-Traumatic Stress Disorder (PTSD)[138], and substance use disorders[139,140]. By aggregating clinical, population-based, and self-reported cohorts, the PGC achieves unprecedented statistical power. A defining feature of progress in psychiatric genetics has been the steady expansion of GWAS sample sizes. Because psychiatric disorders are highly polygenic, with many variants of small effect, early studies with modest sample sizes identified few robust associations. As consortia matured and biobanks expanded, larger datasets provided sufficient power to detect these effects, leading to a marked increase in the number of genome-wide significant loci across nearly all psychiatric traits.

This trajectory is remarkably consistent across psychiatric disorders. Early GWAS for psychiatric traits like schizophrenia, OCD or anxiety disorders identified few, if any, robust associations. As sample sizes expanded through successive consortium meta-analyses, locus discovery increased steadily, with recent studies now reporting dozens to hundreds of genome-wide significant loci[133,137,141]. Importantly, there is no evidence that this growth is slowing. Even at current sample sizes, additional participants continue to yield new associations, suggesting that genetic risk remains distributed across many variants with progressively smaller effects. This pattern reflects the fundamentally polygenic architecture of psychiatric disorders. Rather than being driven by a small number of large-effect variants, liability arises from the cumulative contribution of thousands of common variants, each with minimal effects. Such architectures place inherent limits on locus discovery, interpretation, and prediction, and help explain why even the largest GWAS capture only a fraction of SNP-based heritability[78]. While incorporation of rare variants through whole-genome sequencing has been shown to partially reduce this gap for some well-measured polygenic traits, such effects appear weaker for behavioral phenotypes and have not yet been systematically evaluated in psychiatric disorders[142]. Polygenicity therefore implies that locus discovery does not saturate with increasing sample size, and that continued gains in power primarily reveal ever-smaller effects rather than exhaust the underlying genetic architecture. It was also suggested that reaching 90% heritability in highly polygenic traits would require tens of millions of participants[143]. Large-scale studies provide concrete illustrations of this diffuse genetic architecture. The insomnia GWAS identified 554 loci in over 2.3 million individuals, yet these loci together explained only 17.3% of SNP-based heritability[144]. OCD shows similar features: despite the identification of 30 genome-wide significant loci, MiXeR modeling estimates that approximately 11,500 common variants are required to account for 90% of its SNP-based heritability[137].

Because psychiatric disorders are highly polygenic, it is unsurprising that many genetic variants contributing to risk for one condition also influence other psychiatric traits. This phenomenon, known as pleiotropy, has emerged as a defining feature of psychiatric genetics and is recognized as pervasive across the genome[145]. Rather than representing isolated genetic architectures, most psychiatric disorders share substantial portions of their underlying genetic liability, reflecting overlapping biological processes and developmental pathways[146,147].

A growing number of methods have been developed to explicitly model pleiotropy and characterize this shared architecture. Approaches such as genomic structural equation

modeling[148] and MiXeR[149] allow decomposition of genetic signal into shared and trait-specific components, enabling more accurate estimation of genetic correlation and overlap. These methods have demonstrated that many psychiatric disorders share a large fraction of their causal variants, even when genome-wide significant loci differ, reinforcing the view that psychiatric disease risk is organized along partially overlapping dimensions rather than discrete categories[145]. Recent large-scale genetic studies applying these approaches have highlighted the extent of shared genetic architecture across psychiatric traits. Studies consistently show high genetic correlations, indicating that common variants contribute jointly to liability for related disorders. This overlap is particularly pronounced among internalizing disorders such as anxiety and depression[147], but extends more broadly across psychiatric disease categories. As GWAS sample sizes have grown, these shared signals have become increasingly apparent, revealing that apparent diagnostic boundaries often do not correspond to distinct genetic architectures.

At the same time, the recognition of pervasive pleiotropy has exposed important methodological challenges. Traditional GWAS typically analyze psychiatric disorders independently using case-control designs, implicitly assuming genetic independence between traits. When psychiatric disorders share a substantial proportion of their genetic basis, this assumption becomes problematic. Closely related phenotypes may contribute shared genetic signal and individuals affected by one disorder may be included as controls for another. Such overlap can dilute trait-specific associations and complicate interpretation, highlighting the limitations of standard case-control designs in the presence of widespread pleiotropy[146].

2.2.3 Study Design in Psychiatric Genetics

Since psychiatric disorders were historically known to be difficult to analyse in genetic analysis, different approaches were considered to improve the genetic signal of psychiatric disorders. Early studies with small sample sizes aimed to reduce heterogeneity from comorbidity and pleiotropy by excluding other psychiatric traits from the control group. The idea was that, due to high comorbidity among psychiatric conditions, the presence of related traits in controls would attenuate the case-control signal and obscure true associations. This approach became known as the use of super-normal controls, where, for example, in a study of depression (trait A), individuals with closely related conditions such as anxiety (trait B) were also excluded from the control group[150].

In subsequent work, important critiques emerged. It was noted that if controls are screened more strictly than cases, their relatives will appear artificially healthier, which can falsely suggest familial clustering of disorders A and B even when none exists[151]. Some researchers attempted to address this by excluding trait B not only from the controls but also from the case group to restore balance[152,153]. However, later work showed that excluding trait B from controls can itself introduce bias: it may generate an apparent association between traits A and B even if no true relationship exists[154], and the same logic applies directly to GWAS, where such exclusions can produce false signals for trait A that actually reflect liability to trait B[155].

In addition to biases introduced by control screening, misclassification within the control group can further distort genetic signals. Under certain conditions, removing misclassified

cases from the control group may increase power, despite a reduction in sample size[14]. A comprehensive investigation of both misclassification and super-normal controls, referred to as double-screening controls, showed that both practices can bias genetic analysis[85]. A recent contribution reinforced this conclusion, noting that misclassification and super-normal designs can bias genetic analysis, although the effect is modest when the prevalence of trait B is low[156].

The preceding sections show that although increasing sample sizes has substantially advanced locus discovery in psychiatric genetics, scale alone cannot resolve key challenges arising from diagnostic noise, pervasive polygenicity, and widespread pleiotropy. Misclassification in EHR-based phenotypes remains common, and comorbidity frequently blurs the distinction between cases and controls, limiting the interpretability of standard case-control analyses. In Study II, I therefore investigated how diagnostic misclassification and super-normal control designs interact. Using simulations, I demonstrate that these factors are interdependent and that, under conditions of substantial misclassification, stricter control definitions can partially mitigate misclassification bias. I then evaluated this framework empirically across ten psychiatric disorders in FinnGen.

2.3 The Genetic Link between Psychiatric Disorders and Pain

Pain and psychiatric traits share structural characteristics that complicate their analysis in large-scale biobank data. Both are highly polygenic, clinically heterogeneous, and lack objective biomarkers, such that diagnosis-based phenotypes provide only an incomplete representation of underlying liability. Pain varies substantially between individuals and over time, is influenced by biological, psychological, and contextual factors, and arises from diverse causes affecting multiple anatomical sites[157]. As a result, EHR-based diagnostic codes tend to capture specific conditions or locations rather than overall burden, chronicity, or treatment intensity, while differences in healthcare access and clinical thresholds further contribute to heterogeneity in recorded pain phenotypes[158,159].

Beyond these structural similarities, pain exhibits substantial genetic, epidemiological, and clinical overlap with psychiatric disorders and is among the phenotypes most consistently implicated in shared disease architecture[160,161]. Genetic analyses have shown that, with the exception of joint pain, most pain phenotypes are positively correlated with depression and neuroticism, and the number of shared loci is higher between pain and psychiatric traits than between pain and most other clinical categories[159,160]. Across studies, the strongest and most recurrent overlaps are observed with depression and insomnia, highlighting shared liability rather than coincidental comorbidity[162].

While its overlap with psychiatric traits is substantial, pain also represents a major health burden independent of psychiatric disease. Pain is among the most prevalent and disabling health conditions worldwide. Chronic pain, commonly defined as pain persisting for longer than three months, affects a large proportion of the population and represents a leading cause of years lived with disability[161]. Global Burden of Disease analyses consistently rank pain-related conditions among the top contributors to non-fatal disease burden, with low back pain and migraine ranked first and second globally and neck pain within the top ten[163]. Recurrent tension-type headache alone affects approximately 1.9 billion individuals, making it the most prevalent symptomatic chronic condition worldwide[161]. This data demonstrate that

pain is a leading public health concern independent of its frequent comorbidity with psychiatric disorders.

Pain is also a central driver of healthcare utilization and is managed largely through pharmacological treatment. Analgesic prescriptions are issued repeatedly over time, adjusted to symptom persistence and severity, and systematically recorded in healthcare systems[164]. While non-opioid treatments often provide limited or short-term relief, opioids remain essential for acute, postoperative, cancer-related, and palliative pain[165]. Lifetime opioid exposure varies widely between individuals, ranging from no recorded use to sustained high-dose treatment, for example in the context of cancer pain[166]. Even when appropriately prescribed, opioids carry dose- and duration-dependent risks, including sedation, gastrointestinal complications, respiratory depression, and increased likelihood of prolonged use via central reward pathways[24,25,167]. The liberalization of opioid prescribing in the late 1990s revealed the population-level consequences of these properties, with prescription-driven increases in opioid exposure followed by marked rises in opioid-related morbidity and mortality[22,23,168].

Because of their widespread use and systematic recording, longitudinal opioid prescription data capture treatment intensity and duration in a way that diagnostic labels alone do not. This makes opioids particularly informative as a prescription-derived phenotype, where modest differences in cumulative exposure can reflect substantial differences in clinically relevant pain burden. In Study III, I examined prescription-derived opioid intake as a quantitative phenotype, harmonizing cumulative exposure in FinnGen, UKBB and EstBB and assessing genetic contributions to interindividual variation in opioid use.

2.4 Applications of Machine Learning in Human Genetics

ML methods are widely used in human genetics and underpin several core analytical approaches. As genetic analyses increasingly rely on EHR-linked data, challenges related to noise, missingness, and heterogeneity have become more prominent. In this context, ML provides a flexible framework for integrating large numbers of predictors and modeling complex data structures, making it particularly useful for phenotype refinement and prediction in biobank settings.

2.4.1 Machine Learning Models for Polygenic Scores

PGS construction represents a setting in which a broad range of ML methods are applied within human genetics. Although variant-phenotype associations are estimated using regression-based GWAS, translating these association signals into accurate individual-level scores introduces additional challenges. PGS methods differ primarily in how they model and regularize variant effects when constructing this aggregate score. Threshold-based approaches apply hard selection rules, retaining only variants that exceed a predefined significance threshold and summing their effects after pruning correlated variants. Penalized regression methods instead estimate variant effects jointly, using continuous shrinkage to down-weight weak or redundant signals while allowing many variants to contribute. Bayesian methods adopt a probabilistic framework by placing prior distributions on effect sizes, thereby encoding assumptions about polygenicity and uncertainty and explicitly incorporating LD information. More recent approaches extend this framework by applying general

supervised ML models to the same prediction task. Rather than restricting the relationship between variants and phenotype to a linear form, these methods use models such as tree-based algorithms, neural networks, or ensemble learners to capture more complex dependency structures when supported by the data. Across all these approaches, PGS construction can be understood as a supervised learning problem that seeks to compress high-dimensional genetic information into a single, predictive measure of genetic liability[98].

Beyond its role in PGS construction, ML is increasingly used to complement polygenic scores with additional sources of information in applied prediction settings. Integrating PGS with family history has been shown to improve prediction for 11 complex diseases, with gains varying across diseases and illustrating the complementary nature of genetic and familial risk[169]. In some settings, models that directly leverage detailed clinical data outperform polygenic scores alone. A continuous coronary artery disease risk score, derived using random forest models trained on EHR data from the BioMe[170] and UKBB[68] cohorts, achieved an area under the receiver operating characteristic curve (AUROC) of up to 0.95, outperforming both PGS and traditional clinical risk equations while identifying large numbers of previously undiagnosed cases[171]. This highlights how ML models can capture disease risk that is not fully reflected by genetic variation alone. More general ML frameworks extend risk prediction beyond single diseases by leveraging routinely recorded medical history at scale. Neural network models trained on large health-record datasets can predict risk across thousands of diseases, substantially improving discrimination over baseline predictors and generalizing well to independent cohorts, especially if genetic data are unavailable[172].

Together, these studies show that while PGS summarize genetic liability, predictive performance in applied settings often depends strongly on the availability and resolution of phenotypic information, particularly for highly polygenic traits. In such situations prediction models leveraging EHR records or other additional information can add valuable insight in polygenic traits.

2.4.2 Machine Learning for Phenotype Refinement

Not only PGS methods and predictive applications illustrate the growing role of ML in improving accuracy and transferability, but ML has also been increasingly applied to address inherent limitations of EHR data itself. In particular, challenges such as diagnostic uncertainty, heterogeneous coding, and missingness have motivated the development of ML-based approaches for phenotype construction and refinement.

Imputation of Missing Data and Expansion of Effective Sample Sizes

Missingness is one of the most pervasive limitations of EHR-derived phenotypes and directly reduces the effective sample size available for genetic analyses. Early efforts to address this problem relied on classical multiple-imputation approaches such as MICE[173]. While effective in smaller settings, these methods do not scale to biobank data with thousands of variables and millions of individuals, and they struggle to capture higher-order correlations across phenotypes. To overcome these limitations, more recent approaches have applied ML models that explicitly learn structure across the phenome. AutoComplete introduced a deep-learning-based encoder-decoder framework to simultaneously impute thousands of

missing phenotypes by leveraging cross-phenotype correlations[27]. Applied to UKBB data, this approach substantially increased effective sample sizes across many traits and enabled the discovery of dozens of additional genome-wide significant loci, demonstrating that recovering missing phenotype information alone can markedly improve GWAS power. Building on this idea, PIXANT uses a fast, mixed random-forest framework optimized for large-scale, multi-phenotype EHR imputation[29]. By explicitly targeting extreme sparsity, PIXANT enables scalable imputation of high-dimensional EHR panels and further increases GWAS discovery. Together, these studies show that ML-based imputation is an effective strategy for expanding usable sample sizes in biobank-scale data and for recovering genetic associations that are missed when analyses are restricted to complete or narrowly defined phenotypes.

Disease-Specific Phenotype Refinement Applications

While approaches such as AutoComplete focus on general phenotype imputation across the EHR, many studies instead target the refinement of specific diseases, often leveraging additional domain-specific information to recover more informative phenotypes. Phenotype refinement is particularly important in psychiatric genetics, where diagnostic boundaries are often blurred and objective biomarkers are largely absent. As a result, clinical labels derived from EHR or diagnostic codes can be noisy and heterogeneous, limiting the ability of genetic studies to detect meaningful signal. ML methods provide a way to extract latent structure from such data and to reconstruct phenotypes that better reflect underlying disease liability.

Depression provides a clear example of how EHR data can be leveraged for disease-specific phenotype refinement. Using EHR-derived features such as symptom severity, comorbidities, demographics, and behavioral indicators, ML models have been applied to identify clinically meaningful variation within depression. In one study, a random forest was trained to detect rare psychotic symptoms in depressive disorders, despite severe class imbalance. The model identified clinically relevant predictors, illustrating that ML can recover nuanced psychiatric features that are missed by ICD-based coding alone[174]. Beyond routine EHR data, additional sources of phenotypic information can further refine depression definitions. Self-reported questionnaire data in large biobanks provide a valuable source of information for refining depression phenotypes, although such data are often incompletely observed. In the UKBB, a depression phenotype derived from the LifetimeMDD questionnaire was available only for a subset of participants. This information was leveraged to impute LifetimeMDD responses for individuals who had not completed the questionnaire using AutoComplete, thereby extending the phenotype to the full cohort[16]. This approach effectively doubled the usable sample size and increased the number of GWAS loci from one to up to forty. The imputed phenotype showed very high genetic correlation with the observed LifetimeMDD questionnaire phenotype, and polygenic score accuracy also doubled, highlighting the downstream benefits of ML-assisted phenotype reconstruction. In addition to clinical and self-reported data, biological measurements have also been explored to gain insight into mechanisms of depression. Magnetic resonance imaging (MRI) data, including structural MRI, functional MRI, and diffusion MRI, have been used to capture anatomical and functional variation in the brain with the aim of identifying objective biomarkers for depression. ML methods, including decision-tree-based approaches, have been used to extract imaging features that differ between depression cases and controls, with the goal of improving disease characterization. However, substantial heterogeneity in

imaging modalities, feature definitions, and modeling strategies across studies has limited convergence and hindered the identification of robust biomarkers. Despite these challenges, imaging-based approaches represent an important attempt to gain biological insight into depression[175].

Just as ML-based phenotype refinement has proven valuable in psychiatric disorders, similar approaches have been applied to other diseases characterized by underdiagnosis, heterogeneous presentation, or limited biological characterization. In some settings, ML is used to model disease risk directly from EHR data when conventional case definitions are sparse or incomplete. In endometriosis, an Extreme Gradient-Boosted Decision Trees (XGBoost) model integrated over one thousand clinical, lifestyle, and genetic predictors from EHR-linked biobank data to predict disease risk. SHAP-based interpretation showed that clinical variables dominated prediction, while polygenic scores contributed little, highlighting substantial heterogeneity in prediagnostic risk[176]. In other diseases, ML has been used to recover underdiagnosed cases by reconstructing phenotypes from sparse clinical coding. Stuttering provides a clear example, as affected individuals are often poorly captured in EHR. Using enriched comorbidity patterns, a decision-tree-based model inferred stuttering status in the BioVU cohort, expanding the number of identified cases from 142 ICD-coded individuals to over 9,000 reconstructed cases. This enabled ancestry-specific GWAS, revealed new loci, and supported validation of the resulting polygenic score[30]. Beyond EHR-derived phenotypes, ML has also been applied to redefine disease traits using richer biological measurements. In respiratory disease, a deep-learning framework trained on raw spirograms generated a continuous liability score for chronic obstructive pulmonary disease that outperformed traditional spirometry-based definitions. Using this ML-derived phenotype as the GWAS outcome led to the identification of 356 loci, including 67 novel associations, illustrating how feature extraction from high-resolution physiological data can yield more informative phenotypes than conventional clinical summaries[31]. Collectively, these disease-focused studies demonstrate how ML-based phenotype refinement can expose latent structure in clinical data, rescue underdiagnosed cases, and transform noisy labels into phenotypes that are more informative for genetic discovery and prediction.

Systematic Phenotype Refinement Frameworks

Beyond disease-specific applications, several methodological frameworks aim to improve phenotype definitions in a systematic way by incorporating additional information into genetic analyses. These approaches vary in the data they leverage, the degree to which phenotypes are modelled explicitly, and whether refined phenotypes are used to augment conventional case-control GWAS or to replace them with continuous liability estimates.

One strategy uses additional, curated information to probabilistically refine disease status in controls while retaining binary GWAS frameworks. Using XGBoost ensembles trained on clinical biomarkers, plasma proteins, and covariates, disease-state probabilities were estimated for thousands of traits. These probabilities were then used to define a series of case-control definitions with increasing stringency by reclassifying selected individuals from the control group into cases. This strategy increased effective case numbers and improved locus discovery across 3,213 ICD-10 phenotypes, illustrating how integrating measured biomarkers with EHR-derived information can refine phenotype definitions and enhance genetic power.

A related but more structured line of work is based on explicit liability modeling. LT-FH method models disease liability, conditional on observed case-control status and family history, yielding posterior mean liability estimates for each individual. Individuals are then included in the case cohort if their estimated liability exceeds a population-wide threshold defined by disease prevalence[177]. LT-FH++ extends this framework by incorporating age-, sex-, and birth-year-specific prevalence and allowing the liability threshold itself to vary across individuals. By accounting for heterogeneity in baseline risk, LT-FH++ further refines case-control definitions and yields additional gains in association power in diverse populations[178]. More recent methods further relax predefined modeling assumptions by learning relevant phenotype structure directly from EHR data. LTPI extends liability modeling by automatically selecting correlated traits to include in a multitrait liability model, rather than relying on predefined predictor sets. Using the selected traits, LTPI estimates individual-level liabilities and performs association testing directly on these continuous values. This approach achieves higher power than both standard GWAS and LT-FH in simulations and biobank analyses, but at the cost of increased nonspecific signal in pleiotropy analyses, illustrating a trade-off between power and phenotypic specificity.[179].

At the most data-driven end of this spectrum, several studies extend the liability-based GWAS paradigm by learning continuous phenotypes directly from high-dimensional EHR data without predefined predictor selection. Using models such as XGBoost and deep neural networks trained on large sets of diagnoses, biomarkers, and environmental variables, these approaches estimate individual-level liabilities that are used directly as GWAS outcomes, analogous to LTPI. In contrast to LTPI, however, predictors are included broadly rather than selected based on trait correlations, allowing models to absorb a wider range of signals. While this strategy can increase power and has been shown to improve locus discovery relative to binary phenotypes, it likely also captures more nonspecific or pleiotropic signal. Formal pleiotropy analyses were not reported in these studies, limiting assessment of phenotypic specificity despite gains in association power[28,180].

Limitations of Machine Learning based Models

A key limitation of ML-based phenotype refinement is the potential introduction of noise into downstream GWAS. Predictive phenotypes may improve classification accuracy while capturing genetic signal that is not specific to the target disease. Pleiotropy analyses have therefore been introduced to assess whether ML-derived phenotypes show increased sharing of genetic signal across traits[16], but these are not consistently applied. Studies that do report pleiotropy often find it to be elevated for prediction-derived phenotypes, particularly when continuous liability scores are used and when models are trained on broad predictor sets with limited feature selection[179]. While pleiotropy analyses provide a useful diagnostic, they offer only a coarse indication of noise and do not fully distinguish disease-relevant signal from spurious associations. Methods such as POP-GWAS attempt to address this by evaluating whether associations discovered using refined phenotypes are supported by power considerations from the original binary GWAS[181]. However, this approach implicitly treats the conventional GWAS as a reference standard, despite its own susceptibility to phenotype misclassification. As a result, validation often relies on replication in large external GWAS, limiting the potential for novel discovery and underscoring the need for improved frameworks to assess signal specificity in ML-derived phenotypes.

All studies included in this thesis involve the application of ML, albeit with different objectives and levels of integration. In Study I, ML concepts are reflected primarily in the evaluation and comparison of diverse PGS methods across phenotypes and biobanks, including the construction of an ensemble PGS using an elastic net, with a focus on predictive performance and generalizability. Study II applies a XGBoost model more directly to phenotype refinement in psychiatric genetics by implementing a strict supernormal control definition and using predictive models to identify likely cases among individuals excluded from the control group. In Study III, a XGBoost model is used to construct an entirely new phenotype by predicting opioid intake patterns from medical history, enabling the identification of individuals with elevated use or early onset.

3 Aims of the Study

The increasing availability of large, genotyped biobanks has enabled genetic analyses at unprecedented scale. While this expansion has led to major advances, it has also introduced challenges related to data heterogeneity, phenotype noise, and limited transferability across cohorts. Many biobank phenotypes are derived from EHRs, which reflect clinical practice rather than research design, and therefore vary substantially across healthcare systems and populations. Addressing these challenges is essential for improving the robustness and generalizability of genetic findings.

This thesis focuses on understanding and mitigating sources of variability and bias in biobank-based genetic studies. Across all studies, analyses are conducted in multiple biobanks and leverage ML approaches where appropriate, with an emphasis on evaluating robustness across cohorts rather than within a single dataset.

Study I aims to systematically assess variability in PGS performance across methods and biobanks. Seven PGS methods are compared across sixteen diseases in five biobanks, including analyses in two ancestry groups. The primary goal is to characterize how much variation in predictive performance arises from methodological choices versus differences between biobanks, and to identify strategies that yield more stable performance across cohorts.

Study II addresses bias introduced by phenotype definition in EHR-based psychiatric genetics. Using simulations, this study examines how misclassification interacts with control selection and evaluates whether refined control definitions can improve the informativeness and cross-biobank consistency of psychiatric phenotypes. In addition, ML-based imputation is explored as a way to incorporate excluded individuals and recover power while maintaining robustness across datasets.

Study III focuses on heterogeneity in prescription-derived phenotypes using opioid use data from three biobanks. Rather than relying on binary exposure definitions, this study aims to leverage continuous prescription information harmonized through oral morphine equivalents (OME). ML models are used to account for medical context and identify patterns such as early onset or disproportionately high intake, enabling more nuanced phenotypic definitions that can be evaluated across biobanks.

4 Methods

4.1 Study Cohorts

4.1.1 FinnGen

The FinnGen study is a nationwide research project that integrates genome-wide genotype data with longitudinal health register information from Finland. The cohort comprises 520,149 anonymized participants, representing approximately 10% of the Finnish adult population. Participants were recruited through hospital biobanks, disease-specific studies, and blood donor registries, with intentional enrichment for disease cases to increase statistical power for genetic discovery. The median age was 63 years in 2023, with 57% female and 43% male participants[65].

FinnGen links genetic data to comprehensive nationwide health and social care registers, enabling long-term follow-up across a wide range of clinical outcomes. Available data include hospitalizations, outpatient visits, cancer diagnoses, medication purchases, mortality, and demographic information such as smoking status, occupational history, and number of children. Clinical phenotypes are defined using harmonized algorithms developed by expert working groups and applied consistently across registers. All three studies rely on phenotype definitions derived from FinnGen, with Study I using Data Freeze 9 (DF9) and Studies II and III using Data Freeze 12 (DF12).

Data on approximately 70,000 individuals were derived from legacy collections and genotyped using a heterogeneous set of non-custom genome-wide genotyping arrays from multiple platforms and array generations[65]. Genotyping of newly collected samples was performed using custom ThermoFisher Axiom arrays. Approximately 121,000 samples were genotyped using array version 1 (657,675 markers), and approximately 328,000 samples using array version 2 (664,510 markers). The arrays include a standard GWAS backbone as well as variants targeting coding regions, pharmacogenomics, immune-related loci, and Finnish-enriched alleles[182].

Extensive quality control was applied at both the sample and variant level. Samples were excluded based on genotype missingness greater than 2%, excess heterozygosity, discordance between genetically inferred and registry-reported sex, and excess relatedness. Related individuals were removed in two stages to reduce cryptic relatedness. Variant-level quality control excluded markers with non-ATCG alleles, poor batch performance, significant Hardy-Weinberg disequilibrium, or substantial allele frequency discrepancies relative to the reference panel.

Genotypes were pre-phased using Eagle and imputed using the Finnish-specific SISu v4 reference panel, constructed from high-coverage WGS of Finnish individuals and mapped to the GRCh38 reference genome[183]. The use of a population-specific imputation panel implicitly accounts for ancestry, and population outliers were removed during quality control

to limit residual population structure. FinnGen served as the replication cohort in Study I and as the primary analysis cohort in Studies II and III, where it was used for phenotype definition, GWAS, and the construction of ML-derived phenotypes.

4.1.2 UK Biobank (UKBB)

The UKBB is a large population-based prospective cohort that integrates genome-wide genotype data with longitudinal health record information from the United Kingdom. The cohort comprises approximately 500,000 anonymized participants recruited between 2006 and 2010, with individuals aged 40-69 years at the time of recruitment. Participants provided biological samples and informed consent for long-term follow-up through linkage to national health records.

UKBB links genetic data to extensive EHR, enabling longitudinal follow-up across a broad range of clinical outcomes. Available data include hospital inpatient records, primary care data, death registries, cancer registries, and prescription records, covering diagnoses, procedures, medications, and lifestyle-related variables. Medication data are coded using the BNF system.

Genotyping was performed using two closely related array platforms: the UK BiLEVE Axiom array and the UKBB Axiom array. Together, these arrays provide approximately >800,000 markers per individual and were designed to support genome-wide imputation and trait discovery. Genotype data underwent centralized sample- and variant-level quality control by UKBB.

Genotypes were imputed using a combined reference panel from the Haplotype Reference Consortium[184] (HRC), UK10K[185], and the 1000 Genomes Project[58]. Ancestry inference and population structure assessment were performed centrally by UKBB as part of the quality control and imputation pipeline, and population outliers were excluded to limit residual population stratification[68]. UKBB served as the primary cohort for polygenic score training and tuning in Study I, with predictions evaluated in the remaining cohorts. In Study II, UKBB was used for phenotype quality control and replication using polygenic scores, and in Study III it was used to construct opioid phenotypes for meta-analysis.

4.1.3 Estonian Biobank (EstBB)

The EstBB is a population-based biobank managed by the Institute of Genomics at the University of Tartu. It includes over 200,000 participants, representing approximately 20% of the Estonian adult population. The cohort overrepresents younger women and underrepresents individuals from the Russian-speaking minority relative to the general population[67].

Participants provided informed consent for long-term follow-up through linkage to national health registries, EHR, and prescription databases. These linkages provide near-complete longitudinal coverage of diagnoses, treatments, and medication use. Health data are harmonized to common vocabularies, including ICD-10 diagnosis codes and ATC medication codes. The data support large-scale analyses and international research and regulatory collaborations. In addition, extensive questionnaire data capture lifestyle, sociodemographic

factors, mental health, and medication side effects, with repeated collections enabling longitudinal analyses[66].

Genotyping was performed at the University of Tartu Core Genotyping Laboratory using multiple versions of the Illumina Global Screening Array, yielding data for over 212,000 DNA samples. Sample-level quality control excluded individuals with low call rate or discordance between genetically inferred and registry-reported sex. Variant-level quality control excluded markers with low call rate, Hardy-Weinberg disequilibrium, poor clustering metrics, or inconsistent allele frequencies across genotyping batches. Variants were mapped to the GRCh38 reference genome, and strand-ambiguous SNPs, structural variants, and invariant sites were removed.

Genotype data were pre-phased using Eagle and imputed using Beagle with a population-specific reference panel constructed from WGS of Estonian individuals. Ancestry inference and population structure assessment were performed using genotype-based methods, and individuals with non-European ancestry, duplicates, or monozygotic twins were excluded as part of quality control to limit residual population stratification. The EstBB was used for polygenic score evaluation in Study I and for construction of opioid phenotypes for meta-analysis in Study III.

4.1.4 Other Cohorts

Trøndelag Health Study (HUNT)

HUNT is a population-based health study conducted in Trøndelag County, Norway. The cohort was established through repeated population health surveys with three main recruitment waves: HUNT1 (1984-1986), HUNT2 (1995-1997), and HUNT3 (2006-2008). Participants provided health-related information through surveys and clinical examinations, enabling longitudinal follow-up across multiple decades.

HUNT links survey-based health data to clinical diagnoses derived from registry sources. Disease information is captured using ICD codes, with both primary and secondary diagnoses available for phenotypic analyses[70].

Genotyping was performed for 71,860 HUNT participants using the Illumina HumanCoreExome array. Sample-level quality control excluded individuals based on low call rates, sample contamination, chromosomal abnormalities, or discordance between genetically inferred and reported sex. Genotype calling followed standardized GenomeStudio quality control protocols. Variant-level quality control excluded markers with mapping errors, poor cluster separation, deviation from Hardy-Weinberg equilibrium, or low call rates.

Genotype imputation was conducted for 69,716 individuals of recent European ancestry using Minimac with default parameters and a customized HRC (HRC v1.1) reference panel. The customized panel incorporated low-coverage WGS data from HUNT participants in combination with the HRC v1.1 reference. Imputed variants with low imputation quality (Rsq < 0.3) were excluded, resulting in a final dataset comprising more than 24.9 million variants[186,187]. HUNT was used for polygenic score prediction and evaluation in Study I.

Genes & Health (GNH)

The GNH is a population-based cohort integrating questionnaire data, EHR, and genetic data. The cohort includes volunteer participants, predominantly of British-Bangladeshi and British-Pakistani ancestry, who consented to long-term follow-up through linkage to national health records. Health outcomes are derived from EHR using standardized coding systems, including ICD diagnosis codes.

The July 2021 data release comprised 44,190 individuals. Genotyping was performed on saliva-derived DNA using the Illumina Infinium Global Screening Array v3, which includes 730,059 variants. Genotype calling and initial quality control were conducted using GenomeStudio.

Sample-level quality control excluded duplicate samples and individuals who withdrew consent, prioritizing samples with higher call rates and valid National Health Service identifiers. Variant-level quality control excluded markers with low call rates, structural variant tags, or deviations from Hardy-Weinberg equilibrium. Additional filtering removed monomorphic variants, non-ACGT and palindromic SNPs, chromosome Y variants, and rare variants with MAF below 1×10^{-4}.

Genotype imputation was performed using the TOPMed-r2 reference panel via the Minimac4 imputation server, following additional quality control with TOPMed pipelines[188]. Variants with imputation quality Rsq ≥ 0.3 were retained for downstream analyses[71,189]. The GNH cohort was used for polygenic score prediction and evaluation in Study I.

4.2 Phenotype Definitions and Machine Learning Refinement

4.2.1 Phenotype Definitions for Polygenic Score Analyses (Study I)

Phenotypes in Study I were defined to enable a systematic comparison of PGS methods across multiple population-based biobanks. We included both binary disease traits and continuous quantitative traits, harmonizing definitions across cohorts where possible while respecting cohort-specific data structures and measurement protocols.

Phenotypes were selected to cover a broad range of disease classes and genetic architectures, including highly polygenic traits, autoimmune diseases involving the HLA region, cardiometabolic traits, and cancer outcomes. Selection emphasized medical relevance and the availability of large, publicly accessible GWAS summary statistics, preferably from discovery studies that did not include any of the five target biobanks. The final phenotype set comprised stroke, Alzheimer's disease, rheumatoid arthritis, type 2 diabetes, breast cancer, prostate cancer, type 1 diabetes, inflammatory bowel disease, chronic kidney disease (CKD) (evaluated via estimated glomerular filtration rate, eGFR), height, gout (evaluated via serum urate levels), body mass index (BMI), high-density lipoprotein cholesterol (HDL), and glycated hemoglobin (HbA1c).

Binary disease phenotypes were defined using ICD-based case definitions where available or using biobank-specific curated phenotypes when provided. For Alzheimer's disease and height, the selected GWAS summary statistics included UKBB European participants;

therefore, the UKBB-EUR sample was excluded from tuning and evaluation for these traits to avoid sample overlap.

Continuous phenotypes were defined within each biobank to accommodate differences in data collection while targeting the same underlying traits. In UKBB, continuous traits were defined using cohort-specific data fields. In GNH, all available measurements were extracted from primary and secondary health records; trait-specific outliers were removed using a ±6 SD threshold, and the mean value per individual was used.

In HUNT, the most recent measurement was selected when traits were assessed at multiple baseline examinations across recruitment waves, following standard cohort-specific quality-control procedures. BMI was calculated from measured height and weight, creatinine and HDL cholesterol were measured from serum samples obtained in non-fasting individuals, and HbA1c values were reported according to the IFCC standard (mmol/mol).

In the EstBB, the earliest available measurements were selected for BMI and height to reduce potential reverse causation, and extreme values were excluded using predefined thresholds (height <100 cm or >260 cm; BMI <10 kg/m² or >200 kg/m²). For a subset of participants, metabolic traits including HDL cholesterol and creatinine were obtained using nuclear magnetic resonance-based analysis.

For all cohorts with available creatinine measurements, eGFR was calculated using the Modification of Diet in Renal Disease (MDRD) equation without an ethnicity multiplier, as analyses were restricted to ancestry-matched populations. During evaluation, all continuous traits were standardized to mean zero and unit variance.

4.2.2 Psychiatric Traits and SuperControl Framework (Study II)

Psychiatric traits in Study II were defined in the FinnGen cohort using curated FinnGen endpoint definitions derived from nationwide health registers. Ten major psychiatric traits were included to capture a broad range of psychiatric disorders with differing prevalence, comorbidity patterns, and genetic architectures. These comprised sleep disorders (SLD), ADHD, bipolar disorder (BD), schizophrenia (SCZ), OCD, PTSD, anxiety disorders (ANX), major depressive disorder (MDD), substance use disorder (SUD), and a composite cross-disorder trait capturing any mental health disorder (MHD). Each psychiatric trait corresponded to a predefined FinnGen endpoint (F5_*), with case definitions curated by FinnGen clinical expert groups and diagnoses ascertained from all available register sources. For reference analyses, we used the original FinnGen phenotype definitions, in which cases were defined by at least one relevant diagnosis and controls comprised all remaining individuals, with additional exclusions for selected highly comorbid diagnoses for some endpoints as specified in FinnGen documentation. Full endpoint definitions and inclusion and exclusion criteria are publicly available through the FinnGen Risteys portal (https://risteys.finngen.fi/).

We defined a SuperControl phenotype for each psychiatric trait. Case definitions were identical to those used in the original FinnGen phenotypes, while control definitions were refined through additional exclusion criteria (Figure 4). Specifically, individuals were excluded from the control group if they had any recorded diagnosis of a psychiatric or

neurodegenerative disorder, based on a predefined list of 489 FinnGen traits. Individuals were also excluded if they had records of prescriptions for psychiatric or neurologically active medications, covering 53 drug categories defined using ATC codes. These exclusions were applied to reduce comorbidity and potential misclassification among controls.

Individuals excluded from the SuperControl definition were retained as a separate cohort and reintroduced via a ML-derived phenotype termed PRISMA (PRedicted excluded Individuals and SuperControls via Multi-Trait Analysis of GWAS-based Analysis). A ML model trained in the SuperControl cohort was applied to the excluded cohort to predict psychiatric liability, and summary statistics from the SuperControl and predicted phenotypes were jointly analyzed using Multi-Trait Analysis of GWAS (MTAG) to generate the PRISMA phenotype[190].

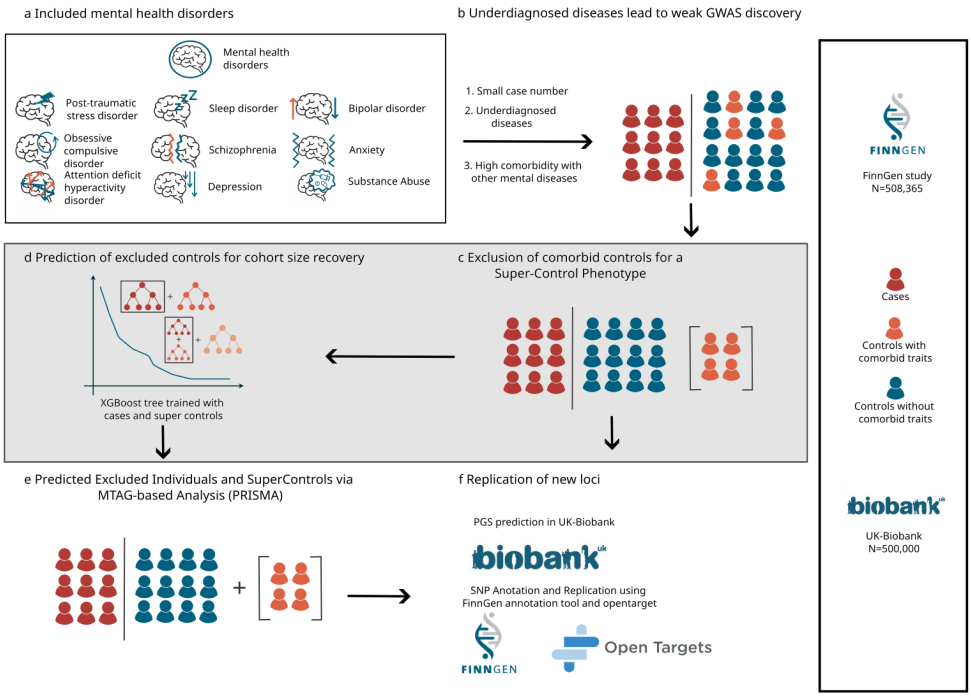


Figure 4 | Study overview and phenotype refinement workflow
(a) Overview of the psychiatric traits included in the analysis. (b) Conventional case-control GWAS using EHR-based diagnoses. (c) Definition of a super-control phenotype by excluding individuals with psychiatric comorbidity. (d) ML-based prediction of disease status in excluded individuals. (e) Integration of predicted cases with super-control GWAS results to construct the PRISMA phenotype. (f) Replication and downstream biological characterization of associated loci.

Because psychiatric traits are heterogeneous and often limited by low case numbers, additional phenotype harmonization was performed in UKBB for downstream PGS analyses. Psychiatric traits were defined using combined ICD-10 diagnoses together with self-reported mental health information from data field 20544. This resulted in case counts ranging from 105 to 96,478 across traits, with four traits (ANX, MDD, SUD, and MHD) exceeding 2,000 cases and therefore providing sufficient power for PGS analyses.

4.2.3 Opioid Prescription Phenotypes and Opioid Intake Harmonization (Study III)

Opioid exposure phenotypes were derived from longitudinal prescription and purchase records in FinnGen DF12, the EstBB, and the UKBB. In FinnGen and EstBB, opioid data were obtained from nationwide drug purchase registries, whereas in UKBB prescription information was extracted from EHR. In FinnGen, the full available prescription history was used, corresponding to nationwide registry availability from 1995 onward through the end of follow-up, while in the EstBB and UKBB prescription histories were used according to cohort-specific data availability.

Opioid medications were identified primarily using ATC codes starting with N02A. In EstBB, ATC code N07BC02 was additionally included to capture methadone prescriptions previously recorded under N02AC02. ATC codes N02AB02 (pethidine), N02AB01 (ketobemidone), and N02AX06 (tapentadol) were excluded in EstBB due to low prescription counts and incomplete opioid information. In UKBB, where ATC coding is not directly available, opioid medications were identified manually based on active substance names, guided by the ATC summary tables to ensure consistency with FinnGen and EstBB.

Prescription records were harmonized using national drug catalogues to ensure comparability across cohorts. Product strength and package size were standardized by converting composite entries into single numeric quantities, harmonizing salt forms and fixed-dose combinations, and resolving duplicate or inconsistent records[191]. When strength or package size information was missing after harmonization, values were completed for rare outliers using alternative national formularies. Transdermal formulations recorded as release rates were converted to total milligram amounts based on labeled wear times.

For each prescription record, opioid exposure was quantified by multiplying the normalized drug strength (in milligrams) by the total dispensed quantity. Each opioid molecule and route of administration was assigned an opioid factor (OF) reflecting its potency relative to morphine, primarily based on published comparative guidelines[192]. When route-specific conversion OF were unavailable, pharmacologically equivalent routes were grouped, and conservative assumptions were applied in the absence of published estimates. Parenteral tramadol, for example, was assumed to be approximately twice as potent as oral tramadol, consistent with clinical conversion charts (Gloucestershire Hospitals NHS Foundation Trust)[193]. This procedure yielded an OME value expressed in milligrams for each prescription.

For each individual i, cumulative opioid exposure was calculated as the sum of prescription-level OME values across all opioid prescriptions:

$$OME_i = \sum_{ATC_a}^{ATC_z} st_a * qu_a * pn_a * OF_a$$

where st is opioid strength, qu is package quantity, pn is the number of prescriptions, and opioid factor is the OF. This definition combines drug potency with prescription volume,

providing a continuous measure of exposure. where the sum is taken over all opioid prescriptions for individual i.

Based on cumulative OME values, three opioid-related phenotypes were constructed:

- **RxExpPop**, a binary phenotype indicating whether an individual ever received an opioid prescription.
- **RxDoseUsers**, a continuous phenotype defined as total cumulative OME among individuals with at least one opioid prescription.
- **RxDosePop**, a continuous phenotype defined as total cumulative OME across the full population, assigning a value of zero to individuals without prescriptions.

Cumulative OME values exhibited strong right skewness and were therefore log-transformed prior to downstream analyses. In all cohorts, the minimum observed OME value among exposed individuals exceeded 1 mg, such that no offset was required before transformation. Log-transformed OME values were used directly and were not further standardized.

4.2.4 FinnGen Harmonization for Machine Learning Application

FinnGen integrates data from multiple nationwide registers with heterogeneous temporal coverage. To ensure consistent data availability, ML analyses in Studies II and III were restricted to periods in which all relevant registries were simultaneously available. Because nationwide prescription data are available from 1995 onward, all phenotypic information, covariates, and event histories were restricted to records dated 1995 or later. This temporal alignment resulted in the exclusion of approximately 2.5% of recorded events and 2,056 individuals (0.4%) whose endpoints occurred exclusively before 1995.

In addition to diagnostic and prescription data, covariates derived from national registers included BMI, smoking status, and birth registry information. Socioeconomic variables comprised highest attained educational degree, field of education, and occupational category. Educational variables were included only for individuals aged ≥32 years, corresponding to the mean age of highest educational attainment in FinnGen, and occupational information was restricted to records coded using the PAMKO classification. All categorical covariates were encoded using one-hot encoding prior to analysis.

4.2.5 Psychiatric Risk Prediction and PRISMA Phenotype Construction

To infer psychiatric liability among individuals excluded from the SuperControl cohort in Study II, we trained a supervised ML model using XGBoost[194] in Python. The model was trained on the SuperControl cohort using harmonized FinnGen data, retaining for each disease trait the year of first occurrence and total event count. Medication features were derived from ATC codes supplemented with package-level information to capture prescription counts and approximate intake. The final feature set comprised 4,006 binary disease traits, 516 medication variables, and demographic, lifestyle, and socioeconomic covariates. All traits and medication variables used to define the SuperControl phenotype

were explicitly excluded from the feature set to prevent information leakage from variables inherently linked to case-control status.

Model training was performed exclusively within the SuperControl cohort. Data were randomly split into 80% training and 20% validation sets, stratified by case-control status. Hyperparameter optimization was conducted using Bayesian optimization with BayesSearchCV in scikit-learn, applied to a nested 20% split of the training data. The final model was retrained on the full training set using the optimal hyperparameters and evaluated on the held-out validation set.

After training, the fitted classifier was applied to individuals excluded from the SuperControl cohort. Model outputs were interpreted as predicted probabilities of psychiatric disease liability. Based on the observed bimodal distribution of prediction scores, a binary classification threshold of 0.5 was applied. Individuals exceeding this threshold were labeled as predicted cases and used to construct a synthetic case group within the excluded cohort.

To increase discovery power while accounting for genetic correlation between observed and predicted phenotypes, summary statistics from the SuperControl phenotype and the predicted phenotype were jointly analyzed using MTAG. This yielded a combined phenotype termed PRISMA. Genetic correlations between the SuperControl and excluded-cohort phenotypes were estimated using LDSC (v1.0.1)[195], with most traits showing moderate to high correlation ($r_g \approx 0.6$-0.7), exceeding thresholds previously considered sufficient for joint GWAS analysis[196] (Figure 4).

4.2.6 Opioid Intake Prediction Model

For opioid-related prediction analyses in Study III, ML models were trained using opioid values derived from the opioid medication class (N02A) across 520,149 FinnGen participants. For prediction, all available harmonized FinnGen information was used, excluding only the six opioid medication classes used to define the outcome variables, to preserve detailed prescription and phenotype histories.

The predictor space comprised 9,617 phenotypes and covariates, including disease phenotypes with age-of-onset information (n = 4,006), prescription records covering 516 medication categories, and other harmonized FinnGen data.

Predictive models were trained using the XGBoost[194] algorithm in Python. Continuous outcomes were modeled using squared-error loss, and binary outcomes were modeled using logistic loss. Hyperparameters were optimized using RandomizedSearchCV in scikit-learn, using 5% of the training data for tuning. For each prediction task, we trained three models using independent 80/20 train-test splits. The resulting predictions were combined into an ensemble by averaging model outputs. Model stability was assessed by examining the mean and variance of predictions across the three runs. Model performance was assessed using mean squared error (MSE), the proportion of variance in the dependent variable explained by the regression model (r^2), and AUROC. Feature importance was quantified using SHAP[197] values. Two opioid overuse phenotypes were derived from the trained prediction models. The first phenotype, RxOverUse_Amount, was based on predicted opioid intake among individuals with at least one recorded opioid prescription. Three independent dosage

prediction models were trained, and their predictions were averaged to reduce variance. Individuals with an average predicted intake exceeding the minimum observed opioid dose in FinnGen (log□(39)) were classified as over users, ensuring that classified individuals exceeded at least one unit of OME above the minimum observed dose. Controls were defined as opioid users with predicted intake at or below this threshold, excluding individuals with prediction variance greater than 0.25. This yielded 3,163 over users and 277,176 controls.

The second phenotype, RxOverUse_Onset, was based on exposure misclassification from binary models of opioid use. Three independent binary models were trained, and predictions were averaged. The optimal classification threshold was determined using the Youden Index derived from the receiver operating characteristic curve. Over users were defined as individuals with at least one opioid prescription who were incorrectly predicted as non-users. Controls consisted of individuals with no opioid prescriptions, excluding those predicted as users. This resulted in 67,307 individuals with early onset and 175,397 controls (Figure 5).

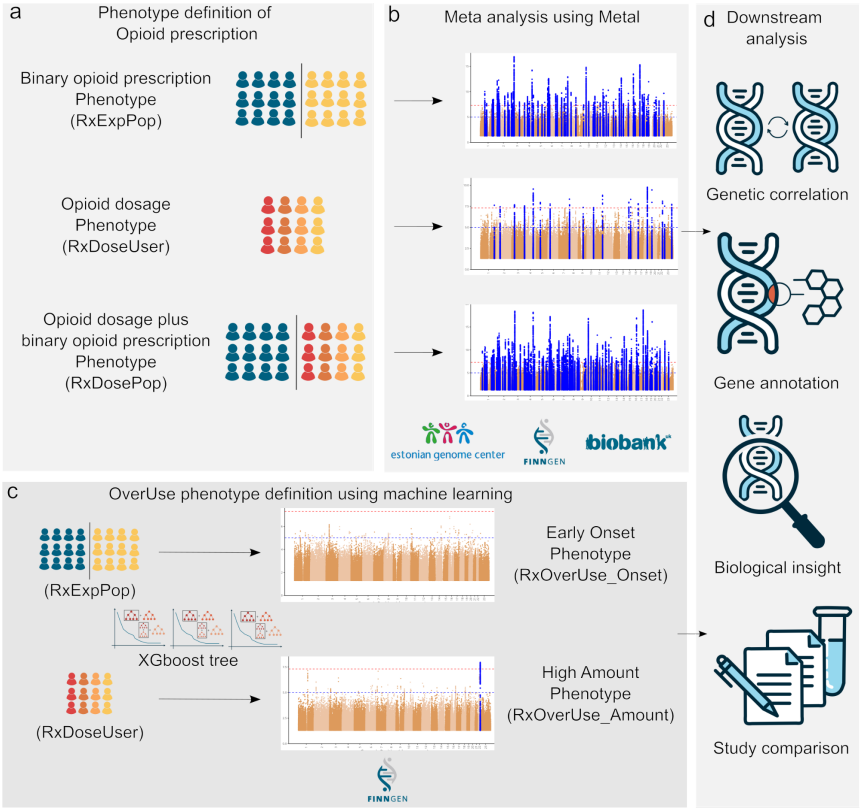


Figure 5 | Phenotype definitions, meta-analysis, and ML-based refinement of opioid prescription traits.
(a) Three opioid prescription phenotypes were defined: RxExpPop, RxDoseUser, RxDosePop. (b) GWAS were performed across multiple biobanks and meta-analyzed using METAL. (c) ML was applied to refine overuse phenotypes, resulting in early onset (RxOverUse_Onset) and high dose (RxOverUse_Amount) subtypes. (d) Downstream analyses included genetic correlation, gene annotation, biological interpretation, and cross-study comparisons.

4.3 Genetic Analysis

4.3.1 Simulation of SuperControl Design and Misclassification

To assess the impact of control selection and diagnostic misclassification on GWAS, we performed simulations modeling two genetically correlated binary traits, denoted trait A (the target GWAS trait) and trait B (a correlated trait relevant for control definition). Simulations were conducted in a cohort of 10,000 individuals.

We simulated four groups of SNPs: (i) 500 SNPs with shared causal effects on both traits A and B, (ii) 500 SNPs with causal effects specific to trait A, (iii) 500 SNPs with causal effects specific to trait B, and (iv) 500 non-causal SNPs. Genotypes were simulated independently with a MAF of 0.3. Effect sizes for causal variants were drawn from a standard normal distribution.

Trait liability was modeled using a liability threshold framework with heritability set to 0.6. Individual genetic risk was computed as the aggregated polygenic risk across causal variants, modeled as a genetic liability with variance equal to the assumed heritability. Binary trait status was assigned by applying a threshold corresponding to a specified prevalence, which was varied across simulations ($p \in \{0.1, 0.3, 0.5\}$).

To model diagnostic misclassification, two misdiagnosis probabilities were introduced: PB0, the proportion of individuals with A=1, B=0 misclassified as A=0, resulting in classification as pure controls, and PB1, the proportion of individuals with A=1, B=1 misclassified as A=0, resulting in classification as controls with comorbid trait B. PB1 was constrained to be greater than or equal to PB0, reflecting higher misclassification in comorbid cases. PB0 was varied in {0.05, 0.1, 0.2} and PB1 in {0.3, 0.4, 0.5, 0.6, 0.7}.

Four GWAS designs were evaluated: standard GWAS with and without misclassification, and SuperControl GWAS with and without misclassification. Performance was assessed based on power to detect shared and trait-specific causal variants and type I error rates. All simulations were repeated ten times, and results were summarized using mean and standard deviation across runs.

4.3.2 Genome-wide Association Analyses (GWAS)

GWAS were conducted using REGENIE[198]. In FinnGen, analyses were performed with REGENIE v2.2.4 (https://github.com/FINNGEN/regenie-pipelines). Binary phenotypes were analyzed using logistic regression, while continuous phenotypes were analyzed using linear regression. All models were adjusted for sex, birth year, age at death or end of follow-up (2023; including a squared term), the first ten genetic principal components (PCs), and biobank-specific genotyping batch and chip effects.

Variants were filtered to those with MAF greater than 0.1% and imputation information score greater than 0.8. GWAS were performed separately for each phenotype definition, including the original FinnGen phenotype, the SuperControl phenotype, and the PRISMA phenotype where applicable.

For opioid-related continuous phenotypes, cumulative dose measures were log-transformed prior to association testing to reduce skewness. The RxDosePop phenotype exhibited a highly skewed distribution driven by the mixture of non-users and a long right tail of positive dose values. Although linear regression is generally robust to deviations from normality in large sample sizes[199], this distribution raised concerns that a log transformation alone might not adequately capture the underlying structure. We therefore additionally evaluated inverse-rank normalization as a robustness check. GWAS results obtained using inverse-rank-normalized values were nearly identical to those obtained from the log-transformed phenotype (genetic correlation $r_g = 0.9993$, $p = 0$), indicating that association results were stable with respect to the chosen transformation. Given this concordance and the greater interpretability of dosage differences, we proceeded with the log-transformed phenotype for downstream analyses.

4.3.3 Post-GWAS Downstream Analysis

GWAS Harmonization and Metaanalysis

To ensure comparability across studies and cohorts, GWAS were harmonized prior to downstream analyses. For external GWAS used for comparison or replication, genomic coordinates were converted to the GRCh38 reference build when required using the UCSC LiftOver tool[200].

In Study III, opioid dose phenotypes were transformed differently across cohorts. UKBB analyses used a $\log_{10}$ transformation, whereas FinnGen and the EstBB applied a natural logarithm transformation. To ensure effect sizes were comparable across cohorts, UKBB effect estimates were rescaled by multiplying by ln(10) prior to meta-analysis.

Meta-analysis of opioid GWAS summary statistics was conducted across FinnGen, the EstBB, and the UKBB using METAL (release 2020-05-05)[82]. Fixed-effects inverse-variance weighted models were used. Between-study heterogeneity was assessed using the ANALYZE HETEROGENEITY option with Cochran's Q statistic and the I-squared heterogeneity metric (I^2). Variants with heterogeneity $P < 0.05$ were excluded from downstream analyses.

SNP-based Heritability, Genetic Correlation and Partitioned Heritability

SNP-based heritability and genetic correlations were estimated using LD Score Regression (LDSC v1.0.1)[48,195]. For FinnGen analyses, a Finnish-specific LD reference panel (fin_ldsc) was used to account for population-specific LD structure. For European ancestry external datasets, LD scores derived from the 1000 Genomes European reference were used[58].

For binary traits, effective sample sizes were calculated based on trait prevalence and total sample size. Heritability estimates were reported on the observed scale. Cross-trait genetic correlations were computed using LDSC. For FinnGen-wide phenotype correlation analyses, multiple testing was controlled using Storey's q-value procedure[201].

Partitioned SNP heritability was estimated using stratified LDSC. Three complementary annotation models were applied: a cell-type baseline model, a brain-specific baseline model, and a tissue-specific model incorporating chromatin and genetic annotations[83]. For all

models, summary statistics were restricted to HapMap3 SNPs. Multiple testing across annotations was controlled using Bonferroni correction[202].

Fine-Mapping of GWAS Loci

To identify independent association signals and enable comparison of loci across phenotype definitions, conditional analyses were performed using the COJO framework[203] in GCTA[204]. In Study II, conditional analysis was used to determine whether association signals detected for the SuperControl and PRISMA phenotypes represented independent loci relative to the original FinnGen phenotype, and to support locus-level comparison against larger external GWAS of the corresponding traits.

In Study III, association loci were defined around genome-wide significant lead variants ($P < 5 \times 10^{-8}$) using ±1.5 Mb windows (3 Mb total). This conservative locus definition was chosen because the meta-analysis combined multiple cohorts with differing LD structures, for which LD-based methods can be challenging to apply consistently. Using a fixed physical window minimized the risk of overestimating associations in the meta-analysis while ensuring harmonized locus definitions across cohorts.

Where LD information was available, fine-mapping was performed to identify putative causal variants within each locus. Fine-mapping in FinnGen was conducted using SuSiE[88], which models association signals as a sum of single effects and allows for multiple causal variants within a region. For both Study II and Study III, the FinnGen SuSiE fine-mapping pipeline (https://github.com/FINNGEN/finemapping-pipeline) was applied, allowing up to ten causal variants per locus ($L = 10$). In-sample LD matrices were computed from genotype dosage data and used as input to fine-mapping. Fine-mapping results were summarized using posterior inclusion probabilities, from which 99% credible sets were constructed. These credible sets were used for downstream functional annotation and to functional annotation of associated loci.

Functional Annotation and Replication Analyses

In Study II, replication was assessed using both locus-based and gene-based approaches. For locus-based replication (LBR), lead variants were compared to previously published GWAS of the same traits. A locus was classified as having strong prior evidence if a genome-wide significant association ($P < 5 \times 10^{-8}$) in the published GWAS was reported within the locus. If no such signal was present but the lead variant reached $P < 5 \times 10^{-5}$ in a prior GWAS, the locus was classified as having moderate prior evidence. Loci without signals meeting either threshold were classified as having no supporting evidence. Gene-based replication (GBR) was evaluated using the Open Targets Genetics platform. A locus was assigned strong prior evidence if the lead variant itself had a prior GWAS association with the same trait. Moderate prior evidence was assigned if a gene mapped to the variant was previously implicated in a GWAS of the same trait, but without variant-level support. Loci without variant- or gene-level evidence were classified as having no supporting evidence.

In Study III, lead variants were annotated using fine-mapping and functional evidence from Open Targets Genetics. Available annotations included posterior inclusion probabilities, locus-to-gene scores, and mapped genes[205]. When fine-mapping results were unavailable,

variants were annotated using the Ensembl Variant Effect Predictor (VEP)[206]. In these cases, the annotation with the highest predicted impact and closest distance to the transcription start site was selected.

4.4 Polygenic Score (PGS) Analyses

4.4.1 Target Genotype Processing and Ancestry Matching

In Study I, we constructed a custom HapMap3 variant set. HapMap3 rsIDs were retrieved, and reference genotypes from the 1000 Genomes Project (phase 3, v5)[207] were downloaded. Variants with an allele frequency of at least 1% in any 1000 Genomes superpopulation were retained, yielding a HapMap3-1KG reference set of 1,330,821 variants. Target genotype data were intersected with this reference based on genomic position and allele codes, renamed accordingly, and converted to PLINK[208] format. These harmonized genotypes served as input for ancestry matching and polygenic scoring and were generated as part of the prspipe workflow (https://github.com/intervene-EU-H2020/prspipe).

For ancestry inference, both target and 1000 Genomes reference data were further restricted to shared HapMap3-1KG variants with allele frequency greater than 5%, genotype missingness below 2%, and Hardy-Weinberg equilibrium p-values above 1×10^{-6}. Variants were LD-pruned using PLINK based on the 1000 Genomes reference panel, and genomic regions with long LD were excluded.

Rather than assigning ancestry labels directly, individuals were matched to 1000 Genomes superpopulations based on genetic similarity. Using the GenoPred framework[6], target genotypes were projected into the principal-component space of the 1000 Genomes reference panel. An elastic net classifier, trained on the first 100 PCs of the 1000 Genomes samples using 5-fold cross-validation (CV), was then used to assign each individual to one of five superpopulations: Middle Eastern (MID), South Asian (SAS), East Asian (EAS), African (AFR), admixed American (AMR), European (EUR).

After superpopulation assignment, extreme population outliers were removed within each group. PCs were recomputed within assigned superpopulations using the first eight PCs, and up to ten centroids were defined in PC space using k-means clustering with Euclidean distance, as implemented in NbClust. Individuals with distances exceeding the 75th percentile plus 30 interquartile ranges from the nearest centroid were excluded. UKBB was the only cohort with more than one well-represented ancestry group. Downstream analyses therefore focused on ancestry groups replicated across biobanks, namely EUR (UKBB, EstBB, HUNT, FinnGen) and SAS (UKBB, GNH).

4.4.2 Polygenic Score Construction and Ensemble Modeling

To compare PGS methods in Study I, we constructed polygenic scores using a diverse set of established methods, including pT+clump, lassosum[92], PRS-CS[94], SBayesR[96] (robust parameterization), LDpred2[93], DBSLMM[95], and MegaPRS[97]. For each method, we evaluated both automatically selected models based solely on GWAS summary statistics and, where supported, models tuned using target data. Rather than selecting a single best-performing

approach, we used the resulting set of scores to build an ensemble polygenic score that integrates information across methods and tuning strategies.

Each method was first applied using its automatic or default model selection strategy, in which model parameters are inferred directly from GWAS summary statistics without reference to target phenotype data. For pT+clump, the score defined by a p-value threshold of 1×10^{-8} was treated as the automatically tuned model. SBayesR and DBSLMM each produce a single set of weights under their default settings and therefore were not subject to additional tuning. MegaPRS was run using the author-recommended *--model mega* setting with the BLD-LDAK heritability model[209] and inclusion of the HLA region. This procedure generates multiple candidate models internally, which are treated as automatically selected MegaPRS scores.

For methods that support explicit target-data-based model selection (pT+clump, lassosum, PRS-CS, LDpred2, and MegaPRS), we additionally performed hyperparameter tuning using 10-fold CV. Tuning was conducted in UKBB EUR samples, with 80% of individuals used for training and 20% held out for evaluation. For each method, the model achieving the highest correlation with the target trait in the validation data was selected. Automatically tuned scores were retained and included alongside CV models in all downstream analyses, ensuring that both tuning strategies were evaluated consistently.

All analyses used EUR LD reference panels. Author-provided LD references were used for DBSLMM, lassosum, LDpred2, SBayesR, and PRS-CS (DBSLMM and lassosum based on 1000 Genomes. LDpred2, SBayesR, and PRS-CS based on UKBB). For pT+clump and MegaPRS, LD was computed using the 1000 Genomes EUR subset. PGS scoring files compatible with PLINK2 were generated using GenoPred scripts within the prspipe workflow, and single-method scores were computed using PLINK2 and GenoPred.

To construct an ensemble PGS, we combined all scores across methods, tuning strategies, and hyperparameter settings using an elastic net model implemented in caret. Each individual PGS was treated as an input feature, allowing the model to estimate ensemble coefficients that quantify the contribution of each score while accounting for correlations between methods. Shrinkage parameters were optimized using 10-fold CV. For pT+clump, variants were partitioned into ten non-overlapping p-value bins, and a separate score was constructed for each bin to reduce correlation between inputs and allow the ensemble to capture signal across different GWAS significance levels.

To apply the UKBB-trained ensemble model to other biobanks, the fitted elastic net was collapsed into a single scoring file. This was achieved by multiplying the variant-level weights from each individual PGS by their corresponding ensemble coefficients and summing these contributions across methods. The resulting file contains one unified set of variant weights, enabling consistent and portable application of the ensemble PGS across cohorts.

4.4.3 Performance Evaluation of Polygenic Score Methods

To estimate method performance in Study I, all PGS were standardized to mean 0 and unit standard deviation separately within each biobank and ancestry group prior to evaluation. After standardization, PGS values indicate how many standard deviations above or below the average PGS an individual is within that biobank and ancestry group. As a result, effect sizes are on a common scale and can be compared directly across PGS methods, biobanks, and ancestries.

PGS performance was evaluated separately for binary and continuous traits using standard regression-based metrics. For binary traits, we reported four measures of performance. The β coefficient summarizes the association between the PGS and disease risk on the log-odds scale per standard deviation increase in PGS. The corresponding odds ratio $OR = exp(\beta)$ provides an interpretable measure of relative risk. We also report the variance explained on the observed scale (r^2_{obs}), which quantifies how much of the variation in the binary outcome is captured by the PGS, and the AUROC, which measures discrimination between cases and controls.

Observed-scale variance explained was additionally converted to the liability scale (r^2_{liab}) using standard liability-threshold methods, assuming the median prevalence within each ancestry as the population prevalence estimate. Confidence intervals for AUROC were computed using DeLong's method as implemented in pROC[210]. Confidence intervals for (r^2_{obs}) were obtained from 1,000 bootstrap resamples of the observed-scale Pearson correlation.

For continuous traits, performance was summarized by the effect size β, representing the change in standard deviations of the trait per standard deviation increase in PGS, and by the variance explained on the observed scale (r^2_{obs}).

To compare the effect sizes of two PGS methods evaluated in the same biobank, we used a two-sided z-test that accounts for correlation between scores evaluated in the same individuals. This adjustment is necessary because PGS derived from the same sample are not independent.

To directly compare the effect sizes of two PGS methods evaluated within the same biobank, we used a two-sided z-test that accounts for correlation between scores arising from evaluation in the same individuals. The test statistic was defined as

$$z = \frac{\beta_{a,i} - \beta_{b,i}}{\sqrt{\sigma^2_{a,i} + \sigma^2_{b,i} - 2\rho_{(a,b),i}\sigma_{a,i}\sigma_{b,i}}}$$

Here, $\beta_{a,i}$ and $\beta_{b,i}$ are the estimated effect sizes for methods a and b in biobank i, and $\sigma_{a,i}$ and $\sigma_{b,i}$ are their corresponding standard errors. The term $\rho_{(a,b),i}$ denotes the correlation between the two PGS within the same biobank. By incorporating this correlation, the test

provides valid inference when comparing PGS effect sizes derived from the same individuals.

Prior to meta-analysis, selected biobank-trait combinations were excluded when results indicated a lack of informative signal or potential phenotyping issues. Specifically, T1D results from Genes and Health were excluded because no evaluated PGS reached nominal significance and effect sizes were close to zero. HbA1c results in Genes and Health were excluded due to strongly attenuated effects relative to UKBB-SAS, while T1D results from HUNT were also excluded due to consistently lower effect sizes compared with other biobanks. As both datasets deviated substantially from UKBB and other cohorts, we attribute this to underlying phenotyping differences, which complicates meaningful model comparison across biobanks.

4.4.4 Meta-Analysis of Polygenic Score Methods

For the Meta-analyses across methods and biobank in Study I the analysis were performed separately for each trait and ancestry using mixed-effects models in R (version 4.1.1) with the *metafor* package (version 3.8-1)[211]. Models were fitted using restricted maximum likelihood. To account for the fact that multiple PGS were evaluated within the same biobank, we explicitly modeled this dependence in the meta-analysis.

PGS effect sizes were combined across biobanks using models that allow results from the same biobank to be more similar to each other than results from different biobanks. This was done by including a biobank-level random intercept. For each trait, several PGS methods were analyzed together in the same model. The resulting fixed-effect estimates represent the average performance of each PGS method across biobanks, with larger and more precise studies contributing more weight.

Differences between PGS methods were tested using z-tests on these average effect estimates, with additional support from likelihood-based model comparisons. To describe how much effect sizes varied between biobanks, we report the estimated between-biobank variance and its likelihood-based confidence interval. Additionally, AUROC, r^2_{obs}, and r^2_{liab} were meta-analyzed using effective sample sizes.

To better understand where this heterogeneity arises, we further fitted three-level random-effects meta-analysis models that explicitly separate variation due to differences between biobanks from variation due to differences between PGS methods evaluated within the same biobank. In these models, PGS method effects are treated as nested within biobanks.

This framework estimates two distinct variance components. The between-biobank variance component, denoted $\tau^2_{biobank}$, captures how strongly PGS effect sizes differ across biobanks, reflecting differences in study context such as cohort composition, phenotyping, or data structure. The within-biobank, between-method variance component, denoted $\tau^2_{methods}$, captures how strongly PGS effect sizes differ between methods when evaluated in the same biobank, reflecting differences attributable to PGS construction rather than study context.

By separating these two sources of variation, the three-level model allows us to distinguish whether heterogeneity in PGS performance is primarily driven by differences between biobanks or by differences between PGS methods. The relative contribution of each level to the total variability in effect sizes was summarized using three-level extensions of the I^2 statistic.

Because multiple PGS evaluated within the same biobank are based on the same individuals, their effect size estimates are not independent. To account for this dependence, we explicitly modeled correlated uncertainty between PGS in all multivariate meta-analyses.

Within each biobank × ancestry × trait combination, we estimated correlations between PGS using up to 50,000 randomly sampled individuals. These correlations were then combined with the corresponding standard errors to construct a sampling variance-covariance structure that reflects both the uncertainty of each PGS effect estimate and their correlation with one another.

This variance-covariance structure was supplied to the meta-analysis model so that dependence between PGS evaluated within the same biobank was properly accounted for. Dependence was modeled within biobanks, while effect estimates from different biobanks were treated as independent.

For method ranking across traits, analyses were restricted to traits for which both CV-tuned and ensemble PGS were available. To avoid double-counting related phenotypes derived from the same summary statistics, continuous traits were preferred over their binary counterparts when available in the same number of biobanks; otherwise, the binary phenotype was used. This resulted in the use of eGFR for SAS and CKD for EUR, and urate for SAS and gout for EUR.

4.4.5 Cross-Cohort Prediction and Regional Polygenic Score Analyses

As part of Study II, we evaluated cross-cohort predictive performance and cross-trait behavior of the Original, SuperControl, and PRISMA phenotypes using prediction analyses in UKBB. Analyses were restricted to EUR participants with available genetic PCs, resulting in approximately 400,000 individuals.

PGS were generated using the FinnGen PGS-weighting pipeline, which applies PRS-CS for effect size shrinkage and SNP weight estimation. PGS weights derived from FinnGen GWAS of the Original, SuperControl, and PRISMA phenotypes were applied to UKBB genotype data to generate prediction scores. For each target trait, prediction models included the standardized PGS as the main predictor and were adjusted for age, sex, and the first ten genetic PCs.

To assess whether loci that did not replicate across cohorts nevertheless contribute predictive signal, we additionally constructed regional PGS restricted to non-replicated loci. Non-replicated regions were defined using LBR as 3 Mb windows centered on lead variants (±1.5 Mb). Prediction performance and effect sizes derived from these regional PGS were compared across phenotype definitions.

Differences between the Original, SuperControl, and PRISMA phenotypes were evaluated using two-sided z-tests comparing independent effect size estimates derived from their respective GWAS.

4.6 Ethics

Genetic research that relies on large-scale biobank data requires particular attention to ethical standards, given that real individuals contribute sensitive genetic and health information. Participants generously share their data for research purposes, and it is therefore essential that all analyses are conducted under rigorous ethical frameworks to ensure privacy, data protection, and the prevention of harm or disadvantage to participants.

All biobanks used in this thesis collect data under informed consent. Participants provide broad consent allowing their genetic samples and linked health data to be used for research purposes in accordance with national legislation and biobank-specific regulations. This consent framework enables large-scale genomic analyses while maintaining transparency regarding data use.

Across all datasets, participant data are anonymized or pseudonymized prior to researcher access. Identifiable information, such as names, addresses, or national identification numbers, is never shared with researchers. Instead, data are coded using unique study-specific identifiers, ensuring that individual participants cannot be directly identified during analysis.

None of the biobanks included in this work provide unrestricted data access. Researchers must apply for access with a clearly defined research plan and receive approval from the relevant biobank authorities and ethical review boards. In several cases, access is further limited to researchers affiliated with approved institutions or requires collaboration with a local principal investigator. Legal agreements prohibit attempts to re-identify participants and define permissible data use.

Data access is frequently provided through secure research environments specifically designed to minimize the risk of data leakage. These environments implement strict technical and organizational safeguards, including controlled user access, monitored activity, and restrictions on data export. In multiple biobanks, researchers are required to perform analyses within secure platforms rather than downloading individual-level data.

Where data export is permitted, downloads are tightly controlled and limited to aggregated or summary-level results. Statistical disclosure control measures are applied to prevent identification of individuals, including minimum thresholds for reporting results (for example, requiring analyses to include data from multiple individuals before results can be released). These safeguards ensure that no individual-level information is disclosed outside the secure environments.

Together, these measures ensure that the genetic and health data analyzed in this thesis are handled in accordance with high ethical standards, balancing the advancement of scientific knowledge with the protection of participant rights, privacy, and welfare. The official ethical statements and approvals for EstBB, UKBB, and FinnGen are provided in Appendix A.

5 Results

5.1 Variability of Polygenic Prediction Across Methods and Biobanks (Study I)

5.1.1 Performance of Polygenic Score Methods within Biobanks

The rapid growth of genetic data resources has been accompanied by the development of an increasing number of analytical methods for PGS construction. Many of these methods aim to address similar core challenges and often claim incremental improvements over existing approaches. Despite this methodological diversity, robust prediction across biobanks and ancestries remains a major challenge for polygenic scoring.

In Study I, we conducted a systematic comparison of seven PGS methods across 14 traits, five biobanks, and two ancestries. To quantify performance differences in a unified framework, we performed a meta-analysis of PGS performance across biobanks for each trait and ancestry.

PGS weights were generated from GWAS summary statistics, with hyperparameter tuning performed via CV in 80% of UKBB. In addition to individual methods, we constructed ensemble PGS and evaluated a ML-based ensemble approach using elastic net regression to assess whether combining information across methods improves predictive performance. To reduce noise and ensure comparability across cohorts, phenotype harmonization and ancestry matching were applied before performing polygenic scoring in all five biobanks.

Across traits, no single non-ensemble method consistently outperformed all others. CV-tuned LDpred2 and MegaPRS showed highly comparable performance, and effect-size differences between the top single methods were mostly not statistically significant. In contrast, the ensemble PGS demonstrated the most robust performance across biobanks, achieving the largest meta-analyzed effect size in 13 out of 14 traits. Exceptions were observed for T1D in EUR ancestry and stroke in SAS ancestry, where PRS-CS achieved higher predictive accuracy (Figure 6a, b).

Compared to the best-performing single methods, the ensemble PGS showed consistent performance gains. For binary traits, the median relative increase in effect size was 3.7% compared with CV-tuned LDpred2 and 4.5% compared with CV-tuned MegaPRS. For continuous traits, these gains were larger, with median relative increases of 5.2% and 7.9%, respectively. When measured in terms of variance explained, the relative improvements were more pronounced, with median increases of 7.4-9.0% for binary traits on the liability scale and 10.7-16.1% for continuous traits. In EUR ancestry data, ensemble PGS effect sizes were significantly larger than those of all other methods for 6 of 9 binary traits and all 5 continuous traits.

Across methods, CV-based hyperparameter tuning substantially improved performance relative to default parameter settings. In EUR ancestry data, the median relative increase in PGS effect size compared with CV-tuned pT+clump was 29.2% for CV-tuned LDpred2 (mean 30.9% ± 10.3 SD, n = 12 traits) and 29.9% for CV-tuned MegaPRS (mean 31.2% ± 12.6 SD, n = 12 traits), indicating nearly identical average performance between these two approaches. Effect sizes varied considerably between biobanks, demonstrating that the best-performing single method for a given trait was not necessarily consistent across cohorts.

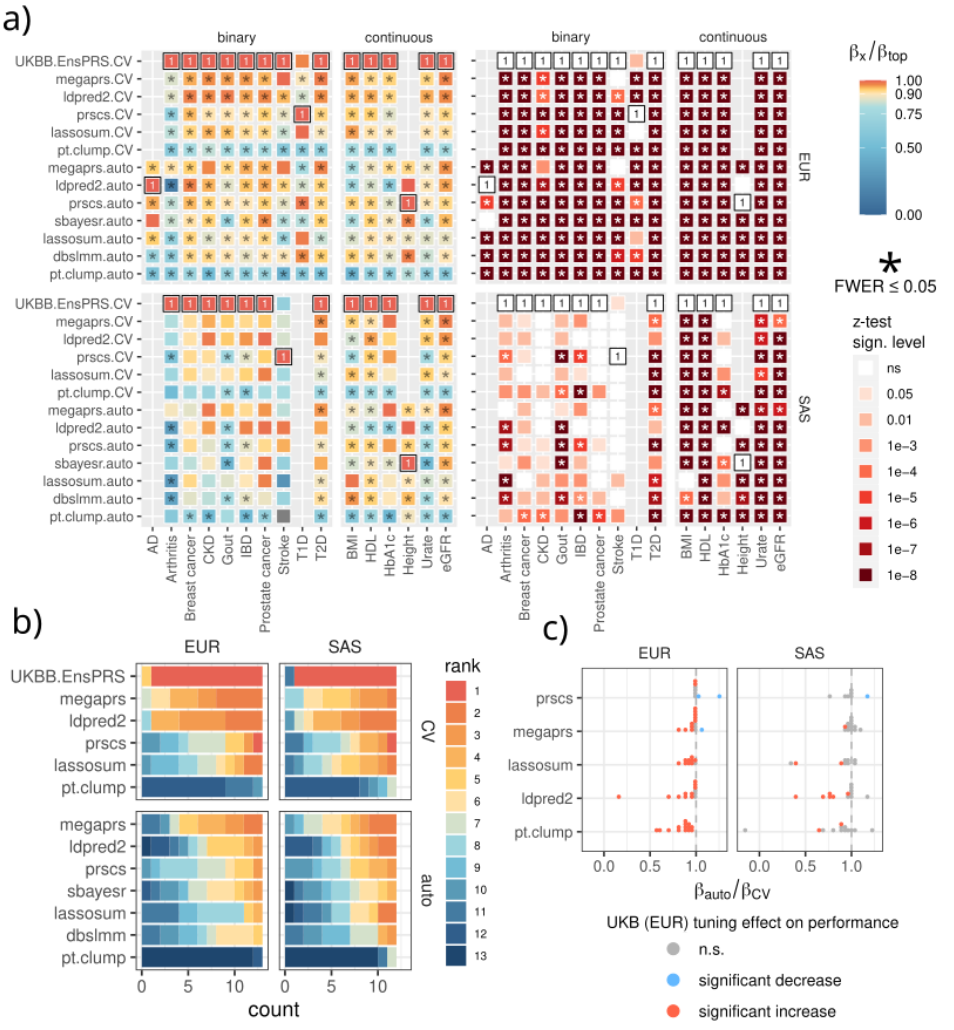


Figure 6 | Relative meta-analyzed PGS effect sizes across 14 traits
(a) Heatmaps showing relative meta-analyzed PGS effect sizes and corresponding significance across traits and ancestries, with the top-performing score per trait highlighted. (b) Distribution of PGS method ranks across traits, stratified by ancestry and tuning strategy. (c) Comparison of automatically tuned and CV PGS effect sizes across methods. (Reprinted from *"Evaluation of polygenic scoring methods in five biobanks shows larger variation between biobanks than methods and finds benefits of ensemble learning"*, with permission from Elsevier (Study I))

5.1.2 Variance in Polygenic Score Performance

Beyond identifying the best-performing PGS methods, this study also provided an opportunity to investigate sources of variability in polygenic score performance. To this end, three-level random-effects meta-analytic models were applied in EUR ancestry target data to partition heterogeneity attributable to differences between methods and differences between biobanks.

Across all 13 traits replicated in at least two biobanks, significant heterogeneity in PGS effect sizes was observed (Figure 7). For the majority of traits, heterogeneity attributable to the target biobank exceeded heterogeneity attributable to the choice of method within biobanks. On average, 82.9% of the total variance in PGS effect sizes was attributable to differences between biobanks, whereas only 12.0% was attributable to differences between methods.

After adjustment for sex, age, and genetic PCs 1-10, biobank-level heterogeneity was modestly reduced, indicating that part of the observed variability reflects differences in cohort composition. Nevertheless, for 10 of 13 traits, heterogeneity between biobanks remained larger than heterogeneity between methods. Exceptions were observed for T1D, T2D, and stroke.

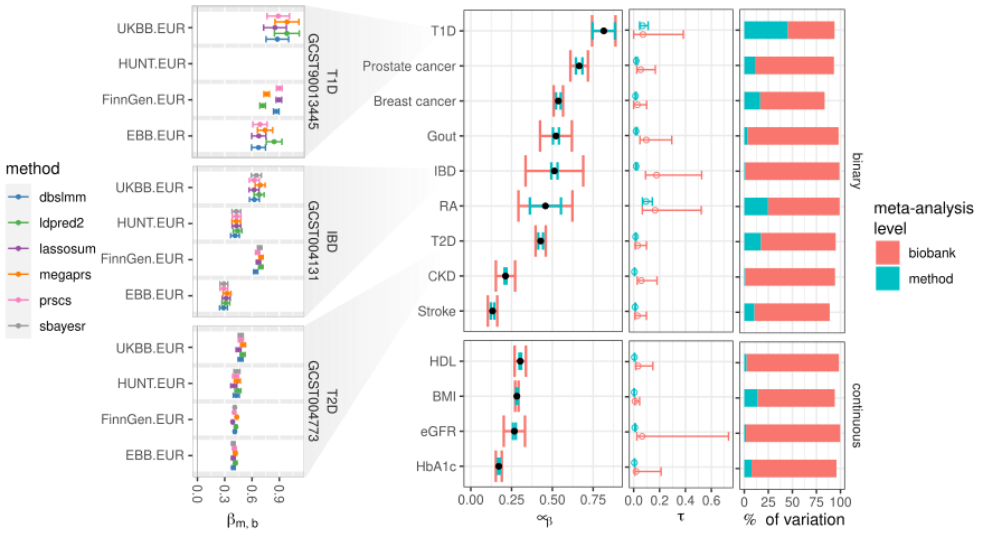


Figure 7 | Three-level meta-analysis of PGS effect sizes in EUR target data
For 13 traits replicated in at least two EUR biobanks and CV-tuned in UKBB, panels show (left to right): (1) PGS effect sizes with 95% confidence intervals across biobanks for three representative traits illustrating high method variability (T1D), high biobank variability (IBD), and low to intermediate variability (T2D); (2) meta-analyzed mean effect sizes across biobanks and methods, with variability at each level; (3) estimates of heterogeneity variance (τ) with 95% confidence intervals; and (4) I^2 estimates quantifying the proportion of variance attributable to biobank- or method-level heterogeneity. (Reprinted from *"Evaluation of polygenic scoring methods in five biobanks shows larger variation between biobanks than methods and finds benefits of ensemble learning"*, with permission from Elsevier (Study I))

Traits with the highest method-related heterogeneity included T1D and RA. Both traits are autoimmune diseases and showed a large fraction of PGS variance attributable to the HLA region, with mean proportions of approximately 70% for T1D and 54% for RA. These findings suggest that complex LD structure and strong locus-specific effects in the HLA region increase sensitivity to methodological differences in PGS construction.

Together, these results demonstrate that, when PGS methods are appropriately tuned, variability in predictive performance is driven more strongly by differences between biobanks than by the choice of method itself, highlighting the importance of cohort characteristics, phenotype definitions, and data-generating processes in cross-biobank genetic prediction.

5.2 Refining Psychiatric Phenotypes to Reduce Misclassification (Study II)

5.2.1 Simulation of Misclassification and SuperControl Bias

After demonstrating in Study I that cross-biobank variability remains a major source of heterogeneity, we focused in Study II on traits that are particularly affected by EHR-based phenotype definitions, highly polygenic genetic architectures, and limited transferability across cohorts. Psychiatric traits pose a particular challenge in this setting due to their highly polygenic and overlapping genetic architecture, as well as often unclear and noisy phenotype definitions.

We therefore revisited an established phenotype definition strategy based on super-normal controls and adapted it to EHR data, referring to this approach as SuperControls. Because super-normal control designs have been justifiably criticized for potentially introducing bias-specifically by amplifying genetic signals related to excluded control traits rather than the target trait[154–156], we first examined the technical properties of this approach using simulations. In particular, we investigated the relationship between SuperControl-induced bias and bias arising from noisy or misclassified phenotypes.

In the simulations, we observed that variants shared between traits consistently gained power under the SuperControl design, as expected. We also observed increased power for variants specific to trait B, corresponding to type I error inflation under the SuperControl phenotype. Importantly, this inflation was reduced in scenarios with increasing phenotype misclassification. At high misclassification rates, type I error for trait B-specific variants was further attenuated. For variants specific to trait A, excluding controls initially reduced power. However, with moderate to high misclassification rates, this loss was mitigated, such that power became comparable to that of the original phenotype definition (Figure 8). These effects were most pronounced when trait A had higher prevalence. Based on these findings, we focused the subsequent empirical analyses on more prevalent psychiatric traits to enable robust estimation of effects and replication.

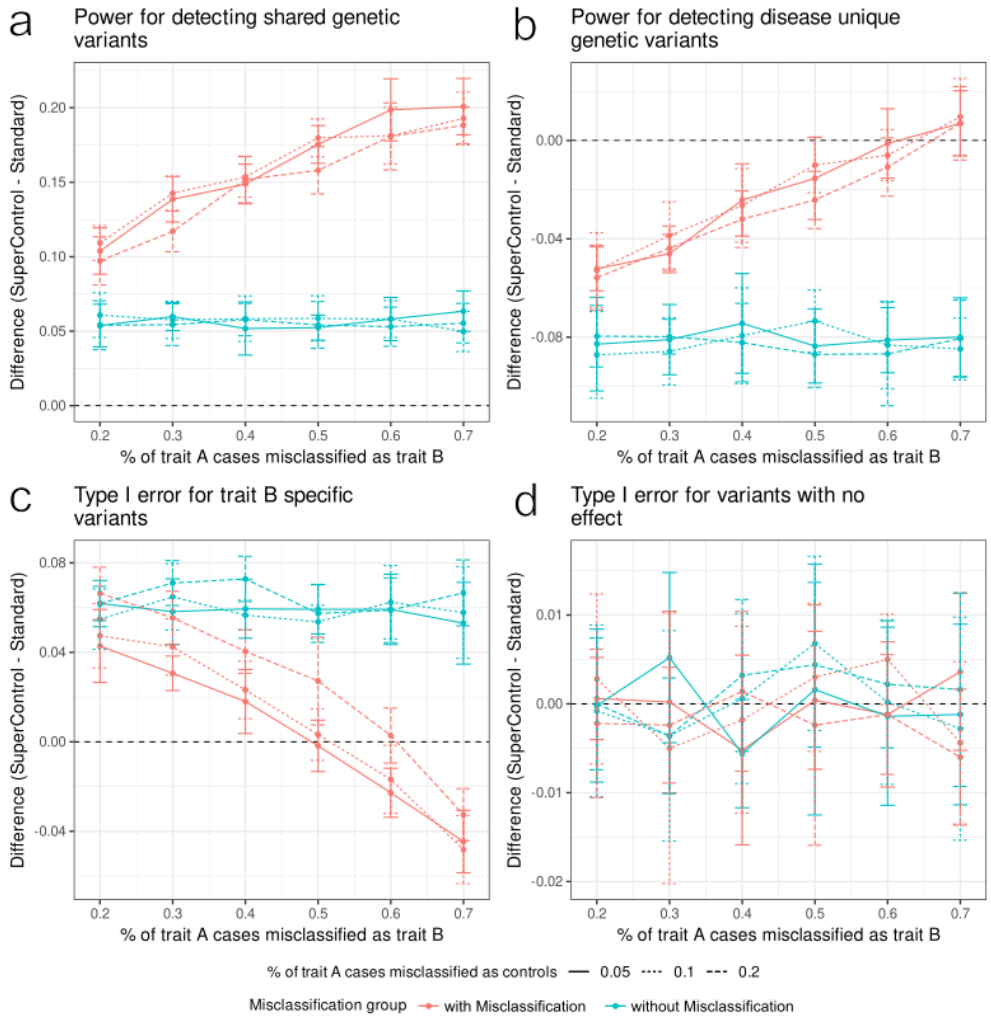


Figure 8 | Simulation of SuperControl and standard phenotypes under varying misclassification (50% trait prevalence)
(a-d) Differences between SuperControl and standard phenotypes in power or type I error across increasing levels of misclassification in trait B and in non-trait B controls.
(a,b) Power for SNPs shared between traits A and B (a) and SNPs specific to trait A (b); positive values indicate higher power under the SuperControl phenotype.
(c,d) Type I error for SNPs causal only for trait B (c) and for non-causal SNPs (d); negative values indicate lower error rates under the SuperControl phenotype.
Curves show the mean of ten independent simulations, with error bars indicating standard deviation.

5.2.2 Empirical Application of SuperControl Phenotypes in EHR

After establishing that phenotype misclassification and SuperControl bias interact, we next evaluated the SuperControl phenotype in empirical data. We applied this approach to four psychiatric traits with sufficient prevalence: MDD, ANX, SUD, and MHD. In addition, we

analyzed six further traits with lower prevalence to assess the impact of prevalence on power and replication: OCD, PTSD, ADHD, SCZ, SLD, and BD.

Case counts ranged from 2,830 individuals (prevalence 0.6%) for OCD to 59,969 individuals (11.8%) for MDD, with 140,342 individuals (27.6%) for MHD. Despite the reduced sample size under the SuperControl definition, GWAS power increased substantially, yielding more than double the number of genome-wide significant loci. After merging overlapping regions, we identified 94 independent loci for the Original FinnGen phenotype and 201 loci for the SuperControl phenotype. Of these, 70 loci were shared between definitions, while 131 loci were unique to the SuperControl phenotype, representing additional discoveries not captured by the original approach (Figure 9).

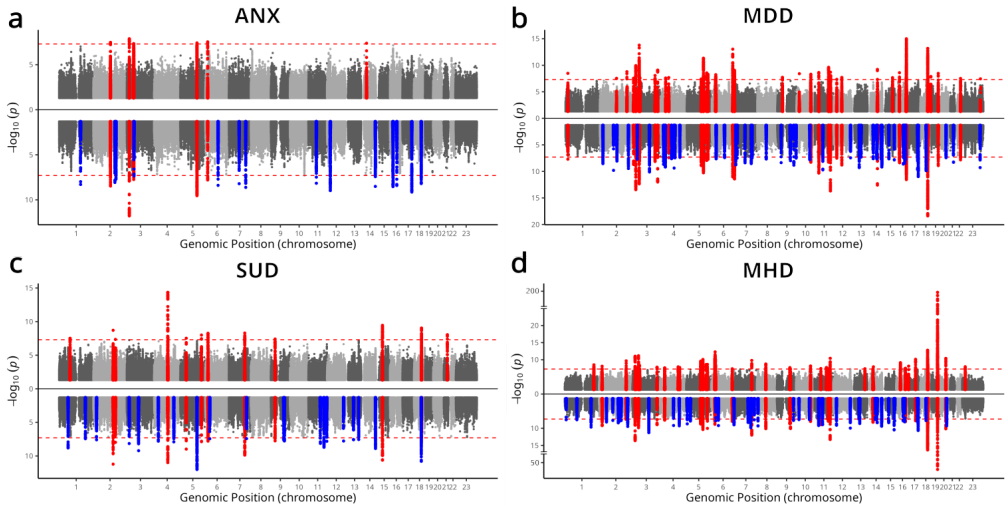


Fig. 9 | Genome-wide significant loci in Original FinnGen and SuperControl phenotypes across psychiatric traits.
a-d, Miami plots for ANX, MDD, SUD, and MHD. In each plot, the upper half represents results from the Original FinnGen phenotype and the lower half from the SuperControl phenotype. Red points denote loci shared between the two phenotypes or unique to the Original FinnGen phenotype, and blue points denote loci unique to the SuperControl phenotype.

We next assessed whether excluding psychiatric cases from the control group reduced genetic distinction between disorders. While overall genetic correlations increased modestly (mean difference = 0.19, $p = 6.5 \times 10^{-13}$), the overall structure of genetic relationships remained highly consistent (Pearson $r = 0.96$, $p < 2.2 \times 10^{-16}$ across pairwise genetic correlations for all ten traits; Figure 10 a, b).

To further assess functional specificity, we performed partitioned heritability analyses across all ten traits. The strongest change was observed for the central nervous system (CNS), where enrichment increased significantly under the SuperControl definition ($p = 1.9 \times 10^{-3}$; FDR-adjusted p = 0.02). Trait-level analyses were consistent with this pattern. Under the SuperControl definition, CNS enrichment reached significance for several disorders that showed no such evidence using the Original FinnGen phenotypes, including SCZ ($p = 1.7 \times 10^{-2}$) and SLD ($p = 3.8 \times 10^{-3}$). In contrast, skeletal muscle enrichment previously observed

for BD under the Original definition ($p = 2.0 \times 10^{-2}$) was no longer significant when using SuperControls ($p = 4.3 \times 10^{-1}$), consistent with an overall reduction in peripheral tissue enrichment. For SLD, adrenal-pancreas enrichment also became significant ($p = 7.6 \times 10^{-3}$), despite the absence of a consistent signal across other traits (Fig. 11 c). This pattern is consistent with the known involvement of sleep on endocrine pathways, including cortisol and other hormonal rhythms[212]. Together, these results suggest that refined control definitions enhance the detection of biologically coherent and trait-specific enrichment signals (Figure 10 c).

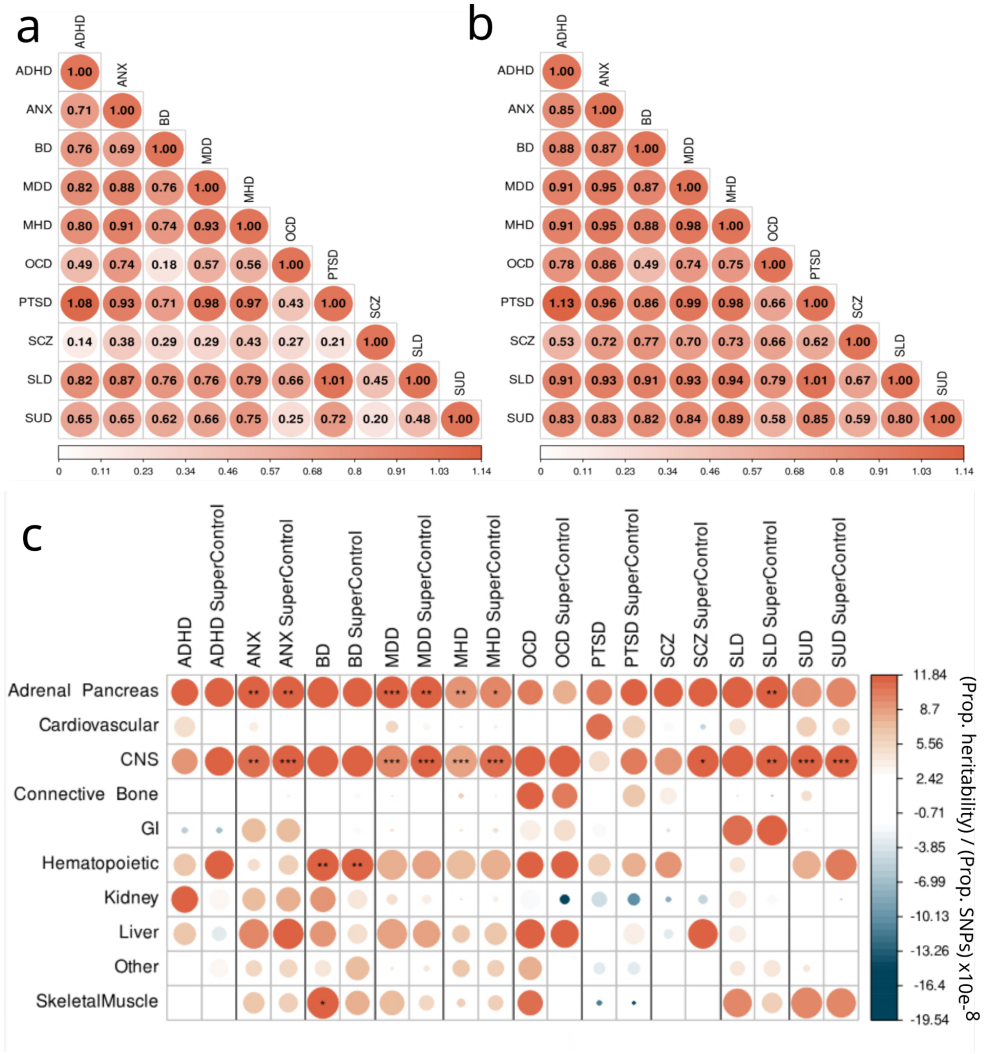


Fig. 10 | Genetic architecture improvements through refined phenotyping.
All panels show data for the 10 psychiatric traits (OCD, PTSD, ADHD, SCZ, MDD, ANX, SUD, SLD, BD, and MHD). (a,b), Heatmaps of pairwise genetic correlations among the traits. (a) shows correlations for the Original FinnGen phenotypes, and (b) for the SuperControl phenotypes. (c), Heatmap of partitioned heritability using the cell-type baseline dataset. Columns represent the 10 traits, and rows represent cell types. Stars indicate significant enrichment.

5.2.3 Recovering Excluded Samples using Machine Learning

Although the SuperControl phenotype increased GWAS power, it also excluded a substantial portion of the cohort, reducing effective sample size. To recover this information, we trained a XGBoost model on the SuperControl cohort using all available biobank features except those required for SuperControl definition. We then used this model to predict trait liability in the excluded cohort. We then applied MTAG to jointly analyze the SuperControl and predicted phenotypes, resulting in PRISMA.

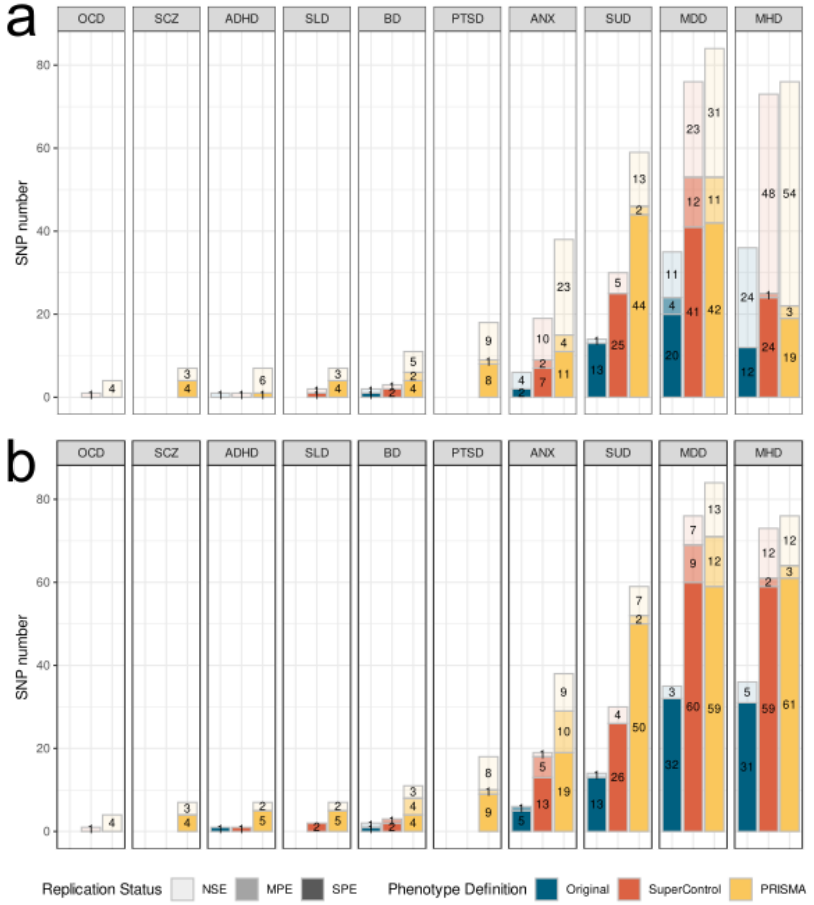


Fig. 11 | Replication of independent loci in Original, SuperControl and PRISMA phenotypes across psychiatric traits.
a-b Barplots showing LBR, (a) and GBR (b) for ten psychiatric traits: SCZ, MDD, BD, ANX, ADHD, PTSD, OCD, SLD, SUD and MHD. Each trait contains three bars representing loci unique to the Original phenotype (left), SuperControl Phenotype (middle), and PRISMA phenotype (right). Bars are further segmented by replication strength: strong prior evidence (SPE, darkest shading), moderate prior evidence (MPE, intermediate shading), and no supporting evidence (NSE, faintest shading)

After merging overlapping regions, we identified 307 independent loci for PRISMA, compared with 201 for the SuperControl and 94 for the Original FinnGen phenotypes. Of these, 168 loci were shared between PRISMA and SuperControl, while 139 loci were unique

to PRISMA and only 33 remained unique to SuperControl. Overall, PRISMA increased discovery by 1.5-fold relative to SuperControl and 3-fold relative to the Original FinnGen phenotype (Figure 11).

5.2.4 Replication and Cross-Biobank Validation

To assess whether the additional signals reflected trait-specific associations or noise introduced by the control design or ML model, we evaluated replication using two complementary approaches. LBR focused on overlap with variants identified in the largest available GWAS for each trait. GBR focused on functional annotation of variants and their linked genes using Open Targets.

LBR rates were modest across phenotypes (55% for Original, 57% for SuperControl, and 51% for PRISMA). For SUD, where multiple reference GWAS were available, replication rates were higher (93%, 83%, and 77%), suggesting that the overall low replication rates partly reflect the conservative replication design. Importantly, replication proportions were similar across phenotype definitions, even though the absolute number of replicated loci increased from Original to SuperControl to PRISMA (52, 114, and 154 loci, respectively; Figure 11a).

Using functional annotation for replication, GBR resulted in higher replication rates (89% for Original, 88% for SuperControl, and 79% for PRISMA). The total number of replicated loci again increased substantially, from 84 to 176 to 244 respectively (Figure 11b), although replication proportions were slightly lower for PRISMA.

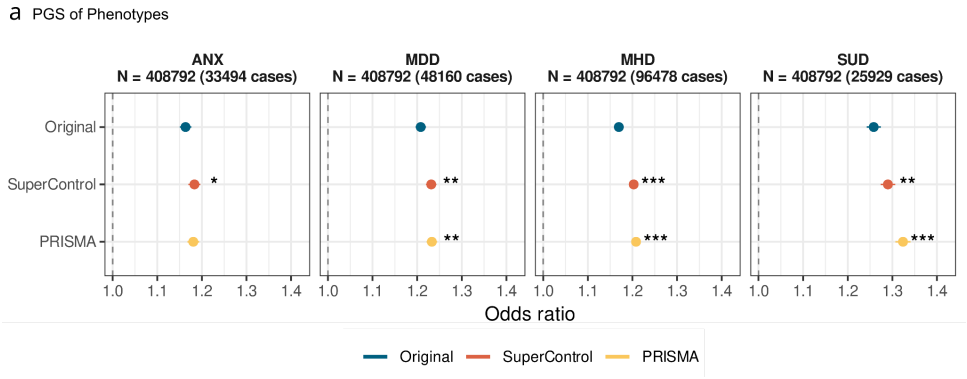


Fig. 12 | Polygenic prediction across biobanks
a, Odds ratios (with standard errors) from PGS prediction for four psychiatric traits, ANX, MDD, MHD, and SUD, based on GWAS from the Original FinnGen, SuperControl, and PRISMA phenotypes. Asterisks indicate traits where the SuperControl or PRISMA PGS showed significantly stronger prediction than the Original FinnGen PGS tested using two-sided *z*-test.

To further evaluate predictive utility, we performed polygenic score analyses in UKBB, focusing on the four main traits, as these were the only traits with more than 2,000 cases. Across all four traits, SuperControl PGS significantly outperformed the Original FinnGen scores, achieving higher odds ratios for ANX ($p = 3.6 \times 10^{-2}$), MDD ($p = 5.4 \times 10^{-3}$), MHD ($p = 9.8 \times 10^{-8}$), and SUD ($p = 6.4 \times 10^{-3}$). Similarly, PRISMA PGS further improved prediction

for MDD ($p = 3.3 \times 10^{-3}$), MHD ($p = 7.2 \times 10^{-10}$), and SUD ($p = 2.9 \times 10^{-8}$) compared with the Original FinnGen phenotype (Figure 12).

In addition to slightly reduced replication rates, PRISMA showed increased cross-trait sharing, indicating modest inflation of pleiotropy under the imputed phenotypes. Partitioned heritability analyses showed that CNS enrichment was largely maintained but somewhat reduced relative to SuperControl, suggesting that PRISMA further increases power while introducing some additional noise and pleiotropy.

5.3 Genetic Architecture of Prescription Opioid Intake (Study III)

5.3.1 Genome-Wide Associations of Opioid Exposure and Cumulative Dose

After examining challenges related to cross-biobank analyses and EHR-based phenotyping, and developing ML-based approaches, we next applied these insights to a concrete use case that leverages the strengths of EHR data. We focused on prescription opioid use, a highly polygenic trait that is difficult to define, similar to psychiatric disorders, and a treatment area that has contributed substantially to the opioid crisis[22,23]. We aimed to determine whether opioid intake captures information beyond binary pain phenotypes and whether it reflects more than pain intensity alone.

To this end, we calculated OME for 859,675 EUR individuals across three biobanks: FinnGen, the EstBB, and UKBB (UKBB). Across cohorts, cumulative OME values ranged from 39 mg to 4,263,860 mg in FinnGen, 3.9 mg to 2,740,780 mg in EstBB, and 2 mg to 6,241,220 mg in UKBB. We defined three phenotypes: prescription exposure (RxExpPop; ever vs. never), cumulative dose among users (RxDoseUsers; log-transformed total cumulative OME), and population-level dose (RxDosePop; log-transformed total cumulative OME across the full cohort, with non-users set to zero). Extensive harmonization of phenotype definitions across the three biobanks resulted in GWAS with genetic correlations typically exceeding 0.8, demonstrating robust reproducibility despite differences in healthcare systems and prescription capture.

Meta-analysis identified a total of 234 independent genome-wide significant variants mapping to 145 unique genomic loci (Figure 13), with high concordance between biobank-specific results (Figure 14b). The largest number of associations was observed for RxDosePop (135 loci), reflecting increased power from integrating both exposure and dosage information. RxExpPop (78 loci) and RxDoseUsers (20 loci) were less powerful but still yielded multiple significant associations. Overlap patterns showed closer relationships between RxExpPop and RxDosePop, whereas RxDoseUsers shared fewer loci with either phenotype (Figure 14a).

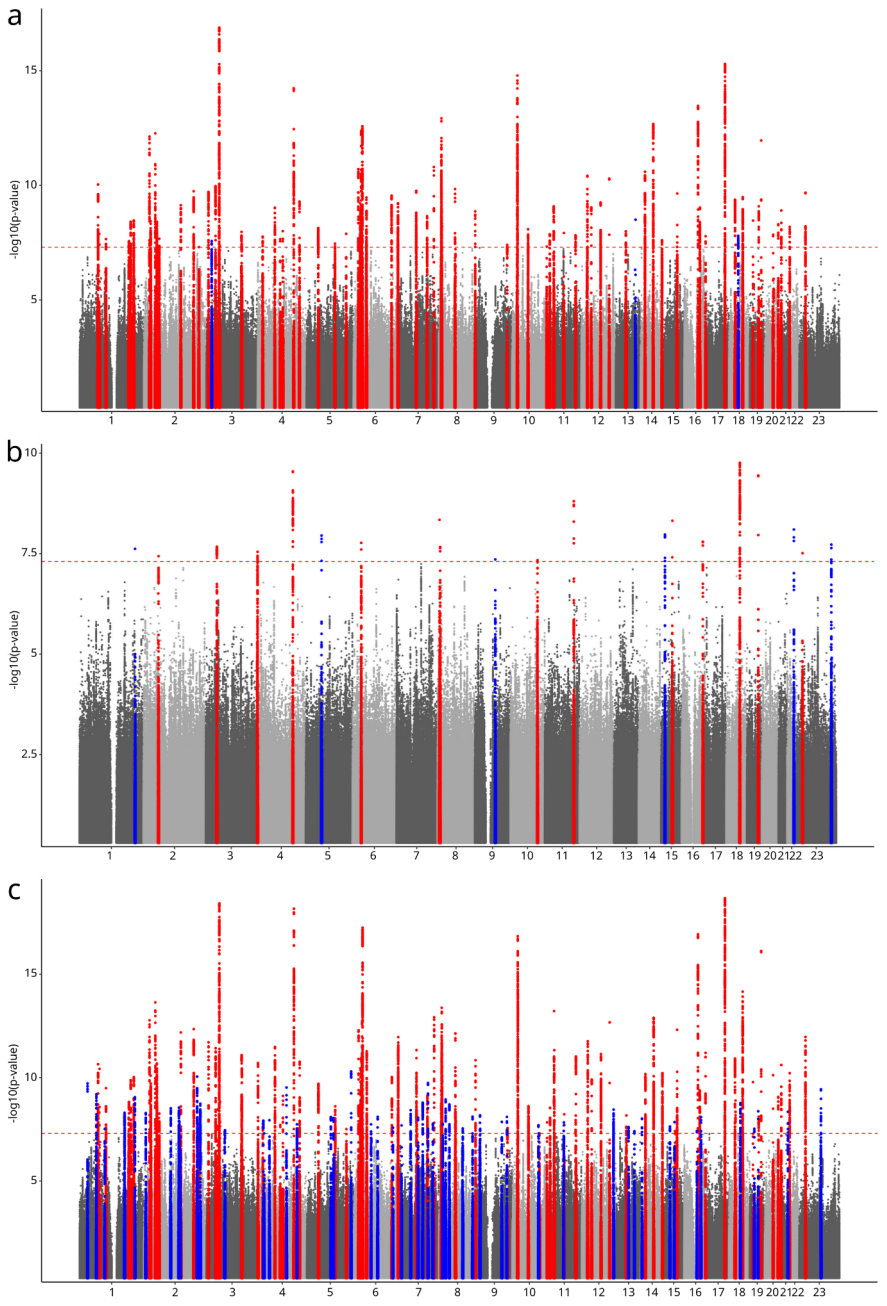


Figure 13 | GWAS results for three opioid prescription phenotypes
Manhattan plots of meta-analyses across biobanks for (a) RxExpPop (78 independent lead SNPs), (b) RxDoseUser (20 independent lead SNPs), and (c) RxDosePop (135 independent lead SNPs). The red line denotes the genome-wide significance threshold ($P < 5 \times 10^{-8}$), Loci in blue are unique to the trait and loci marked in red are shared with at least one other trait.

Annotation revealed previously reported trait associations for 168 of the 234 signals (71.8%). Of these, 65 signals (27.8%) overlapped with prior GWAS of pain, 58 of which were reported by Toikumo et al.[213] (Figure 14d), or analgesic use. Additional signals were associated with weight (21 signals; 9.0%), substance use behavior (10 signals; 4.3%), socioeconomic traits such as education (7 signals; 3.0%), and mental traits (6 signals; 2.6%). Cardiometabolic traits, including hypertension and diabetes, formed the largest remaining category, underscoring pleiotropic genetic links beyond pain-related pathways.

A recent large-scale pain meta-analysis showed strong genetic correlations with all three opioid phenotypes: RxExpPop (r_g = 0.76, p = 0), RxDoseUsers (r_g = 0.76, p = 3.7 × 10^{-280}), and RxDosePop (r_g = 0.77, p = 0). In contrast, correlations with substance use phenotypes, such as alcohol and drug use disorders, were more modest, ranging from 0.3 to 0.5 across phenotypes (Figure 14c), indicating that opioid intake remains most strongly associated with pain-related traits.

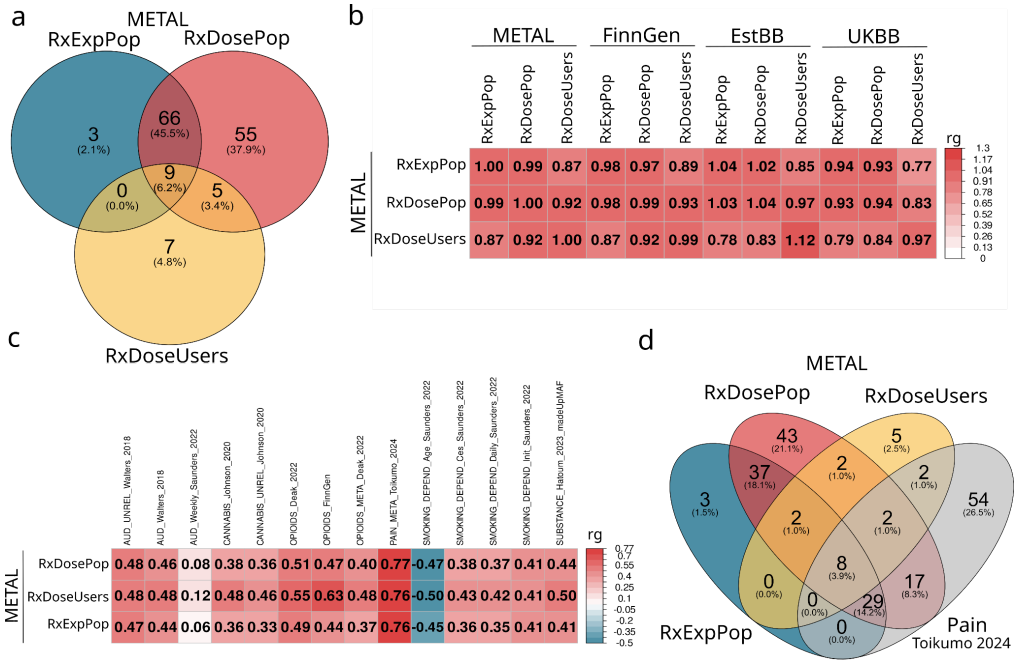


Figure 14 | Overlap and genetic correlations of opioid prescription phenotypes.
(a) SNP overlap among the three METAL phenotypes: RxExpPop, RxDosePop, RxDoseUser.
(b) Genetic correlations between METAL and individual biobank results (FinnGen, EstBB, UKBB).
(c) Genetic correlations of opioid phenotypes with published studies of alcohol use disorder[214,215], cannabis use[216], opioid dependence[139], pain[213], smoking behaviors[215], and substance use[140]. (d) SNP overlap between opioid phenotypes and the recent pain GWAS (Toikumo et al. 2024)[213]

5.3.2 Machine Learning-Derived Overuse and Early-Onset Phenotypes

Building on the ML approaches developed in Studies I and II, we next explored whether ML could be used to derive phenotypes that capture patterns of opioid overuse beyond standard prescription measures. We defined two complementary phenotypes: RxOverUse_Onset, capturing early onset of opioid prescription, and RxOverUse_Amount, reflecting excess dosage among individuals with comparable medical histories.

We applied an ensemble approach similar to Study I, training three XGBoost models for each phenotype. Predictions were averaged across models, and individuals with high prediction variance were excluded to reduce model noise. Overuse status was assigned conservatively when the predicted dosage exceeded the minimum observed opioid value in FinnGen.

GWAS of RxOverUse_Onset did not identify any genome-wide significant associations. In contrast, RxOverUse_Amount yielded a single genome-wide significant locus on chromosome 22 (rs5751229, $p = 1.12 \times 10^{-8}$). Fine-mapping localized this signal to rs58099562 on chromosome 22. Annotation using Open Targets showed that this variant was part of multiple credible sets across published GWAS, including studies of pain and insomnia, and mapped primarily to *CYP2D6/CYP2D7*. This gene plays a central role in drug metabolism and has well-established links to pain and substance use biology[217,218].

To further characterize the relationship between the ML-derived overuse phenotypes and baseline opioid intake measures, we estimated genetic correlations with the three primary opioid phenotypes. RxOverUse_Onset showed strongest correlations with population-wide measures ($r_g = 0.80$ with RxExpPop; $r_g = 0.76$ with RxDosePop) and weaker correlation with RxDoseUsers ($r_g = 0.59$). In contrast, RxOverUse_Amount correlated most strongly with RxDoseUsers ($r_g = 0.77$) and less strongly with RxExpPop ($r_g = 0.38$) and RxDosePop ($r_g = 0.49$) (Figure 15a).

We next compared genetic correlations with external GWAS of pain and substance use (Figure 15b). RxOverUse_Onset displayed patterns similar to the primary prescription phenotypes but with reduced power, showing its strongest correlation with pain ($r_g = 0.52$, $p = 3.0 \times 10^{-8}$) and weaker correlations with other traits. In contrast, RxOverUse_Amount exhibited a shifted profile. Although it remained correlated with pain ($r_g = 0.49$, $p = 4.5 \times 10^{-5}$), it showed stronger correlations with substance use traits, including alcohol use disorder ($r_g = 0.59$, $p = 3.8 \times 10^{-3}$) and the FinnGen opioid abuse phenotype ($r_g = 0.51$, $p = 7.7 \times 10^{-3}$).

To assess whether this shift reflected broader differences in genetic architecture, we estimated genetic correlations with all FinnGen phenotypes. RxOverUse_Onset showed its strongest correlations with pain-related traits, including low back pain ($r_g = 0.54$, $q = 5.5 \times 10^{-10}$), pain ($r_g = 0.64$, $q = 5.7 \times 10^{-14}$), codeine-tramadol prescription ($r_g = 0.73$, $q = 3.9 \times 10^{-14}$), and paracetamol prescription ($r_g = 0.64$, $q = 7.0 \times 10^{-15}$), closely mirroring the pattern observed for the baseline phenotypes (Figure 15c). In contrast, RxOverUse_Amount was dominated by psychiatric and substance use traits, with pain-related traits ranking lower (Figure 15d). The strongest correlations included panic disorder ($r_g = 0.89$, $q = 0.041$; rank 1), suicidal behavior ($r_g = 0.75$, $q = 0.041$; rank 2), suicide-related features ($r_g = 0.70$, q =

0.041; rank 3), alcohol dependence (r_g = 0.68, q = 0.041; rank 4), and substance use disorders (r_g = 0.66, q = 0.041; rank 5). Pain-related traits were less prominent, with only three appearing in the top 20: joint pain (r_g = 0.58, q = 0.041; rank 9), codeine-tramadol prescription (r_g = 0.57, q = 0.041; rank 10), and the aggregated pain phenotype (r_g = 0.50, q = 0.041; rank 14).

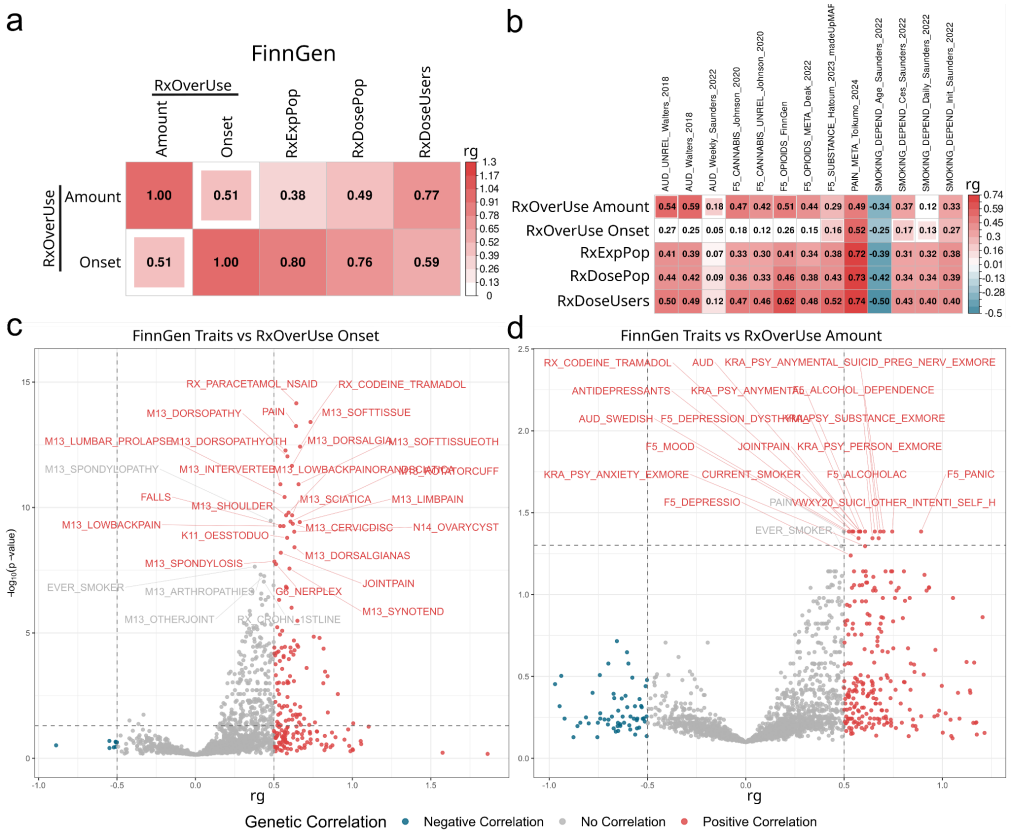


Figure 15 | Genetic correlations of ML-derived opioid overuse phenotypes.
(a) Genetic correlations between FinnGen overuse phenotypes (RxOverUse_Amount, RxOverUse_Onset) and prescription-based phenotypes (RxExpPop, RxDosePop, RxDoseUsers).
(b) Genetic correlations of all FinnGen phenotypes with published studies of alcohol use disorder[214,215], cannabis use[216], opioid dependence[139], pain[213], smoking behaviors[215], and substance use[140].
(c) Genetic correlations of FinnGen traits with RxOverUse_Onset, highlighting strong correlations with musculoskeletal pain, prescription pain medications, and related disorders.
(d) Genetic correlations of FinnGen traits with RxOverUse_Amount, showing significant correlations with psychiatric, substance use, and pain-related traits.

In summary, RxOverUse_Onset closely resembles baseline exposure phenotypes and primarily reflects pain-driven opioid prescription, whereas RxOverUse_Amount shows substantially stronger genetic overlap with psychiatric and substance use traits than with pain, highlighting a distinct genetic architecture underlying opioid overuse.

6 Discussion

Large-scale biobanks have transformed human genetics by enabling analyses at an unprecedented scale. At the same time, they have revealed fundamental limitations that become more visible as sample sizes increase. Across cohorts, genetic prediction and discovery are constrained not only by methodological choices, but more strongly by heterogeneity in data generation and phenotype definition. Differences in healthcare systems, diagnostic practices, and EHR structure introduce noise that reduces heritability, limits transferability, and complicates biological interpretation.

The studies in this thesis jointly examine these limitations and their consequences for genetic prediction. Rather than treating biobank scale as a sufficient solution, they investigate where prediction breaks down, why performance varies across cohorts, and to what extent this variability is driven by phenotype quality rather than modeling choice. Using data from multiple biobanks, the work first quantifies cross-biobank heterogeneity in polygenic score performance. It then shows that phenotype refinement, particularly for psychiatric traits, can recover genetic signal, improve biological specificity, and enhance cross-biobank transferability. Finally, these concepts are extended beyond diagnostic labels to prescription-derived traits, demonstrating how longitudinal treatment data can be leveraged to construct informative and scalable phenotypes for complex behaviors such as opioid use. Across all studies, ML is used not as an end in itself, but as a tool to reveal latent structure in noisy clinical data, integrate heterogeneous sources of information, and test how phenotype refinement reshapes the effective limits of genetic analysis. Taken together, these results indicate that future progress in biobank-based genetics will depend as much on improving phenotype definition and harmonization as on increasing sample size or methodological complexity.

6.1 Sources of Heterogeneity in Cross-Biobank Analysis

For many years, large-scale genetic studies have been motivated by the promise that increasing sample size would yield more robust and transferable results. The rapid expansion of biobanks, enabled by falling genotyping costs, healthcare digitalization, and national cohort initiatives, has made this scale possible. However, as analyses increasingly span multiple cohorts, it has become clear that large sample size alone does not guarantee consistency. Cross-biobank analyses remain challenging due to systematic differences between cohorts, and these differences often dominate variation in genetic prediction performance.

One area where this difficulty is particularly evident is polygenic scoring. Even with large biobanks, polygenic scores often show limited transferability across ancestries and even across cohorts within the same ancestry[5,111]. Performance varies substantially depending on both the target cohort and the phenotype definition[219]. In response, a growing number of polygenic scoring methods have been proposed, many of which claim superior performance.

To investigate this issue and to better understand the origin and impact of heterogeneity, we compared seven polygenic score methods across 14 traits and five biobanks.

This analysis was based on a previously published workflow, which we extended by adding systematic hyperparameter tuning and automating cross-biobank comparisons[6]. Polygenic score tuning was performed in the UKBB using individuals of EUR ancestry. To ensure fair comparability, we followed the recommended settings and LD reference panels proposed by the method developers. Across all analyses, no single method consistently outperformed the others. The two best-performing methods were MegaPRS and LDpred2 when tuned using CV in the UKBB.

Overall, CV methods showed improved performance compared to their default implementations showing that hyperparameter tuning does add valuable information for cross biobank performance, even if only trained in one cohort. In addition to comparing individual methods, we trained an elastic net model to construct an ensemble polygenic score. While this approach is not independent, as it combines the outputs of the individual methods, it outperformed all single methods in 13 out of 14 traits. This result illustrates the advantage of ensemble approaches, which can leverage complementary strengths across different models.

A further opportunity provided by this multi-method, multi-cohort setting was the ability to investigate whether variability in prediction performance is primarily driven by method choice or by differences between biobanks. To address this, we implemented a three-level meta-analytic mixed model to partition variance attributable to methods and to biobanks. Across the majority of traits, variance attributable to differences between biobanks exceeded variance attributable to method choice.

The largest method-related variance was observed for autoimmune diseases such as type 1 diabetes and rheumatoid arthritis. This is likely driven by the human leukocyte antigen (HLA) region, which plays a central role in these diseases. The HLA region is highly polymorphic and inherited as extended haplotypes, resulting in strong LD[220]. These properties make it difficult to disentangle the effects of individual variants, and differences in how methods model and process this region likely contribute to performance variation. In contrast, the overall method variance was relatively low for many traits, possibly reflecting the use of well-powered and high-quality GWAS summary statistics. With lower-quality GWAS, method-dependent differences may become more pronounced.

Before adjustment, prediction variability was dominated by differences between biobanks across all traits. After adjusting for age, sex, and genetic PCs 1-10, biobank-level variance remained larger than method-level variance for 10 out of 13 traits. While these population characteristics accounted for part of the observed variability, a substantial proportion of variance persisted and was attributable to other biobank-specific factors. Additional factors likely include differences in genetic architecture between cohorts, as previous work has shown that polygenic score performance declines with increasing distance between the discovery GWAS and the target population[5,111]. Phenotype definition also appears to play a central role. BMI showed the lowest biobank-level variance, consistent with it being a standardized measurement that is less sensitive to differences in clinical practice and data collection.

This study therefore addressed not only method performance, but also broader sources of prediction variability and their origins. A limiting factor in this work was that, out of the large and rapidly growing number of available polygenic scoring approaches, only seven methods could be included. This restricted set enabled a consistent and systematic comparison but necessarily represents only a small fraction of the current methodological landscape. During the course of this work, updated versions and new recommendations were published even for the included methods, which may influence relative performance. As a result, conclusions regarding method superiority should be interpreted cautiously in a fast-moving field. In addition, although multiple biobanks and two ancestry groups were included, expanding analyses to further ancestries remains essential.

At the same time, the study has several strengths. We provide a standardized, reproducible workflow with containerized version control, which can be executed even in highly secure environments without internet access. This facilitates transparent comparison and reuse across cohorts. Moreover, the ensemble approach, trained only in the UKBB and applied without retuning in other cohorts, consistently outperformed individual methods. This suggests that ensemble methods can improve robustness by integrating complementary information across methods and may be synergistic with approaches that combine multiple GWAS, such as MultiPGS[221]. Most importantly, by explicitly partitioning sources of variance, we show that cross-biobank heterogeneity has a larger impact on polygenic score performance than method choice itself, emphasizing the central role of biobank-specific differences.

6.2 Phenotype Definition in Electronic Health Records: Limits and Opportunities

After establishing the extent of cross-biobank variability and the potential role of phenotyping in driving these differences, we focused on a subset of traits that are known to be particularly noisy in EHR data. Phenotypes derived from EHR are inherently imperfect, as clinical data are collected for care rather than for research purposes. Diagnoses may be incomplete, delayed, or inconsistent across time and settings. In addition, EHR-based case-control definitions often rely on unscreened controls, assigning individuals to the control group who may develop the disease later or who remain undiagnosed[18].

These issues are especially pronounced for psychiatric traits. Psychiatric disorders are difficult to diagnose due to symptom heterogeneity, variable disease trajectories, and frequent diagnostic revisions[125,126]. Social stigma further affects help-seeking behavior and diagnostic accuracy[222]. Many psychiatric conditions are better described as continuous spectra rather than binary states[130], and they are highly intertwined through comorbidity and pleiotropy[131,223]. Together, these factors increase misclassification and dilute genetic signal in standard EHR-based phenotypes.

Concerns about misclassification in psychiatric genetics predate modern biobanks. Earlier studies addressed this problem by screening controls and excluding individuals with related or comorbid psychiatric conditions, an approach commonly referred to as “super-normal” controls[150], here termed SuperControls. While long-standing methodological debate has shown that SuperControl designs introduce bias by altering the relationship between the trait

of interest and excluded conditions[154–156], other work has demonstrated that misclassified controls can substantially reduce power and inflate noise[14]. Although increasing sample size is often proposed as a solution, previous studies suggest that for highly pleiotropic traits, the sample sizes required to fully overcome misclassification remain beyond what is currently achievable[143].

Against this background, Studies II and III examined whether phenotype refinement could mitigate some of these limitations, either through stricter control definitions or by leveraging EHR data to construct continuous phenotypes rather than binary case-control labels. While SuperControl definitions have been justifiably criticized for introducing bias, unscreened EHR controls are also biased through misclassification. This raises the question of whether, in sufficiently noisy settings, SuperControl bias may partially offset misclassification bias rather than exacerbate it.

To address this question, Study II used simulations to explicitly examine the interaction between SuperControl bias and misclassification bias. These simulations showed that for traits with higher prevalence and substantial misclassification, SuperControl definitions could better recover trait-specific genetic signal than standard case-control designs. Shared genetic signal was consistently boosted, and type I error for trait-specific false positives was not increased under high misclassification. Signals unrelated to either trait were not spuriously affected. These results suggest that in highly noisy environments, SuperControl definitions can be beneficial despite their inherent bias.

Guided by these findings, we applied a SuperControl design to ten psychiatric traits in FinnGen. Despite the use of stringent exclusion criteria and a substantial reduction in sample size, this approach led to a marked increase in genome-wide significant loci, with more than a doubling of detected associations compared to the original phenotype definitions. To further increase power, we predicted psychiatric status in excluded individuals and reintegrated them using MTAG in the PRISMA framework. This approach yielded more than three times as many loci as the original phenotypes. While SuperControl definitions provided the largest gains for highly prevalent traits, PRISMA contributed most strongly to traits with moderate or lower prevalence.

Beyond locus discovery, SuperControl phenotypes showed increased biological specificity. Partitioned heritability analyses revealed stronger enrichment in CNS tissues. In SLD, for example, the SuperControl phenotype showed clear enrichment in adrenal-pancreas tissues and the CNS, which was absent in the original FinnGen definition. This pattern is consistent with known links between endocrine, metabolic, and sleep regulation pathways[212,224] and suggests that filtering out misclassified cases can sharpen biologically relevant signals.

At the same time, these gains came with clear trade-offs. Replication analyses at both the locus and gene level showed similar replication rates for the original and SuperControl phenotypes, while PRISMA exhibited nominally reduced replication. Genetic correlations between psychiatric traits increased overall, with SuperControl showing only moderate and PRISMA stronger increase accompanied by reduced separation between traits. This pattern suggests that while PRISMA increases power, it also introduces additional noise and phenotype fusion.

Several limiting factors of this approach should be emphasized. The exclusion criteria applied in Study II were intentionally broad to make effects on power and bias visible, but such brought definitions may not be appropriate for all traits. We recommend more carefully defined exclusion strategies that account for overlapping symptom profiles, comorbidity, and pleiotropy. In addition, SuperControl designs were less effective for very low-prevalence traits with limited baseline power, restricting their general applicability. Genetic distance between traits was reduced under SuperControl definitions, indicating that classic criticisms of SuperControl bias remain valid. The interaction between misclassification and control selection bias is complex, and an optimal balance between the two has yet to be defined. Finally, while PRISMA increased discovery, it also amplified noise and SuperControl-related bias, underscoring that imputation-based refinement must be applied cautiously.

Despite these limitations, the SuperControl phenotype consistently yielded higher power with comparable replication and clearer biological patterns. This indicates that EHR-derived noise is a major limiting factor for psychiatric genetics and that excluding highly comorbid or pleiotropic cases can remove substantial non-specific signal. The observed improvement in cross-biobank transferability further suggests that this approach may be valuable for highly polygenic and pleiotropic traits, particularly where further increases in sample size are not immediately feasible.

A complementary perspective on phenotype refinement was explored in Study III, focusing on pain, a highly heterogeneous trait characterized by symptom variability and diagnostic instability. Despite its diversity, pain is among the most prevalent and disabling conditions worldwide[161,163]. Rather than relying on diagnostic labels, we used opioid prescription as a proxy for pain burden and constructed three opioid intake phenotypes across three biobanks.

Meta-analysis identified 234 genome-wide significant signals mapping to 145 unique loci. Exposure-based phenotypes, RxExpPop and RxDosePop, shared much of their genetic architecture, while cumulative dose among users, RxDoseUsers, captured partially distinct genetic determinants. Age and BMI contributed strongly to both opioid exposure and accumulated dose, reflecting increasing pain burden and multimorbidity with age and higher BMI[161,225,226]. Sex effects were bidirectional, with female sex associated with a higher likelihood of receiving any opioid prescription but a lower cumulative dose among users, consistent with known differences in pain perception, healthcare use, and opioid metabolism[227].

Genetic correlation analyses confirmed extensive overlap with pain-related traits, supporting opioid intake as a meaningful proxy for pain intensity. At the same time, many loci showed pleiotropic associations with obesity, cardiovascular disease, and psychiatric traits, including substance abuse. These findings align with epidemiological evidence linking pain, mental health, and multimorbidity[161,225,228].

To further disentangle opioid use patterns, we constructed early-onset and overuse phenotypes using ensemble-based ML. The early-onset phenotype yielded no genome-wide significant associations and closely mirrored general pain genetics. In contrast, the overuse phenotype identified a significant locus on chromosome 22. Fine-mapping implicated the *CYP2D6/CYP2D7* region, highlighting pharmacokinetic variation as a key contributor. The

lead variant, rs58099562, is in strong LD with rs3892097 in the Finnish population ($r^2 \approx 0.96$, LDlink[229]). This variant reduces *CYP2D6* enzymatic activity, resulting in poor metabolism of opioids such as codeine, which requires *CYP2D6*-mediated conversion to morphine[218,230–232]. Reduced analgesic efficacy in poor metabolizers provides a biologically coherent explanation for increased opioid intake.

Furthermore, opioid dose showed stronger genetic correlations with mental health traits than with pain-related traits, particularly for RxDoseUsers and the overuse phenotype. While prescribed opioid use largely reflects underlying pain, excessive opioid consumption appears more closely linked to neurobiological factors related to pain processing, tolerance, and vulnerability to addictive behaviors. Although the underlying mechanisms remain incompletely understood, these results point to a meaningful biological connection between opioid dosage patterns and mental health.

Using opioid prescriptions as a proxy for pain introduces clear limitations. Opioid intake is highly multimorbid and shares genetic signal with many traits, making it difficult to isolate pain-specific effects. At the same time, this multimorbidity is also informative, as it reflects the complex clinical context in which pain is treated. Importantly, the overuse phenotype yielded biologically interpretable loci, demonstrating that ML-based phenotype construction can uncover meaningful variation in heterogeneous clinical data.

Overall, phenotype refinement emerges as an active design choice that reshapes the effective phenotype used in genetic studies. Rather than serving as a fixed input, phenotype definition directly determines the balance between signal and noise and influences power, transferability, and biological interpretability. This is particularly relevant for EHR-derived traits, where missingness, evolving diagnostic criteria, and limited diagnostic precision are common. Refined control definitions in psychiatric traits and prescription-derived measures of pain and opioid use illustrate how moving beyond nominal diagnostic labels can recover genetic signal and better capture clinically relevant variation. At the same time, pleiotropy and comorbidity limit the suitability of simple binary phenotypes, motivating alternative approaches that incorporate broader or more continuous information.

These gains, however, come with unavoidable trade-offs. Increasing phenotypic specificity often reduces inclusiveness and sample size, while broader or reconstructed phenotypes recover power at the cost of increased pleiotropy and reduced interpretability. Approaches that prioritize statistical power, particularly those leveraging high-dimensional or longitudinal data, may obscure biological insight, whereas more constrained definitions offer clearer interpretation with reduced sensitivity. No single phenotype definition is universally optimal. Instead, phenotype refinement shifts the balance between power, specificity, and interpretability in predictable ways. Explicitly recognizing and managing these trade-offs is therefore essential for robust genetic analysis in EHR-linked biobanks and for meaningful cross-biobank comparison and biological interpretation.

6.3 Machine Learning as an Opportunity for Signal and a Potential Source of Noise

ML methods are increasingly applied in human genetics. Across the studies in this thesis, ML was used to enhance predictive performance, increase power, and open new perspectives on complex phenotypes. At the same time, these applications also revealed clear limitations and risks, particularly related to noise amplification and bias.

In Study I, we applied an elastic net model to construct an ensemble polygenic score in the UKBB. This ensemble outperformed all individual methods in 13 out of 14 traits. This result reflects a general property of ensemble approaches, which often perform better than single models because individual methods are optimized for specific assumptions or datasets, and combining them can capture complementary signal[233]. A limiting factor of this approach is its complexity. Ensemble methods require substantially more computational resources than single models and depend on the availability and compatibility of their constituent methods. As individual polygenic scoring methods evolve rapidly and are not always backward compatible, ensemble solutions may become outdated quickly. As a result, such approaches may be most appropriate when maximal predictive performance is required, for example in clinical or high-stakes settings. Nevertheless, combining ensemble methods with aggregated GWAS information, as done in approaches such as MultiPGS[221], appears to offer a particularly robust strategy.

ML has also been widely applied to improve phenotyping in EHR-based studies. Recent work has used predictive models to address missingness[27], impute diagnoses[29], broaden narrow phenotype definitions[16], integrate additional data sources such as clinical records[171] or family history[178], apply liability-scale modeling[28], or predict disease onset[172]. These approaches aim to recover information that is not explicitly encoded in structured diagnostic labels.

In Studies II and III, we used XGBoost for phenotype refinement. XGBoost is well suited to EHR data for several reasons. It is inherently an ensemble method, combining multiple decision trees where each new tree focuses on residual error from previous ones. This structure allows it to capture complex non-linear relationships and often achieve performance comparable to more complex deep learning models[234]. XGBoost can also handle missing values directly, which is particularly important for EHR data where missingness is often informative and not at random. In addition, as a tree-based method, it allows partial interpretability through feature attribution approaches such as SHAP[197], which is important in medical contexts where opaque models may reinforce existing biases.

In Study II, XGBoost was used to predict psychiatric status among individuals excluded under the SuperControl definition. To reduce potential noise, we did not use continuous liability scores but instead applied a binary threshold and integrated predicted individuals into GWAS using MTAG. This approach increased locus discovery by approximately 3-fold compared to the original phenotype. At the same time, it also led to increased genetic correlations between psychiatric traits and a nominal reduction in replication rates. This indicates that while prediction can amplify true signal, it can also amplify shared or non-specific signal, introducing additional noise. Similar observations have been reported in previous work, where predicted phenotypes tend to exhibit increased noisiness[180,181]. One

contributing factor may be that predictors lacking disease-specific information introduce correlated structure unrelated to the target trait, resulting in false-positive associations.

A commonly proposed mitigation strategy is to restrict model inputs to predictors that are strongly informative for the trait of interest and to avoid training on all available variables[180,181]. In the context of SuperControl designs, however, this is challenging. Traits used to define SuperControls cannot be included as predictors without reintroducing bias, as they are explicitly excluded from the control group. This constraint limits the information available for prediction and highlights the need for further refinement of SuperControl definitions to balance power gains against bias and noise.

Lessons from Studies I and II informed a more cautious application of ML in Study III. For opioid phenotypes, we trained three independent XGBoost models, leaning further into ensemble principles. Predictions were averaged, and individuals with high variance across models were excluded to reduce uncertainty in early-onset and overuse classifications. For the overuse phenotype, we applied an additional threshold, requiring at least one unit of FinnGen OME, to avoid marginal predictions. Because pain is a highly multimorbid trait and the task was closer to reclassification than de novo prediction, we included all available predictors, as the true labels were already observed and many predictors were potentially meaningful.

This approach enabled the construction of a novel phenotype that would not have been accessible without ML. In FinnGen, it revealed a biologically interpretable locus linking opioid overuse to impaired drug metabolism. The identified association supports a plausible mechanism whereby individuals with reduced opioid metabolism require higher doses to achieve analgesic effects, leading to increased cumulative intake.

A broader challenge in ML-augmented genetic analysis is verifying that newly identified associations reflect true biological signal rather than noise. One approach is to assess biological plausibility through pathway analysis or prior evidence, as done in Study III. However, this strategy is limited for highly polygenic traits, where hundreds of loci may be implicated, or for traits where biological mechanisms remain poorly understood. Other approaches compare ML-derived GWAS to results from original phenotypes and filter heterogeneous signals, such as POP-GWAS[181]. This strategy, however, assumes that the original phenotype is a reliable reference. When misclassification is a central problem, using the original phenotype as a benchmark may constrain discovery rather than improve it. We therefore applied multiple complementary validation strategies, including comparison with the largest available trait-specific GWAS and annotation using resources such as Open Targets. While effective for well-studied traits, these approaches are less informative when prior knowledge is limited. Pleiotropy-based metrics offer another perspective by assessing whether gains in predictability are trait-specific or reflect general genetic signal[16]. These measures also have limitations, particularly when the reference trait is itself noisy or underpowered.

An additional concern is that ML models inevitably learn biases present in the data. Since fully unbiased clinical data are currently unattainable, ML must be applied with caution in medical genetics. Improved performance does not guarantee improved validity, and gains in power may come at the cost of increased bias or reduced interpretability.

Taken together, these findings suggest that ML can be a powerful tool for phenotype refinement and genetic discovery, but only when applied in a carefully designed and transparent manner. Ensemble learning and exclusion of high-uncertainty predictions can improve robustness. When predicting binary traits, applying thresholds rather than using continuous liabilities can reduce unintended capture of shared structure. Finally, results should be validated using multiple complementary strategies rather than relying on a single metric. ML expands what is possible in biobank genetics, but its benefits depend critically on how models are constructed, constrained, and interpreted.

7 Conclusion and Future Directions

While large-scale biobank data enable powerful genetic analyses, they also reveal persistent limitations related to heterogeneity, phenotype definition, and data structure. Rather than being resolved through scale alone, these limitations become more apparent in cross-biobank settings, where genetic prediction varies across cohorts depending on methodological and phenotypic choices. This thesis therefore focuses on cross-biobank genetic analysis, examining both the gains enabled by increasing sample size and the challenges that arise when analyses are extended across cohorts.

In Study I, we showed that no single polygenic scoring method consistently outperformed others across traits and biobanks. Instead, an ensemble approach combining multiple methods achieved the highest performance in 13 out of 14 traits. At the same time, variance decomposition analyses revealed that differences between biobanks accounted for a larger proportion of performance variability than differences between methods. This finding emphasizes that cross-biobank heterogeneity remains a dominant factor in genetic prediction and highlights the importance of harmonization across cohorts.

Building on this observation, Study II examined phenotype definition as a potential contributor to cross-biobank variability, focusing on psychiatric traits derived from EHR data. By applying a SuperControl definition that excluded individuals with other psychiatric diagnoses, we observed substantial increases in power despite reduced sample size, with replication rates comparable to the original phenotypes. Reintegration of excluded individuals using ML further increased discovery but also led to increased genetic correlations between traits and reduced replication. These results illustrate that gains in power are often accompanied by trade-offs in specificity and noise, and that phenotype refinement reshapes genetic signal rather than uniformly improving it. The findings underscore that decisions about phenotype construction require careful consideration of the balance between power, specificity, and interpretability.

In Study III, these concepts were extended beyond diagnostic labels to prescription-derived phenotypes of opioid intake across three biobanks. Opioid use emerged as a highly multimorbid trait influenced by a wide range of genetic and clinical factors. By constructing an overuse phenotype using ML, we identified a locus linked to impaired opioid metabolism, providing a biologically plausible explanation for increased dosage in poor metabolizers. This study demonstrates that ML can be used to extract meaningful signal from heterogeneous longitudinal data, while also highlighting that such phenotypes capture complex mixtures of biological, behavioral, and clinical processes.

Taken together, the studies in this thesis show that large biobanks substantially increase statistical power, but do not remove limitations arising from heterogeneity in data generation and phenotype definition. These limitations are particularly pronounced for highly polygenic and pleiotropic traits, where shared genetic architecture, comorbidity, and diagnostic uncertainty dilute signal and complicate interpretation. Across all analyses, phenotype definition emerged as an active design choice that fundamentally shapes genetic results,

influencing power, transferability, and biological specificity. Methodological advances and ML can enhance signal and enable new analyses, but they cannot substitute for careful phenotype construction and evaluation.

Rather than identifying a single optimal strategy, this work highlights the need to explicitly manage trade-offs inherent in biobank-based genetic analysis. Different research goals require different balances between inclusiveness and specificity, discovery and interpretability, and robustness and sensitivity. Recognizing these trade-offs, and making them explicit, is essential for meaningful cross-biobank comparison and for responsible interpretation of genetic findings.

Future Directions

While EHR records and genetic data continue to grow rapidly, this thesis highlights underlying limitations that cannot be addressed through increasing sample size alone. This is particularly true for highly polygenic and pleiotropic traits, where the amount of data required to achieve high heritability and stable prediction remains beyond what is currently available[143]. As a result, future progress will depend not only on data growth, but also on improvements in data harmonization, phenotype design, and interpretation.

A central priority is the harmonization of data across biobanks. In clinical practice, international efforts already exist to standardize diagnostic and prescription coding, such as ICD-10 and ATC codes recommended by the WHO. Despite increasing adoption, countries often retain local coding systems or apply shared codes differently due to differences in healthcare structure, reimbursement policies, and clinical practice[18]. Even when identical coding systems are used, diagnostic thresholds, prescription habits, and healthcare access can vary across settings[235]. These differences introduce systematic heterogeneity that complicates cross-biobank analysis. In Study I, we showed that some phenotypes exhibit substantially larger cross-biobank variance than others. A more systematic investigation of phenotype-specific variance across cohorts could identify traits that are particularly sensitive to system-level differences and may also provide insight into differences in clinical practice across healthcare systems.

Harmonization challenges extend beyond phenotypes to the methods and data formats used in genetic analysis. Agreement on standardized inputs, reporting conventions, and containerized analysis pipelines would improve transparency, reproducibility, and comparability across studies. During this work, we encountered substantial heterogeneity even in GWAS summary statistics, including inconsistent allele definitions, genome builds, variant identifiers, and missing positional information. While efforts to standardize outputs using rsIDs and shared formats are ongoing, data heterogeneity remains common across EHR-derived phenotypes, polygenic scoring methods, and GWAS results. Addressing these inconsistencies would reduce avoidable sources of noise and lower barriers to cross-cohort analysis.

Phenotype refinement and construction represent another important direction. By leveraging a broader range of EHR information, refined phenotypes can provide deeper insight into complex diseases, but this thesis also demonstrates that such approaches introduce new challenges. Simulations in Study II showed that misclassification and control selection bias interact in non-trivial ways. Future work could extend these simulations to explore how trait

prevalence, comorbidity structure, and diagnostic uncertainty determine when phenotype refinement is beneficial and when it primarily introduces bias. As the genetic architecture of psychiatric disorders becomes better understood, it will also be important to reassess whether excluding overlapping phenotypes remains appropriate, or whether such exclusions remove biologically meaningful signal. Further empirical studies and simulation-based frameworks could help define practical guidelines for phenotype refinement rather than ad hoc strategies.

A largely untapped opportunity lies in the range of EHR data that are already collected but not fully utilized. Beyond diagnoses, EHR include longitudinal prescription records, laboratory measurements, clinical procedures, socioeconomic information, and questionnaire data. Integrating these data sources offers opportunities to study disease trajectories, treatment response, and cumulative burden in ways that are not captured by binary diagnostic labels. Prescription-based phenotypes, as explored in Study III, represent one example of how such data can be leveraged, but many similar applications remain unexplored.

At the same time, improving biological interpretability remains a major challenge. Many approaches successfully increase statistical power and identify additional associations but often struggle to distinguish true biological signal from noise or shared pleiotropic effects. Methods such as pleiotropy analysis, replication metrics, and phenotype-specific validation provide partial insight, but they do not fully resolve the gap between statistical association and biological mechanism. Bridging this gap will require integrating genetic results with multiple layers of biological information, including pathways, regulatory context, and molecular interactions.

In this context, ML approaches such as knowledge graphs, large language models, and other deep learning frameworks may offer useful tools, not primarily for prediction, but for organizing, contextualizing, and interpreting genetic findings. These models have the potential to integrate heterogeneous sources of biological and clinical knowledge and to support structured interpretation of complex genetic architectures. However, their application in genetics will require careful validation, transparency, and awareness of inherited biases in training data.

Overall, the future of biobank-based genetics will depend on coordinated advances in data harmonization, phenotype design, and interpretation rather than on scale alone. ML and increasingly rich EHR data provide powerful tools, but their value depends on careful application and explicit acknowledgment of trade-offs. Progress will come from combining methodological innovation with principled study design, transparent reporting, and a clear understanding of the limitations inherent in large, heterogeneous clinical datasets.

Acknowledgements

I conducted my PhD studies within the Doctoral Programme in Population Health (DOCPOP) and completed this thesis at the Institute for Molecular Medicine Finland (FIMM), University of Helsinki. FIMM provided an excellent environment for research and collaboration, offering many opportunities to engage with fellow PhD students. I was fortunate to form close friendships during this time, which made the experience particularly rewarding. The institute also actively supports its doctoral researchers through various events and initiatives, and I appreciated the opportunity to contribute to this community as a member of the PhD and Postdoc Council. I thank FIMM and everyone working at the institute for creating such a supportive and inspiring place to work. I would also like to acknowledge Professor Samuli Ripatti, Director of FIMM, for his continuous efforts to integrate new ideas and to foster a positive and evolving research environment.

I acknowledge the University of Helsinki Research Foundation and the Vilho, Yrjö and Kalle Väisälä Fund of the Finnish Academy of Science and Letters for funding my PhD. Their support provided me with the freedom and independence to shape my research and was instrumental in allowing me to complete this work. Without this independence, it would have been significantly more challenging to carry the project through to completion.

I would like to thank everyone who supervised and supported me throughout my PhD, for their patience, kindness, and for providing me with ideas and projects that shaped my work. I thank Andrea Ganna for giving me the opportunity to conduct my doctoral thesis in his group, for serving as my custos at the defence, and for encouraging a high degree of independence. I would like to express my sincere gratitude to my second supervisor, Zhiyu Yang, who took over my supervision at a particularly difficult stage of my PhD, when my project had reached a challenging point. At a time when others might have stepped back, she chose to invest in my work and in my development as a researcher. She provided me with a new project that became an important part of this thesis, and through our close collaboration I learned a great deal about the structure, planning, and execution of research, as well as how to approach and interpret unexpected results. Her ability to challenge my thinking in a constructive and supportive way helped me grow both scientifically and personally. Her patience, honesty, and encouragement fundamentally changed my experience of research and were central to my ability to continue and complete this work.
I would also like to thank Tuomo Kiiskinen for providing a project that I was able to include in this thesis. This opportunity came at a time when I was in urgent need of direction, and I greatly appreciate his willingness to work with me during that period.

I thank Mikko Arvas and Sandra Sanchez-Roige for the examination of my thesis and for their valuable feedback. I am additionally grateful to Sandra Sanchez-Roige for agreeing to serve as my opponent, and to Iris Horvatta for serving as my faculty representative. I also thank my thesis committee members, Caroline Heckman, Eija Pirinen, and Pekka Marttinen, for their guidance, patience, and thoughtful advice throughout my PhD. Their support, especially during challenging periods, made me feel less alone in the process. I would like to give special thanks to Caroline Heckman and Esa Pitkänen, in their roles within the thesis

and doctoral training committee, for their exceptional support. They consistently offered their time and guidance without any expectation of benefit. This generosity and patience, particularly, had a lasting impact on my PhD journey.

I am thankful to all collaborators, without whom this work would not have been possible. For the project conducted with Tuomo, I would especially like to acknowledge our international collaborators who made this large-scale study possible, in particular Laura Luitva, Kristi Krebs, and Lili Milani for their support with the Estonian Biobank analyses. I also thank Remo Monti and Christoph Lippert for including me in their project and for working closely with me during the revision process. I learned a great deal from this collaboration, which benefited my other projects as well. I am grateful to the entire Intervene consortium, and especially to Georgi Hudjashov, Kristi Läll, Stavroula Kanoni, and Brooke Wolford, who worked closely with me on the project. Finally, I recognize the privilege of working with large-scale data, including FinnGen, UK Biobank, and Estonian Biobank. I am deeply thankful to the organizations behind these resources and, in particular, to all participants whose contributions make this research possible.

I had the privilege of conducting my research within a working group of truly wonderful people. I greatly appreciated the warm, kind, and enjoyable environment that made everyday work both productive and fun. I would therefore like to thank everyone in the DSGE lab, as well as former members with whom I had the pleasure of working. I thank Anne for bringing much-needed organization to the often chaotic nature of research. I am especially grateful to Matteo, who sat next to me and was always willing to provide input, no matter how often I asked, and to Kira, with whom I shared many lunches and conversations, both work-related and beyond. I would also like to thank Essi and Feiyi, my fellow PhD students, who understood the challenges of this journey and were always there with support and encouragement. Additionally I would like to thank Sara, whom I met during my time in the DSGE lab. Although you moved on from academia during my PhD, I was fortunate that we stayed in close contact and developed a lasting friendship. I truly appreciated the many relaxed evenings and dance practices we shared.

But I was also fortunate enough to meet incredible scientists and friends in neighbouring working groups. I thank Arina and Ankita, with whom I began this journey during the COVID-19 period. Sharing those early experiences created a strong sense of connection, and I greatly valued the advice, conversations, and moments we shared along the way. A special thanks goes to Anniina Tervi, who has been an incredible source of guidance throughout my PhD. Although you were a fellow PhD student, you brought the insight and perspective of a much more senior researcher, supporting me both academically and personally. I feel fortunate to have shared many lunches, tea breaks, and relaxed evenings with you, during which you were always willing to listen and offer thoughtful advice. Your dedication, reliability, and genuine care for others are truly remarkable. I am especially grateful for your valuable input to this thesis and for your help in writing the *Tiivistelmä*.

Another neighbouring working group I would like to thank consists of Joseph, Nemo, June, and Rhiannon. Thank you for all the shared hangouts, the holidays and birthdays we celebrated together, and all the fun moments and milestones along the way. I thank Nemo for being my reliable gym buddy and someone I could always turn to. I especially appreciated your seemingly unlimited wisdom, and I feel very lucky to have you as a friend.

June and Nemo, thank you for the years of Tuska and for giving me the opportunity to join you for so many great concerts and hangouts. Thanks to you, my music taste has become more refined than ever. You all have been like a family to me here.

Joseph, I want to thank you in particular. You are an incredibly energetic and generous person, always willing to support those around you. Your excitement is contagious, you bring people together, make them laugh, and create a sense of belonging, whether through dancing or simply spending time together. Your support during my PhD meant more than I can fully express. At times when I struggled, you helped me regain perspective and move forward. Without you, I am not sure I would have finished this PhD. Where I felt stuck or overwhelmed, you helped me grow instead. Thank you for your patience, for always being there, and for creating space where I could both laugh and be vulnerable. You never made me feel like a burden, even when I felt that way myself, and that made a lasting difference. You helped me better understand both the world and myself. I am deeply grateful to have you in my life. You have shaped this journey in a fundamental way, and I look forward to many more years and experiences together.

Elli, you were the first person I knew in Finland, even before moving here. When we first met, I was hoping for some guidance and tips on how to start a life in a new country, but instead we built a wonderful friendship as we explored museums across Helsinki and shared our passion for sewing. I look forward to continuing to share sewing projects and life moments with you.

Ronja, after more than 10 years of knowing me, I am still surprised you have not had enough. A friendship lasting this long and across so many kilometers is truly special. You are one of the most impressive people I know. You approach things I struggle with in a way that seems effortless, while remaining humble and kind, always reminding me of my strengths and encouraging me along the way. I admire how you navigate life, with focus and determination, but without losing sight of what truly matters. Thank you for always making the effort to stay close despite the distance, for our regular catch-ups, and for being there during difficult times. It means a great deal to me that you always make space for our friendship. I hope you know how important you are to me, and that I will always make room for you in my life.

Maike, I have known you since 7th grade, and since then we have been part of each other's lives. We have seen each other through ups and downs, through victories and setbacks, always supporting one another and coming out stronger on the other side. You are someone I can always laugh with and who takes me as I am. Thank you for being such a constant friend over all these years, for making the effort to stay present in my life and for including me in yours. I am excited to see what the future holds for us.

To my family, Hanna, Ingo, Eva, Charlotta, Hans-Christian, Jonas, Dina, and Kanto, thank you for being there every step of the way. First and foremost, to my parents, Hanna and Ingo. The older I get, the more I realize what incredible parents you are. You mastered the balance of encouraging me to step out of my comfort zone (to study, to explore the world, and to start new in another country) while always providing such a strong sense of safety and belonging that no challenge ever felt overwhelming. This balance is almost impossible to achieve. I do not think I will ever be able to fully express what it means to have you by my

side, cheering me on, and to have parents I can completely rely on. You have made my life so much easier. Thank you for always being there.

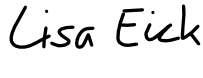
Eva and Charly, my sisters, having you in my life is a privilege I would never want to miss, and I am grateful for it every day. You know me, sometimes better than I know myself, and I always feel that I can talk to you about anything, no matter how complex or abstract. We helped each other through many hardships, and I hope I have been there for you as you have been for me. You have paved the way for me in many ways, and I have learned so much from you over the years. You are both incredibly strong, and I am deeply proud of who you have become and what you have achieved. Thank you for being in my life.

Jonas, even though we sometimes struggle to stay in touch across distances, I am always amazed at how easily we pick up exactly where we left off. I am grateful for all the time we have shared and look forward to all that is still to come. As your twin, I am always rooting for you, and I love you more than I sometimes manage to express.

Last but not least, Nils. When I first met you, I told you I wanted to do a PhD in a foreign country, and then I almost immediately left the country to do so. Despite that you stayed by my side for all those years. I have never met anyone like you. You are the most supportive person I know, with strengths I deeply admire, especially your patience and consistency. You have been a rock in rough seas, someone I can always rely on. Growing together with you, through both difficult and good times, has given me a sense of stability that nothing else could. I feel grateful every day to have you in my life. Seeing you grow alongside me makes me proud and even more excited for what the future holds for us. I love you.

No matter the hardships, I feel fortunate to have such amazing and wonderful people in my life. This is truly all I need. Thank you all for walking every step of this journey with me.

Helsinki
April 23rd, 2026
Lisa Eick



Appendix A

Study subjects in FinnGen provided informed consent for biobank research, based on the Finnish Biobank Act. Alternatively, separate research cohorts, collected prior the Finnish Biobank Act came into effect (in September 2013) and start of FinnGen (August 2017), were collected based on study-specific consents and later transferred to the Finnish biobanks after approval by Fimea (Finnish Medicines Agency), the National Supervisory Authority for Welfare and Health. Recruitment protocols followed the biobank protocols approved by Fimea. The Coordinating Ethics Committee of the Hospital District of Helsinki and Uusimaa (HUS) statement number for the FinnGen study is Nr HUS/990/2017.
The FinnGen study is approved by Finnish Institute for Health and Welfare (permit numbers: THL/2031/6.02.00/2017, THL/1101/5.05.00/2017, THL/341/6.02.00/2018, THL/2222/6.02.00/2018, THL/283/6.02.00/2019, THL/1721/5.05.00/2019 and THL/1524/5.05.00/2020), Digital and population data service agency (permit numbers: VRK43431/2017-3, VRK/6909/2018-3, VRK/4415/2019-3), the Social Insurance Institution (permit numbers: KELA 58/522/2017, KELA 131/522/2018, KELA 70/522/2019, KELA 98/522/2019, KELA 134/522/2019, KELA 138/522/2019, KELA 2/522/2020, KELA 16/522/2020), Findata permit numbers THL/2364/14.02/2020, THL/4055/14.06.00/2020, THL/3433/14.06.00/2020, THL/4432/14.06/2020, THL/5189/14.06/2020, THL/5894/14.06.00/2020, THL/6619/14.06.00/2020, THL/209/14.06.00/2021, THL/688/14.06.00/2021, THL/1284/14.06.00/2021, THL/1965/14.06.00/2021, THL/5546/14.02.00/2020, THL/2658/14.06.00/2021, THL/4235/14.06.00/2021, THL/4990/14.02.00/2023 Statistics Finland (permit numbers: TK-53-1041-17 and TK/143/07.03.00/2020 (earlier TK-53-90-20) TK/1735/07.03.00/2021, TK/3112/07.03.00/2021) and Finnish Registry for Kidney Diseases permission/extract from the meeting minutes on 4th July 2019.

The Biobank Access Decisions for FinnGen samples and data utilized in FinnGen Data Freeze 13 include: THL Biobank BB2017_55, BB2017_111, BB2018_19, BB_2018_34, BB_2018_67, BB2018_71, BB2019_7, BB2019_8, BB2019_26, BB2020_1, BB2021_65, BB22-0025-A01, BB22-0025-A03, BB23-0222-A01, BB22-0025-A04, BB22-0025-A06, BB22-0025-A08, THLBB2024_30. Finnish Red Cross Blood Service Biobank 7.12.2017, 13.11.2023, 001-2023, Helsinki Biobank HUS/359/2017, HUS/248/2020, HUS/430/2021 §28, §29, HUS/150/2022 §12, §13, §14, §15, §16, §17, §18, §23, §58, §59, HUS/128/2023 §18, BB22-0025-A01, BB22-0025-A02, BB22-0025-A05, BB22-0025-A07, BB22-0025-A09, BB22-0025-A10, BB22-0025-A03, BB23-0222-A01, BB22-0025-A04, BB22-0025-A06, BB22-0025-A08, Amendment_BB22-0025-A05, Decision allowing to continue data processing until 31st Aug 2027: BB_2021-0140, HUS/150/2022 §12, BB_2021-0139, HUS/150/2022 §13, BB_2021-0161,HUS/150/2022 §14, BB_2021-0164, HUS/150/2022 §15, BB_2021-0169, HUS/150/2022 §16, BB_2021-0170, HUS/150/2022 §17, BB_2021-0179, HUS/150/2022 §18, BB_2022-0262, HUS/150/2022 §58, BB22-0067, HUS/150/2022 §59, Auria Biobank AB17-5154 and amendment #1 (August 17 2020) and amendments BB_2021-0140, BB_2021-0156 (August 26 2021, Feb 2 2022), BB_2021-0169, BB_2021-0179, BB_2021-0161, AB20-5926 and amendment #1 (April 23 2020) and it´s modifications (Sep 22 2021), BB_2022-0262, BB_2022-0256, BB22-0025-A01, BB22-0025-A02, BB22-0025-A03, BB23-0222_A01, BB22-0025-A02, BB22-0025-A05, BB22-0025-A07, BB22-0025-A09, BB22-0025-A10, BB22-0025-A03, BB23-0222-A01, BB22-0025-A04, BB22-0025-A06, BB22-0025-A08, Decision allowing to continue data processing until 31st Aug 2027: AB20-5926, BB_2021- 0140, BB_2021-0156, BB_2021-0161, BB_2021- 0161, BB_2021-0164, BB_2021-0169, BB_2021-0179, BB_2022-0262, Biobank Borealis of Northern Finland_2017_1013, 2021_5010, 2021_5010 Amendment, 2021_5018, 2021_5018 Amendment, 2021_5015, 2021_5015 Amendment, 2021_5015 Amendment_2, 2021_5023, 2021_5023 Amendment, 2021_5023 Amendment_2, 2021_5017, 2021_5017 Amendment, 2022_6001, 2022_6001 Amendment, 2022_6006 Amendment, 2022_6006

Amendment_2, BB22-0067, 2022_0262, 2022_0262 Amendment, BB22-0025-A01, BB22-0025-A02, BB22-0025-A05, BB22-0025-A07, BB22-0025-A09, BB22-0025-A10, BB22-0025-A03, BB23-0222-A01, BB22-0025-A04, BB22-0025-A06, BB22-0025-A08, Decision allowing to continue data processing until 31st Aug 2027: BB/2021/5015, BB/2021/5017, BB/2021/5018, BB/2021/5023, BB/2022/6006, BB/2022/6001, BB/2022-0262, BB/2021/5010, Biobank of Eastern Finland 1186/2018 and amendment 22§/2020, 53§/2021, 13§/2022, 14§/2022, 15§/2022, 27§/2022, 28§/2022, 29§/2022, 33§/2022, 35§/2022, 36§/2022, 37§/2022, 39§/2022, 7§/2023, 32§/2023, 33§/2023, 34§/2023, 35§/2023, 36§/2023, 37§/2023, 38§/2023, 39§/2023, 40§/2023, 41§/2023, BB22-0025-A01, BB22-0025-A02, BB22-0025-A05, BB22-0025-A07, BB22-0025-A09, BB22-0025-A10, BB22-0025-A03, BB23-0222-A01, BB22-0025-A04, BB22-0025-A06, BB22-0025-A08, Decision allowing to continue data processing until 31st Aug 2027: MO-BB_2021-0179-A0, MO-BB_2021-0156_PRE-A01, BB_2021-0140, MO-BB_2021-0170_PRE-A0, MO-BB_2021-0169-A01, MO-BB_2022-0256-A01, MO-BB_2021-0161-A01, MO-BB_2021-0161-A02, BB22-0067-A01, MO-BB_2022-0262-A0, Finnish Clinical Biobank Tampere MH0004 and amendments (21.02.2020 & 06.10.2020), BB2021-0140 8§/2021, 9§/2021, §9/2022, §10/2022, §12/2022, 13§/2022, §20/2022, §21/2022, §22/2022, §23/2022, 28§/2022, 29§/2022, 30§/2022, 31§/2022, 32§/2022, 38§/2022, 40§/2022, 42§/2022, 1§/2023, BB2021-0140, BB22-0025-A01, BB_2021-0161, BB22-0025-A02, BB22-0025-A05, BB22-0025-A07, BB22-0025-A09, BB22-0025-A10, BB22-0025-A03, BB23-0222-A01, BB22-0025-A04, BB22-0025-A06, BB22-0025-A08, Decision allowing to continue data processing until 31st Aug 2027: BB_2021- 0140, BB_ 2021- 0161, BB_ 2021- 0179, BB_ 2021- 0156, BB_ 2021- 0169, BB_ 2021- 0170, BB22- 0067-A01, Central Finland Biobank 1-2017, BB_2021-0169, BB_2021-0179, BB_2022-0256, BB_2022-0262, Decision allowing to continue data processing until 31st Aug 2027 for projects: BB_2021-0179, BB22-0067,BB_2022-0262, BB_2021-0170, BB_2021-0164, BB_2021-0161, and BB_2021-0169, BB22-0025-A01, BB22-0025-A02, BB22-0025-A05, BB22-0025-A07, BB22-0025-A09, BB22-0025-A10, BB22-0025-A03, BB23-0222-A01, BB22-0025-A04, BB22-0025-A06, BB22-0025-A08, Terveystalo Biobank STB 2018001 and amendment 25th Aug 2020, Finnish Hematological Registry and Clinical Biobank decision 18th June 2021, Amendment 2nd January 2024 and Arctic biobank P0844: ARC_2021_1001, ARC_2023_3003 (BB22-0025-A01), BB22-0025-A03, BB23-0222-A01, BB22-0025-A04, BB22-0025-A06, BB22-0025-A08. The application “F_2020_036” and “F_2022_001” provided access to the FinnGen data used in this study.

The activities of the EstBB are regulated by the Human Genes Research Act, which was adopted in 2000 specifically for the operations of the EstBB. Individual level data analysis in the EstBB was carried out under ethical approval "1.1-12/624" from the Estonian Committee on Bioethics and Human Research (Estonian Ministry of Social Affairs), using data according to release application "6-7/GI/28554” and “6-7/GI/16259” from the EstBB.

Ethics approval for the UKBB study was obtained from the North West Centre for Research Ethics Committee (11/NW/0382). UKBB data used in this study were obtained under application 22627 and 78537.

References

1. Mars, N. *et al.* Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. *Nat. Med.* **26**, 549–557 (2020).
2. Mars, N. *et al.* Systematic comparison of family history and polygenic risk across 24 common diseases. *Am. J. Hum. Genet.* **109**, 2152–2162 (2022).
3. Lewis, C. M. & Vassos, E. Polygenic risk scores: from research tools to clinical instruments. *Genome Med.* **12**, 44 (2020).
4. Polygenic Risk Score Task Force of the International Common Disease Alliance. Responsible use of polygenic risk scores in the clinic: potential benefits, risks and gaps. *Nat. Med.* **27**, 1876–1884 (2021).
5. Ding, Y. *et al.* Polygenic scoring accuracy varies across the genetic ancestry continuum. *Nature* 1–8 (2023) doi:10.1038/s41586-023-06079-4.
6. Pain, O. *et al.* Evaluation of polygenic prediction methodology within a reference-standardized framework. *PLoS Genet.* **17**, e1009021 (2021).
7. Ni, G. *et al.* A comparison of ten polygenic score methods for psychiatric disorders applied across multiple cohorts. *Biol. Psychiatry* **90**, 611–620 (2021).
8. Pendergrass, S. A. & Crawford, D. C. Using Electronic Health Records To Generate Phenotypes For Research. *Curr. Protoc. Hum. Genet.* **100**, e80 (2019).
9. Al-Sahab, B., Leviton, A., Loddenkemper, T., Paneth, N. & Zhang, B. Biases in Electronic Health Records Data for Generating Real-World Evidence: An Overview. *J. Healthc. Inform. Res.* **8**, 121–139 (2023).
10. Verheij, R. A., Curcin, V., Delaney, B. C. & McGilchrist, M. M. Possible Sources of Bias in Primary Care Electronic Health Record Data Use and Reuse. *J. Med. Internet Res.* **20**, e9134 (2018).

11. Sinnott, J. A. *et al.* Improving the Power of Genetic Association Tests with Imperfect Phenotype Derived from Electronic Medical Records. *Hum. Genet.* **133**, 1369–1382 (2014).
12. Newby, D., Taylor, N., Joyce, D. W. & Winchester, L. M. Optimising the use of electronic medical records for large scale research in psychiatry. *Transl. Psychiatry* **14**, 232 (2024).
13. Smoller, J. W. The Use of Electronic Health Records for Psychiatric Phenotyping and Genomics. *Am. J. Med. Genet. Part B Neuropsychiatr. Genet. Off. Publ. Int. Soc. Psychiatr. Genet.* **177**, 601–612 (2018).
14. Hodge, S. E., Subaran, R. L., Weissman, M. M. & Fyer, A. J. Designing case-control studies: decisions about the controls. *Am. J. Psychiatry* **169**, 785–789 (2012).
15. Jensen, K. H. R. *et al.* Deep phenotyping towards precision psychiatry of first-episode depression — the Brain Drugs-Depression cohort. *BMC Psychiatry* **23**, 151 (2023).
16. Dahl, A. *et al.* Phenotype integration improves power and preserves specificity in biobank-based genetic studies of major depressive disorder. *Nat. Genet.* **55**, 2082–2093 (2023).
17. Guan, J. *et al.* Family-based genome-wide association study designs for increased power and robustness. *Nat. Genet.* **57**, 1044–1052 (2025).
18. Tang, A. S. *et al.* Harnessing EHR data for health research. *Nat. Med.* **30**, 1847–1855 (2024).
19. Giacomini, K. M. *et al.* Genome-wide association studies of drug response and toxicity: an opportunity for genome medicine. *Nat. Rev. Drug Discov.* **16**, 70–70 (2017).
20. Sherman, C. A., Claw, K. G. & Lee, S. Pharmacogenetic analysis of structural variation in the 1000 genomes project using whole genome sequences. *Sci. Rep.* **14**, 22774 (2024).
21. Sadler, M. C. *et al.* Leveraging large-scale biobank EHRs to enhance pharmacogenetics of cardiometabolic disease medications. *Nat. Commun.* **16**, 2913 (2025).

22. Kolodny, A. *et al.* The Prescription Opioid and Heroin Crisis: A Public Health Approach to an Epidemic of Addiction. *Annu. Rev. Public Health* **36**, 559–574 (2015).
23. Roberts, E., Copeland, C., Humphreys, K. & Shover, C. L. Drug-related deaths among housed and homeless individuals in the UK and the USA: comparative retrospective cohort study. *Br. J. Psychiatry* **223**, 562–568 (2023).
24. de Boer, H. D., Detriche, O. & Forget, P. Opioid-related side effects: Postoperative ileus, urinary retention, nausea and vomiting, and shivering. A review of the literature. *Best Pract. Res. Clin. Anaesthesiol.* **31**, 499–504 (2017).
25. Imam, M. Z., Kuo, A., Ghassabian, S. & Smith, M. T. Progress in understanding mechanisms of opioid-induced gastrointestinal adverse effects and respiratory depression. *Neuropharmacology* **131**, 238–255 (2018).
26. Kalinin, A. A. *et al.* Deep Learning in Pharmacogenomics: From Gene Regulation to Patient Stratification. *Pharmacogenomics* **19**, 629–650 (2018).
27. An, U. *et al.* Deep learning-based phenotype imputation on population-scale biobank data increases genetic discoveries. *Nat. Genet.* 1–8 (2023) doi:10.1038/s41588-023-01558-w.
28. Yang, L., Sadler, M. C. & Altman, R. B. Genetic association studies using disease liabilities from deep neural networks. *Am. J. Hum. Genet.* **112**, 675–692 (2025).
29. Gu, L.-L. *et al.* Rapid and accurate multi-phenotype imputation for millions of individuals. *Nat. Commun.* **16**, 387 (2025).
30. Shaw, D. M. *et al.* Phenome risk classification enables phenotypic imputation and gene discovery in developmental stuttering. *Am. J. Hum. Genet.* **108**, 2271–2283 (2021).
31. Cosentino, J. *et al.* Inference of chronic obstructive pulmonary disease with deep learning on raw spirograms identifies new genetic loci and improves risk models. *Nat. Genet.* **55**, 787–795 (2023).
32. Cheng, Y., Zhai, S., Zhong, W., West, R. M. & Shen, J. Improving polygenic risk score based drug response prediction using transfer learning. *Npj Genomic Med.* **10**, 74 (2025).

33. *An Owner's Guide to the Human Genome*. (2023).

34. Alexander, R. P., Fang, G., Rozowsky, J., Snyder, M. & Gerstein, M. B. Annotating non-coding regions of the genome. *Nat. Rev. Genet.* **11**, 559–571 (2010).

35. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

36. Non-coding DNA – a brief review. *J. Appl. Biol. Biotechnol.* https://doi.org/10.7324/JABB.2017.50507 (2017) doi:10.7324/JABB.2017.50507.

37. Portela, A. & Esteller, M. Epigenetic modifications and human disease. *Nat. Biotechnol.* **28**, 1057–1068 (2010).

38. Petronis, A. Epigenetics as a unifying principle in the aetiology of complex traits and diseases. *Nature* **465**, 721–727 (2010).

39. Feuk, L., Carson, A. R. & Scherer, S. W. Structural variation in the human genome. *Nat. Rev. Genet.* **7**, 85–97 (2006).

40. Frazer, K. A., Murray, S. S., Schork, N. J. & Topol, E. J. Human genetic variation and its contribution to complex traits. *Nat. Rev. Genet.* **10**, 241–251 (2009).

41. Beichman, A. C., Zhu, L. & Harris, K. The Evolutionary Interplay of Somatic and Germline Mutation Rates. *Annu. Rev. Biomed. Data Sci.* **7**, 83–105 (2024).

42. Campbell, C. D. & Eichler, E. E. Properties and rates of germline mutations in humans. *Trends Genet.* **29**, 575–584 (2013).

43. Lehner, B. Genotype to phenotype: lessons from model organisms for human genetics. *Nat. Rev. Genet.* **14**, 168–178 (2013).

44. Mendel, G. *Versuche Über Pflanzen-Hybriden*. (Im Verlage des Vereines, Brünn, 1866).

45. van Dijk, P. J., Jessop, A. P. & Ellis, T. H. N. How did Mendel arrive at his discoveries? *Nat. Genet.* **54**, 926–933 (2022).

46. Abdellaoui, A. & Verweij, K. J. H. Dissecting polygenic signals from genome-wide association studies on human behaviour. *Nat. Hum. Behav.* **5**, 686–694 (2021).

47. Novembre, J. *et al.* Addressing the challenges of polygenic scores in human genetic research. *Am. J. Hum. Genet.* **109**, 2095–2100 (2022).

48. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).

49. Hunter, D. J. Gene–environment interactions in human diseases. *Nat. Rev. Genet.* **6**, 287–298 (2005).

50. Paaby, A. B. & Rockman, M. V. The many faces of pleiotropy. *Trends Genet.* **29**, 66–73 (2013).

51. Mackay, T. F. C. & Anholt, R. R. H. Pleiotropy, epistasis and the genetic architecture of quantitative traits. *Nat. Rev. Genet.* **25**, 639–657 (2024).

52. Schork, N. J., Murray, S. S., Frazer, K. A. & Topol, E. J. Common vs. Rare Allele Hypotheses for Complex Diseases. *Curr. Opin. Genet. Dev.* **19**, 212–219 (2009).

53. Norio, R. Finnish Disease Heritage I: *Hum. Genet.* **112**, 441–456 (2003).

54. Minton, K. The FinnGen study: disease insights from a 'bottlenecked' population. *Nat. Rev. Genet.* **24**, 207–207 (2023).

55. Read, J. & Sharma, S. Hypothesis-driven science in large-scale studies: the case of GWAS. *Biol. Philos.* **36**, 46 (2021).

56. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).

57. Schneider, V. A. *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* **27**, 849–864 (2017).

58. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

59. International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).

60. International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).

61. Verlouw, J. A. M. *et al.* A comparison of genotyping arrays. *Eur. J. Hum. Genet.* **29**, 1611–1624 (2021).

62. Zhao, S. *et al.* Strategies for processing and quality control of Illumina genotyping arrays. *Brief. Bioinform.* **19**, 765–775 (2018).

63. A broad overview of genotype imputation: Standard guidelines, approaches, and future investigations in genomic association studies. *Biocell* **47**, 1225–1241 (2023).

64. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11**, 499–511 (2010).

65. Kurki, M. I. *et al.* FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **613**, 508–518 (2023).

66. Milani, L. *et al.* The Estonian Biobank's journey from biobanking to personalized medicine. *Nat. Commun.* **16**, 3270 (2025).

67. Leitsalu, L. *et al.* Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. *Int. J. Epidemiol.* **44**, 1137–1147 (2015).

68. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

69. Carss, K. *et al.* Whole-genome sequencing of 490,640 UK Biobank participants. *Nature* **645**, 692–701 (2025).

70. Åsvold, B. O. *et al.* Cohort Profile Update: The HUNT Study, Norway. *Int. J. Epidemiol.* **52**, e80–e91 (2023).

71. Finer, S. *et al.* Cohort Profile: East London Genes & Health (ELGH), a community-based population genomics and health study in British Bangladeshi and British Pakistani people. *Int. J. Epidemiol.* **49**, 20–21i (2020).

72. Jafari, E. *et al.* Using electronic health records for clinical pharmacology research: Challenges and considerations. *Clin. Transl. Sci.* **17**, e13871 (2024).

73. Kim, E. *et al.* The Evolving Use of Electronic Health Records (EHR) for Research. *Semin. Radiat. Oncol.* **29**, 354–361 (2019).

74. Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium. Electronic address: andrew.mcintosh@ed.ac.uk & Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium. Trans-ancestry genome-wide study of

depression identifies 697 associations implicating cell types and pharmacotherapies. *Cell* **188**, 640-652.e9 (2025).

75. Howard, D. M. *et al.* Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nat. Neurosci.* **22**, 343–352 (2019).
76. Shorter, J. R. *et al.* Genome-wide association analyses identify distinct genetic architectures for early-onset and late-onset depression. *Nat. Genet.* 1–8 (2025) doi:10.1038/s41588-025-02396-8.
77. Yengo, L. *et al.* A saturated map of common genetic variants associated with human height. *Nature* **610**, 704–712 (2022).
78. Tam, V. *et al.* Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* **20**, 467–484 (2019).
79. Uffelmann, E. *et al.* Genome-wide association studies. *Nat. Rev. Methods Primer* **1**, 1–21 (2021).
80. Abdellaoui, A., Yengo, L., Verweij, K. J. H. & Visscher, P. M. 15 years of GWAS discovery: Realizing the promise. *Am. J. Hum. Genet.* **110**, 179–194 (2023).
81. Evangelou, E. & Ioannidis, J. P. A. Meta-analysis methods for genome-wide association studies and beyond. *Nat. Rev. Genet.* **14**, 379–389 (2013).
82. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
83. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
84. Finucane, H. K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621–629 (2018).
85. van Rheenen, W., Peyrot, W. J., Schork, A. J., Lee, S. H. & Wray, N. R. Genetic correlations of polygenic disease traits: from theory to practice. *Nat. Rev. Genet.* **20**, 567–581 (2019).

86. Benner, C. *et al.* FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).

87. Li, Z. & Zhou, X. Towards improved fine-mapping of candidate causal variants. *Nat. Rev. Genet.* **26**, 847–861 (2025).

88. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A Simple New Approach to Variable Selection in Regression, with Application to Genetic Fine Mapping. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **82**, 1273–1300 (2020).

89. McCreight, A. *et al.* SuSiE 2.0: improved methods and implementations for genetic fine-mapping and phenotype prediction. 2025.11.25.690514 Preprint at https://doi.org/10.1101/2025.11.25.690514 (2025).

90. Choi, S. W., Mak, T. S.-H. & O'Reilly, P. F. Tutorial: a guide to performing polygenic risk score analyses. *Nat. Protoc.* **15**, 2759–2772 (2020).

91. Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* **19**, 581–590 (2018).

92. Mak, T. S. H., Porsch, R. M., Choi, S. W., Zhou, X. & Sham, P. C. Polygenic scores via penalized regression on summary statistics. *Genet. Epidemiol.* **41**, 469–480 (2017).

93. Privé, F., Arbel, J. & Vilhjálmsson, B. J. LDpred2: better, faster, stronger. *Bioinforma. Oxf. Engl.* **36**, 5424–5431 (2021).

94. Polygenic prediction via Bayesian regression and continuous shrinkage priors | Nature Communications. https://www.nature.com/articles/s41467-019-09718-5.

95. Yang, S. & Zhou, X. Accurate and Scalable Construction of Polygenic Scores in Large Biobank Data Sets. *Am. J. Hum. Genet.* **106**, 679–693 (2020).

96. Lloyd-Jones, L. R. *et al.* Improved polygenic prediction by Bayesian multiple regression on summary statistics. *Nat. Commun.* **10**, 5086 (2019).

97. Zhang, Q., Privé, F., Vilhjálmsson, B. & Speed, D. Improved genetic prediction of complex traits from individual-level data or summary statistics. *Nat. Commun.* **12**, 4192 (2021).

98. Benoumhani, S. *et al.* A review of methods and software for polygenic risk score analysis. *PeerJ Comput. Sci.* **11**, e3039 (2025).

99. Lachance, J. & Tishkoff, S. A. SNP ascertainment bias in population genetic analyses: Why it is important, and how to correct it. *BioEssays* **35**, 780–786 (2013).

100. Nielsen, R. Population genetic analysis of ascertained SNP data. *Hum. Genomics* **1**, 218 (2004).

101. Johnson, E. O. *et al.* Imputation across genotyping arrays for genome-wide association studies: assessment of bias and a correction strategy. *Hum. Genet.* **132**, 509–522 (2013).

102. Pattillo Smith, S. *et al.* Discovering non-additive heritability using additive GWAS summary statistics. *eLife* **13**, e90459 (2024).

103. Stringer, S., Wray, N. R., Kahn, R. S. & Derks, E. M. Underestimated Effect Sizes in GWAS: Fundamental Limitations of Single SNP Analysis for Dichotomous Phenotypes. *PLOS ONE* **6**, e27964 (2011).

104. Atwood, L. D. & Heard-Costa, N. L. Limits of fine-mapping a quantitative trait. *Genet. Epidemiol.* **24**, 99–106 (2003).

105. Wang, J. *et al.* Fine-mapping methods for complex traits: essential adaptations for samples of related individuals. *Brief. Bioinform.* **26**, bbaf614 (2025).

106. Zaitlen, N. & Kraft, P. Heritability in the genome-wide association era. *Hum. Genet.* **131**, 1655–1664 (2012).

107. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).

108. Ma, Y. & Zhou, X. Genetic prediction of complex traits with polygenic scores: a statistical review. *Trends Genet.* **37**, 995–1011 (2021).

109. Schwarzerova, J. *et al.* A perspective on genetic and polygenic risk scores—advances and limitations and overview of associated tools. *Brief. Bioinform.* **25**, bbae240 (2024).

110. Jayasinghe, D., Eshetie, S., Beckmann, K., Benyamin, B. & Lee, S. H. Advancements and limitations in polygenic risk score methods for genomic prediction: a scoping review. *Hum. Genet.* **143**, 1401–1431 (2024).
111. Kachuri, L. *et al.* Principles and methods for transferring polygenic risk scores across global populations. *Nat. Rev. Genet.* **25**, 8–25 (2024).
112. Savitz, S. T., Savitz, L. A., Fleming, N. S., Shah, N. D. & Go, A. S. How much can we trust electronic health record data? *Healthcare* **8**, 100444 (2020).
113. Chen, Y. *et al.* A Systematic Review of Coding Systems Used in Pharmacoepidemiology and Database Research. *Methods Inf. Med.* **57**, 1–42 (2018).
114. de Lusignan, S. *et al.* A survey to identify the clinical coding and classification systems currently in use across Europe. *Stud. Health Technol. Inform.* **84**, 86–89 (2001).
115. Gallagher, C. S., Ginsburg, G. S. & Musick, A. Biobanking with genetics shapes precision medicine and global health. *Nat. Rev. Genet.* **26**, 191–202 (2025).
116. Fry, A. *et al.* Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am. J. Epidemiol.* **186**, 1026–1034 (2017).
117. Schoeler, T. *et al.* Participation bias in the UK Biobank distorts genetic associations and downstream analyses. *Nat. Hum. Behav.* **7**, 1216–1227 (2023).
118. Pirastu, N. *et al.* Genetic analyses identify widespread sex-differential participation bias. *Nat. Genet.* **53**, 663–671 (2021).
119. Brayne, C. & Moffitt, T. E. The limitations of large-scale volunteer databases to address inequalities and global challenges in health and aging. *Nat. Aging* **2**, 775–783 (2022).
120. Czeisler, M., Wiley, J., Czeisler, C., Rajaratnam, S. & Howard, M. Uncovering Survivorship Bias in Longitudinal Mental Health Surveys During the COVID-19 Pandemic. *MedRxiv Prepr. Serv. Health Sci.* 2021.01.28.21250694 (2021) doi:10.1101/2021.01.28.21250694.
121. Anderson, C. D. *et al.* The Effect of Survival Bias on Case-Control Genetic Association Studies of Highly Lethal Diseases. *Circ. Cardiovasc. Genet.* **4**, 188–196 (2011).

122. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).

123. Fatumo, S. *et al.* A roadmap to increase diversity in genomic studies. *Nat. Med.* **28**, 243–250 (2022).

124. Alfaro-Almagro, F. *et al.* Confound modelling in UK Biobank brain imaging. *NeuroImage* **224**, 117002 (2021).

125. Madden, J. M., Lakoma, M. D., Rusinak, D., Lu, C. Y. & Soumerai, S. B. Missing clinical and behavioral health data in a large electronic health record (EHR) system. *J. Am. Med. Inform. Assoc. JAMIA* **23**, 1143–1149 (2016).

126. Kariotis, T. C., Prictor, M., Chang, S. & Gray, K. Impact of Electronic Health Records on Information Practices in Mental Health Contexts: Scoping Review. *J. Med. Internet Res.* **24**, e30405 (2022).

127. Dueñas, H. R., Seah, C., Johnson, J. S. & Huckins, L. M. Implicit bias of encoded variables: frameworks for addressing structured bias in EHR–GWAS data. *Hum. Mol. Genet.* **29**, R33–R41 (2020).

128. Jablensky, A. The Nature of Psychiatric Classification: Issues Beyond ICD-10 and DSM-IV. *Aust. N. Z. J. Psychiatry* **33**, 137–144 (1999).

129. Darrel A. Regier, M. D., William E. Narrow, M. D., Emily A. Kuhl, P. D. & David J. Kupfer, M. D. The Conceptual Development of DSM-V. *Am. J. Psychiatry* https://doi.org/10.1176/appi.ajp.2009.09020279 (2009) doi:10.1176/appi.ajp.2009.09020279.

130. Kendler, K. S. The dappled nature of causes of psychiatric illness: replacing the organic–functional/hardware–software dichotomy with empirically based pluralism. *Mol. Psychiatry* **17**, 377–388 (2012).

131. McGrath, J. J. *et al.* Comorbidity within mental disorders: a comprehensive analysis based on 145 990 survey respondents from 27 countries. *Epidemiol. Psychiatr. Sci.* **29**, e153 (2020).

132. Sullivan, P. F. *et al.* Psychiatric Genomics: An Update and an Agenda. *Am. J. Psychiatry* https://doi.org/10.1176/appi.ajp.2017.17030283 (2018) doi:10.1176/appi.ajp.2017.17030283.

133. Trubetskoy, V. *et al.* Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature* **604**, 502–508 (2022).

134. Strom, N. I. *et al.* Genome-wide association study of major anxiety disorders in 122,341 European-ancestry cases identifies 58 loci and highlights GABAergic signaling. *Nat. Genet.* **58**, 275–288 (2026).

135. O'Connell, K. S. *et al.* Genomics yields biological and phenotypic insights into bipolar disorder. *Nature* **639**, 968–975 (2025).

136. Demontis, D. *et al.* Genome-wide analyses of ADHD identify 27 risk loci, refine the genetic architecture and implicate several cognitive domains. *Nat. Genet.* **55**, 198–208 (2023).

137. Strom, N. I. *et al.* Genome-wide analyses identify 30 loci associated with obsessive–compulsive disorder. *Nat. Genet.* **57**, 1389–1401 (2025).

138. Nievergelt, C. M. *et al.* Genome-wide association analyses identify 95 risk loci and provide insights into the neurobiology of post-traumatic stress disorder. *Nat. Genet.* **56**, 792–808 (2024).

139. Deak, J. D. *et al.* Genome-wide association study in individuals of European and African ancestry and multi-trait analysis of opioid use disorder identifies 19 independent genome-wide significant risk loci. *Mol. Psychiatry* **27**, 3970–3979 (2022).

140. Hatoum, A. S. *et al.* Multivariate genome-wide association meta-analysis of over 1 million subjects identifies loci underlying multiple substance use disorders. *Nat. Ment. Health* **1**, 210–223 (2023).

141. Friligkou, E. *et al.* Gene discovery and biological insights into anxiety disorders from a large-scale multi-ancestry genome-wide association study. *Nat. Genet.* **56**, 2036–2045 (2024).

142. Wainschtein, P. *et al.* Estimation and mapping of the missing heritability of human phenotypes. *Nature* **649**, 1219–1227 (2026).

143. Watanabe, K. *et al.* A global overview of pleiotropy and genetic architecture in complex traits. *Nat. Genet.* **51**, 1339–1348 (2019).

144. Watanabe, K. *et al.* Genome-wide meta-analysis of insomnia prioritizes genes associated with metabolic and psychiatric pathways. *Nat. Genet.* **54**, 1125–1132 (2022).

145. Fabbri, C. Genetics in psychiatry: Methods, clinical applications and future perspectives. *Psychiatry Clin. Neurosci. Rep.* **1**, e6 (2022).

146. Cai, N. *et al.* Assessment and ascertainment in psychiatric molecular genetics: challenges and opportunities for cross-disorder research. *Mol. Psychiatry* **30**, 1627–1638 (2025).

147. Grotzinger, A. D. *et al.* Mapping the genetic landscape across 14 psychiatric disorders. *Nature* 1–15 (2025) doi:10.1038/s41586-025-09820-3.

148. Grotzinger, A. D. *et al.* Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nat. Hum. Behav.* **3**, 513–525 (2019).

149. van der Meer, D. *et al.* Mapping the Genetic Landscape of Psychiatric Disorders With the MiXeR Toolset. *Biol. Psychiatry* **98**, 455–465 (2025).

150. Weissman, M. M. Depression and Anxiety Disorders in Parents and Children: Results From the Yale Family Study. *Arch. Gen. Psychiatry* **41**, 845 (1984).

151. Kendler, K. S. The super-normal control group in psychiatric genetics: Possible artifactual evidence for coaggregation. *Psychiatr. Genet.* **1**, 45–53 (1990).

152. Chen, T. J. H. *et al.* Are dopaminergic genes involved in a predisposition to pathological aggression? Hypothesizing the importance of ‘super normal controls’ in psychiatricgenetic research of complex behavioral disorders. *Med. Hypotheses* **65**, 703–707 (2005).

153. Moskvina, V., Holmans, P., Schmidt, K. M. & Craddock, N. Design of case-controls studies with unscreened controls. *Ann. Hum. Genet.* **69**, 566–576 (2005).

154. Schwartz, S. & Susser, E. Genome-Wide Association Studies: Does Only Size Matter? *Am. J. Psychiatry* **167**, 741–744 (2010).

155. Schwartz, S. & Susser, E. The use of well controls: an unhealthy practice in psychiatric research. *Psychol. Med.* **41**, 1127–1131 (2011).

156. Kendler, K. S., Chatzinakos, C. & Bacanu, S.-A. The impact on estimations of genetic correlations by the use of super-normal, unscreened, and family-history screened controls in genome wide case-control studies. *Genet. Epidemiol.* **44**, 283–289 (2020).

157. Kohoutová, L. *et al.* Individual variability in brain representations of pain. *Nat. Neurosci.* **25**, 749–759 (2022).

158. Steingrímsdóttir, Ó. A., Landmark, T., Macfarlane, G. J. & Nielsen, C. S. Defining chronic pain in epidemiological studies: a systematic review and meta-analysis. *PAIN* **158**, 2092 (2017).

159. Mocci, E. *et al.* Genome wide association joint analysis reveals 99 risk loci for pain susceptibility and pleiotropic relationships with psychiatric, metabolic, and immunological traits. *PLOS Genet.* **19**, e1010977 (2023).

160. Meng, W. *et al.* Genetic correlations between pain phenotypes and depression and neuroticism. *Eur. J. Hum. Genet.* **28**, 358–366 (2020).

161. Mills, S. E. E., Nicolson, K. P. & Smith, B. H. Chronic pain: a review of its epidemiology and associated factors in population-based studies. *Br. J. Anaesth.* **123**, e273–e283 (2019).

162. Yao, C. *et al.* Exploring the bidirectional relationship between pain and mental disorders: a comprehensive Mendelian randomization study. *J. Headache Pain* **24**, 82 (2023).

163. Vos, T. *et al.* Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990–2016: a systematic

analysis for the Global Burden of Disease Study 2016. *The Lancet* **390**, 1211–1259 (2017).

164. Dowell, D. CDC Clinical Practice Guideline for Prescribing Opioids for Pain — United States, 2022. *MMWR Recomm. Rep.* **71**, (2022).
165. Nicol, A. L., Hurley, R. W. & Benzon, H. T. Alternatives to Opioids in the Pharmacologic Management of Chronic Pain Syndromes: A Narrative Review of Randomized, Controlled, and Blinded Clinical Trials. *Anesth. Analg.* **125**, 1682 (2017).
166. Rodin, R. *et al.* The 2016 CDC Opioid Guideline and Analgesic Prescribing Patterns in Older Adults With Cancer. *JAMA Netw. Open* **8**, e259043 (2025).
167. Koob, G. F. & Volkow, N. D. Neurocircuitry of Addiction. *Neuropsychopharmacology* **35**, 217–238 (2010).
168. Volkow, N. D. & McLellan, A. T. Opioid Abuse in Chronic Pain — Misconceptions and Mitigation Strategies. *N. Engl. J. Med.* **374**, 1253–1263 (2016).
169. Lu, T., Forgetta, V., Richards, J. B. & Greenwood, C. M. T. Capturing additional genetic risk from family history for improved polygenic risk prediction. *Commun. Biol.* **5**, 595 (2022).
170. Wenger, B. M. *et al.* A genotype-first approach to exploring Mendelian cardiovascular traits with clear external manifestations. *Genet. Med.* **23**, 94–102 (2021).
171. Forrest, I. S. *et al.* Machine learning-based marker for coronary artery disease: derivation and validation in two longitudinal cohorts. *The Lancet* **401**, 215–225 (2023).
172. Steinfeldt, J. *et al.* Medical history predicts phenome-wide disease onset and enables the rapid response to emerging health threats. *Nat. Commun.* **16**, 585 (2025).
173. Buuren, S. van & Groothuis-Oudshoorn, K. mice: Multivariate Imputation by Chained Equations in R. *J. Stat. Softw.* **45**, 1–67 (2011).
174. Kim, K. *et al.* A Machine-Learning-Algorithm-Based Prediction Model for Psychotic Symptoms in Patients with Depressive Disorder. *J. Pers. Med.* **12**, 1218 (2022).

175. Gao, S., Calhoun, V. D. & Sui, J. Machine learning in major depression: From classification to treatment outcome prediction. *CNS Neurosci. Ther.* **24**, 1037–1052 (2018).

176. Blass, I. *et al.* Revisiting the Risk Factors for Endometriosis: A Machine Learning Approach. *J. Pers. Med.* **12**, 1114 (2022).

177. Hujoel, M. L. A., Gazal, S., Loh, P.-R., Patterson, N. & Price, A. L. Liability threshold modeling of case-control status and family history of disease increases association power. *Nat. Genet.* **52**, 541–547 (2020).

178. Pedersen, E. M. *et al.* Accounting for age of onset and family history improves power in genome-wide association studies. *Am. J. Hum. Genet.* **109**, 417–432 (2022).

179. Lee, C. H. *et al.* Liability threshold model-based disease risk prediction based on electronic health record phenotypes. *Nat. Genet.* **57**, 2872–2881 (2025).

180. Chen, R. *et al.* Genetic analyses of eight complex diseases using predicted continuous representations of disease. *Cell Rep. Methods* **5**, (2025).

181. Miao, J. *et al.* Valid inference for machine learning-assisted genome-wide association studies. *Nat. Genet.* **56**, 2361–2369 (2024).

182. Genetic data | FinnGen. https://www.finngen.fi/en/genetic_data.

183. Lim, E. T. *et al.* Distribution and Medical Impact of Loss-of-Function Variants in the Finnish Founder Population. *PLOS Genet.* **10**, e1004494 (2014).

184. Loh, P.-R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).

185. Walter, K. *et al.* The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015).

186. Brumpton, B. M. *et al.* The HUNT study: A population-based cohort for genetic research. *Cell Genomics* **2**, 100193 (2022).

187. Alam, M. S. *et al.* Population screening of adults identifies novel genetic variants associated with celiac disease. *Sci. Rep.* **15**, 19764 (2025).

188. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).

189. Heng, T. H. *et al.* Widespread recessive effects on common diseases in a cohort of 44,000 British Pakistanis and Bangladeshis with high autozygosity. *Am. J. Hum. Genet.* **112**, 1316–1329 (2025).

190. Turley, P. *et al.* Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat. Genet.* **50**, 229–237 (2018).

191. Front page. *Fimea* https://www.fimea.fi/documents/542809/838272/13740_Q4.2008_vaihtokelpoiset.txt, https://spc.fimea.fi/indox/nam/html/nam/humspc/0/259390.pdf, https://www.fimea.fi/documents/160140/1187305/GS-ladattava.txt, https://www.fimea.fi/documents/160140/744738/29472_Varastointivelvoitteen_piriin_kuuluvat_laakevalmisteet_2015-05-19.txt, https://www.fimea.fi/documents/160140/744738/27145_Varastointivelvoitteen_piriin_kuuluvat_laakevalmisteet_2014-08-25.txt.

192. Nielsen, S., Degenhardt, L., Hoban, B. & Gisev, N. Comparing opioids: A guide to estimating oral morphine equivalents (OME) in research.

193. Opioid Equivalence Chart. *Gloucestershire Hospitals NHS Foundation Trust* https://www.gloshospitals.nhs.uk/healthcare-professionals/treatment-guidelines/opioid-equivalence-chart/.

194. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (Association for Computing Machinery, New York, NY, USA, 2016). doi:10.1145/2939672.2939785.

195. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).

196. Muiño, E. *et al.* Identification of Genetic Loci Associated With Intracerebral Hemorrhage Using a Multitrait Analysis Approach. *Neurology* **103**, e209666 (2024).

197. Lundberg, S. M. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. in *Advances in Neural Information Processing Systems* (eds Guyon, I. et al.) vol. 30 (Curran Associates, Inc., 2017).

198. Mbatchou, J. *et al.* Computationally efficient whole-genome regression for quantitative and binary traits. *Nat. Genet.* **53**, 1097–1103 (2021).

199. Lumley, T., Diehr, P., Emerson, S. & Chen, L. The Importance of the Normality Assumption in Large Public Health Data Sets. *Annu. Rev. Public Health* **23**, 151–169 (2002).

200. Genovese, G. *et al.* BCFtools/liftover: an accurate and comprehensive tool to convert genetic variants across genome assemblies. *Bioinformatics* **40**, btae038 (2024).

201. Storey, J. D. The positive false discovery rate: a Bayesian interpretation and the q-value. *Ann. Stat.* **31**, 2013–2035 (2003).

202. Armstrong, R. A. When to use the Bonferroni correction. *Ophthalmic Physiol. Opt.* **34**, 502–508 (2014).

203. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–375, S1-3 (2012).

204. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).

205. Ghoussaini, M. *et al.* Open Targets Genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. *Nucleic Acids Res.* **49**, D1311–D1320 (2021).

206. Open Targets Platform: facilitating therapeutic hypotheses building in drug discovery | Nucleic Acids Research | Oxford Academic. https://academic.oup.com/nar/article/53/D1/D1467/7917960?login=false.

207. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).

208. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

209. Speed, D., Holmes, J. & Balding, D. J. Evaluating and improving heritability models using summary statistics. *Nat. Genet.* **52**, 458–462 (2020).

210. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach on JSTOR. https://www.jstor.org/stable/2531595?origin=crossref&seq=1.

211. Viechtbauer, W. Conducting Meta-Analyses in R with the metafor Package. *J. Stat. Softw.* **36**, 1–48 (2010).

212. Jiao, Y. *et al.* Sleep disorders impact hormonal regulation: unravelling the relationship among sleep disorders, hormones and metabolic diseases. *Diabetol. Metab. Syndr.* **17**, 305 (2025).

213. Toikumo, S. *et al.* A multi-ancestry genetic study of pain intensity in 598,339 veterans. *Nat. Med.* **30**, 1075–1084 (2024).

214. Walters, R. K. *et al.* Transancestral GWAS of alcohol dependence reveals common genetic underpinnings with psychiatric disorders. *Nat. Neurosci.* **21**, 1656–1669 (2018).

215. Saunders, G. R. B. *et al.* Genetic diversity fuels gene discovery for tobacco and alcohol use. *Nature* **612**, 720–724 (2022).

216. Johnson, E. C. *et al.* A large-scale genome-wide association study meta-analysis of cannabis use disorder. *Lancet Psychiatry* **7**, 1032–1045 (2020).

217. Reizine, N. *et al.* Impact of CYP2D6 Pharmacogenomic Status on Pain Control Among Opioid-Treated Oncology Patients. *The Oncologist* **26**, e2042–e2052 (2021).

218. Tyndale, R. F., Droll, K. P. & Sellers, E. M. Genetically deficient CYP2D6 metabolism provides protection against oral opiate dependence. *Pharmacogenetics* **7**, 375–379 (1997).

219. Wang, Y. *et al.* Global Biobank analyses provide lessons for developing polygenic risk scores across diverse cohorts. *Cell Genomics* **3**, (2023).

220. Evseeva, I., Nicodemus, K. K., Bonilla, C., Tonks, S. & Bodmer, W. F. Linkage disequilibrium and age of HLA region SNPs in relation to classic HLA gene alleles within Europe. *Eur. J. Hum. Genet.* **18**, 924–932 (2010).

221. Albiñana, C. *et al.* Multi-PGS enhances polygenic prediction by combining 937 polygenic scores. *Nat. Commun.* **14**, 4702 (2023).

222. Knaak, S., Mantler, E. & Szeto, A. Mental illness-related stigma in healthcare: Barriers to access and care and evidence-based solutions. *Healthc. Manage. Forum* **30**, 111–116 (2017).

223. Bray, N. J. & O'Donovan, M. C. The genetics of neuropsychiatric disorders. *Brain Neurosci. Adv.* **2**, 2398212818799271 (2018).

224. Briançon-Marjollet, A. *et al.* The impact of sleep disorders on glucose metabolism: endocrine and molecular mechanisms. *Diabetol. Metab. Syndr.* **7**, 25 (2015).

225. Fayaz, A., Croft, P., Langford, R. M., Donaldson, L. J. & Jones, G. T. Prevalence of chronic pain in the UK: a systematic review and meta-analysis of population studies. *BMJ Open* **6**, e010364 (2016).

226. Okifuji, A. & Hare, B. D. The association between chronic pain and obesity. *J. Pain Res.* **8**, 399–408 (2015).

227. Bartley, E. J. & Fillingim, R. B. Sex differences in pain: a brief review of clinical and experimental findings. *BJA Br. J. Anaesth.* **111**, 52–58 (2013).

228. McIntosh, A. M. *et al.* Genetic and Environmental Risk for Chronic Pain and the Contribution of Risk Variants for Major Depressive Disorder: A Family-Based Mixed-Model Analysis. *PLoS Med.* **13**, e1002090 (2016).

229. Machiela, M. J. & Chanock, S. J. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics* **31**, 3555–3557 (2015).

230. Crews, K. R. *et al.* Clinical Pharmacogenetics Implementation Consortium Guideline for CYP2D6, OPRM1, and COMT Genotypes and Select Opioid Therapy. *Clin. Pharmacol. Ther.* **110**, 888–896 (2021).

231. Dean, L. & Kane, M. Codeine Therapy and CYP2D6 Genotype. in *Medical Genetics Summaries* (eds Pratt, V. M. et al.) (National Center for Biotechnology Information (US), Bethesda (MD), 2012).

232. Smith, D. M. *et al.* CYP2D6-guided opioid therapy improves pain control in CYP2D6 intermediate and poor metabolizers: a pragmatic clinical trial. *Genet. Med.* **21**, 1842–1850 (2019).

233. Rincy, T. N. & Gupta, R. Ensemble Learning Techniques and its Efficiency in Machine Learning: A Survey. in *2nd International Conference on Data, Engineering and Applications (IDEA)* 1–6 (2020). doi:10.1109/IDEA49133.2020.9170675.

234. Romagnoni, A., Jégou, S., Van Steen, K., Wainrib, G. & Hugot, J.-P. Comparative performances of machine learning methods for classifying Crohn Disease patients using genome-wide genotyping data. *Sci. Rep.* **9**, 10351 (2019).

235. Otero Varela, L., Doktorchik, C., Wiebe, N., Quan, H. & Eastwood, C. Exploring the differences in ICD and hospital morbidity data collection features across countries: an international survey. *BMC Health Serv. Res.* **21**, 308 (2021).

Original Publication