Capturing additional genetic risk from family history for improved polygenic risk prediction

Tianyuan Lu[1,2], **Vincenzo Forgetta**[1], **J Brent Richards**[1,3,4,5], **and Celia MT Greenwood**[1,3,4,6,*]

[1]Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, Canada
[2]Quantitative Life Sciences Program, McGill University, Montreal, Canada
[3]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Canada
[4]Department of Human Genetics, McGill University, Montreal, Canada
[5]Department of Twin Research and Genetic Epidemiology, King's College London, London, United Kingdom
[6]Gerald Bronfman Department of Oncology, McGill University, Montreal, Canada
[*]**Correspondence to** Celia MT Greenwood (celia.greenwood@mcgill.ca); Address: Jewish General Hospital, Room H-414, 3755 Chemin de la Côte-Sainte-Catherine, Montréal, Québec, H3T 1E2, Canada; Tel: 514-340-8222 ext. 28397

ABSTRACT

(150 words; currently 150)
Family history of complex traits may reflect transmitted rare pathogenic variants, intrafamilial shared exposures to environmental and lifestyle factors, as well as a common genetic predisposition. We developed a latent factor model to quantify trait heritability in excess of that captured by a common variant-based polygenic risk score, but inferable from family history. We applied our model to predict adult height for 941 children in the Avon Longitudinal Study of Parents and Children cohort as well as 11 complex diseases for ~400,000 European ancestry participants in the UK Biobank. Parental history brought consistent significant improvements in the predictive power of polygenic risk prediction. For instance, a joint predictor was able to explain ~55% of the total variance in sex-adjusted adult height z-scores, close to the estimated heritability. Our work showcases an innovative paradigm for risk calculation, and supports incorporation of family history into polygenic risk score-based genetic risk prediction models.

1 Introduction

Predicting phenotypic values of complex traits, or the risks and outcomes of complex diseases have strong implications in health care and biomedical research[1,2]. In recent years, large-scale genome-wide association studies (GWASs) have characterized the genetic architecture of many complex traits and diseases[3]. Developing polygenic risk scores aggregating the effects of well-profiled genetic determinants has become possible[4,5], and polygenic risk scores have demonstrated the potential to improve risk stratification in large populations[6–9], assist diagnosis and clinical differentiation[10,11], and refine risk management and treatment strategies[12,13].

However, most polygenic risk scores only capture linear additive effects of common genetic variants. Currently, most GWASs – even those based on the largest biobank studies – restrict consideration to genetic variants with a minor allele frequency $> 0.1\%$ or higher[3,14]. Furthermore, extremely rare pathogenic variants with high penetrance are rarely detected with array-based genotyping and imputation[3,15]. Furthermore, powerful approaches to accurately model more complex non-linear effects (i.e. dominance effects) and interaction effects (i.e. gene-by-environment effects and gene-by-gene effects) in a high dimensional setting are scarce, particularly since these effects are in general weaker than the linear additive effects[3,16]. Despite continuing methodological innovations in mining hidden heritability, these under-captured genetic effects likely prevent polygenic risk scores from achieving a further-improved predictive performance.

In contrast, family history information, such as parental measures of phenotypic values and disease records, provides an indirect measure of the overall genetic predisposition amongst relatives[17]. Although traditionally considered as a crucial risk factor for Mendelian diseases, family history has also shown added value in polygenic risk prediction[18–22]. For instance, individuals at an elevated level of risk for various types of cancer and cardiovascular diseases are more likely to be identified by assessments combining the family history of disease with polygenic risk scores[18–22], compared to using polygenic risk scores or family history alone.

Nevertheless, creating accurate joint predictors may be challenging, since it requires modelling individual-level data on phenotypes, genotypes, and family history information. Such regression models may not lead to effective prediction models if datasets containing all the required information are too small, particularly for diseases with a low prevalence in the population.

Therefore, in this work, we demonstrate both by theory and with examples the improved predictions associated with a new scheme for combining association test summary statistics and parental disease histories. We approach this goal by inquiring what proportion of trait variance is captured by parental information and not by existing polygenic risk scores. We construct joint predictors – both conceptually and concretely – incorporating both a polygenic risk score and the parental trait or disease information. Then evaluate the performance of these joint predictors in predicting adult height amongst 2,397 European ancestry children in the Avon Longitudinal Study of Parents and Children (ALSPAC) cohort[23,24], as well as in predicting risk for 11 complex diseases amongst ~400,000 European ancestry participants in the UK

Biobank[25].

2 Results

2.1 Inference of under-captured genetic components in family history

We propose a conceptual latent factor model to account for genetic components that are not modelled by polygenic risk scores but could be inferred from parental history (**Methods**). Briefly, for a continuous polygenic trait (**Figure 1A**), we assume that its genetic determinants can be partitioned into two orthogonal genetic components: one component captured by a polygenic risk score used for prediction, and the other component representing under-captured genetic effects. The under-captured genetic component could include the effects of unmeasured common variants, rare variants, gene-by-environment interactions, epistasis, inherited risk factor exposures, etc. We suppose that these two genetic components are independently passed on from the parents to the children, and parental measures of the trait may partially inform the under-captured genetic component (**Methods and Figure 1A**). This model can be adapted for binary diseases, wherein the parental disease history may inform the underlying genetic liability (**Figure 1B**).

Based on this model, it follows that prediction of a continuous trait or the genetic liability for a disease can be improved by combining a polygenic risk score and family history, requiring only (1) the magnitude of association between a polygenic risk score and the target trait or disease, and (2) the magnitude of association between parents' trait measures or disease history and the target trait or disease amongst children (**Methods**). Importantly, these estimates can be obtained from separate well-powered cohort studies, without the need to access individual-level information.

2.2 Improved height prediction amongst children by incorporating parental height measures

We assessed the performance of this joint predictor in predicting children's adult height. A polygenic risk score for age and sex-adjusted adult height z-score was recently developed using resources from the UK Biobank and the Genetic Investigation of ANthropometric Traits consortium[7,25,26]. On an out-of-sample test dataset in the UK Biobank, this score explained 36.7% of the total variance in height z-score[7]. To complement this score, an observational study found that the mid-parental height z-score was able to

explain 44.9% of the total variance in offspring height z-scores in the Erasmus Rucphen Family Study[27]. Based on these estimates, we derived that 58.1% of the total height z-score variance was under-captured and could be partially inferred from parental height measures (**Methods**).

A joint predictor combining this polygenic risk score and mid-parental height was then obtained for European ancestry children in the ALSPAC cohort (**Methods and Supplementary Table S1**). Consequently, among 941 genotyped children who had both biological parents' height measures, this joint predictor was able to explain 55.3% of the total variance in their sex-adjusted adult height z-score (**Figure 2A**). In contrast, similar to metrics obtained in literature, the polygenic risk score alone explained 38.2% of the total variance, while the mid-parental height z-score alone explained 43.1% (**Figure 2A**). As expected, the joint predictor also achieved the highest prediction accuracy with a root-mean-square error (RMSE) of 4.212 cm, compared to 4.955 cm by the polygenic risk score and 4.752 cm by the mid-parental height z-score (**Figure 2B**).

Furthermore, polygenic prediction could also be improved when only one parent's height was measured (**Methods**). Specifically, among 2,246 children with maternal height measures, a joint predictor incorporating the polygenic risk score and maternal height z-score explained 47.1% of the total variance and achieved an RMSE of 4.584 cm (**Figure 2C and D**). Among 1,092 children with paternal height measures, a joint predictor incorporating the polygenic risk score and paternal height z-score explained 47.0% of the total variance and achieved an RMSE of 4.607 cm (**Figure 2E and F**). Evidently, these joint predictors demonstrated superior predictive performance over the polygenic risk score or parental height alone.

Notably, all joint predictors demonstrated almost identical predictive performance as an in-sample combination of the polygenic risk score and parental height z-scores obtained from multivariate linear regression using individual-level data (**Figure 2**).

2.3 Improved complex disease risk prediction by incorporating parental disease history

Next, we tested whether a joint predictor could improve polygenic risk prediction for 11 complex diseases in the UK Biobank (**Supplementary Table S2**). Polygenic risk scores have been recently developed for each of these diseases and are documented in the PGS Catalog (**Supplementary Table**

S3)[28]. Estimation of under-captured genetic effects was conducted based on a training dataset consisting of 10% of the participants from the UK Biobank (**Methods and Supplementary Figure S1**). For most of these complex diseases, individuals having a parental disease history were significantly more likely to have the corresponding disease (**Supplementary Table S4**). The magnitudes of association reflect that a substantial proportion of genetic influence is not captured by the polygenic risk scores (**Supplementary Table S4**).

Incorporation of parental disease history into polygenic risk predictions led to a re-stratification of the predicted risks (**Figure 3A**). Specifically, individuals with a parental disease history would be considered at an elevated level of risk compared to those with a similar polygenic risk score but without a parental disease history. The discriminative power of polygenic risk scores in identifying individuals who developed the corresponding diseases was significantly improved (**Table 1 and Supplementary Table S5**). For instance, combined with age, sex, genotyping array, recruitment centre, and the first 10 genetic principal components, a joint predictor for myocardial infarction achieved an area under the receiver operating characteristic curve (AUROC) of 0.7625 (**Figure 3B**) and an area under the precision-recall curve (AUPRC) of 0.0834 (**Figure 3C**), based on a test dataset consisting of 90% of the participants from the UK Biobank (**Methods and Supplementary Figure S1**). In contrast, prediction based on the polygenic risk score (with other covariates but without parental history) had an AUROC of 0.7567 (DeLong's test p-value = 5.3×10^{-13}) and an AUPRC of 0.0800; prediction based on the parental disease history (with other covariates but without the polygenic risk score) had an AUROC of 0.7375 (DeLong's test p-value = 3.0×10^{-126}) and an AUPRC of 0.0671 (**Figure 3B and C**).

Meanwhile, the joint predictor could more accurately assign individuals into higher-vs-lower risk groups, indicated by positive net-reclassification indices (NRI) across different percentile cutoffs and positive integrated discrimination improvement (IDI) indices (**Table 1**). For example, at the 95-th percentile cutoff (i.e. 5% of the population to be considered at high risk), the joint predictor for Alzheimer's disease had the highest NRI of 5.76% over the polygenic risk score for Alzheimer's disease, followed by the joint predictor for chronic obstructive pulmonary disease (COPD), with an NRI of 5.07% (**Table 1**). Notably, the joint predictor for Alzheimer's disease was significantly associated with carrying one or two *APOE* e4 alleles (**Methods and Supplementary Figure S2A and B**), which is a well-known genetic risk factor

for Alzheimer's disease[29–31] but is not included in the polygenic risk score for this analysis. In addition, individuals with a higher score in the joint predictor for COPD were slightly yet significantly more likely to be ever-smokers, while the the polygenic risk score alone did not capture the genetic predisposition to smoking (**Methods and Supplementary Figure S2C**).

For most diseases under investigation, the joint predictors consistently achieved comparable predictive performance to gold-standard predictors obtained by fitting multivariate logistic regression models based on individual-level data from the training datasets (**Methods and Supplementary Figure S1**). For breast cancer, lung cancer, stroke and Alzheimer's disease, the joint predictors even achieved a marginally higher AUROC than the data-driven predictors (**Table 1 and Supplementary Figure S3**), though the joint predictors for ischemic heart disease and type 2 diabetes did not appear to be ideal (**Table 1 and Supplementary Figure S3**). However, as expected, all of the individual-level data-driven predictors demonstrated sensitivity to sample size (**Figure 3D and Supplementary Figure S3**). For instance, for myocardial infarction, if a regression-based predictor were to be derived using $\leq 20\%$ of the individuals included in the current training dataset, its discriminative power would almost certainly be worse than the polygenic risk score alone (**Figure 3D**).

3 Discussion

Polygenic risk scores for complex traits are effective research tools and may greatly improve personalized health care in clinical practice[1,2]. Development of polygenic risk scores relies on accurate characterization of genetic determinants in large-scale GWASs. Due to limitation of resources, most existing polygenic risk scores are restricted to modelling linear additive effects of common genetic variants, thus may benefit from being combined with predictors that are able to capture more elusive genetic effects. In this study, we have proposed a simple latent factor model to quantify and extract heritability not captured by polygenic risk scores but inferable from family history. We have systematically investigated the utility of adding family history into polygenic risk score-based polygenic risk prediction models. For the 12 complex traits we examined, the combination of parental information and polygenic risk scores brought important, though sometimes marginal improvements in predictive performance.

In developing this prediction scheme, we partitioned trait or disease heritability onto two orthogonal

genetic components, where family history is assumed to partially inform a latent genetic component despite being correlated with the polygenic risk score. To account for the polygenicity and complexity of the underlying genetic effects, the genetic components were assumed to be normally distributed. We recognize that while these assumptions have strength in facilitating model specification and construction of joint predictors, they may not be theoretically optimal. For example, if strong gene-by-environment interaction effects exist, the under-captured genetic component should, at least in part, be correlated with the polygenic risk score. However, correlation or interaction between the two genetic components is unidentifiable without accessing individual-level data. Encouragingly, despite these strong assumptions, we found in practice that polygenic prediction could consistently be improved through our models built with this point of view, regardless of the underlying genetic architecture of complex traits, disease prevalence, and the methods adopted to develop polygenic risk scores[28]. In fact, the proportion of variance explained in adult height z-scores by our joint predictor in the ALSPAC cohort was similar to the estimated total SNP heritability of height z-scores[32–35]. In contrast to multivariate regression-based joint predictors, our method can leverage results of association tests obtained from separate cohort studies with a high statistical power, thus this method does not suffer when large individual-level datasets are unavailable.

Our findings have important implications for developing polygenic risk predictors, since the under-captured genetic components may have known sources. For instance, a large proportion of Alzheimer's disease risk heritability is conferred by the common allele of *APOE* e4[29–31]. Not surprisingly, the parental disease history of Alzheimer's disease appeared to be more predictive of the disease risk compared to a polygenic risk score not including this allele, while the joint predictor demonstrated clear advantages in risk stratification. Furthermore, by design, inherited risk factor exposure was also included in the under-captured genetic component. Smoking is one of the most important risk factors for COPD[36] and is heritable[37,38]. In our model, family disease history of COPD may partially capture the genetic predisposition to smoking as well as other risk factors that is not fully represented in the polygenic risk score. As a result, the corresponding joint predictor aggregating additional genetic risks demonstrated a prominent improvement over the polygenic risk score. These results not only support the utility of family history in enhancing predictive power, but also encourage explicit modelling of large effects such as monogenic causes or significant risk factor exposures in polygenic risk prediction.

We note possible model mis-specification for ischemic heart disease and type 2 diabetes because the joint predictors we constructed displayed compromised predictive performance compared to the gold-standard predictors. We hypothesize that this may be due to errors in the definitions of disease in the parental histories as well as in the medical histories of the participants. Specifically, sub-classifications of diseases were not available for parental “heart disease” and "diabetes", where the former may include various types of diseases affecting the cardiovascular system, and the latter may include both type 1 and type 2 diabetes. Consistent disease definitions and comprehensive phenotyping are thus required for validating our findings in clinical practice and in research. Furthermore, we anticipate that the predictive performance of the joint predictors could be further enhanced with additional knowledge of family history, such as the disease history of relatives other than the parents, as well as age of disease onset for each affected relative. However, we expect that information gained from second-degree or more distant relatives would be less significant compared to first-degree relatives.

Last, our findings should be considered population-specific. Participants in our study cohorts are predominantly of European ancestries. It has been widely recognized that polygenic risk scores can have largely attenuated predictive performance when applied to populations of different genetic ancestries[39,40]. Hence, we recommend extensive validations of our results in diverse populations as well as development of ancestry-specific polygenic risk scores. Adding to this restriction, we also note that participants in the UK Biobank have been shown to be slightly healthier, less obese, and less likely to smoke and drink alcohol than the general population in the United Kingdom[41]. In particular, the average age of the study cohort was substantially younger than the average age of onset for Alzheimer’s disease and Parkinson’s disease[42,43], thus the corresponding disease prevalence was low. Therefore, if a joint predictor were constructed for other populations, we strongly recommend estimating model parameters based on reference cohort studies with similar demographic characteristics and prevalences of disease.

In summary, we have developed a novel risk calculation scheme by incorporating family history into polygenic risk scores. We found a substantial proportion of complex trait heritability was under-captured and could be partly inferred from family history. Our findings support the utility of combining family history with polygenic risk scores as well as investigations into complex genetic effects not captured by existing polygenic risk scores for improved genetic risk prediction.

4 Methods

4.1 A latent factor model for polygenic inheritance in parent-child trios

We first consider a simple latent factor model for a normally distributed polygenic trait

$$Y_M = \alpha P_M + \beta G_M + \varepsilon_M \tag{1}$$

$$Y_F = \alpha P_F + \beta G_F + \varepsilon_F \tag{2}$$

$$Y_C = \alpha P_C + \beta G_C + \varepsilon_C \tag{3}$$

where the subscripts M, F, and C stand for the populations of mothers, fathers, and children, respectively. The mothers and the fathers are assumed to be unrelated. P represents the genetic component already captured by a polygenic risk score, while G represents the under-captured genetic component, including under-captured linear additive effects, non-additive effects, rare pathogenic variant effects, gene-gene or gene-environment interaction effects, inherited exposure to risk factors, etc. Importantly, G is assumed to be independent of P, and these two components are independently passed on from the parents to the children. We realize that this may be an unrealistic assumption, particularly for common genetic effects that were not captured in existing polygenic risk scores, but this assumption enables conceptual development of our latent model. Results in **Figure 2** and **Table 1** justify the usefulness of the perspective. ε captures the non-genetic residual variance in Y, where $\varepsilon \perp (P, G)$. We assume both P and G are normally distributed because P, under a polygenic model, consists of multiple independent effects, each having an infinitesimal effect size, while G has a multifactorial nature. Without loss of generality, we suppose that Y, P, and G are scaled to have zero mean and unit variance.

We further assume that the effects of the modelled polygenic component and the under-captured genetic component, α and β, are invariant in the populations of parents and children. Naturally, the overall heritability of this trait is $\alpha^2 + \beta^2$.

Next, we describe the joint distribution of $\boldsymbol{P}$, $\boldsymbol{G}$, and $\boldsymbol{Y}$, where $\boldsymbol{P} = (P_M, P_F, P_C)^\top$, $\boldsymbol{G} = (G_M, G_F, G_C)^\top$, and $\boldsymbol{Y} = (Y_M, Y_F, Y_C)^\top$:

$$\begin{pmatrix} \boldsymbol{P} \\ \boldsymbol{G} \\ \boldsymbol{Y} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \Sigma_{PP} & \Sigma_{PG} & \Sigma_{PY} \\ \Sigma_{GP} & \Sigma_{GG} & \Sigma_{GY} \\ \Sigma_{YP} & \Sigma_{YG} & \Sigma_{YY} \end{pmatrix} \right) \tag{4}$$

Based on the polygenic model assumptions, the covariance between the children's genetic components and their parents' corresponding genetic components is expected to be $\frac{1}{2}$, as each child inherits half of the trait-determining alleles from their mother, and the other half from their father. Hence, we specify empirical covariance matrices for Σ_{PP} and Σ_{GG}:

$$\Sigma_{PP} = \Sigma_{GG} = \begin{pmatrix} 1 & 0 & \frac{1}{2} \\ 0 & 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 1 \end{pmatrix} \tag{5}$$

It is noteworthy that by specifying these variance-covariance matrices, we assume no consanguinity and no assortative mating.

Since $\boldsymbol{P}$ and $\boldsymbol{G}$ are assumed to be independent of each other, we set:

$$\Sigma_{PG} = \Sigma_{GP} = \boldsymbol{0} \tag{6}$$

For associations between the trait and genetic components, we can derive that

$Cov(P_k, Y_k) = Cov(P_k, \alpha P_k + \beta G_k + \varepsilon_k) = \alpha$ for $k \in \{M, F, C\}$ and $Cov(P_C, Y_{M/F}) = Cov(P_C, \alpha P_{M/F} +$

$\beta G_{M/F} + \varepsilon_{M/F}) = \frac{\alpha}{2}$ (the subscript M/F represents maternal or paternal component), thus

$$\Sigma_{PY} = \Sigma_{YP} = \begin{pmatrix} \alpha & 0 & \frac{\alpha}{2} \\ 0 & \alpha & \frac{\alpha}{2} \\ \frac{\alpha}{2} & \frac{\alpha}{2} & \alpha \end{pmatrix} \tag{7}$$

Similarly,

$$\Sigma_{GY} = \Sigma_{YG} = \begin{pmatrix} \beta & 0 & \frac{\beta}{2} \\ 0 & \beta & \frac{\beta}{2} \\ \frac{\beta}{2} & \frac{\beta}{2} & \beta \end{pmatrix} \tag{8}$$

Furthermore, the covariance between the children's trait and the parents' trait $Cov(Y_C, Y_{M/F}) = Cov(\alpha P_C + \beta G_C + \varepsilon_C, \alpha P_{M/F} + \beta G_{M/F} + \varepsilon_{M/F}) = \frac{\alpha^2+\beta^2}{2}$. Therefore,

$$\Sigma_{YY} = \begin{pmatrix} 1 & 0 & \frac{\alpha^2+\beta^2}{2} \\ 0 & 1 & \frac{\alpha^2+\beta^2}{2} \\ \frac{\alpha^2+\beta^2}{2} & \frac{\alpha^2+\beta^2}{2} & 1 \end{pmatrix} \tag{9}$$

Equation 9 implies that if a combination of the parental trait measures, usually the mid-parental trait measure, is used to predict the children's trait, the expected proportion of variance explained is

$$Corr^2\left(Y_C, \frac{Y_M + Y_F}{2}\right) = \left(\frac{\frac{1}{2}Cov(Y_C, Y_M) + \frac{1}{2}Cov(Y_C, Y_F)}{\sqrt{Var(Y_C)Var(\frac{Y_M+Y_F}{2})}}\right)^2 = \frac{(\alpha^2+\beta^2)^2}{2} \tag{10}$$

In practice, an estimate of α can be obtained from studies developing polygenic risk scores, i.e. $\hat{\alpha}^2 = \widehat{Corr}^2(P, Y)$. Meanwhile, observational studies reporting the magnitude of associations between

parental trait measures and the children's trait measures can inform the under-captured heritability β^2, e.g.

$$\hat{\beta}^2 = \sqrt{2\hat{Corr}^2(Y_C, \frac{Y_M + Y_F}{2}) - \hat{\alpha}^2} \tag{11}$$

4.2 Continuous trait prediction incorporating parental trait measures

Based on the distributional assumptions in Equation 4, we are able to infer Y_C with the parental trait measures (Y_M and/or Y_F) and the children's polygenic risk score (P_C), if we have estimates of α and β. Specifically, given the properties of the multivariate normal distribution,

$$\begin{pmatrix} Y_C \\ Y_M \\ Y_F \\ P_C \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \frac{\alpha^2+\beta^2}{2} & \frac{\alpha^2+\beta^2}{2} & \alpha \\ \frac{\alpha^2+\beta^2}{2} & 1 & 0 & \frac{\alpha}{2} \\ \frac{\alpha^2+\beta^2}{2} & 0 & 1 & \frac{\alpha}{2} \\ \alpha & \frac{\alpha}{2} & \frac{\alpha}{2} & 1 \end{pmatrix} \right) \tag{12}$$

For the j-th individual with a polygenic risk score $p_{C,j}$ and parental measures of the trait $y_{M,j}$, $y_{F,j}$, we use the conditional expectation as the predictor for the trait:

$$\bar{Y}_{C,j} = E(Y_C | Y_M = y_{M,j}, Y_F = y_{F,j}, P_C = p_{C,j}) = \begin{pmatrix} \frac{\hat{\alpha}^2+\hat{\beta}^2}{2} & \frac{\hat{\alpha}^2+\hat{\beta}^2}{2} & \hat{\alpha} \end{pmatrix} \begin{pmatrix} 1 & 0 & \frac{\hat{\alpha}}{2} \\ 0 & 1 & \frac{\hat{\alpha}}{2} \\ \frac{\hat{\alpha}}{2} & \frac{\hat{\alpha}}{2} & 1 \end{pmatrix}^{-1} \begin{pmatrix} y_{M,j} \\ y_{F,j} \\ p_{C,j} \end{pmatrix} \tag{13}$$

Alternatively, we can create a predictor using the polygenic risk score with one parental measure if only the maternal or the paternal measure is available:

$$\bar{Y}_{C,j} = E(Y_C | Y_{M/F} = y_{M/F,j}, P_C = p_{C,j}) = \begin{pmatrix} \frac{\hat{\alpha}^2+\hat{\beta}^2}{2} & \hat{\alpha} \end{pmatrix} \begin{pmatrix} 1 & \frac{\hat{\alpha}}{2} \\ \frac{\hat{\alpha}}{2} & 1 \end{pmatrix}^{-1} \begin{pmatrix} y_{M/F,j} \\ p_{C,j} \end{pmatrix} \tag{14}$$

This framework is generalizable to any degree of family relationships, with modification of the variance-covariance matrices. However, because information upon more distantly related family members is more difficult to obtain in large cohort studies and because parental information is the most relevant to risk prediction, we focus on parental trait measures or disease history in this work.

4.3 Modelling latent genetic components for binary diseases

For binary diseases, we combine a liability model with the above latent factor model and explicitly model measurable covariate effects (such as age and sex) since they are non-trivial in this case. We denote the mothers', fathers' and children's diseases of interest as Z_M, Z_F, and Z_C, respectively. We assume that

$$(Z_k|P_k = p, G_k = g, Q_k = \mathbf{q}) \sim Bernoulli(\frac{e^{\mu_0+\alpha p+\beta g+\gamma \mathbf{q}}}{1+e^{\mu_0+\alpha p+\beta g+\gamma \mathbf{q}}}) \tag{15}$$

for $k \in \{M, F, C\}$, where Q represents measured (single or multiple) covariates with effects γ, and μ_0 denotes the baseline odds (on the logarithmic scale) of the disease in the target population. P and G are assumed to be independently distributed with $P \sim \mathcal{N}(0,1)$ and $G \sim \mathcal{N}(0,1)$.

If we introduce $H = \alpha P + \beta G$, where ($H \sim \mathcal{N}(0, \alpha^2 + \beta^2)$) includes both genetic components, then equation 15 simplifies to

$$(Z_k|H_k = h, Q_k = \mathbf{q}) \sim Bernoulli(\frac{e^{\mu_0+h+\gamma \mathbf{q}}}{1+e^{\mu_0+h+\gamma \mathbf{q}}}) \tag{16}$$

With the Bernoulli distribution assumptions, we adopt a logit-link in modelling the genetic liability, instead of the probit-link which is more commonly used for "liability threshold models"[17]. This choice facilitates estimation of the model parameters from the results of widely-implemented logistic regression models. Nevertheless, heritability estimates can still be obtained. As shown previously[44], an approximation

to obtain the effect size estimate on the liability scale is

$$\tau_\alpha \approx \frac{\Phi^{-1}[F_{Logistic}(\mu_0+\alpha)] - \Phi^{-1}[F_{Logistic}(\mu_0)]}{\sqrt{1+\{\Phi^{-1}[F_{Logistic}(\mu_0+\alpha)] - \Phi^{-1}[F_{Logistic}(\mu_0)]\}^2 Var(P)}} \tag{17}$$

and

$$\tau_\beta \approx \frac{\Phi^{-1}[F_{Logistic}(\mu_0+\beta)] - \Phi^{-1}[F_{Logistic}(\mu_0)]}{\sqrt{1+\{\Phi^{-1}[F_{Logistic}(\mu_0+\beta)] - \Phi^{-1}[F_{Logistic}(\mu_0)]\}^2 Var(G)}} \tag{18}$$

where Φ is the standard normal cumulative distribution function, $F_{Logistic}$ is the cumulative distribution function of a logistic distribution with mean of 0 and variance of $\frac{\pi^2}{3}$, i.e. $F_{Logistic}(x) = \frac{1}{1+exp\{-\frac{\pi}{\sqrt{3}}x\}}$, and $Var(P) = Var(G) = 1$.

Consequently, heritability captured by the polygenic risk score can be estimated as τ_α^2, while the under-captured heritability can be estimated as τ_β^2. Notably, unlike the model for continuous traits (Section 4.1) where α and β are automatically bounded ($0 \leq \alpha^2 + \beta^2 \leq 1$), here, the effects of the genetic components P and G are unconstrained.

Estimates of the unknown parameters can be obtained from summary statistics of existing observational studies. That is, we can directly obtain $\hat{\alpha}$ based on the magnitude of association between a polygenic risk score and the disease risk, i.e. odds ratio (OR) per one standard deviation increase in the polygenic risk score, together with $\hat{\mu}_0$ and $\hat{\gamma}$. Furthermore, we can empirically estimate β if we have an estimate of the association between a parental disease history and the disease risk amongst the children.

Specifically, from the joint distribution

$$\begin{aligned} f(Z_C, Z_{M/F}) &= \int_{H_C, H_{M/F}} f(Z_C, Z_{M/F}, H_C, H_{M/F}) \\ &= \int_{H_C, H_{M/F}} f(Z_C|H_C) f(Z_{M/F}|H_{M/F}) f(H_C, H_{M/F}) \end{aligned} \tag{19}$$

where

$$f(Z_k|H_k = h) = (\frac{e^{\mu_0+h+\gamma\mathbf{q}}}{1+e^{\mu_0+h+\gamma\mathbf{q}}})^{z_k}(\frac{1}{1+e^{\mu_0+h+\gamma\mathbf{q}}})^{1-z_k} \tag{20}$$

for $k \in \{M, F, C\}$, and

$$\begin{pmatrix} H_C \\ H_{M/F} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \alpha^2+\beta^2 & \frac{\alpha^2+\beta^2}{2} \\ \frac{\alpha^2+\beta^2}{2} & \alpha^2+\beta^2 \end{pmatrix}\right) \tag{21}$$

we have an explicit expression for the OR based on either the maternal or the paternal disease history as a function of the unknown β:

$$\widetilde{OR}_{M/F} = \frac{f(Z_C = 1, Z_{M/F} = 1)/f(Z_C = 0, Z_{M/F} = 1)}{f(Z_C = 1, Z_{M/F} = 0)/f(Z_C = 0, Z_{M/F} = 0)} \tag{22}$$

Therefore, we perform a numerical line search to obtain an empirical estimate of β such that the theoretical $\widetilde{OR}_{M/F}$ is close to the observed $\widehat{OR}_{M/F}$:

$$\hat{\beta} = argmin_{\beta}\{|(log(\widehat{OR}_M) - log(\widetilde{OR}_M)) + (log(\widehat{OR}_F) - log(\widetilde{OR}_F))|\} \tag{23}$$

Note that the OR associated with the maternal disease history may differ from that based on the paternal disease history, because sex is included as a covariate (in Q) in this model.

4.4 Binary disease risk prediction incorporating parental disease history

Similar to predicting a continuous trait (Section 4.2), we aim to infer $f(H_C|Z_M, Z_F, P_C)$. since the distribution of H_C naturally informs $f(Z_C)$. Then we can use the conditional expectation as the predictor.

We observe that

$$f(H_C|Z_M,Z_F,P_C) = f(H_C|H_M,H_F,P_C)f(H_M|Z_M)f(H_F|Z_F) \tag{24}$$

has no closed-form solution for its expectation, since

$$\begin{pmatrix} H_C \\ H_M \\ H_F \\ P_C \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \alpha^2+\beta^2 & \frac{\alpha^2+\beta^2}{2} & \frac{\alpha^2+\beta^2}{2} & \alpha \\ \frac{\alpha^2+\beta^2}{2} & \alpha^2+\beta^2 & 0 & \frac{\alpha}{2} \\ \frac{\alpha^2+\beta^2}{2} & 0 & \alpha^2+\beta^2 & \frac{\alpha}{2} \\ \alpha & \frac{\alpha}{2} & \frac{\alpha}{2} & 1 \end{pmatrix}\right) \tag{25}$$

and

$$f(H_k|Z_k=z_k) = \frac{f(Z_k=z_k|H_k)f(H_k)}{f(Z_k=z_k)} \propto \left(\frac{e^{\mu_0+h_k+\gamma\mathbf{q}}}{1+e^{\mu_0+h_k+\gamma\mathbf{q}}}\right)^{z_k}\left(\frac{1}{1+e^{\mu_0+h_k+\gamma\mathbf{q}}}\right)^{(1-z_k)}\left(e^{-\frac{h_k^2}{2(\alpha^2+\beta^2)}}\right) \tag{26}$$

for $k \in \{M,F,C\}$. Therefore, we implement an importance sampling scheme to approximate this conditional distribution and derive its expectation:

- For the j-th individual with a polygenic risk score $p_{C,j}$ and parental disease history $z_{M,j} \in \{0,1\}$, $z_{F,j} \in \{0,1\}$, we randomly generate L (a large number, e.g. 1,000,000) samples of $H_{M,j}$ and $H_{F,j}$ based on $f(H_M|Z_M=z_{M,j})$ and $f(H_F|Z_F=z_{F,j})$ in Equation 26, respectively;

- For the l-th sample of $H_{M,j}$ and $H_{F,j}$, denoted as $h_{M,j,l}$ and $h_{F,j,l}$, we derive

$$E_{(l)}(H_C|H_M=h_{M,j,l},H_F=h_{F,j,l},P_C=p_{C,j}) = \begin{pmatrix} \frac{\hat{\alpha}^2+\hat{\beta}^2}{2} & \frac{\hat{\alpha}^2+\hat{\beta}^2}{2} & \hat{\alpha} \end{pmatrix} \begin{pmatrix} \hat{\alpha}^2+\hat{\beta}^2 & 0 & \frac{\hat{\alpha}}{2} \\ 0 & \hat{\alpha}^2+\hat{\beta}^2 & \frac{\hat{\alpha}}{2} \\ \frac{\hat{\alpha}}{2} & \frac{\hat{\alpha}}{2} & 1 \end{pmatrix}^{-1} \begin{pmatrix} h_{M,j,l} \\ h_{F,j,l} \\ p_{C,j} \end{pmatrix} \tag{27}$$

- We repeat for all L samples and obtain the predictor for the j-th individual as

$$\bar{H}_{C,j} = \frac{1}{L}\sum_{l=1}^{L} E_{(l)}(H_C|H_M = h_{M,j,l}, H_F = h_{F,j,l}, P_C = p_{C,j}) \tag{28}$$

Alternatively, we can utilize the disease history of only one parent, particularly when a disease is highly sex-specific, where Equation 27 is modified as

$$E_{(l)}(H_C|H_{M/F} = h_{M/F,j,l}, P_C = p_{C,j}) = \begin{pmatrix} \frac{\hat{\alpha}^2+\hat{\beta}^2}{2} & \hat{\alpha} \end{pmatrix} \begin{pmatrix} \hat{\alpha}^2+\hat{\beta}^2 & \frac{\hat{\alpha}}{2} \\ \frac{\hat{\alpha}}{2} & 1 \end{pmatrix}^{-1} \begin{pmatrix} h_{M/F,j,l} \\ p_{C,j} \end{pmatrix} \tag{29}$$

and Equation 28 is modified correspondingly as

$$\bar{H}_{C,j} = \frac{1}{L}\sum_{l=1}^{L} E_{(l)}(H_C|H_{M/F} = h_{M/F,j,l}, P_C = p_{C,j}) \tag{30}$$

4.5 Predicting adult height for children in the Avon Longitudinal Study of Parents and Children

4.5.1 Study cohort

From 1991 to 1992, the ALSPAC cohort recruited 14,541 pregnancies in the Bristol and Avon areas in the United Kingdom. In addition, 913 pregnancies were enrolled in later phases of the study. The total sample size for analyses using data collected after the age of seven is therefore 15,454 pregnancies, resulting in 15,589 foetuses. Of these, 14,901 were alive at 1 year of age. Details of this prospective cohort have been described previously[23,24]. Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees. Consent for biological samples has been collected in accordance with the Human Tissue Act (2004). Informed consent for the use of data collected via questionnaires and clinics was obtained from participants following the recommendations of the ALSPAC Ethics and Law Committee at the time.

Genotyping of the children initially recruited into this cohort was conducted using the Illumina

HumanHap550 quad genotyping platforms. The genotypes were imputed to the 1000 Genomes Phase 3 reference panel[45]. Height of the biological parents was measured during clinical visits. Children's adult height was measured at age 24. Measurement of height was performed using a Harpenden stadiometer (Holtain Ltd). Among the genotyped children of European ancestries who had measured adult standing height, 941 had both maternal and paternal height measures of biological parents, 1,305 only had maternal height measures of biological mothers, and 151 only had paternal height measures of biological fathers (**Supplementary Table S1**). We generated height z-scores by separately standardizing the children's, mothers' and fathers' measured height to have zero mean and unit variance, where the standardization for the population of children was stratified by sex.

4.5.2 Parameter estimation and prediction

From studies that did not involve the ALSPAC cohort, we obtained an empirical estimate for α based on the association between a polygenic risk score for height z-score and the measured height z-score[7]. We further obtained an empirical estimate for β based on the association between the mid-parental height z-score and the measured height z-score[27] based on Equation 11. Subsequently, following Section 4.2 we derived predicted height z-score for each child by combining their calculated polygenic risk score with one or both parental height z-scores (Equation 13 for children with mid-parental height measures and Equation 14 for children with only maternal or paternal height measures).

We evaluated the proportion of variance explained and the RMSE (after transforming the predictor of z-scores back to the scale of absolute height measures) of the joint predictors, and compared these metrics of predictive performance with those based on the polygenic risk score, the parental height z-score alone.

Furthermore, a gold-standard predictor was obtained by fitting multivariate linear regression including both the polygenic risk score and parental height z-score as predictors on the test samples. This represents the upper-bound of predictive performance achievable by any linear predictors.

4.6 Predicting risk of complex diseases in the UK Biobank

4.6.1 Study cohort

From 2006 to 2010, the UK Biobank study recruited approximately 500,000 participants who were aged between 40-69 years, at multiple recruitment centers in the United Kingdom[25]. Ethics approval

for the UK Biobank study was obtained from the North West Centre for Research Ethics Committee (11/NW/0382). The UK Biobank ethics statement is available at https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/about-us/ethics. All UK Biobank participants provided informed consent at recruitment.

Upon recruitment, demographic and anthropometric information were collected. Genotyping of more than 480,000 participants were conducted using the Applied Biosystems™ UK BiLEVE Axiom™ Array or UK Biobank Axiom™ Array. The genotypes were imputed to the Haplotype Reference Consortium reference panel[46].

4.6.2 Ascertainment of disease outcomes and parental disease history

Participants who had any of following complex diseases were identified based on inpatient International Classification of Diseases (ICD-10) diagnosis codes, Office of Population Censuses and Surveys (OPCS-4) procedure codes, or self-reported medical history during an interview with a trained nurse (**Supplementary Table S2**). ICD-10 codes for cancer diagnoses were retrieved by the UK Biobank through the national cancer registries. These disease outcomes included both prevalent cases identified upon initial recruitment and incident cases identified in more recent follow-up data collection.

- Breast cancer: ICD-10 code C50 (malignant neoplasm of breast), specific to women;
- Prostate cancer: ICD-10 code C61 (malignant neoplasm of prostate), specific to men;
- Colorectal cancer: ICD-10 codes C18 (malignant neoplasm of colon), C19 (malignant neoplasm of rectal sigmoid junction), or C20 (malignant neoplasm of rectum);
- Lung cancer: ICD-10 code C34 (malignant neoplasm of bronchus and lung);
- Myocardial infarction: ICD-10 code I21 (acute myocardial infarction);
- Ischemic heart disease: ICD-10 codes I20 (angina pectoris), I21 (acute myocardial infarction), I22 (subsequent myocardial infarction), I23 (complications following acute myocardial infarction), I24 (other acute ischemic heart diseases), or I25 (chronic ischemic heart disease), or OPCS-4 codes for coronary artery bypass grafting or coronary angioplasty with or without stenting;

- Stroke: ICD-10 codes I60 (subarachnoid haemorrhage), I61 (intracerebral haemorrhage), I62 (other nontraumatic intracranial haemorrhage), I63 (cerebral infarction), or I64 (stroke, not specified as haemorrhage or infarction);

- Type 2 diabetes: ICD-10 codes E11 (non-insulin-dependent diabetes mellitus), E13 (other specified diabetes mellitus), or E14 (unspecified diabetes mellitus), self-reported physician-made diagnosis, or self-reported use of anti-diabetic medications, excluding ICD-10 code E10 (insulin-dependent/type 1 diabetes mellitus);

- Alzheimer's disease: ICD-10 code G30 (Alzheimer's disease);

- Parkinson's disease: ICD-10 code G20 (Parkinson's disease);

- COPD: ICD-10 code J44 (chronic obstructive pulmonary disease), self-reported physician-made diagnosis, or self-reported use of medications for COPD.

Upon recruitment, a questionnaire inquired whether a participant had a parental history of "breast cancer", "prostate cancer", "bowel cancer", "lung cancer", "heart disease", "stroke", "high blood pressure", "diabetes", "Alzheimer's disease or dementia", "Parkinson's disease", or "chronic bronchitis or emphysema". Further classifications of these diseases were not available, e.g. "heart disease" may include various types of diseases affecting the cardiovascular system, and "diabetes" may include both type 1 and type 2 diabetes. Participants who responded "do not know" or "prefer not to answer" were considered missing data. No participant reported paternal history of breast cancer or maternal history of prostate cancer. We matched the participants' diseases with these parental records of diseases that had the same or a similar clinical definition.

Notably, while disease history of siblings was also available in the UK Biobank, it lacked information upon whether the sibling was a full-, half-, or adopted sibling, and how many siblings were affected by the disease. Because these details were essential for correctly specifying the model, we refrained from incorporating sibling disease history in this study.

4.6.3 Parameter estimation and prediction

We retrieved polygenic risk scores for the above complex diseases from the PGS catalog[28]. These polygenic risk scores were developed using different computational approaches based on source populations that did not strongly overlap with the UK Biobank (**Supplementary Table S3**). We used the same well-powered polygenic risk score for predicting myocardial infarction and ischemic heart disease, as myocardial infarction is a manifestation of ischemic heart disease.

Because the model parameters of interest likely have a strong population-specificity, we first used a training dataset, comprising randomly selected 10% of the UK Biobank participants for deriving these parameters (**Supplementary Table S2 and Supplementary Figure S1**). Specifically, for each disease, we fitted logistic regression models separately for the corresponding polygenic risk score and parental disease history (maternal disease history and paternal disease history as two independent variables), while including covariate effects of age, sex (except for breast cancer and prostate cancer), recruitment centre, genotyping array, and the first 10 genetic principal components. These two logistic regression models led to empirical estimates of $\hat{\alpha}$, $\hat{\beta}$, the baseline odds of disease $\hat{\mu}_0$, as well as the covariate effects $\hat{\gamma}$ (Section 4.3; **Supplementary Table S4**).

Next, we leveraged these parameters to obtain predictors of disease risk (at the liability scale; Section 4.4) for the rest 90% of the UK Biobank participants (**Supplementary Table S2 and Supplementary Figure S1**). We evaluated the discriminative power of these joint predictors in identifying individuals at an elevated level of disease risk by AUROC and AUPRC. We tested whether the joint predictors could more accurately quantify the genetic risk than the polygenic risk score alone through comparing their AUROC by DeLong's test[47]. We evaluated whether the risk stratification of the population could be improved by the joint predictors compared to the polygenic risk score by calculating NRI and IDI.

Last, we compared our joint predictors with a gold-standard numerical solution by jointly modelling the polygenic risk score and the parental disease history in a multivariate logistic regression model based on the training dataset (**Supplementary Figure S1**), and evaluating its discriminative power on the test dataset. To assess the robustness of this data-driven approach when individual-level data were insufficient, multivariate logistic regression models were fitted on randomly sampled subsets of the training dataset of smaller sample sizes (10%, 20%, and 50% of the original training dataset, each having 100 replicates).

Predictive performance of these models was also evaluated on the test dataset.

4.6.4 Identifying sources of under-captured genetic risk

We attempted to identify sources of the under-captured genetic component for Alzheimer's disease and COPD, for which the joint predictors demonstrated the most improvements in risk stratification.

For Alzheimer's disease, we determined the *APOE* genotype for each individual in the test dataset based on genotyping data of two SNPs: rs429358 and rs7412. We tested whether the polygenic risk score and the joint predictor were associated with carrying at least one e4 allele or carrying two e4 alleles, respectively. For COPD, we tested whether the polygenic risk score and the joint predictor were associated with self-reported smoking status (ever-smokers vs. never-smokers), respectively. All association tests were based on logistic regression, adjusted for the effects of age, sex, genotyping array, recruitment centre, and the first 10 genetic principal components.

Data availability

Individual genotype and pheontype data from the UK Biobank (https://www.ukbiobank.ac.uk/) and the ALSPAC (http://www.bristol.ac.uk/alspac/) are available through successful applications to the research committees. The ALSPAC website contains details of all the data that is available through a fully searchable data dictionary and variable search tool (http://www.bristol.ac.uk/alspac/researchers/our-data/). A computational toolkit implementing the latent factor model developed in this study is available at https://github.com/tianyuan-lu/PRS-FH-Prediction.

Disclosures

J.B.R. is the founder of 5 Prime Sciences, and has served as a consultant to GlaxoSmithKline and Deerfield Capital for their genetics programs. The other authors have no relevant disclosures.

Funding

C.M.T.G. is supported by a Canadian Institutes of Health Research grant (CIHR; PJT-148620). The Richards research group is supported by the Canadian Institutes of Health Research (365825; 409511),

the Lady Davis Institute of the Jewish General Hospital, the Canadian Foundation for Innovation, the NIH Foundation, Cancer Research UK, Genome Québec, the Public Health Agency of Canada and the Fonds de Recherche Québec Santé (FRQS). J.B.R. is supported by a FRQS Clinical Research Scholarship Merite. T.L. has been supported by a Vanier Canada Graduate Scholarship, an FRQS Doctoral Training Fellowship and a McGill University Faculty of Medicine Scholarship. The UK Medical Research Council and Wellcome (217065/Z/19/Z) and the University of Bristol provide core support for ALSPAC. A comprehensive list of grants funding is available on the ALSPAC website (http://www.bristol.ac.uk/alspac/external/documents/grant-acknowledgements.pdf). This research was specifically funded by Wellcome Trust and the Medical Research Council (076467/Z/05/Z).

Acknowledgements

This research has been conducted using the UK Biobank resource under Application Number 27449 and 60755, and the ALSPAC resource under Application Number B3359. ALSPAC children were genotyped using the Illumina HumanHap550 quad chip genotyping platforms by Sample Logistics and Genotyping Facilities at Wellcome Sanger Institute and LabCorp (Laboratory Corporation of America), using support from 23andMe. We thank all the families who took part in the ALSPAC study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists, and nurses. This study was enabled in part by support provided by Calcul Québec and Compute Canada. T.L. thanks Wenmin Zhang for helpful discussion about the latent factor models and prediction schemes.

References

1. Lambert, S. A., Abraham, G. & Inouye, M. Towards clinical utility of polygenic risk scores. *Hum. molecular genetics* **28**, R133–R142 (2019).

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

3. Visscher, P. M. *et al.* 10 years of gwas discovery: biology, function, and translation. *The Am. J. Hum. Genet.* **101**, 5–22 (2017).

4. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. genetics* **50**, 1219–1224 (2018).

5. 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. medicine* **26**, 549–557 (2020).

6. Inouye, M. *et al.* Genomic risk prediction of coronary artery disease in 480,000 adults: implications for primary prevention. *J. Am. Coll. Cardiol.* **72**, 1883–1893 (2018).

7. Lu, T. *et al.* A polygenic risk score to predict future adult short stature among children. *The J. Clin. Endocrinol. & Metab.* **106**, 1918–1928 (2021).

8. Lu, T. *et al.* Improved prediction of fracture risk leveraging a genome-wide polygenic risk score. *Genome medicine* **13**, 1–15 (2021).

9. Mavaddat, N. *et al.* Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *The Am. J. Hum. Genet.* **104**, 21–34 (2019).

10. Lu, T. *et al.* Individuals with common diseases but with a low polygenic risk score could be prioritized for rare variant screening. *Genet. Medicine* **23**, 508–515 (2021).

11. Oram, R. A. *et al.* A type 1 diabetes genetic risk score can aid discrimination between type 1 and type 2 diabetes in young adults. *Diabetes care* **39**, 337–344 (2016).

12. Turnwald, B. P. *et al.* Learning one's genetic risk changes physiology independent of actual genetic risk. *Nat. human behaviour* **3**, 48–56 (2019).

13. Kullo, I. J. *et al.* Incorporating a genetic risk score into coronary heart disease risk estimates: effect on low-density lipoprotein cholesterol levels (the mi-genes clinical trial). *Circulation* **133**, 1181–1188 (2016).

14. Canela-Xandri, O., Rawlik, K. & Tenesa, A. An atlas of genetic associations in uk biobank. *Nat. genetics* **50**, 1593–1599 (2018).

15. Povysil, G. *et al.* Rare-variant collapsing analyses for complex traits: guidelines and applications. *Nat. Rev. Genet.* **20**, 747–759 (2019).

16. Zhu, Z. *et al.* Dominance genetic variation contributes little to the missing heritability for human complex traits. *The Am. J. Hum. Genet.* **96**, 377–385 (2015).

17. Hujoel, M. L., 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. genetics* **52**, 541–547 (2020).

18. Weigl, K. *et al.* Strongly enhanced colorectal cancer risk stratification by combining family history and genetic risk score. *Clin. epidemiology* **10**, 143 (2018).

19. Hindieh, W. *et al.* Association between family history, a genetic risk score, and severity of coronary artery disease in patients with premature acute coronary syndromes. *Arter. thrombosis, vascular biology* **36**, 1286–1292 (2016).

20. Chen, H. *et al.* Adding genetic risk score to family history identifies twice as many high-risk men for prostate cancer: results from the prostate cancer prevention trial. *The Prostate* **76**, 1120–1129 (2016).

21. Li, H. *et al.* Breast cancer risk prediction using a polygenic risk score in the familial setting: a prospective study from the breast cancer family registry and kconfab. *Genet. Medicine* **19**, 30–35 (2017).

22. Shieh, Y. *et al.* Breast cancer risk prediction using a clinical risk model and polygenic risk score. *Breast cancer research treatment* **159**, 513–525 (2016).

23. Boyd, A. *et al.* Cohort profile: the 'children of the 90s'—the index offspring of the avon longitudinal study of parents and children. *Int. journal epidemiology* **42**, 111–127 (2013).

24. Fraser, A. *et al.* Cohort profile: the avon longitudinal study of parents and children: Alspac mothers cohort. *Int. journal epidemiology* **42**, 97–110 (2013).

25. Bycroft, C. *et al.* The uk biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

26. Wood, A. R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. genetics* **46**, 1173–1186 (2014).

27. Aulchenko, Y. S. *et al.* Predicting human height by victorian and genomic methods. *Eur. J. Hum. Genet.* **17**, 1070–1075 (2009).

28. Lambert, S. A. *et al.* The polygenic score catalog as an open database for reproducibility and systematic evaluation. *Nat. Genet.* **53**, 420–425 (2021).

29. Green, R. C. *et al.* Disclosure of apoe genotype for risk of alzheimer's disease. *New Engl. J. Medicine* **361**, 245–254 (2009).

30. Verghese, P. B., Castellano, J. M. & Holtzman, D. M. Apolipoprotein e in alzheimer's disease and other neurological disorders. *The Lancet Neurol.* **10**, 241–252 (2011).

31. Stocker, H., Möllers, T., Perna, L. & Brenner, H. The genetic risk of alzheimer's disease beyond apoe $\varepsilon 4$: systematic review of alzheimer's genetic risk scores. *Transl. psychiatry* **8**, 1–9 (2018).

32. Yengo, L. *et al.* Meta-analysis of genome-wide association studies for height and body mass index in 700000 individuals of european ancestry. *Hum. molecular genetics* **27**, 3641–3649 (2018).

33. Yang, J. *et al.* Common snps explain a large proportion of the heritability for human height. *Nat. genetics* **42**, 565–569 (2010).

34. Yang, J. *et al.* Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat. genetics* **47**, 1114–1120 (2015).

35. Yang, J., Zeng, J., Goddard, M. E., Wray, N. R. & Visscher, P. M. Concepts, estimation and interpretation of snp-based heritability. *Nat. genetics* **49**, 1304–1310 (2017).

36. Fletcher, C. & Peto, R. The natural history of chronic airflow obstruction. *Br Med J* **1**, 1645–1648 (1977).

37. Liu, M. *et al.* Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nat. genetics* **51**, 237–244 (2019).

38. Erzurumluoglu, A. M. *et al.* Meta-analysis of up to 622,409 individuals identifies 40 novel smoking behaviour associated genetic loci. *Mol. psychiatry* **25**, 2392–2409 (2020).

39. Martin, A. R. *et al.* Human demographic history impacts genetic risk prediction across diverse populations. *The Am. J. Hum. Genet.* **100**, 635–649 (2017).

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

41. Fry, A. *et al.* Comparison of sociodemographic and health-related characteristics of uk biobank participants with those of the general population. *Am. journal epidemiology* **186**, 1026–1034 (2017).

42. Jacobs, D. *et al.* Age at onset of alzheimer's disease: relation to pattern of cognitive dysfunction and rate of decline. *Neurology* **44**, 1215–1215 (1994).

43. Pagano, G., Ferrara, N., Brooks, D. J. & Pavese, N. Age at onset and parkinson disease phenotype. *Neurology* **86**, 1400–1407 (2016).

44. Gillett, A. C., Vassos, E. & Lewis, C. M. Transforming summary statistics from logistic regression to the liability scale: application to genetic and environmental risk scores. *Hum. heredity* **83**, 210–224 (2018).

45. Delaneau, O. & Marchini, J. Integrating sequence and array data to create an improved 1000 genomes project haplotype reference panel. *Nat. communications* **5**, 1–9 (2014).

46. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. genetics* **48**, 1279 (2016).

47. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 837–845 (1988).

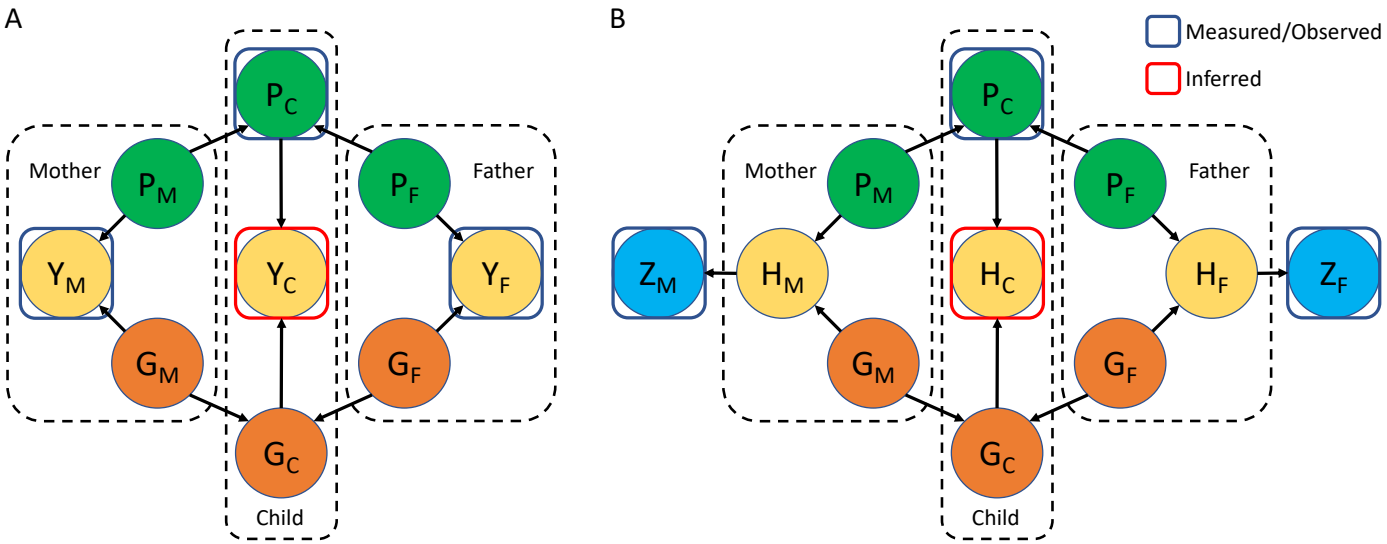

Figure 1. Causal diagrams representing latent factor models. (A) For the parents and the children separately, a continuous trait *Y* is determined by a polygenic component *P* captured by a polygenic risk score, and a latent genetic component *G* independent of *P*. (B) A binary trait *Z* is determined by the underlying genetic liability *H*. Analogous to *Y* in (A), *H* is jointly determined by *P* and *G*. Prediction combining a polygenic risk score and parental information is equivalent to inference of the distribution of *Y* in (A) and *H* in (B) for the children. The two genetic components, *P* and *G*, are assumed to be independently passed on from the parents to the offspring. Trait heritability is assumed to be constant across different generations.

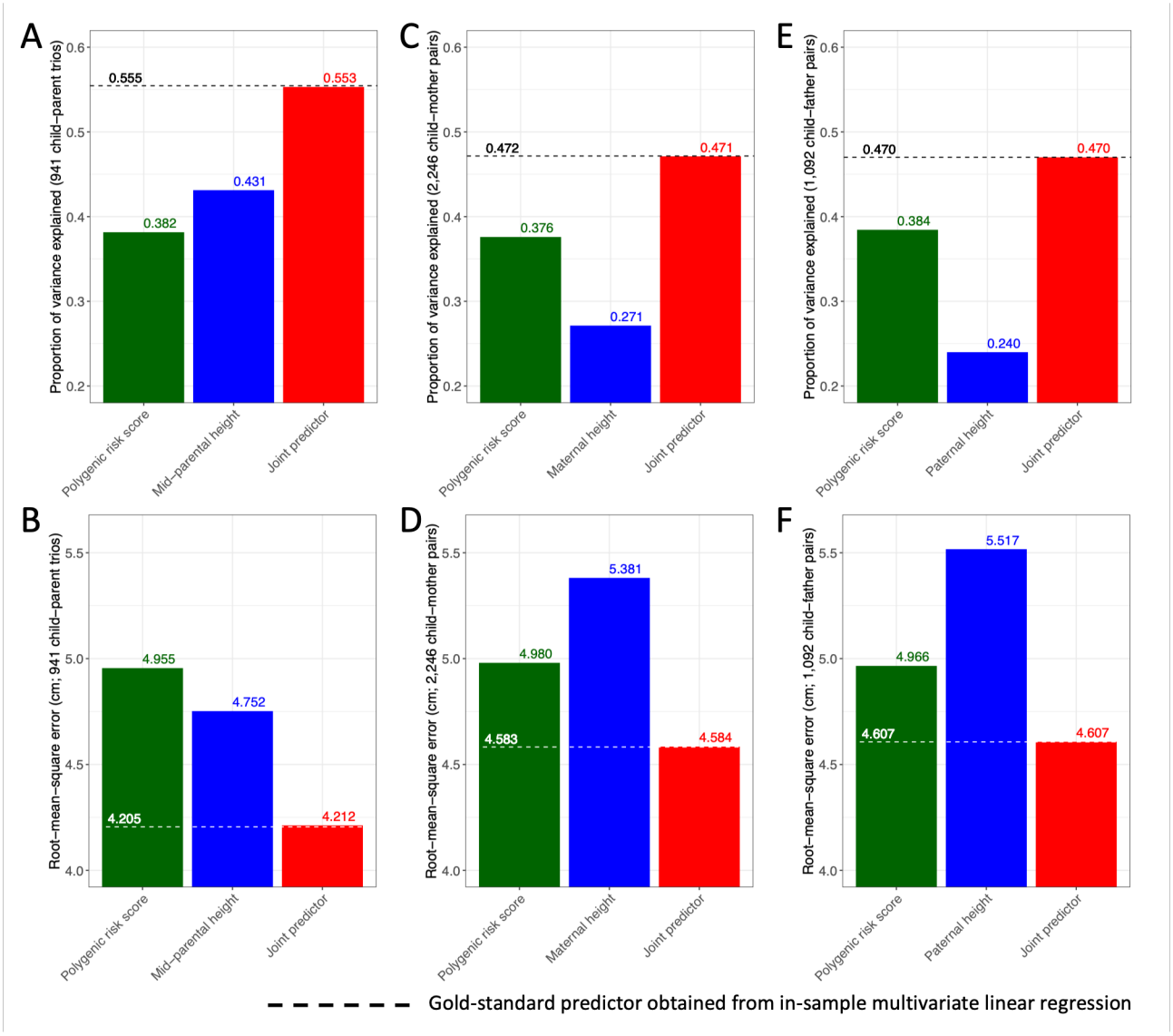

Figure 2. Comparison of predictive performance in predicting children's sex-adjusted adult height z-scores for a polygenic risk score, parental height measures, and a joint predictor. A joint predictor combining the polygenic risk score and the mid-parental height predictor achieves (A) the highest proportion of variance explained, and (B) the lowest root-mean-square error based on 941 child-parent trios. A joint predictor incorporating (C) and (D) only the maternal height for 2,246 child-mother pairs, or (E) and (F) only the paternal height for 1,092 child-father pairs also outperforms the polygenic risk score and the single parental height predictor alone. These joint predictors have similar predictive performance as the corresponding gold-standard in-sample best linear unbiased predictors, indicated by dashed lines.

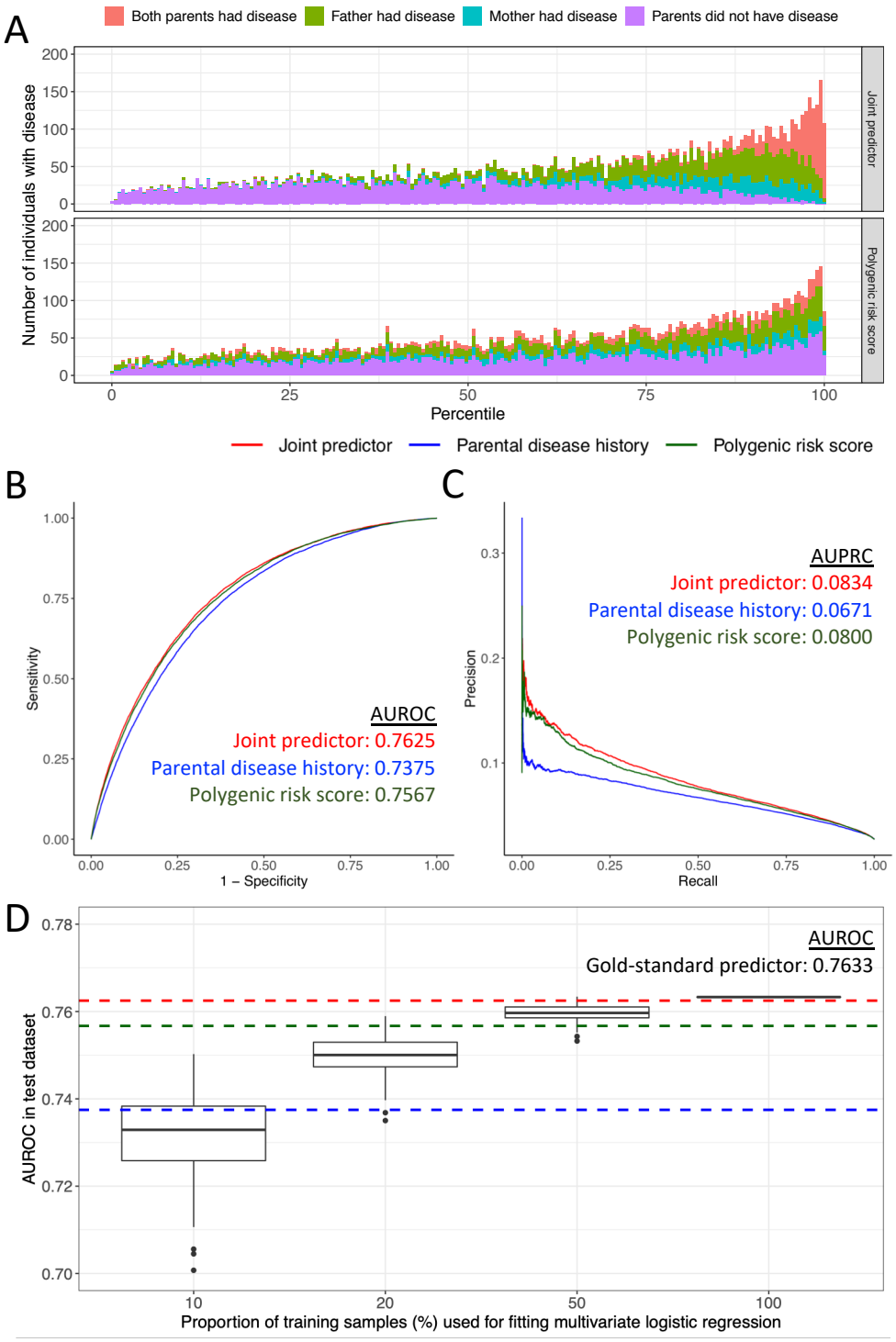

Figure 3. A joint predictor improves risk stratification for myocardial infarction in the UK Biobank. (A) Incorporation of parental disease history leads to calibration of polygenic risk scores. Individuals with a parental disease history are more likely to be considered at risk. The joint predictor achieves higher (B) area under the receiver operating characteristic curve (AUROC) and (C) are under the precision-recall curve (AUPRC), compared to the polygenic risk score and the parental disease history. (D) Performance of a data-driven joint predictor obtained from multivariate logistic regression is sensitive to sample size when using individual-level data. AUROC of the joint predictor, polygenic risk score, and parental disease history are indicated by red, green, and blue dashed lines, respectively.

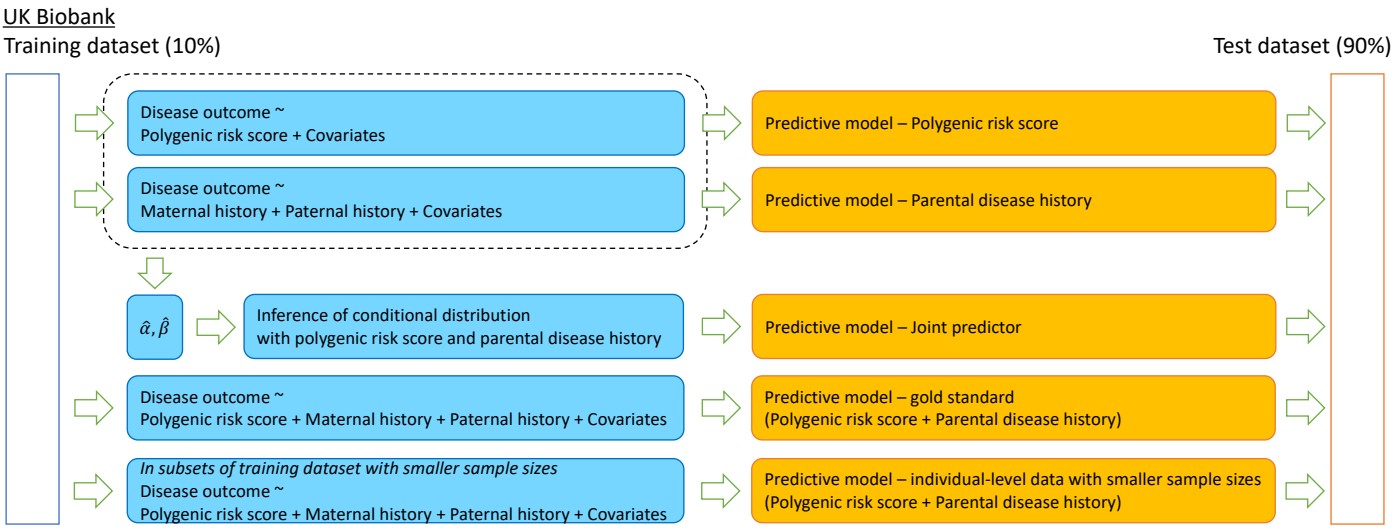

Figure S1. Development and evaluation of predictors in the UK Biobank. The UK Biobank was randomly split into a training dataset comprising 10% of the samples and a test dataset comprising 90% of the samples. For each disease, on the test dataset, the magnitude of association between the polygenic risk score as well as the parental disease history and the disease risk was estimated. These estimates were utilized to obtain a joint predictor with the latent factor model. A multivariate logistic regression model including the polygenic risk score and the parental disease history was fitted to obtain the gold-standard predictor. Furthermore, 10%, 20%, or 50% of the training dataset was randomly selected and used to fit multivariate logistic regression models, each repeating 100 times. These generate data-driven predictors based on individual-level data of smaller sample sizes. All predictors were tested on the test dataset.

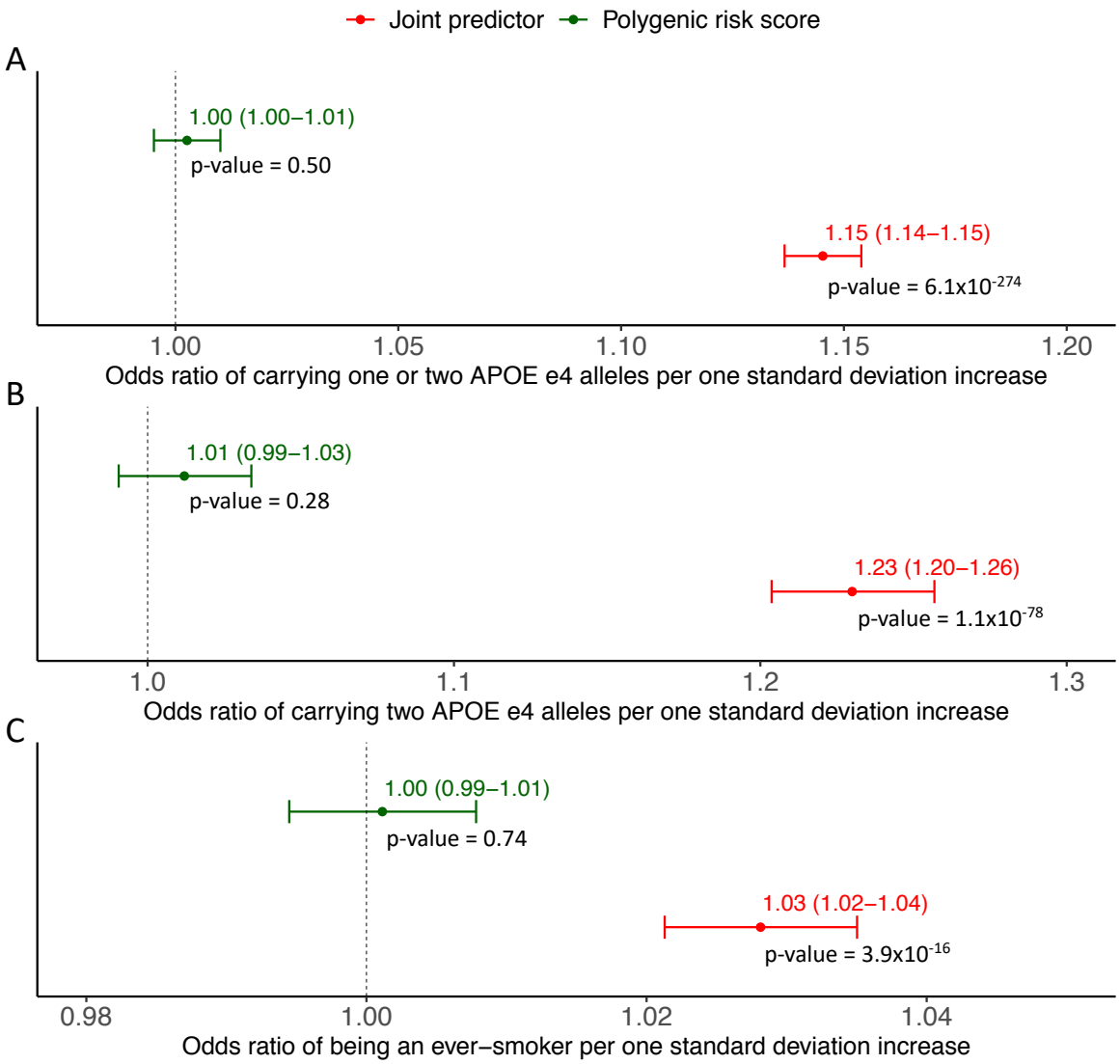

Figure S2. Association between genetic risk predictors and known risk factors. (A) Associations between genetic risk predictors for Alzheimer's disease and carrying at least one *APOE* e4 allele. (B) Associations between genetic risk predictors for Alzheimer's disease and carrying two *APOE* e4 alleles. In the UK Biobank test dataset for Alzheimer's disease, there were 8,616 e4/e4 homozygotes, 85,613 e3/e4 heterozygotes, 209,466 e3/e3 homozygotes, 44,293 e2/e3 heterozygotes, and 2,280 e2/e2 homozygotes. In addition, 9,113 individuals had ambiguous genotypes due to heterozygosity at both the rs429358 and rs7412 SNPs, and were not included in these association tests. (C) Associations between genetic risk predictors for COPD and being ever-smokers. All association tests were based on individuals in the test dataset, adjusted for the effects of age, sex, genotyping array, recruitment center, and the first 10 genetic principal components.

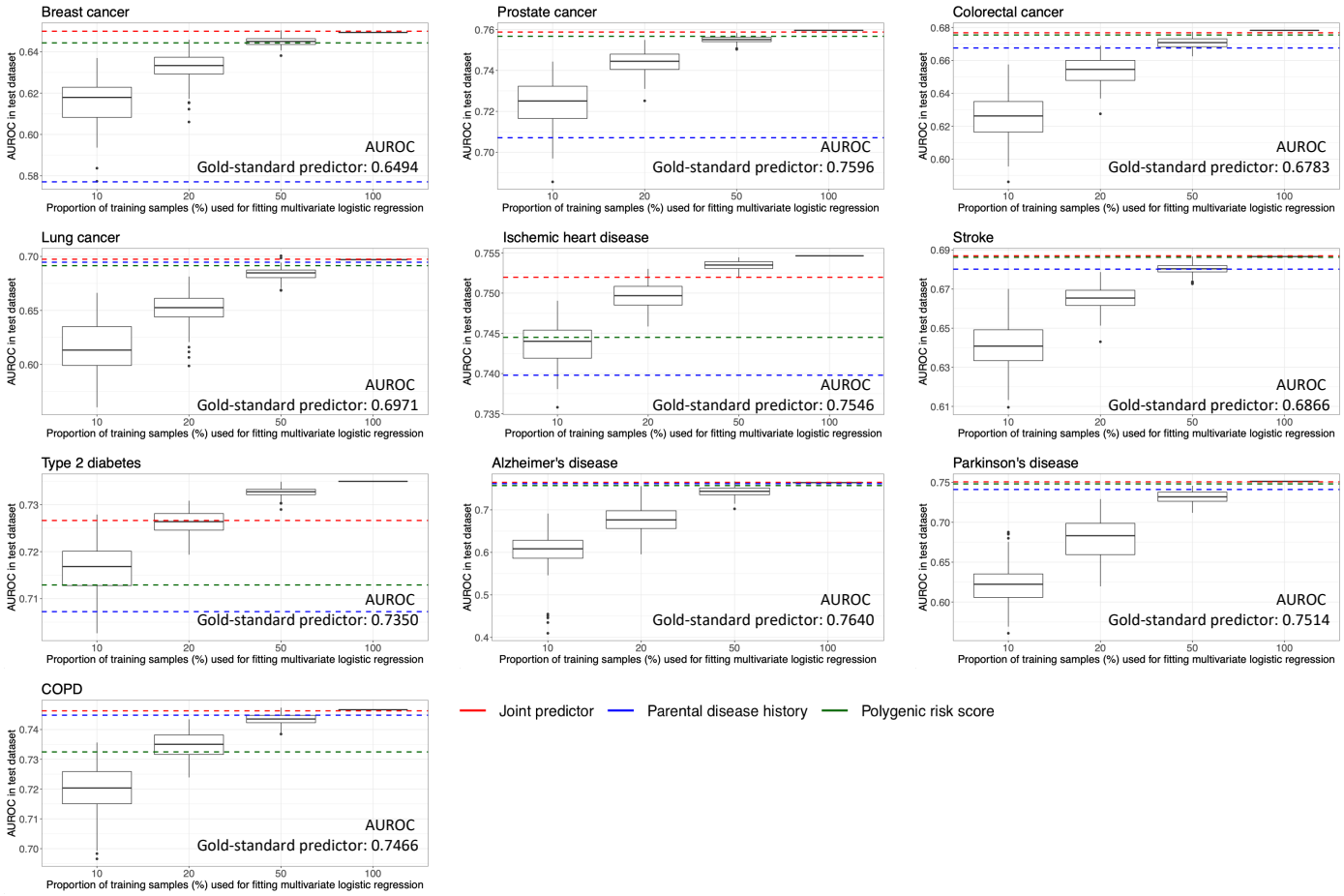

Figure S3. Performance of data-driven joint predictors obtained from multivariate logistic regression for 10 complex diseases. AUROC of the corresponding joint predictor, polygenic risk score, and parental disease history are summarized in Table 1, and are indicated by red, green, and blue dashed lines, respectively. Apart from ischemic heart disease and type 2 diabetes, the joint predictors obtained from the latent factor model have comparable discriminative power to the gold-standard predictors obtained from multivariate logistic regression based on all samples in the training dataset, and the joint predictors outperform other data-driven predictors based on subsets of the training dataset of smaller sample sizes.