Aalborg University

DENMARK

Multi-Trait Genomic Risk Stratification for Type 2 Diabetes

Rohde, Palle Duun; Nyegaard, Mette; Kjølby, Mads Fuglsang; Sørensen, Peter

Link to publication from Aalborg University

Multi-Trait Genomic Risk Stratification for Type 2 Diabetes

Palle Duun Rohde[1,2]*, Mette Nyegaard[2,3], Mads Kjolby[3,4,5,6] and Peter Sørensen[7]

[1] *Department of Chemistry and Bioscience, Aalborg University, Aalborg, Denmark,* [2] *Department of Health Science and Technology, Aalborg University, Aalborg, Denmark,* [3] *Department of Biomedicine, Aarhus University, Aarhus, Denmark,* [4] *Department of Population Health and Genomics, University of Dundee, Dundee, United Kingdom,* [5] *Department of Clinical Pharmacology, Aarhus University Hospital, Aarhus, Denmark,* [6] *Steno Diabetes Center Aarhus, Aarhus University Hospital, Aarhus, Denmark,* [7] *Centre for Quantitative Genetics and Genomics, Aarhus University, Aarhus, Denmark*

Type 2 diabetes mellitus (T2DM) is continuously rising with more disease cases every year. T2DM is a chronic disease with many severe comorbidities and therefore remains a burden for the patient and the society. Disease prevention, early diagnosis, and stratified treatment are important elements in slowing down the increase in diabetes prevalence. T2DM has a substantial genetic component with an estimated heritability of 40–70%, and more than 500 genetic loci have been associated with T2DM. Because of the intrinsic genetic basis of T2DM, one tool for risk assessment is genome-wide genetic risk scores (GRS). Current GRS only account for a small proportion of the T2DM risk; thus, better methods are warranted for more accurate risk assessment. T2DM is correlated with several other diseases and complex traits, and incorporating this information by adjusting effect size of the included markers could improve risk prediction. The aim of this study was to develop multi-trait (MT)-GRS leveraging correlated information. We used phenotype and genotype information from the UK Biobank, and summary statistics from two independent T2DM studies. Marker effects for T2DM and seven correlated traits, namely, height, body mass index, pulse rate, diastolic and systolic blood pressure, smoking status, and information on current medication use, were estimated (i.e., by logistic and linear regression) within the UK Biobank. These summary statistics, together with the two independent training summary statistics, were incorporated into the MT-GRS prediction in different combinations. The prediction accuracy of the MT-GRS was improved by 12.5% compared to the single-trait GRS. Testing the MT-GRS strategy in two independent T2DM studies resulted in an elevated accuracy by 50–94%. Finally, combining the seven information traits with the two independent T2DM studies further increased the prediction accuracy by 34%. Across comparisons, body mass index and current medication use were the two traits that displayed the largest weights in construction of the MT-GRS. These results explicitly demonstrate the added benefit of leveraging correlated information when constructing genetic scores. In conclusion, constructing GRS not only based on the disease itself but incorporating genomic information from other correlated traits as well is strongly advisable for obtaining improved individual risk stratification.

Keywords: UK Biobank, genetic risk scores, GRS, multi-trait analysis, precision medicine

INTRODUCTION

Type 2 diabetes mellitus (T2DM) is a chronic disease with severe comorbidities, such as myocardial infarction, loss of kidney function, blindness, and risk of amputations (1). Globally, the prevalence of T2DM is expected to increase exponentially in developing countries (2, 3), and it is a disease that places a severe economic burden on health systems. Accurate disease risk assessment is important for early disease diagnosis for initiating lifestyle changes early in the disease progression or prompt the clinician to treat high-risk patients more aggressively, which is expected to slow down disease progression, reduce disease symptoms, and prevent severe morbidity and mortality. Thus, methods for accurate disease risk assessment are absolutely critical for reducing morbidity and mortality.

Studies have unambiguously shown that T2DM is a complex, multifactorial disease, where an individual's risk of developing the disease is influenced by a combination of genetic variation at multiple sites across the genome acting in concert with environmental factors (4–6). The heritability of T2DM has been estimated to be 40–70% (7, 8), and more than 500 distinct genetic loci have been implicated with T2DM risk (6, 9–12). As T2DM is greatly impacted by genetics, genomic information has the potential to not only aid with early disease diagnosis but importantly also to stratify patients across disease subtypes (13) to initiate treatment intervention and lifestyle changes early in the disease progression.

During the last decade, an enormous effort has been in method development and construction of disease risk scores based on genomic information (14–17). However, until recently, these genome-wide genetic risk scores (GRS) have mainly been constructed using a single-trait approach. Because much of the variation within the human genome contributes to a large number of different complex traits and diseases (18), the accuracy of risk stratification can be improved by developing multi-trait (MT)-GRS accounting for the genetic correlation *among* traits. Using correlated information to construct GRS has theoretically—and to a minor extend empirical—been shown to increase the accuracy of disease risk prediction (6–8). T2DM is strongly correlated with a range of complex diseases and traits, such as overweight (19), cardiovascular disease (1, 19–21), hypertension (19, 22), and chronic kidney disease (19, 23); hence, T2DM is an excellent case for developing accurate GRS by leveraging correlated information.

The objective of the current study was to investigate the predictive performance of a MT-GRS model that combines marker effects from genome-wide association studies (GWAS) of T2DM and a number of correlated traits. The types of information included in this study were body mass index (BMI), height, smoking status, pulse rate, diastolic and systolic blood pressure, and a quantity of current medication use, as the total count of different prescription and over-the-counter medications is a proxy for general health and disease status. The aim of the present study was to investigate whether a MT-GRS model based on loci for multiple correlated traits had increased predictive discriminative power compared with a traditional single-trait (ST)-GRS model. This strategy was first applied within the UK Biobank (UKB) (24), and then extended to include information on two UKB-independent GWAS summary statistics and, finally, a combined model incorporating information from the UKB and the two independent T2DM GWAS data sets.

MATERIALS AND METHODS

Phenotype and Genotype Data

Only unrelated British Caucasian individuals from the UKB (24) ($n = 335,652$ subjects) were used in the current study (excluding individuals with more than 5,000 missing genotype values or if having chromosomal aneuploidy). T2DM status was determined based on in-hospital records (by ICD-10 E.11, UKB data field 41270, which contains both main and secondary diagnoses) and self-reported disease state (UKB data field 20002) counting a total of 18,809 individuals. Seven additional phenotypes were also included: standing height, BMI, diastolic and systolic blood pressure, pulse rate, smoking status, and current medication use (measured as the number of different prescription and over-the-counter medications taken). These phenotypes were all adjusted for sex, age, UKB assessment center, and the first 10 genetic principal components (to account for any cryptic relatedness that were not accounted for by restricting to unrelated Caucasian British individuals), following inverse rank normalization to approximate normality.

Genotyped variants with minor allele frequency <0.01, genotype missingness >5%, or variants within the major histocompatibility complex were excluded from the analyses, resulting in a total of 599,297 genetic variants.

Prediction of Diabetes Risk

T2DM risk was determined using GRS based on either summary statistics obtained within the UKB cohort and other T2DM-related GWAS studies (**Table 1**). The overall workflow is depicted in **Figure 1** and is described in detail below.

UKB Summary Statistics

The White-British UKB cohort of unrelated individuals (335,652 subjects) was split into 10 folds with no overlap of samples within each fold, and for each fold, the marker effects for T2DM, standing height, BMI, diastolic and systolic blood pressure, pulse rate, smoking status, and current medication use, were estimated using logistic or linear regression as implemented in PLINK2 (26). In all analyses, the same set of covariates were included as those used during phenotypic adjustment as this has been shown to increase statistical power (27).
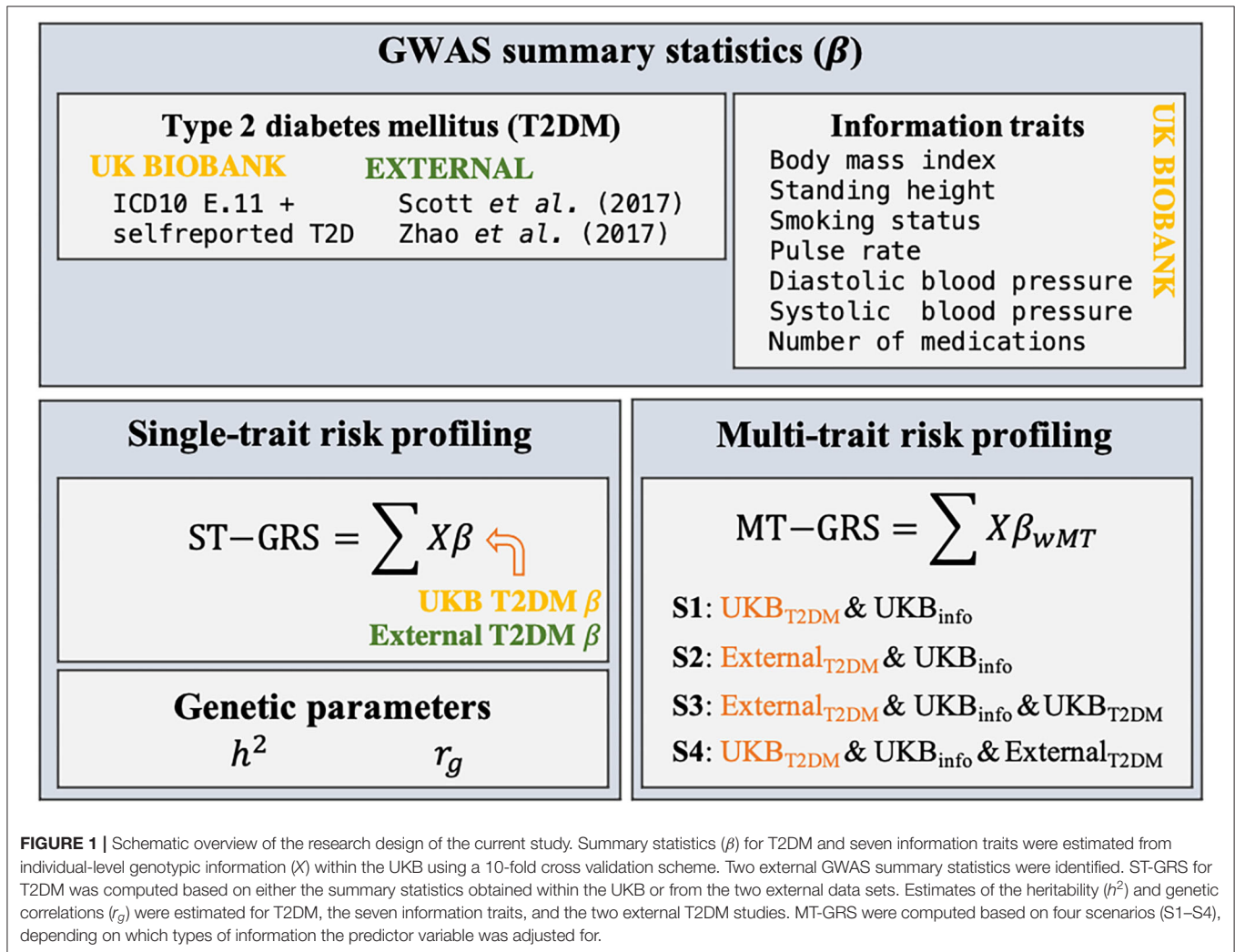
Publicly Available Type 2 Diabetes Summary Statistics

Two recently published GWAS for T2DM were identified (**Table 1**). Common for the studies were that they did not include UKB data, and therefore provide an independent training set. The regression coefficients were flipped such that the marker effect of the effect allele matched the effect allele within the UKB data.

Study	References	n_{total}	n_{case}	m_{total}	m_{UKB}
Scott et al. (2017)	(10)	159,208	26,676	12,056,346	595,528
Zhao et al. (2017)	(25)	265,678	73,337	8,796,184	558,105

n_{total} is the sample size of the listed study. n_{case} is the number of T2DM cases within the listed study. m_{total} is the number of genetic variants used in the GWAS of the listed study. m_{UKB} is the number of variants in the listed study that were among the 599,297 quality-controlled genotyped variants in the UKB.

FIGURE 1 | Schematic overview of the research design of the current study. Summary statistics (β) for T2DM and seven information traits were estimated from individual-level genotypic information (X) within the UKB using a 10-fold cross validation scheme. Two external GWAS summary statistics were identified. ST-GRS for T2DM was computed based on either the summary statistics obtained within the UKB or from the two external data sets. Estimates of the heritability (h^2) and genetic correlations (r_g) were estimated for T2DM, the seven information traits, and the two external T2DM studies. MT-GRS were computed based on four scenarios (S1–S4), depending on which types of information the predictor variable was adjusted for.

Estimation of Genetic Parameters

Linkage disequilibrium (LD) between the genotyped variants was estimated as the squared Pearson's correlation coefficient (r^2) between two genetic variants adjusted for sample size (N) as the standard estimator of the Pearson's correlation coefficient has an upward bias (28). The adjusted squared Pearson's correlation coefficient ($\tilde{r}^2$) is obtained as (28):

$$\tilde{r}^2 = r^2 - \frac{1 - r^2}{N - 2}, \qquad (1)$$

which was computed with the R package qgg (29). LD scores (l) for all variants within a window size of 5,000 markers (2,500 markers around the i-th variant) were computed as

$$l_i = \sum_{k=1}^{m=5000} \tilde{r}^2_{i,k}. \qquad (2)$$

The MT-GRS model relies on selection index theory to obtain marker weights that require estimates of genetic parameters (30). The heritability (h^2) and the genetic correlation (r_g) between traits can be computed based on GWAS summary statistics using LD score regression (28). The heritability was estimated as the regression of the summary statistics on the LD score:

TABLE 2 | UKB cohort description ($n = 335,652$) of T2DM cases and controls (count (%) or mean ± standard deviation).

Characteristics	Controls	T2DM cases	Information trait
N	316,935	18,809	
Age (years)	56.4 ± 8.0	60.5 ± 6.7	
Sex, male	144,070 (45.5%)	11,693 (62.2%)	
BMI (kg/m^2)	27.1 ± 4.5	31.9 ± 5.8	X
Height (cm)	168.8 ± 9.2	170.0 ± 9.3	X
Pulse rate (BPM)	69.1 ± 11.1	73.6 ± 13.1	X
Systolic blood pressure (mmHg)	138.0 ± 18.6	142.6 ± 18.0	X
Diastolic blood pressure (mmHg)	82.3 ± 10.1	82.3 ± 10.3	X
Smoking status			X
Never	175,002 (55.4%)	7,687 (41.1%)	
Former	109,007 (34.5%)	8,663 (46.3%)	
Current	31,867 (10.1%)	2,345 (12.6%)	
Number of medications	2.3 ± 2.4	5.7 ± 3.7	X

$$\hat{h}^2 = \left(\mathbf{Z}'\mathbf{Z}\right)^{-1} \left(\mathbf{Z}'\mathbf{y}\right), \tag{3}$$

where $\mathbf{Z} = \mathrm{n}_{eff} \times l/m$, with l being the LD score (see Equation 2), m is the number of genetic variants, and n_{eff} is the effective number of individuals and is $\mathrm{n}_{eff} = median\left(1/2 \times af \times (1-af) \times SE\left(\hat{b}\right)^2\right)$, where af is the allele frequency, and $SE\left(\hat{b}\right)$ is the estimated standard error of the marker regression estimate. The response variable is $\mathbf{y} = \left(\frac{\hat{b}}{SE(\hat{b})}\right)^2$, where $\hat{b}$ is the estimated regression coefficient for the genetic variants [for binary traits, the odds ratios (ORs) were converted to $\hat{b} = \log(OR)$, and $SE\left(\hat{b}\right) = \left|\hat{b}/P(X < (1-p)/2)\right|$, where $P(X < (1-p)/2)$ is the normal cumulative distribution given the marker P-value, p (31)]. Similarly, the genetic correlation between traits 1 and 2 can be estimated as:

$$\hat{r}_g = \frac{\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\left(\mathbf{Z}'\mathbf{y}\right)}{\sqrt{\hat{h}_1^2}\sqrt{\hat{h}_2^2}}, \tag{4}$$

where $\mathbf{Z} = \sqrt{n_1}\sqrt{n_2} \times \frac{l}{m}$, and $\mathbf{y} = \frac{\hat{b}_1}{SE(\hat{b}_1)} \times \frac{\hat{b}_2}{SE(\hat{b}_2)}$. LD score regression was implemented in the R package qgg (29) and was computed for each of the 10-folds of random data subdivisions for T2DM and the seven information traits (**Table 2**), and among the information traits and the publicly available T2DM summary statistics (**Table 1**).

ST-GRS
The ST-GRS was computed as,

$$\mathrm{ST} - \mathrm{GRS} = \sum_{i=1}^{m} \mathbf{X}_i \hat{b}_i, \tag{5}$$

where $\mathbf{X}_i$ is the i-th column of the genotype matrix containing allelic counts, $\hat{b}_i$ is the estimated marker effect for the i-th marker, and m is the number of variants left after LD pruning ($r^2 < 0.1$, <0.5, or <0.9) and P-value thresholding ($P < 0.001, 0.01, 0.05, 0.1, 0.2, 0.3, 0.5, 0.7, 0.9$, and 0.99). The genetic scoring was performed with the R package qgg (29).

MT-GRS
The accuracy of GRS can be improved by leveraging information from correlated traits by adjusting the marker effects ($\hat{\mathbf{b}}$) (30). The adjustment of the marker effects for the focal trait (f, i.e., T2DM) is obtained by computing index weights for each marker ($\mathbf{w}_i'$)

$$\hat{\mathbf{b}}_{wMT_i} = \mathbf{w}_i' \hat{\mathbf{b}}_i. \tag{6}$$

From quantitative genetic theory, selection indices have been developed for MT selection, in which many ST individual genetic effects (i.e., breeding values) are combined with an index weight allowing selection of the individuals with the best MT phenotype (32, 33). The optimal weights can be derived as $\mathbf{w} = \mathbf{V}^{-1}\mathbf{C}$, where $\mathbf{C}$ is a $k \times 1$ column vector of covariances between the $\hat{\mathbf{b}}$ values of the k traits and the true marker effects of the focal trait ($\mathbf{b}_f$), and $\mathbf{V}$ is a $k \times k$ variance–covariance **matrix** of the $\hat{\mathbf{b}}$ values:

$$\mathbf{w} = \begin{bmatrix} var(\hat{\mathbf{b}}_1) & \dots & cov(\hat{\mathbf{b}}_1, \hat{\mathbf{b}}_k) \\ \vdots & \ddots & \vdots \\ cov(\hat{\mathbf{b}}_k, \hat{\mathbf{b}}_1) & \dots & var(\hat{\mathbf{b}}_k) \end{bmatrix}^{-1} \begin{bmatrix} cov(\mathbf{b}_f, \hat{\mathbf{b}}_1) \\ \dots \\ cov(\mathbf{b}_f, \hat{\mathbf{b}}_k) \end{bmatrix}. \tag{7}$$

The diagonal elements of variance–covariance matrix, $\mathbf{V}$, are

$$var\left(\hat{\mathbf{b}}_k\right) = \frac{h_k^2}{M} + \frac{1}{N_k}, \tag{8}$$

where M is the effective number of chromosomal segments [here $M = 60,000$ (30, 34)] and N_k is the number of observations for trait k. The off-diagonal elements of $\mathbf{V}$ for trait k and l are

$$cov\left(\hat{\mathbf{b}}_k, \hat{\mathbf{b}}_l\right) = \frac{r_g h_k h_l}{M}, \tag{9}$$

which is the same for the elements of $\mathbf{C}$. Combining Equations (8) and (9), Equation (7) becomes

$$\mathbf{w} = \begin{bmatrix} \frac{h_1^2}{M} + \frac{1}{N_1} & \dots & \frac{r_g h_1 h_k}{M} \\ \vdots & \ddots & \vdots \\ \frac{r_g h_k h_1}{M} & \dots & \frac{h_k^2}{M} + \frac{1}{N_k} \end{bmatrix}^{-1} \begin{bmatrix} \frac{h_1^2}{M} \\ \dots \\ \frac{r_g h_1 h_k}{M} \end{bmatrix}. \tag{10}$$

The MT-GRS is then obtained as the sum of adjusted marker effects,

$$\mathrm{MT} - \mathrm{GRS} = \sum_{i=1}^{m} \mathbf{X}_i \hat{\mathbf{b}}_{wMT_i}. \tag{11}$$

MT-GRS was computed by applying LD pruning ($r^2 < 0.1$, <0.5, or <0.9) and P-value thresholding ($P < 0.001, 0.01, 0.05, 0.1,$

0.2, 0.5, 0.75, and 0.99) based on UKB genotypes and T2DM summary statistics; thus, the same LD pruning and *P*-value thresholding were applied across traits.

Four MT scenarios were applied, resulting in four different predictors (**Figure 1**): (1) UKB T2DM summary statistics combined with the seven UKB information traits; (2) external T2DM summary statistics [i.e., results from Scott et al. (10) and Zhao et al. (25)] combined with the seven UKB information traits; (3) external T2DM summary statistics combined with the seven UKB information traits and UKB T2DM summary statistics; and (4) UKB T2DM summary statistics combined with the seven UKB information traits and the two external T2DM summary statistics.

GRS Accuracy

The accuracy of ST-GRS and MT-GRS was determined using Nagelkerke's variance explained (R^2),

$$R^2 = \frac{1 - e^{-LR/n}}{1 - e^{-(-2L_0)/n}} \quad (12)$$

where LR is the likelihood ratio comparing two nested logistic regression models, L_0 is the log-likelihood of a model neglecting the GRS, and n is the number of observations. The full model included sex, age, UKB assessment center, the first 10 genetic principal components, and the GRS, whereas the reduced model did not contain the GRS effect. For visualization, the GRS were divided into percentiles, and the disease prevalence within each bin was computed; the OR for each percentile was computed adjusting for sex, age, UKB assessment center, and the first 10 genetic principal components, and the OR was expressed relative to the 50-th percentile.

RESULTS

ST Prediction and Genetic Parameters

The analysis of T2DM was performed using 335,662 unrelated individuals from UKB with more than 18,000 T2DM cases (**Table 2**). A larger proportion of T2DM cases were males and smokers; on average, T2DM cases were older than individuals without T2DM, had higher BMI, and on average used more medications than non-diabetic individuals (**Table 2**).

The UKB cohort was split into 10 training and validation sets, and within-cohort marginal marker effects of common genotyped variants were estimated for each training set. After LD pruning and *P*-value thresholding, ST-GRS were computed for individuals within the validation sets. The maximum prediction accuracy for ST-GRS was $R^2 = 0.032$ when using variants with LD $r^2 < 0.9$ and $P < 0.05$ (**Figure 2**; **Supplementary Table 2**).

Across the 10 training sets, the average heritability for T2DM on the observed scale was 0.07 (0.31 on the liability scale). Seven information traits were included and used in the MT genetic risk scoring (**Table 2**). All seven traits showed non-zero heritability estimates (**Figure 3A**), and the strongest genetic correlation was observed between diastolic and systolic blood pressure (**Figure 3B**). Current medication use was the trait that showed the highest genetic correlation to most of the other traits,

and only standing height showed negative genetic correlation to the other traits (**Figure 3B**).

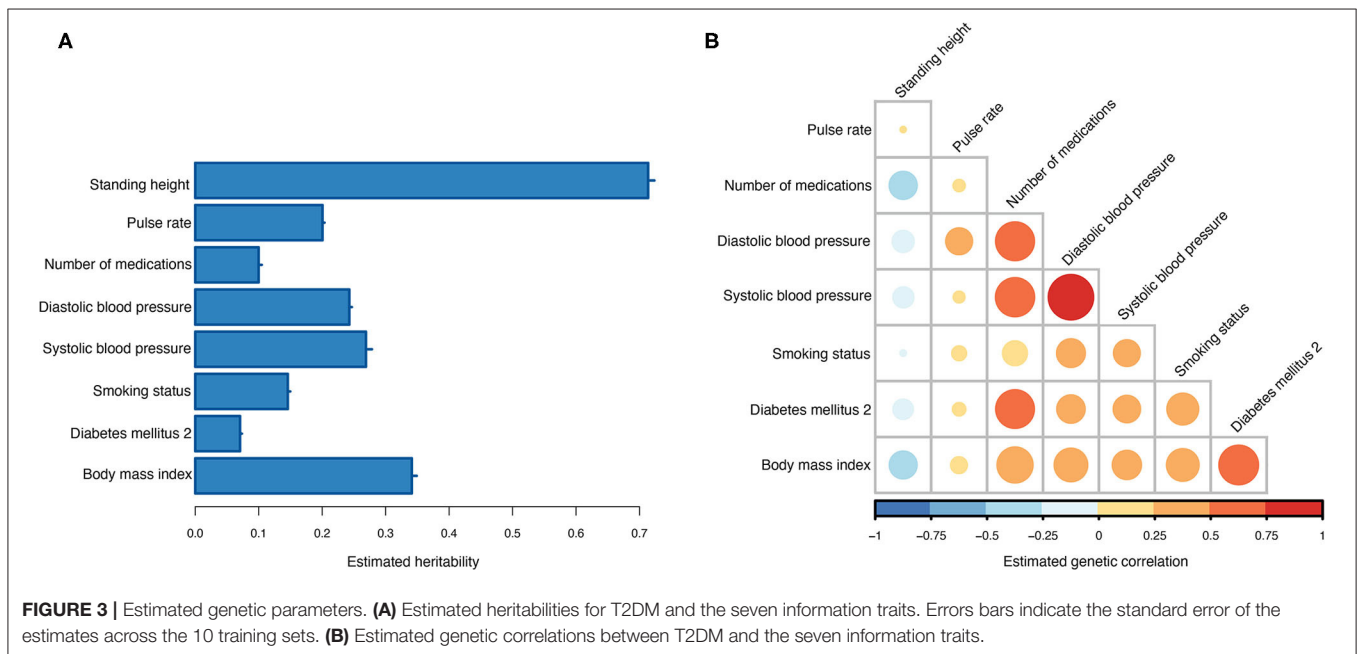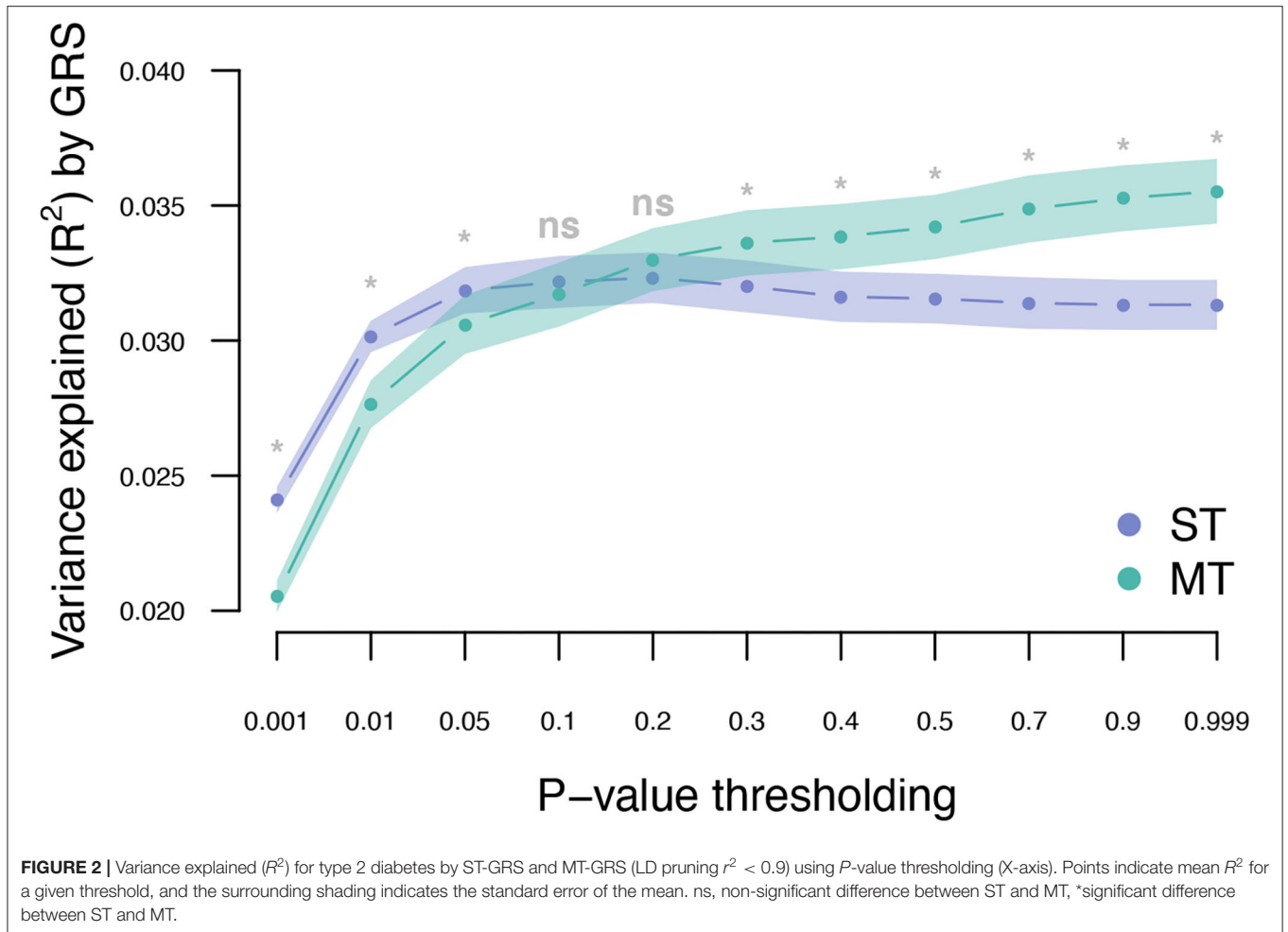## Leveraging Correlated Information for MT Prediction

The T2DM marginal effects were adjusted using the estimated genetic parameters to compute MT-GRS (Scenario 1; **Figure 1**). Across the three levels of LD pruning, the predictive ability was generally improved when the marginal SNP effects were adjusted by the seven information traits (**Supplementary Figure 1**; **Supplementary Table 2**). The highest prediction accuracy ($R^2 = 0.036$) was obtained at LD $r^2 < 0.9$ and $P < 0.999$ (**Figure 2**; **Supplementary Table 2**), which corresponds to an improved prediction accuracy by 12.5%
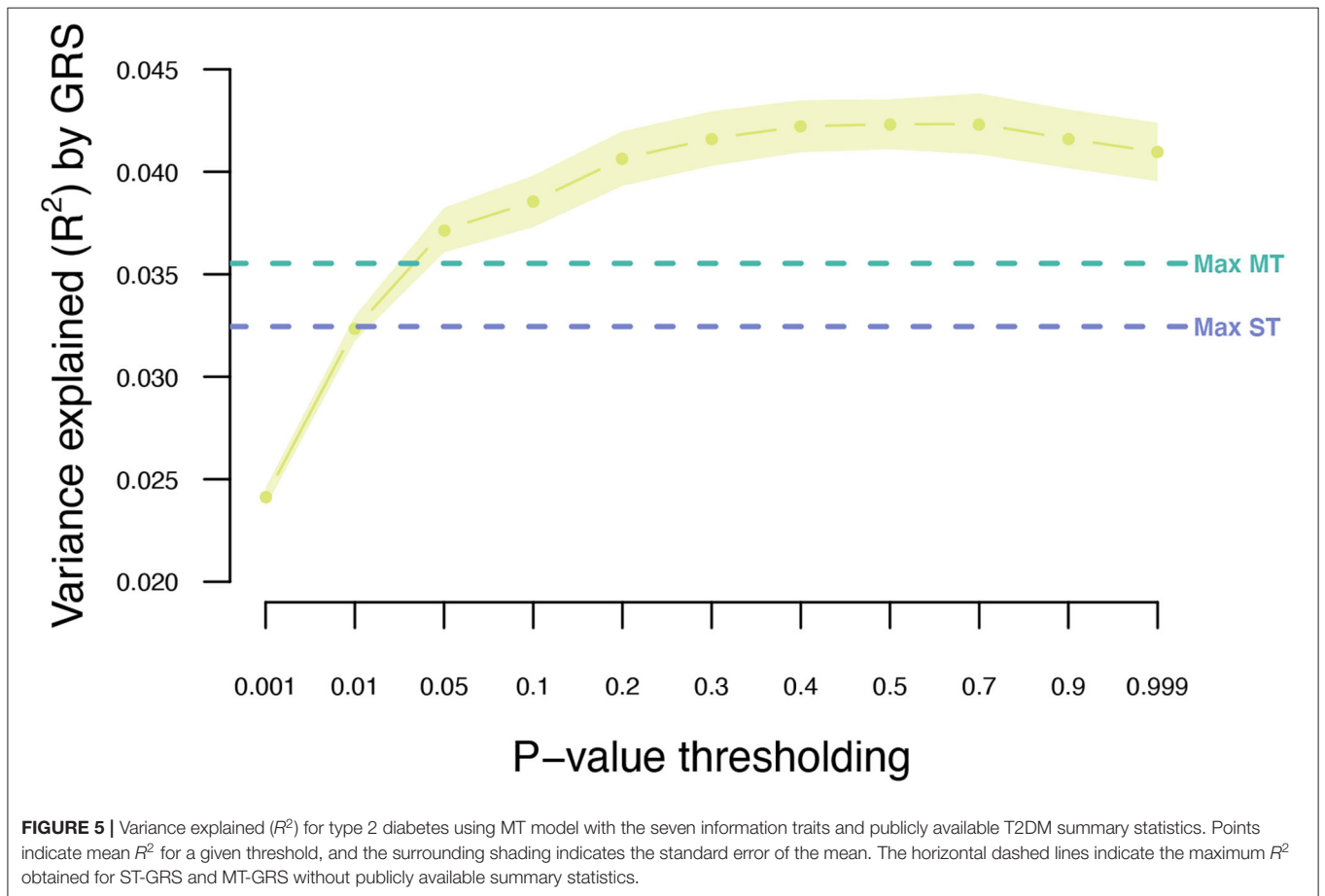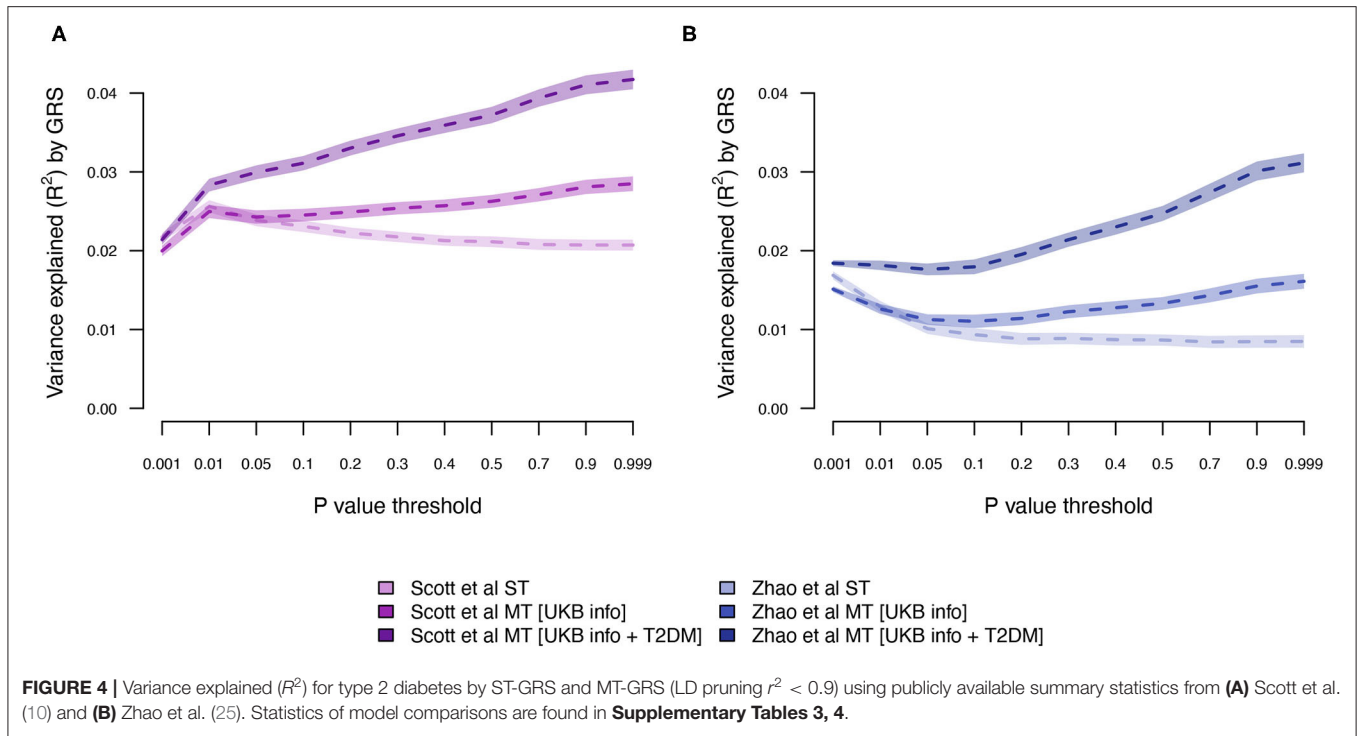
Next, we estimated the T2DM risk within the UKB using summary statistics from two independent external sets of summary statistics (**Figure 1**). Both external data sets [Scott et al. (10) and Zhao et al. (25)] showed low prediction accuracy when the GRS solely were computed using T2DM summary statistics [Scott et al. (10): $R^2 = 0.026$ at LD $r^2 = 0.9$ and $P < 0.01$; and Zhao et al. (25): $R^2 = 0.017$ at LD $r^2 = 0.9$ and $P < 0.001$; **Figure 4**; **Supplementary Tables 3, 4**; **Supplementary Figure 2**]. The external T2DM summary statistics were adjusted using summary statistics from the seven information traits obtained from the UKB (Scenario 2; **Figure 1**; **Supplementary Table 1**; **Supplementary Figure 3**), which for the summary statistics from Scott et al. (10) increased the prediction accuracy by 8%, but for Zhao et al. (25), a marginal drop in accuracy was observed when comparing the local maximum for ST-GRS with the local maximum for MT-GRS [$R^2 = 0.017$ ($r^2 = 0.9$, $P < 0.001$)] vs. 0.016 [$R^2 = 0.016$ ($r^2 = 0.9$, $P < 0.999$); **Supplementary Table 4**]; however, comparing the accuracy within the *P*-value threshold, the accuracy of the MT-GRS model was superior over the ST (**Supplementary Table 4**). Extending the MT model to also include UKB T2DM summary statistics (Scenario 3, **Figure 1**), the accuracy was further increased by 50% (from 0.028 to 0.042; **Figure 4**) and 94% (from 0.016 to 0.031; **Figure 4**) using the summary statistics of Scott et al. (10) and Zhao et al. (25), respectively.

The MT model trained within the UKB was further extended to also include summary statistics from the two independent T2DM GWAS data sets (Scenario 4; **Figure 1**). Adjusting the UKB T2DM summary statistics by the seven information traits and the two independent T2DM GWAS data sets resulted in an increase in prediction accuracy from 0.032 to 0.043 (**Figure 5**; **Supplementary Table 2**), which is an increase of 34%.

T2DM Risk Stratification

Stratifying UKB participants based on their T2DM genetic risk showed that a larger proportion of individuals with a T2DM diagnosis were among the top 10% of individuals with highest genetic score when applying the MT strategy (**Figure 6**). The MT-GRS that in addition to the seven information traits also included information from the independent testing data gave a better stratification of cases by distributing a larger proportion of T2DM cases within the top risk (**Figure 6**), which also was apparent

FIGURE 2 | Variance explained (R^2) for type 2 diabetes by ST-GRS and MT-GRS (LD pruning $r^2 < 0.9$) using P-value thresholding (X-axis). Points indicate mean R^2 for a given threshold, and the surrounding shading indicates the standard error of the mean. ns, non-significant difference between ST and MT, *significant difference between ST and MT.

FIGURE 3 | Estimated genetic parameters. **(A)** Estimated heritabilities for T2DM and the seven information traits. Errors bars indicate the standard error of the estimates across the 10 training sets. **(B)** Estimated genetic correlations between T2DM and the seven information traits.

FIGURE 4 | Variance explained (R^2) for type 2 diabetes by ST-GRS and MT-GRS (LD pruning $r^2 < 0.9$) using publicly available summary statistics from **(A)** Scott et al. (10) and **(B)** Zhao et al. (25). Statistics of model comparisons are found in **Supplementary Tables 3, 4**.

FIGURE 5 | Variance explained (R^2) for type 2 diabetes using MT model with the seven information traits and publicly available T2DM summary statistics. Points indicate mean R^2 for a given threshold, and the surrounding shading indicates the standard error of the mean. The horizontal dashed lines indicate the maximum R^2 obtained for ST-GRS and MT-GRS without publicly available summary statistics.

FIGURE 6 | Comparison of T2DM risk gradient within the UKB according to GRS percentile for **(A)** ST model, **(B)** MT model using the seven information traits, and **(C)** MT model with the seven information traits and the T2DM testing data. Each point indicates the average T2DM prevalence within each percentile of GRS across then 10 validation sets. Horizontal lines indicate the prevalence at the top 10 GRS percentile, and percentage indicates the prevalence among the top 10% with the highest genetic risk.

with a large OR of the top 10% compared to the remaining (**Supplementary Figure 4**).

DISCUSSION

Precision medicine is predicted to change the way we prevent, diagnose, risk stratify individuals, and treat medical conditions (35, 36) through development of targeted preventive or treatment approaches based on the genetic background, biomarkers, environmental exposures, and lifestyle of the individual. Diagnosis and treatment plans based on genetic testing has been effectively applied to several monogenic disorders (37); however, for common complex diseases, genomic information has been far less incorporated. One reason for the lack of incorporating genomic information in disease prevention and diagnosis for complex diseases is because a large proportion of the underlying genetic variation remains unexplained (38, 39). In the current study, we investigated whether an MT-GRS approach provided more accurate risk stratification than traditional ST genetic scoring approaches.

Adjusting the UKB T2DM marker effects by the genomic correlation of the seven information traits increased the prediction accuracy from $R^2 = 0.032$ to 0.036, and further adjusted by the two UKB-independent T2DM studies increased the accuracy to $R^2 = 0.042$. The great improvement in prediction accuracy (31%) is achieved as a consequence of abundant genomic pleiotropy (18, 30) and the apparent genomic correlation with the selected traits. In comparison, Khera et al. (14) reported a prediction accuracy of ST-GRS of $R^2 = 0.028$ (14), and Maier et al. (30) obtained an accuracy of $R^2 < 0.01$ for both ST-GRS and MT-GRS (30). Although Maier et al. (30) showed increased prediction accuracy by combining the marker effects of selected traits (30), our reported prediction accuracies were greatly elevated compared with Maier et al. (30), most likely driven by differences in the included traits, and thereby in the optimal weights caused by differences in genomic correlation among the traits.

One of the information traits we included in the MT-GRS was the genetic liability to current medication use, which is the number of different medications the UKB participants have taken at the time of the verbal interview. Because most individuals that suffers from temporary or chronic diseases will undergo medical intervention and because of comorbidity many individuals will have multiple medical conditions, those individuals will be treated with a range of different medicines. Consequently, the total set of prescription and over-the-counter drugs is potentially an informative index of the current medical and health status of an individual. Wu et al. (40) performed genetic analysis of self-reported medication use within the UKB and found that categories of different types of medication were strongly genetically associated with a range of different diseases and traits (40). We found that the genetic correlation between T2DM and medication use was $r_g = 0.55$ (only the correlation between T2DM and BMI had higher estimate, $r_g = 0.58$). This is also evident by investigating the optimal weights (Equation 7), where BMI and medication use were the two information traits with the largest weights (**Supplementary Figure 5A**), besides T2DM itself. Including summary statistics from the two published T2DM association studies only marginally affected the optimal weights (**Supplementary Figure 5B**).

Although the exact level of prediction accuracy of T2DM was considerably lower when using external data from Zhao et al. (25) compared to data from Scott et al. (10) (**Figure 4**), the percentage increase when extending ST-GRS to the MT-GRS was higher for Zhao et al. (25) (82%) compared with Scott et al. (10) (62%), despite the much greater sample size by Zhao et al. (25) (**Table 1**). The discrepancy in prediction accuracy is most likely a consequence of different ancestries of the two external T2DM studies (10, 25), where the ancestry of the individuals in the study by Scott et al. (10) is more similar to the ancestry of the UKB (European) than the study by Zhao et al. (25) (mixed ancestry). It is well-established that across ancestry, risk prediction is very difficult because the LD between populations is very diverse (41–43).

The last decade has shown us that the sample size of human genetic association studies keeps increasing (44, 45), not only entailing more association signals but also providing more accurate effect estimates. This in conjunction with the increasingly accessibility of publicly available GWAS summary statistics (46, 47) implies that genomic prediction of complex diseases will continually improve, in particular if multivariate predictors are created by integrating information across studies. Although we have demonstrated increased prediction accuracy by constructing MT-GRS, our work has several limitations. Firstly, as our training data were the UKB and with a 10-fold cross-validation scheme, the number of cases became limited, meaning less accurate marker effect estimation and thereby less accurate risk stratification. Secondly, although we in addition to the UKB summary statistics from the 10-fold cross-validation obtained T2DM summary statistics from two independent studies (**Table 1**), we only had access to genotype information from the UKB and no other T2DM cohorts. Thirdly, we restricted the number of information traits to seven (**Table 2**), based on the criterion that it should be a type of information that is easy and accurate to measure and obtain; height, BMI, pulse rate, and diastolic and systolic blood pressure are things that we easily and accurately can measure, and smoking status and current medication use can easily be obtained by asking the participants. Accurate observations lead to more accurate estimation of marker effects and thereby better prediction accuracies. It is compelling to speculate whether other types of information traits would improve prediction accuracy even more, and additional studies are warranted for developing methods for identifying the set of information traits most important for a particular disease.

Genomic information has the potential to change the way we diagnose and treat individuals today and will be central for implementing preventive healthcare in the clinics. An important aspect of precision medicine is accurate prediction of genetic risk toward common diseases, as it may guide the general practitioners to better and earlier identify those individuals who have an inherent genetically lifetime high disease risk, and then to initiate lifestyle changes potentially before disease outcome. Moreover, precise stratification of T2DM patients not only based on their pathophysiological symptoms (13) but also on their genetic makeup may help the general practitioners to treat high-risk patients more aggressively, which has the potential to slow down disease progression, reduce symptoms, and prevent severe morbidity and mortality.

In conclusion, by incorporating information traits and two previously published T2DM GWAS results, the prediction accuracy for T2DM was increased by 31% (from $R^2 = 0.032$ to $R^2 = 0.042$), clearly demonstrating the added benefit of incorporating correlated information in the construction of GRS. Thus, incorporating genomic information on correlated traits and disease is advisable for obtaining improved individual genetic risk stratification.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: The genetic and phenotypic data were obtained from the UK Biobank Resource (ID 31269). Researchers can apply for access through: https://www.ukbiobank.ac.uk/registerapply/. Summary statistics for T2DM were obtained from published studies.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by The Ethics and Governance Framework (EGF) sets standards for the UK Biobank project so that all necessary safeguards are in place to ensure that the data and samples are only used for scientifically and ethically approved research. Participants of the UK Biobank have given their consent to participate which will apply throughout the lifetime of the UK Biobank unless the participants withdraw. Their consent involves the collection and storage of biological material (blood, saliva, urine samples) as well as collection of electronic health records (GP, hospitals, dental and prescription records). Information on the individual data level is anonymised for the researchers, and every research project has its own anonymised data. The ethics committee waived the requirement of written informed consent for participation.

AUTHOR CONTRIBUTIONS

FUNDING

ACKNOWLEDGMENTS

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2021.711208/full#supplementary-material

REFERENCES

1. The Emerging Risk Factors Collaboration, Sarwar N, Gao P, Kondapally Seshasai SR, Gobin R, Kaptoge S, et al. Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. *The Lancet* (2010) 375:2215–22. doi: 10.1016/S0140-6736(10)60484-9

2. Guariguata L, Whiting DR, Hambleton I, Beagley J, Linnenkamp U, Shaw JE. Global estimates of diabetes prevalence for 2013 and projections for 2035. *Diabetes Res Clin Pract.* (2014) 103:137–49. doi: 10.1016/j.diabres.2013.11.002

3. Cho NH, Shaw JE, Karuranga S, Huang Y, da Rocha Fernandes JD, Ohlrogge AW, et al. IDF Diabetes Atlas: global estimates of diabetes prevalence for 2017 and projections for 2045. *Diabetes Res Clin Pract.* (2018) 138:271–81. doi: 10.1016/j.diabres.2018.02.023

4. Kolb H, Martin S. Environmental/lifestyle factors in the pathogenesis and prevention of type 2 diabetes. *BMC Med.* (2017) 15:1–11. doi: 10.1186/s12916-017-0901-x

5. Flannick J, Florez JC. Type 2 diabetes: genetic data sharing to advance complex disease research. *Nat Rev Genet.* (2016) 17:535–49. doi: 10.1038/nrg.2016.56

6. Fuchsberger C, Flannick J, Teslovich TM, Mahajan A, Agarwala V, Gaulton KJ, et al. The genetic architecture of type 2 diabetes. *Nature.* (2016) 536:41–7. doi: 10.1038/nature18642

7. Poulsen P, Ohm Kyvik K, Vaag A, Beck-Nielsen H. Heritability of type II (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance - a population-based twin study. *Diabetologia.* (1999) 42:139–45. doi: 10.1007/s001250051131

8. Willemsen G, Ward KJ, Bell CG, Christensen K, Bowden J, Dalgård C, et al. The concordance and heritability of Type 2 diabetes in 34,166 twin pairs from international twin registers: The Discordant Twin (DISCOTWIN) Consortium. *Twin Res Hum Genet.* (2015) 18:762–71. doi: 10.1017/thg.2015.83

9. Morris AP, Voight BF, Teslovich TM, Ferreira T, Segrè A V., Steinthorsdottir V, et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat Genet.* (2012) 44:981–90. doi: 10.1038/ng.2383

10. Scott RA, Scott LJ, Mägi R, Marullo L, Gaulton KJ, Kaakinen M, et al. An expanded genome-wide association study of Type 2 diabetes in Europeans. *Diabetes.* (2017) 66:2888–902. doi: 10.2337/db16-1253

11. Mahajan A, Taliun D, Thurner M, Robertson NR, Torres JM, Rayner NW, et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat Genet.* (2018) 50:1505–13. doi: 10.1038/s41588-018-0241-6

12. Vujkovic M, Keaton JM, Lynch JA, Miller DR, Zhou J, Tcheandjieu C, et al. Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis. *Nat Genet.* (2020) 52:680–91. doi: 10.1101/19012690

13. Ahlqvist E, Storm P, Käräjämäki A, Martinell M, Dorkhan M, Carlsson A, et al. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. *Lancet Diabetes Endocrinol.* (2018) 6:361–9. doi: 10.1016/S2213-8587(18)30051-2

14. Khera A V., Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet.* (2018) 50:1219–24. doi: 10.1038/s41588-018-0183-z

15. Vilhjálmsson BJ, Yang J, Finucane HK, Gusev A, Lindström S, Ripke S, et al. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am J Hum Genet.* (2015) 97:576–92. doi: 10.1016/j.ajhg.2015.09.001

16. Choi SW, Mak TSH, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. *Nat Protoc.* (2020) 15: doi: 10.1038/s41596-020-0353-1

17. Euesden J, Lewis CM, O'Reilly PF. PRSice: polygenic risk score software. *Bioinformatics.* (2015) 31:1466–8. doi: 10.1093/bioinformatics/btu848

18. Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh PR, et al. An atlas of genetic correlations across human diseases and traits. *Nat Genet.* (2015) 47:1236–41. doi: 10.1038/ng.3406

19. Iglay K, Hannachi H, Joseph Howie P, Xu J, Li X, Engel SS, et al. Prevalence and co-prevalence of comorbidities among patients with type 2 diabetes mellitus. *Curr Med Res Opin.* (2016) 32:1243–52. doi: 10.1185/03007995.2016.1168291

20. Goodarzi MO, Rotter JI. Genetics insights in the relationship between Type 2 diabetes and coronary heart disease. *Circ Res.* (2021) 126:1526–48. doi: 10.1161/CIRCRESAHA.119.316065

21. Danaei G, Lawes CM, Vander Hoorn S, Murray CJ, Ezzati M. Global and regional mortality from ischaemic heart disease and stroke attributable to higher-than-optimum blood glucose concentration: comparative risk assessment. *Lancet.* (2006) 368:1651–9. doi: 10.1016/S0140-6736(06)69700-6

22. Coresh J, Astor BC, Greene T, Eknoyan G, Levey AS. Prevalence of chronic kidney disease and decreased kidney function in the adult US population: third national health and nutrition examination survey. *Am J Kidney Dis.* (2003) 41:1–12. doi: 10.1053/ajkd.2003.50007

23. Dean J. Organising care for people with diabetes and renal disease. *J Ren Care.* (2012) 38:23–9. doi: 10.1111/j.1755-6686.2012.00272.x

24. Bycroft C, Elliott LT, Young A, Vukcevic D, Effingham M, Marchini J, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature.* (2018) 562:203–9. doi: 10.1038/s41586-018-0579-z

25. Zhao W, Rasheed A, Tikkanen E, Lee JJ, Butterworth AS, Howson JMM, et al. Identification of new susceptibility loci for type 2 diabetes and shared etiological pathways with coronary heart disease. *Nat Genet.* (2017) 49:1450–7. doi: 10.1038/ng.3943

26. Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience.* (2015) 4:1–16. doi: 10.1186/s13742-015-0047-8

27. Sofer T, Zheng X, Gogarten SM, Laurie CA, Grinde K, Shaffer JR, et al. A fully adjusted two-stage procedure for rank-normalization in genetic association studies. *Genet Epidemiol.* (2019) 43:263–75. doi: 10.1002/gepi.22188

28. Bulik-Sullivan B, Loh PR, Finucane HK, Ripke S, Yang J, Patterson N, et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet.* (2015) 47:291–5. doi: 10.1038/ng.3211

29. Rohde PD, Fourie Sørensen I, Sørensen P. qgg: an R package for large-scale quantitative genetic analyses. *Bioinformatics.* (2020) 36:2614–5. doi: 10.1093/bioinformatics/btz955

30. Maier RM, Zhu Z, Lee SH, Trzaskowski M, Ruderfer DM, Stahl EA, et al. Improving genetic prediction by leveraging genetic correlations among human diseases and traits. *Nat Commun.* (2018) 9:989. doi: 10.1038/s41467-017-02769-6

31. Hu D, Wang C, O'Connor AM. A method of back-calculating the log odds ratio and standard error of the log odds ratio from the reported group-level risk of disease. *PLoS ONE.* (2020) 15:e0222690. doi: 10.1371/journal.pone.0222690

32. Hazel LN. The genetic basis for constructing selection indexes. *Genetics.* (1943) 28:476–90. doi: 10.1093/genetics/28.6.476

33. Wientjes YCJ, Bijma P, Veerkamp RF, Calus MPL. An equation to predict the accuracy of genomic values by combining data from multiple traits, populations, or environments. *Genetics.* (2016) 202:799–823. doi: 10.1534/genetics.115.183269

34. Yang J, Weedon MN, Purcell S, Lettre G, Estrada K, Willer CJ, et al. Genomic inflation factors under polygenic inheritance. *Eur J Hum Genet.* (2011) 19:807–12. doi: 10.1038/ejhg.2011.39

35. Ginsburg GS, McCarthy JJ. Personalized medicine: revolutionizing drug discovery and patient care. *Trends Biotechnol.* (2001) 19:491–6. doi: 10.1016/S0167-7799(01)01814-5

36. Ashley EA. Towards precision medicine. *Nat Rev Genet.* (2016) 17:507–22. doi: 10.1038/nrg.2016.86

37. Katsanis SH, Katsanis N. Molecular genetic testing and the future of clinical genomics. *Nat Rev Genet.* (2013) 14:415–26. doi: 10.1038/nrg3493

38. Chatterjee N, Shi J, García-Closas M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat Rev Genet.* (2016) 17:392–406. doi: 10.1038/nrg.2016.27

39. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. *Nature.* (2009) 461:747–53. doi: 10.1038/nature08494

40. Wu Y, Byrne EM, Zheng Z, Kemper KE, Yengo L, Mallett AJ, et al. Genome-wide association study of medication-use and associated disease in the UK Biobank. *Nat Commun.* (2019) 10:1891. doi: 10.1038/s41467-019-09572-5

41. Kerminen S, Martin AR, Koskela J, Ruotsalainen SE, Havulinna AS, Surakka I, et al. Geographic variation and bias in the polygenic scores of complex diseases and traits in Finland. *Am J Hum Genet.* (2019) 104:1169–81. doi: 10.1016/j.ajhg.2019.05.001

42. Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, et al. Human demographic history impacts genetic risk prediction across diverse populations. *Am J Hum Genet.* (2017) 100:635–49. doi: 10.1016/j.ajhg.2017.03.004

43. Duncan L, Shen H, Gelaye B, Meijsen J, Ressler K, Feldman M, et al. Analysis of polygenic risk score usage and performance in diverse human populations. *Nat Commun.* (2019) 10:3328. doi: 10.1038/s41467-019-11112-0

44. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. *Am J Hum Genet.* (2012) 90:7–24. doi: 10.1016/j.ajhg.2011.11.029

45. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 Years of GWAS Discovery: biology, function, and translation. *Am J Hum Genet.* (2017) 101:5–22. doi: 10.1016/j.ajhg.2017.06.005

46. Pasaniuc B, Price AL. Dissecting the genetics of complex traits using summary association statistics. *Nat Rev Genet.* (2017) 18:117–27. doi: 10.1038/nrg.2016.142

47. Buniello A, Macarthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* (2019) 47:D1005–12. doi: 10.1093/nar/gky1120

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.