Acta Diabetologica (2021) 58:467–474
https://doi.org/10.1007/s00592-020-01650-1

ORIGINAL ARTICLE

Genomic risk score provides predictive performance for type 2 diabetes in the UK biobank

Xiaolu Chen[1] · Congcong Liu[1] · Shucheng Si[1] · Yunxia Li[1] · Wenchao Li[1] · Tonghui Yuan[1] · Fuzhong Xue[1,2]

Received: 29 July 2020 / Accepted: 1 December 2020 / Published online: 3 January 2021

Abstract

Aims Type 2 diabetes (T2D) is affected by a combination of genetic and environmental factors. However, the comprehensive genomic risk scores (GRSs) for T2D prediction have not been evaluated.

Methods Using a meta-scoring approach, we developed a metaGRS for T2D; T2D-related traits consist of 1,692 genetic variants in the UK Biobank training set (n = 40,423 + 7,558 events) and evaluate this score in the validation set (n = 303,053).

Results The hazard ratio (HR) for T2D was 1.32 (95% confidence interval [CI]: 1.29–1.35) per standard deviation of metaGRS and was larger than previously published T2D-GRS. Individuals, in the top 25% of metaGRS, have an HR of 2.08 (95%CI: 1.93–2.23) compared with those in the bottom 25%. The addition of metaGRS to all conventional risk factors significantly increased the AUC ($P < 0.001$). Adding metaGRS to all conventional risk factors significantly improved the reclassification accuracy (continuous net reclassification improvement = 11.8%, 95%CI: 9.2%–14.2%). All analyses adjusted for age, sex, and 10PCs.

Conclusions The metaGRS significantly improves T2D prediction ability.

Keywords Type 2 diabetes · Genomic risk scores · Risk factors · Competing risk model · Net reclassification improvement

Introduction

The increasing prevalence of type 2 diabetes (T2D) is one of the greatest challenges in published health, causing enormous costs and a decreasing in the quality of life [1]. In Europe, T2D accounted for 80%–90% of all diabetes and decreased life expectancy by 5–10 years [2]. The risk of T2D is determined by a complex interplay of genetic and environmental factors and can be partly acted by changes in lifestyles [3].

It is well known that the heritability for T2D is 26%–69% [4]. Genome-wide association studies (GWAS) have enabled major advances in the identification of single-nucleotide polymorphisms (SNPs) associated with T2D risk [5–8]. Recently, some researches highlighted the potential of genomic risk scores (GRSs) for risk prediction of common diseases. They identified 8.0, 6.1, 3.5, 3.2, and 1.5% of the population at greater than threefold increased risk for coronary artery disease, atrial fibrillation, type 2 diabetes, inflammatory bowel disease, and breast cancer, respectively [9, 10]. For predicting disease risk, GRS has a notable advantage over conventional clinical factors as it could identify high-risk individuals from birth [11].

A recent T2D-GRS using 66 SNPs derived from European-ancestry participants indicated that individuals with high GRS and unhealthy lifestyle factors increased the risk of T2D. The research concluded that compared with participants with the healthiest lifestyle and low T2D-GRS, the relative risk of T2D for participants with the least healthy lifestyle and high T2D-GRS was 8.72 [12]. Nevertheless, the clinical utility of GRS depends on the ability to predict

Managed by Massimo Federici .

Supplementary information The online version contains supplementary material available at (https://doi.org/10.1007/s00592-020-01650-1)

✉ Fuzhong Xue
xuefzh@sdu.edu.cn

[1] Department of Biostatistics, School of Public Health, Cheeloo College of Medicine, Shandong University, No.44 Wenhuaxi Road, Jinan 250012, People's Republic of China

[2] Institute for Medical Dataology, Shandong University, No.12550 Erhuandong Road, Jinan 250002, People's Republic of China

future T2D events, not on the strength of the association with T2D. Several studies have examined the utility of the combination of susceptibility variants for T2D prediction [13, 14]. These results showed that the additive effect of GRS on clinical factors was marginal, although statistically significant. In previous studies on coronary artery disease prediction, there is an approach that combined multiple GRSs into one meta-score (metaGRS) to improve the predictive performance [15]. But this comprehensive metaGRS construction method has not been applied to the prediction of T2D.

Here, we extend the metaGRS method to predict T2D risk by incorporating summary GWAS statistics for T2D and its risk factors. The metaGRS is constructed and validated using UK Biobank database (UKB), and compared with previously published T2D-GRS [6]. Most previous diabetes predictions were based on a conventional risk model that included clinically available indicators, including lipids, blood pressure, family history, etc. Moreover, we examine the improvement in prediction performance and reclassification accuracy after adding metaGRS to the conventional risk model.

Methods

Study design and participants

Study design and methods of UKB have been reported previously [16]. In brief, UKB is a large-scale prospective study with 502, 527 participants aged 37–73 years and recruited in 2006–2010. At recruitment, detailed information was collected via a standardized socio-demographic questionnaire, as well as health status and physician-diagnosed medical conditions, family history, and lifestyle factors. Several physical measurements were obtained, including height, weight, waist–hip ratio (WHR), systolic blood pressures (SBP), and diastolic blood pressures (DBP). Individual data were linked to Hospital Episode Statistics (HES) data, and national death and cancer registries. HES uses International Classification of Diseases (ICD)-9th and -10th Revisions to record diagnosis information. Death registries include all death in the UK with both primary and secondary causes of death coded in ICD-10. The UKB study has approval from the North West Multicentre Research Ethical Committee. All participants provided written informed consent.

For diabetes cases, we used HES/death data (ICD-10: E10–E14, ICD-9: 250), diabetes diagnosed by doctor (data files #2443), and glucose medication (data files #6153, #6177, #20,003); type 2 diabetes case was defined by ICD-10 E11.

We defined risk factors at the first assessment, including body mass index (BMI, data files #21,001), smoking status (data files #20,116), triglycerides (TG, data files #30,870), high-density lipoprotein (HDL, data files #30,760), low-density lipoprotein (LDL, data files #30,780), glucose (data files #30,740), hypertension, and family history of diabetes. For hypertension, we used an expanded definition: blood pressure medication (data files #6153, #6177), SBP > 140 mm Hg (the mean of two SBP measurements, data files #4080), DBP > 90 mm Hg (the mean of two DBP measurements, data files #4079), or HES/death data (ICD-10: I10–I15). For the family history of diabetes, we considered history in any first-degree relative (father, mother, sibling; data files #20,107, #20,110, #20,111, respectively).

The design of this study and detailed inclusion and exclusion criteria are shown in Fig. 1. We randomly divided UKB British white data set into training ($n = 40{,}423$) and validation set ($n = 303{,}053$). In order to increase statistical power in metaGRS generating phase, we enriched the training set with 7,558 T2D cases, those excluded from the validation set due to T2D diagnosis at baseline, leading to 9,913 T2D cases and 38,068 controls.

Generation of metaGRS

The training data set was used to construct GRSs and was excluded from further analysis. The genotyping process and arrays used in the UK Biobank study have been described previously [17]. We constructed 17 genetic risk scores (GRSs) for phenotypes associated with T2D: T2D [6], HbA1c [18], 2-h blood glucose (2hGlu), fasting glucose (FG), fasting insulin (FI) [19], HDL, LDL, total cholesterol (TC), TG [20], SBP, DBP [21], waist, hip, WHR [22], BMI [23], height [24], and smoking [25]. Totally, we selected 1,692 SNPs associated with corresponding phenotypes with genome-wide significance ($P < 5\text{e-}8$) in populations of European descent (details in supplementary table). We generated these GRSs based on $r^2 < 0.1$ with PLINK [26] LD thinning. To control for population structure, we used the genetic principal components (PCs) supplied by UKB [27].

On the basis of the selected SNPs, the 17 GRSs were calculated separately, using a weighted method (the sum of the risk allele dosages of each variant multiplied by its marginal effect size):

$$\text{Weighted GRS}_i = \sum\nolimits_{j=1}^{m} x_{ij}\beta_j \tag{1}$$

where $x_{ij} \in \{0, 1, 2\}$is the count of risk alleles for the jth variant in the ith individual, and β_{j} is the marginal effect size for the jth variant obtained from the reported GWAS data.

Each GRS was standardized to zero mean and unit standard deviation over the entire data set. Next, we employed elastic-net logistic regression [28] using the R package "glmnet" [29] to model the associations between the 17 GRSs and T2D, adjusting for sex and ten PCs. A range of models with different

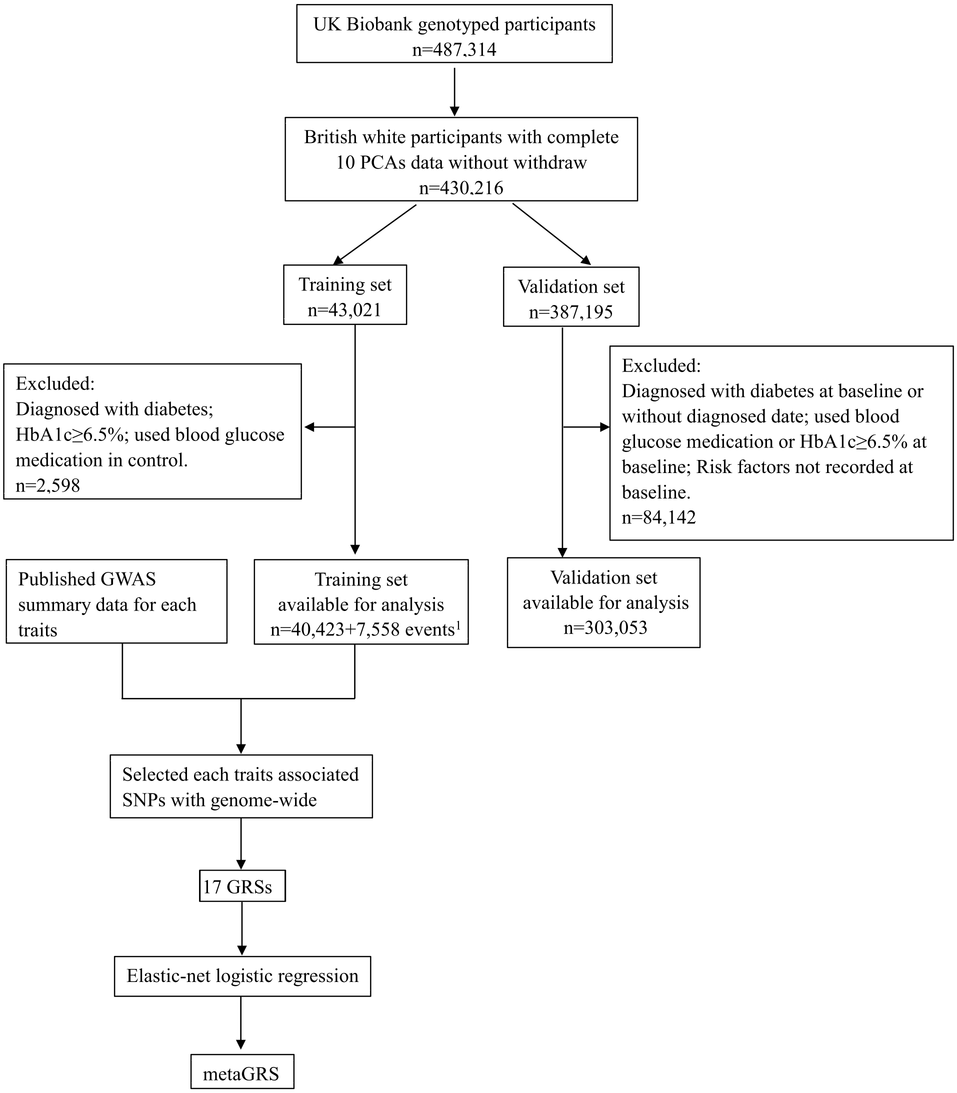

UK Biobank genotyped participants
n=487,314
British white participants with complete 10 PCAs data without withdraw
n=430,216
Training set
n=43,021
Validation set
n=387,195
Excluded:
Diagnosed with diabetes; HbA1c≥6.5%; used blood glucose medication in control.
n=2,598
Excluded:
Diagnosed with diabetes at baseline or without diagnosed date; used blood glucose medication or HbA1c≥6.5% at baseline; Risk factors not recorded at baseline.
n=84,142
Published GWAS summary data for each traits
Training set available for analysis
n=40,423+7,558 events1
Validation set available for analysis
n=303,053
Selected each traits associated SNPs with genome-wide
17 GRSs
Elastic-net logistic regression
metaGRS

Fig. 1 Study design. GRS, genomic risk score, GWAS, genome-wide association studies, SNP, single-nucleotide polymorphisms. [1]These 7,558 events were cases excluded from the validation set due to T2D diagnosis at baseline

penalties was evaluated using tenfold cross-validation. We selected the best model, in terms of the highest cross-validated area under the receiving-operating characteristic curve (AUC), as the final model to generate metaGRS and evaluated it in UKB validation set.

We generated metaGRS consisting of a weighted average of the standardized scores

$$\mathrm{GRS}_i^{\mathrm{meta}} = \frac{\alpha_1 GRS_{i1} + \cdots + \alpha_{17} GRS_{i17}}{\alpha_1 + \cdots + \alpha_{17}} \quad (2)$$

where $GRS_{i1}, \ldots, GRS_{i17}$are the 17 zero mean and unit variance standardized GRSs for the ith individual; $\alpha_1,\ldots, \alpha_{17}$ are the coefficients (log odds ratio) for each of the 17 GRSs.

Statistical analysis

The demographic and clinical characteristics of the UKB validation subset were described using median with interquartile for continuous variables and the frequency and percent for categorical variables. The subsequent analysis focused only on incident T2D events. Taking into account the existence of death, we evaluated the metaGRS in the validation subset using competing risk model proposed by Fine and Gray [30] in the R package “cmprsk,” estimated hazard ratios (HR) and 95% confidence interval (CI), and computed five-year T2D event probabilities. Based on the predicted five-year incidence probabilities of T2D for each individual and whether it actually occurs, we calculated the C-index and AUC under the presence of competitive risk. We employed the competing risk model to estimate the cumulative incidence of T2D, stratified by sex, with Gray test for comparison between groups. AUC and C-index were based on a five-year follow-up window previously described [31] using the R package "pROC" and "Hmisc," respectively. Between the two groups, the difference in AUC was estimated using the "roc.test" and the other difference in effect size was estimated using the two-sample z test [32]. All analyses were adjusted for age, sex, and ten PCs.

The net reclassification improvement (NRI) was used to assess the potential for improved discrimination between the incident and non-incident cases when added to new factors to T2D [33]. A base model including all conventional risk factors was compared with an alternate model that includes the metaGRS being evaluated. The R package "survIDINRI" was applied for NRI analysis.

All statistical tests were two-sided, and a $P < 0.05$ was considered significant. All analyses are conducted by R, version 3.6.1.

Results

The characteristics of UKB participants in the validation subset are shown in Table 1. The overall UKB validation set consists of 303,528 participants with a median age of 58 years (range 38–73 years). There are 6,724 (2.22%) incident cases of T2D during a median follow-up of 8.90 years (range 0–11.03 years), consisting of 2,714 (1.64%) women and 4,010 (2.92%) men with onset T2D.

Using the independent UKB validation set, we next evaluated the association between metaGRS and T2D via survival analysis adjusted for age, sex, and ten PCs. The metaGRS was associated with incident T2D with an HR of 1.32 (95% CI 1.29–1.35) per standard deviation of metaGRS, which was elevated compared with T2D-GRS (HR = 1.30 [95% CI 1.27–1.33]), but *p* value did not reach significance ($P = 0.12$). Hazard ratios associated with quartiles metaGRS are shown in Table 2. Using the bottom metaGRS quartile as a reference, the top metaGRS quartile of the population was at 2.08-fold increased risk of T2D (95%CI: 1.93–2.23). To investigate the potential role of the metaGRS in earlier life genetic screening, we compared the sex-stratified cumulative incidence of T2D across quartiles of the metaGRS as shown in Fig. 2. The quartile metaGRS showed substantial differences in the cumulative incidence of T2D (the Gray test between four groups: $P < 0.001$). For men, the T2D risk in the top 25% of metaGRS reached 4.15% (95%CI: 3.93%–4.37%) by 10 years of follow-up time. In comparison, the T2D risk in the bottom 25% of metaGRS reached 2.05% (95%CI: 1.89%–2.20%) by 10 years of follow-up time. In UKB women, the results were similar but had a lower T2D risk overall compared with men. For women in the highest metaGRS quartile, T2D risk reached 2.36% (95%CI: 2.21%–2.52%) by 10 years of follow-up time, whereas women in the lowest metaGRS quartile were at extremely low levels of risk, reaching 1.16% (95%CI: 1.05%–1.26%)

Table 2 Hazard ratios associated with incidence type 2 diabetes for metaGRS among validation data set

MetaGRS	Hazard ratio	*p* value
Continuous per unit increment	1.37 (1.33–1.41)	< 0.001
Continuous per SD increment (1SD = 0.889 unit)	1.32 (1.29–1.35)	< 0.001
percentiles		
0%–25%	1.00	
25%–50%	1.33 (1.24–1.44)	< 0.001
50%–75%	1.64 (1.52–1.76)	< 0.001
75%–100%	2.08 (1.93–2.23)	< 0.001

The competing risk model is adjusted for age, sex, and ten PCs

Table 1 Baseline characteristics of UK Biobank validation data set

characteristics	UK Biobank	Female	Male
Participant, N (%)	303,053	165,602 (54.64%)	137,451 (45.36%)
Age, years, median (IQR)	58.00 (13.00)	58.00 (13.00)	58.00 (14.00)
Triglyceride, mmol/L, median (IQR)	1.48 (1.08)	1.33 (0.91)	1.68 (1.25)
HDL, mmol/L, median (IQR)	1.42 (0.50)	1.57 (0.49)	1.25 (0.39)
LDL, mmol/L, median (IQR)	3.57 (1.15)	3.61 (1.16)	3.52 (1.14)
Glucose, mmol/L, median (IQR)	4.91 (0.67)	4.90 (0.65)	4.92 (0.69)
Body mass index, kg/m^2, median (IQR)	26.57 (5.53)	25.96 (6.01)	27.14 (4.86)
Follow-up time, years, median (IQR)	8.90 (1.42)	8.93 (1.39)	8.88 (1.45)
Family history of diabetes, N (%)	81,985 (27.05)	46,459 (28.05)	35,526 (25.85)
Hypertension, N (%)	150,508(49.66)	71,981 (43.47)	78,527 (57.13)
Type 2 diabetes, N (%)	6,724(2.22)	2,714(1.64)	4010(2.92)

IQR: interquartile range, ***HDL:*** high-density lipoprotein, ***LDL:*** low-density lipoprotein

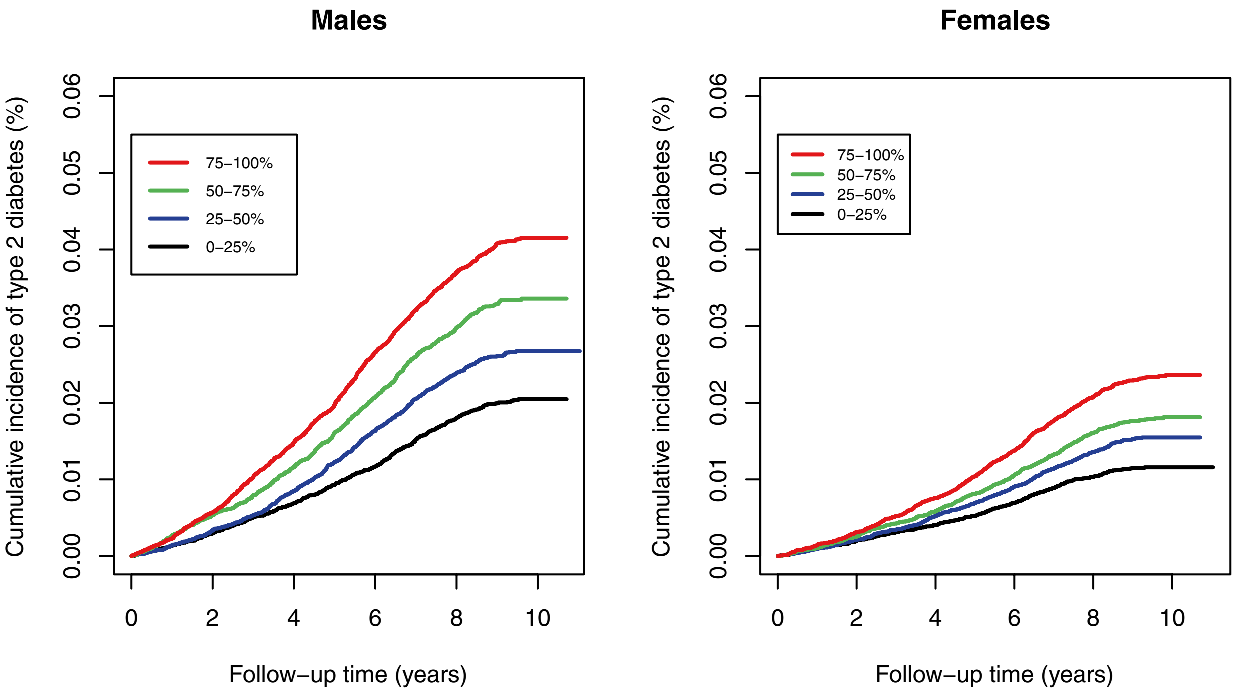
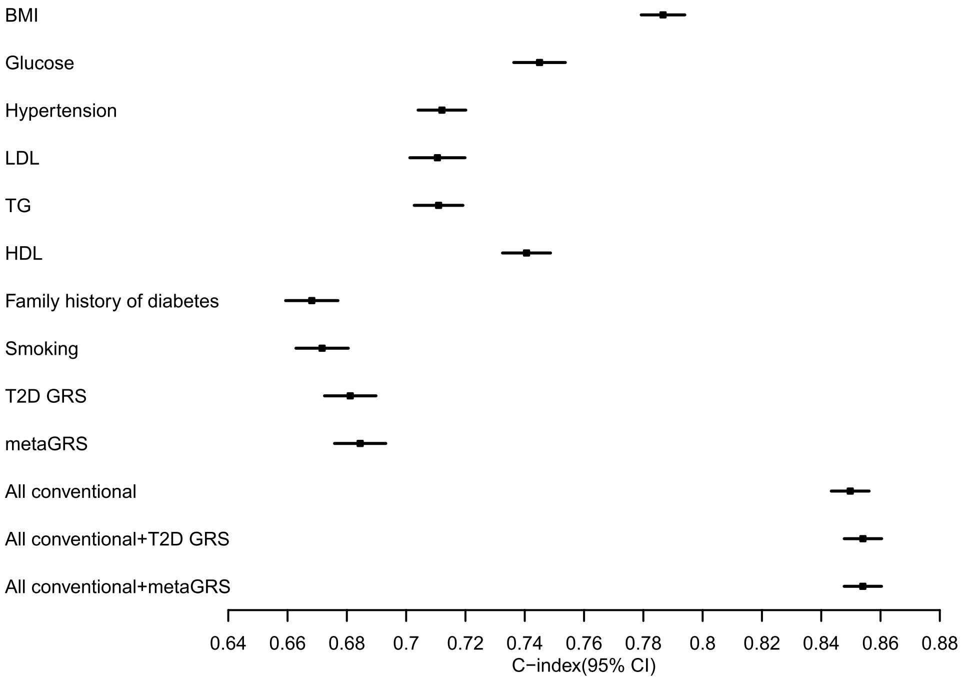


Fig. 2 Cumulative incidence of type 2 diabetes by quartiles of metaGRS in males and females

by 10 years of follow-up time. The p value for the difference between males and females in the highest metaGRS is < 0.001. The result was consistent with a competing risk model of the quartiles metaGRS.

Next, we compared the predictive performance of the metaGRS with conventional risk factors as shown in Fig. 3. We examined eight conventional risk factors at the baseline, consisting of smoking status, glucose, TC, HDL, LDL, BMI, hypertension, and family history of diabetes. BMI had the largest C statistic (0.787, 95%CI: 0.779–0.794). Notably, metaGRS had a higher C statistic and AUC than smoking and family history of diabetes (the AUC of metaGRS vs smoking: 0.684 vs 0.671, $P < 0.001$; vs family history of diabetes: vs 0.668, $P < 0.001$). The C statistic associated

Fig. 3 C-index (95%CI) for incident T2D in UK Biobank validation comparing metaGRS with conventional risk factors. T2D, type 2 diabetes, TG, triglycerides, HDL, high-density lipoprotein, LDL, low-density lipoprotein, BMI, body mass index, GRS, genomic risk score



with metaGRS was 0.684 (95%CI: 0.676–0.693), which was stronger than T2D-GRS of 0.681 (95%CI: 0.672–0.690). The AUC associated with metaGRS had the same value as the above C statistic, so is T2D-GRS, and the p value for AUCs associated with the two GRSs was less than 0.001. The addition of metaGRS to all conventional clinical risk factors modestly but significantly increased the C statistic from 0.850 (95%CI: 0.843–0.856) to 0.854 (95%CI: 0.848–0.860), and the incremental value in C statistic was 0.004; the AUC was increased from 0.851 (95%CI: 0.844–0.857) to 0.855 (95%CI: 0.849–0.862), and the increment in AUC is 0.004 ($P < 0.001$). MetaGRS plus conventional risk factors and T2D-GRS plus conventional risk factors had the same AUC and the same C statistic. The 5-year T2D probability by age groups based on the addition of metaGRS to all conventional risk factors is shown in supplementary figure.

Adding metaGRS to all conventional risk factors significantly improved the reclassification accuracy (continuous NRI = 11.8%, 95%CI: 9.2%–14.2%, $P < 0.001$).

Discussion

In this study, we constructed a metaGRS based on GWAS summary statistics for 17 T2D and its risk factors. We evaluated the predictive power of metaGRS by comparing it to established conventional risk factors. Then, we found that the effects of adding metaGRS to clinical information were significant.

In UKB British white data set, the incidence of type 2 diabetes in men was higher in women (2.92% against 1.64%). First, we showed that the new metaGRS had an elevated association with T2D compared with previously published T2D-GRS, although it did not reach significance (HR: 1.32 against 1.30, $P = 0.12$). The AUC associated with metaGRS was 0.684 (95%CI: 0.676–0.693), which was stronger than T2D-GRS of 0.681 (95%CI: 0.672–0.690), p value was less than 0.001. In addition, individuals in the top metaGRS quartile had a 2.08-fold HR and a higher cumulative incidence for T2D versus the bottom quartile. Next, we found that metaGRS had higher predictive power than the family history of diabetes and smoking status, but lower than other conventional risk factors. Adding metaGRS to all conventional risk factors significantly improved predictive power and 11.8% reclassification accuracy.

Our finding suggested that the addition of genetic factors only marginally improved the C-index and AUC beyond the clinical risk model, although the differences were statistically significant. But metaGRS plus clinical risk factors and T2D-GRS plus these factors had the same AUC and the same C statistics. Similar results were seen with regard to the predictive power of metaGRS, with C statistic in a similar range [11]. This indicated that the modest improvement in C-index was not a result of genetic variant selection. In line with our results, the T2D-GRS constructed from 49 variants showed a limited power to discriminate between susceptible and unsusceptible individuals in a Japanese population [34]. Although gene sequencing is slightly more expensive than clinical indicators, it needs to be tested once in a lifetime; it is helpful for the prediction of diseases from birth. A key measure of the clinical utility of a survival model is its ability to discriminate those who will develop T2D from those who will not. Although C statistic and AUC are the most popular metric, the increase is often very small in magnitude [35]. Therefore, we utilized the NRI to quantify the improvement. We observed that adding metaGRS to conventional factors caused 11.8% improvement in reclassification accuracy. This indicated metaGRS had a certain degree of clinical value to correct reassignment among risk categories.

The advantage of our study is that we examined a new metaGRS constructed with a large number of T2D and T2D-related traits susceptibility loci in three hundred thousand UKB participants whose genotype information was complete and demonstrated its clinical value in predicting T2D. Our study also had some limitations. The UKB participants have a minimum enrollment age of 38 years and have been shown to be healthier than the UK general population [36]. Some diabetes cases are not be distinguished between type 1 and type 2 diabetes. Thus, our study may have underestimated population-level lifetime T2D risk.

In conclusion, the metaGRS, constructed using 1,692 SNPs strongly associated with 17 T2D and T2D-related traits based on GWAS summary statistics, was significantly related to an increased risk of T2D in the European population. This genetic information significantly improved T2D prediction ability. It lays the groundwork for larger GWAS of T2D as well as analyses that leverage the totality of information available for T2D genomic risk prediction.

Author contributions Xiaolu Chen, Shucheng Si, and Fuzhong Xue participated in study conception and design. Xiaolu Chen, Congcong Liu, Yunxia Li, Wenchao Li, and Tonghui Yuan were responsible for the acquisition and cleaning of the data. Xiaolu Chen performed the statistical analyses and drafted the initial manuscript. All authors participated in the interpretation of the results, edited, and reviewed the manuscript. All authors have contributed significantly and approved the submission for publication.

Funding This research was funded by the National Key Research and Development Program (2020YFC2003500). The corresponding author Fuzhong Xue obtained the funding. The funders had no role in this work.

Data availability This research has been conducted using the UK Biobank Resource (application ID: 51,470).

Compliance with ethical standards

Conflict of interest All authors declare no conflict of interests.

Informed consent All participants provided written informed consent.

References

1. Guh DP, Zhang W, Bansback N et al (2009) The incidence of co-morbidities related to obesity and overweight: a systematic review and meta-analysis. BMC Public Health 9:88. https://doi.org/10.1186/1471-2458-9-88
2. Mladovsky P, Allin S, Masseria C, Hernández-Quevedo C, Mossialos E (2009) Health in the European Union: trends and analysis. WHO Regional Office Europe, Denmark
3. Tuomilehto J, Lindström J, Eriksson JG et al (2001) Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance. N Engl J Med 344:1343–1350. https://doi.org/10.1056/nejm200105033441801
4. Almgren P, Lehtovirta M, Isomaa B et al (2011) Heritability and familiarity of type 2 diabetes and related quantitative traits in the Botnia Study. Diabetologia 54:2811–2819. https://doi.org/10.1007/s00125-011-2267-5
5. Pal A, McCarthy MI (2013) The genetics of type 2 diabetes and its clinical relevance. Clin Genet 83:297–306. https://doi.org/10.1111/cge.12055
6. Morris AP, Voight BF, Teslovich TM et al (2012) Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat Genet 44:981–990. https://doi.org/10.1038/ng.2383
7. Voight BF, Scott LJ, Steinthorsdottir V et al (2010) Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. Nat Genet 42:579–589. https://doi.org/10.1038/ng.609
8. Bonnefond A, Froguel P, Vaxillaire M (2010) The emerging genetics of type 2 diabetes. Trends Mol Med 16:407–416. https://doi.org/10.1016/j.molmed.2010.06.004
9. Khera AV, Chaffin M, Aragam KG et al (2018) Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet 50:1219–1224. https://doi.org/10.1038/s41588-018-0183-z
10. Vaxillaire M, Yengo L, Lobbens S et al (2014) Type 2 diabetes-related genetic risk scores associated with variations in fasting plasma glucose and development of impaired glucose homeostasis in the prospective DESIR study. Diabetologia 57:1601–1610. https://doi.org/10.1007/s00125-014-3277-x
11. Abraham G, Malik R, Yonova-Doing E et al (2019) Genomic risk score offers predictive performance comparable to clinical risk factors for ischaemic stroke. Nat Commun 10:5819. https://doi.org/10.1038/s41467-019-13848-1
12. Ming D, Shafqat A, Lu Q et al (2019) Additive and multiplicative interactions between genetic risk score and family history and lifestyle in relation to risk of type 2 diabetes. Am J Epidemiol 189(5):445–460
13. Walford GA, Porneala BC, Dauriz M et al (2014) Metabolite traits and genetic risk provide complementary information for the prediction of future type 2 diabetes. Diabetes Care 37:2508–2514. https://doi.org/10.2337/dc14-0560
14. Qi Q, Li H, Wu Y et al (2010) Combined effects of 17 common genetic variants on type 2 diabetes risk in a Han Chinese population. Diabetologia 53:2163–2166. https://doi.org/10.1007/s00125-010-1826-5
15. Inouye M, Abraham G, Nelson CP et al (2018) Genomic risk prediction of coronary artery disease in 480,000 adults: implications for primary prevention. J Am Coll Cardiol 72:1883–1893. https://doi.org/10.1016/j.jacc.2018.07.079
16. Sudlow C, Gallacher J, Allen N et al (2015) UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med 12:e1001779. https://doi.org/10.1371/journal.pmed.1001779
17. Bycroft C, Freeman C, Petkova D et al (2018) The UK Biobank resource with deep phenotyping and genomic data. Nature 562:203–209. https://doi.org/10.1038/s41586-018-0579-z
18. Wheeler E, Leong A, Liu CT et al (2017) Impact of common genetic determinants of Hemoglobin A1c on type 2 diabetes risk and diagnosis in ancestrally diverse populations: a transethnic genome-wide meta-analysis. PLoS Med 14:e1002383. https://doi.org/10.1371/journal.pmed.1002383
19. Scott RA, Lagou V, Welch RP et al (2012) Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. Nat Genet 44:991–1005. https://doi.org/10.1038/ng.2385
20. Willer CJ, Schmidt EM, Sengupta S et al (2013) Discovery and refinement of loci associated with lipid levels. Nat Genet 45:1274–1283. https://doi.org/10.1038/ng.2797
21. Evangelou E, Warren HR, Mosen-Ansorena D et al (2018) Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits. Nat Genet 50:1412–1425. https://doi.org/10.1038/s41588-018-0205-x
22. Shungin D, Winkler TW, Croteau-Chonka DC et al (2015) New genetic loci link adipose and insulin biology to body fat distribution. Nature 518:187–196. https://doi.org/10.1038/nature14132
23. Locke AE, Kahali B, Berndt SI et al (2015) Genetic studies of body mass index yield new insights for obesity biology. Nature 518:197–206. https://doi.org/10.1038/nature14177
24. Wood AR, Esko T, Yang J et al (2014) Defining the role of common variation in the genomic and biological architecture of adult human height. Nat Genet 46:1173–1186. https://doi.org/10.1038/ng.3097
25. Liu M, Jiang Y, Wedow R et al (2019) Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. Nat Genet 51:237–244. https://doi.org/10.1038/s41588-018-0307-5
26. Chang CC, Chow CC, Tellier LC et al (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4:7. https://doi.org/10.1186/s13742-015-0047-8
27. Bycroft C, Freeman C, Petkova D et al (2017) Genome-wide genetic data on ~500,000 UK Biobank participants. BioRxiv. https://doi.org/10.1101/166298
28. Hui Z, Hastie T (2005) Regularization and variable selection via the elastic net. J Roy Stat Soc 67:301–320
29. Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. J Stat Softw 33:1–22
30. Fine J, Gray R, Jason P (1999) A proportional hazards model for the subdistribution of competing risks in survival analysis. J Am Stat Assoc 94(446):496–509
31. Harrell FE Jr, Califf RM, Pryor DB, Lee KL, Rosati RA (1982) Evaluating the yield of medical tests. JAMA 247:2543–2546
32. Si S, Tewara MA, Ji X et al (2020) Body surface area, height, and body fat percentage as more sensitive risk factors of cancer and cardiovascular disease. Cancer Med 9:4433–4446. https://doi.org/10.1002/cam4.3076
33. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS (2008) Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med 27:157–172. https://doi.org/10.1002/sim.2929

34. Imamura M, Shigemizu D, Tsunoda T et al (2013) Assessing the clinical utility of a genetic risk score constructed using 49 susceptibility alleles for type 2 diabetes in a Japanese population. J Clin Endocrinol Metab 98:E1667–E1673. https://doi.org/10.1210/jc.2013-1642
35. Ridker PM, Buring JE, Rifai N, Cook NR (2007) Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: the reynolds risk score. JAMA 297:611–619. https://doi.org/10.1001/jama.297.6.611
36. Fry A, Littlejohns TJ, Sudlow C et al (2017) Comparison of sociodemographic and health-related characteristics of uk biobank participants with those of the general population. Am J Epidemiol 186:1026–1034. https://doi.org/10.1093/aje/kwx246

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.