Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Real-World Blueprint for Precision Diabetes Care: Integrating Pharmacogenomics into Clinical Decision-Making for A Chinese T2DM Cohort

Ye Liu
Zaozhuang Municipal Hospital
Chun-xiao Yu
chunxiao16@163.com

Zaozhuang Municipal Hospital
Li Lu
Shandong Health Group Zaozhuang Central Hospital

Research Article

Keywords: Pharmacogenomics, Type 2 Diabetes Mellitus, Precision Medicine, Polygenic Risk Score, Oral Antidiabetic Drugs, Chinese Population

Posted Date: March 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8926341/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Aims

To characterize the pharmacogenomic landscape in a Chinese Type 2 Diabetes Mellitus (T2DM) cohort and develop a systematic framework to translate raw genotype data into predicted clinical phenotypes.

Methods

In this retrospective study of 908 Chinese patients, we used a novel 'Interpretation Rulebook' to derive eight predicted phenotypes from seven SNPs. These classifications are genotype-based, not measured clinical outcomes. A weighted polygenic risk score (PRS) for T2DM was constructed and evaluated for its ability to discriminate the genotype-derived risk classification.

Results

Allele frequencies were consistent with East Asian populations. Predefined genetic variants were the primary drivers of their corresponding phenotype classifications, for instance, IRS1 rs1801278 for insulin resistance risk (Adjusted OR = 6450.06) and SLCO1B1 rs4149056 for predicted repaglinide efficacy (OR = 0.00009). The T2DM PRS demonstrated excellent discriminatory power for the genotype-derived risk classification (AUC = 0.990).

Conclusions

This study establishes a transparent and validated framework for standardizing real-world pharmacogenomic data. The high-performance PRS confirms its internal logic, providing a robust blueprint for clinical decision support and implementing pre-emptive pharmacogenomic testing. While acknowledging the need for prospective validation against measured outcomes, this work addresses the translational gap in interpreting heterogeneous genetic data.

1. Introduction

Type 2 diabetes mellitus (T2DM) is a significant and growing global public health challenge, with an estimated 537 million people affected worldwide in 2021. China, in particular, has the largest number of individuals with diabetes, underscoring the urgent need for effective prevention and management strategies. Management is centered on glycemic control through lifestyle interventions and pharmacological therapy [1]. Oral antidiabetic drugs are a cornerstone of this treatment paradigm; however, clinical practice is often marked by significant inter-individual variability in drug efficacy and adverse reactions, motivating the pursuit of pharmacological precision medicine in diabetes [2, 3]. Pharmacogenomics, which studies the influence of genetic variation on drug response, offers a

promising pathway to tailor drug selection and dosage to a patient's unique genetic profile, thereby advancing the goal of precision medicine for T2DM.

A substantial body of evidence has established associations between specific genetic polymorphisms and both T2DM susceptibility and drug response. For disease susceptibility, robust genome-wide association studies have implicated variants in TCF7L2, IRS1, and CDKAL1 among the most reproducible common risk loci [4, 5]. For pharmacogenetics, CYP2C9 polymorphisms alter sulfonylurea metabolism and are associated with variability in glycaemic response and hypoglycaemia risk [6]. Transporter genetics also play a critical role: SLCO1B1 variants markedly influence repaglinide pharmacokinetics—an effect replicated in Han Chinese cohorts—while common variants near ATM have been linked to metformin glycaemic response in large discovery and replication studies [7, 8]. Notably, meta-analytic work indicates that the impact of pharmacokinetic transporter variants on metformin glycaemic response can be modest and context-dependent, underscoring the need for carefully standardized phenotyping [9]. Beyond these loci, TCF7L2 risk alleles have been associated with reduced sulfonylurea effectiveness across multiple studies, consistent with impaired β-cell function and incretin signalling [10].

Despite this wealth of knowledge, the translation of findings from heterogeneous real-world genetic reports into standardized, analyzable phenotypes remains a critical implementation bottleneck. While authoritative resources such as PharmGKB and the Clinical Pharmacogenetics Implementation Consortium (CPIC) provide the essential "evidence base" and high-level clinical recommendations, there remains a distinct "operational gap" in how these guidelines are systematically applied to raw datasets [11]. Current guidelines effectively define what to do when a phenotype is known (e.g., dose reduction for a Poor Metabolizer), but they often lack the deterministic logic required to automatically and consistently batch-process raw genotype data—spanning multiple loci and varying data quality—into discrete, clinically actionable phenotype variables.

Consequently, our proposed framework functions as a translational "middleware." Unlike high-level evidence repositories, our "Interpretation Rulebook" provides the specific, standardized logic layer necessary to convert raw genetic inputs into uniform phenotype codes. This approach mimics the architecture required for Electronic Health Record (EHR)-integrated Clinical Decision Support Systems (CDSS), moving beyond static guidelines to create a dynamic, standardized data model capable of handling the complexities of real-world Chinese cohort data. Furthermore, ancestry-related transferability issues in genetic prediction highlight the importance of generating evidence in underrepresented populations, including East Asians/Chinese [12], to ensure that global guidelines are locally applicable. In parallel, while polygenic risk scores (PRS) have shown promise for stratifying T2DM risk at scale, ensuring equitable performance across ancestries remains an active area of methodological development [13, 14].

To address this gap between genetic knowledge and clinical application, this study was designed to systematically evaluate the pharmacogenomic landscape in a real-world cohort of 908 Chinese patients

with Type 2 diabetes mellitus. To ensure analytical consistency and scientific rigor, we first developed a novel, standardized "Interpretation Rulebook" to convert raw genotype data into eight uniform, predicted clinical phenotypes. Based on this framework, our primary objectives were: (1) to describe the frequency distribution of key genetic polymorphisms; (2) to assess the internal consistency of our "Interpretation Rulebook" by quantifying the contribution of individual SNPs to their corresponding, genotype-derived risk classifications; and (3) to construct and evaluate a Polygenic Risk Score (PRS) to measure the cumulative effect of genetic factors on the composite, rule-based T2DM risk phenotype. Ultimately, we aim to provide a validated methodological framework to support the implementation of genomics-guided precision therapy for diabetes management in the Chinese population, addressing the critical challenge of standardizing heterogeneous real-world genetic data.

2. Methods

2.1 Study design and population

This was a retrospective cohort study analyzing de-identified records from patients who underwent pharmacogenomic testing at the Endocrinology Department of Shandong Health Group Zaozhuang Central Hospital between September 2022 and August 2024. Inclusion criteria required participants to be adults with a clinical diagnosis of T2DM and to have at least one genotyped polymorphism among the pre-specified loci. Patient demographic data (age, sex) and referring department information were extracted directly from the laboratory information system (LIS) associated with the genetic testing service. Patients with incomplete demographic data or indeterminate genotype calls were excluded from the final analysis. The study protocol complied with the Declaration of Helsinki and was approved by the Medical Ethics Committee of Zaozhuang Mining Group Central Hospital (Approval No. 2022-002-001). The requirement for individual patient informed consent was waived for this retrospective analysis of de-identified data. Clinical definitions and treatment targets followed the latest Standards of Care in Diabetes—2025 [15].

2.2 Genotyping, variant curation, and phenotype mapping

Genotyping was conducted in a clinically accredited laboratory using validated platforms. The selected loci were pre-specified based on prior evidence and encompassed disease-risk variants (e.g., TCF7L2, IRS1, CDKAL1) and pharmacogenetic loci relevant to oral antidiabetic therapy (e.g., CYP2C9, SLCO1B1). The framework for evidence appraisal and the development of interpretive rules was aligned with PharmGKB curation standards and informed by guideline frameworks, including the Dutch Pharmacogenetics Working Group (DPWG) [11, 16]. Variant-level quality control excluded SNPs with low call rate or extreme deviation from Hardy–Weinberg equilibrium (HWE); HWE was evaluated using the exact test, which avoids type-I error inflation in modest samples [11].

A central component of our methodology was the transformation of raw genotypes via a pre-specified “Interpretation Rulebook” into eight uniform predicted clinical phenotypes: T2DM Risk, Insulin Resistance Risk, Sulfonylurea (SU) Efficacy, SU Adverse Drug Reaction (ADR) Risk, DPP-4 Inhibitor Efficacy,

Metformin Efficacy, Repaglinide Efficacy, and Rosiglitazone Efficacy. This rule-based mapping was grounded in established pharmacogenomic associations. For example, genotypes indicating decreased CYP2C9 function were mapped to an increased SU-related ADR risk; variation in the SLCO1B1 transporter gene informed the predicted response to repaglinide; and TCF7L2 risk alleles were treated as markers of reduced insulin secretory capacity [17].

2.3 Definition of predicted clinical phenotypes

The primary outcomes of this study were eight "predicted clinical phenotypes," which were derived directly from participant genotype data using the pre-specified "Interpretation Rulebook." These outcomes were defined as binary or ordered categorical variables and included: (i) genetic risk of Type 2 diabetes mellitus (T2DM) and insulin resistance; (ii) predicted efficacy for commonly used oral agents (sulfonylureas (SU), DPP-4 inhibitors, metformin, repaglinide, rosiglitazone); and (iii) SU-related adverse drug reaction (ADR) risk.

It is essential to clarify that these phenotypes represent standardized, rule-based classifications rather than measured clinical outcomes. Consequently, the primary goal of the subsequent association analyses was not to discover novel biological links, but to validate the logic and internal consistency of our classification framework.

The "Interpretation Rulebook" operationalizes the clinical interpretation for each assayed SNP and drug-response locus, based on evidence curated from established resources like PharmGKB [17]. For illustration, the TCF7L2 rs7903146 T allele—linked to reduced insulin secretion in prior studies—was mapped to an “Increased” T2DM risk category; decreased-function CYP2C9 genotypes were mapped to “Increased” SU-ADR risk; and SLCO1B1 variation informed the predicted response for repaglinide. Phenotypes were coded on a standardized ordinal scale designed to function as a logic layer for clinical decision support. The classifications were defined based on the deviation from standard clinical expectations: "Normal" (Reference) indicates that the patient's genetic profile supports standard-of-care management (e.g., standard dosing, routine monitoring), whereas "Increased" or "Reduced" (Risk/Efficacy) indicates a meaningful deviation from the standard trajectory. Clinically, these non-normal classifications serve as flags for potential action, such as enhanced monitoring for adverse reactions or the consideration of alternative agents. Composite risks (e.g., T2DM_Risk) were assigned the highest risk level observed across contributing loci for a given patient. The final, summarized genotype-to-phenotype specifications are presented in Table 1. To ensure full transparency, a comprehensive breakdown is provided in Supplementary Table S1, which details the clinical rationale for each ordinal level and the grading of supporting evidence derived from PharmGKB guidelines.

Table 1
The interpretation rulebook for genotype-to-phenotype mapping

Gene	Locus (rsID)	Genotype	Phenotype Analyzed	Phenotype Code
TCF7L2	rs7903146	CC	T2DM_Risk	Normal
		CT	T2DM_Risk	Increased
		TT	T2DM_Risk	Increased
		CC	SU_Efficacy	Normal
		CT	SU_Efficacy	Reduced
		TT	SU_Efficacy	Reduced
IRS1	rs1801278	CC	T2DM_Risk	Normal
		CT	T2DM_Risk	Increased
		TT	T2DM_Risk	Increased
		CC	Insulin_Resistance_Risk	Normal
		CT	Insulin_Resistance_Risk	Increased
		TT	Insulin_Resistance_Risk	Increased
		CC	SU_Efficacy	Normal
		CT	SU_Efficacy	Reduced
		TT	SU_Efficacy	Reduced
		CC	Metformin_Efficacy	Normal
		CT	Metformin_Efficacy	Reduced
		TT	Metformin_Efficacy	Reduced
CDKAL1	rs7756992	AA	T2DM_Risk	Normal
		AG	T2DM_Risk	Increased
		GG	T2DM_Risk	Increased
		AA	DPP4i_Efficacy	Reduced
		AG	DPP4i_Efficacy	Normal
		GG	DPP4i_Efficacy	Increased
CDKAL1	rs7754840	GG	T2DM_Risk	Normal
		CG	T2DM_Risk	Increased
		CC	T2DM_Risk	Increased

Gene	Locus (rsID)	Genotype	Phenotype Analyzed	Phenotype Code
		GG	DPP4i_Efficacy	Reduced
		CG	DPP4i_Efficacy	Normal
		CC	DPP4i_Efficacy	Increased
CYP2C9	rs1057910	*1/*1 (AA)	SU_ADR_Risk	Normal
		*1/*3 (AC)	SU_ADR_Risk	Increased
		*3/*3 (CC)	SU_ADR_Risk	Increased
SLCO1B1	rs4149056	TT	Repaglinide_Efficacy	Normal
		CT	Repaglinide_Efficacy	Increased
		CC	Repaglinide_Efficacy	Increased
		TT	Rosiglitazone_Efficacy	Normal
		CT	Rosiglitazone_Efficacy	Increased
		CC	Rosiglitazone_Efficacy	Increased

2.4 Statistical analysis

All statistical analyses were performed using Python (version 3.11.4). Logistic regression models were fitted using the statsmodels library (v0.14.0), and predictive performance metrics were evaluated using scikit-learn (v1.3.0). A two-sided P-value < 0.05 was considered statistically significant throughout all analyses.

2.4.1 Descriptive analysis and quality control

Baseline characteristics of the cohort were summarized using descriptive statistics; continuous variables (age) were presented as mean ± standard deviation (SD), while categorical variables (gender, genotypes, clinical phenotypes) were presented as counts and percentages (n, %). Allele and genotype frequencies were calculated by direct counting. As a quality control measure, the conformance of genotype frequencies for each SNP to Hardy-Weinberg equilibrium (HWE) was assessed using a Chi-squared test.

2.4.2 Genotype-phenotype association analysis

The association between individual genetic polymorphisms and the derived clinical phenotypes was evaluated in a stepwise manner. First, unadjusted analyses were performed using Chi-squared or Fisher's exact tests, as appropriate. To assess the independent effect of each polymorphism while adjusting for potential confounders (age and sex), multivariable regression models were constructed. Ordinal logistic regression was employed for phenotypes with three or more ordered levels (e.g., Reduced, Normal,

Increased), while binary logistic regression was used for dichotomous outcomes. Results are presented as odds ratios (ORs) with their corresponding 95% confidence intervals (CIs). Potential gene-gene interactions were explored by including interaction terms in the regression models. All multivariable analyses were conducted using a complete-case approach, excluding patients with missing data for any variable in a given model.

2.4.3 Polygenic risk score (PRS) construction and evaluation

To investigate the cumulative consistency of the rule-based classification, a weighted polygenic risk score (PRS) for T2DM was constructed for each patient. Methodologically, study-derived weights obtained from the coefficients (log-odds) of the multivariable logistic regression model described in Section 2.4.2 were utilized. This approach was selected to validate the internal fidelity of the "Interpretation Rulebook" logic rather than to validate external biological associations.

Genotypes for the T2DM-associated loci (TCF7L2, IRS1, and CDKAL1) were coded additively (0, 1, 2) and weighted by their respective study-derived coefficients. The resulting scores were normalized to a standard scale (Z-score) to facilitate comparison. Allele frequencies were cross-referenced with gnomAD East Asian (EAS) data to ensure ancestry consistency (detailed in Supplementary Table S2). Notably, the dominant contribution of low-frequency variants such as TCF7L2 to the score is a direct consequence of this internal validation design, reflecting the high deterministic weight assigned to this locus in the risk classification rules (as defined in Supplementary Table S1). The predictive utility of the PRS for the composite T2DM_Risk phenotype was quantified using logistic regression, and its discriminatory power was evaluated by calculating the Area Under the Receiver Operating Characteristic Curve (AUC).

3. Results

3.1 Cohort characteristics and genetic profile

The final analysis included 908 patients with a clinical diagnosis of T2DM. The mean age of the cohort was 59.1 ± 12.6 years, and the population was predominantly male (n = 575, 63.3%). Detailed demographic characteristics are presented in FIGURE 1.

The allele frequencies and Hardy-Weinberg equilibrium (HWE) test results for the analyzed SNPs are detailed in Supplementary Table S2. The minor allele frequencies (MAFs) for CYP2C9 rs1057910 (5.0%), TCF7L2 rs7903146 (4.4%), and IRS1 rs1801278 (2.1%) were low. In contrast, the allele frequencies for the two CDKAL1 loci were more balanced (rs7754840 G-allele: 48.7%; rs7756992 A-allele: 44.1%). As a quality control measure, all successfully genotyped SNPs were tested for HWE. The genotype distributions for five of the six primary SNPs were in Hardy-Weinberg equilibrium (HWE) (all $P > 0.05$). A borderline deviation was noted for CDKAL1 rs7754840 (P = 0.048); however, given that CDKAL1 is a validated susceptibility locus for T2DM, this slight deviation is likely attributable to the selection bias

inherent in a case-only diabetes cohort rather than genotyping error. The FGG rs2066865 locus was excluded from all association analyses due to a small sample size (n = 4) that rendered its HWE test statistically unreliable. These genetic characteristics are visualized in FIGURE 2.

3.2 Internal validation of the genotype-phenotype mapping logic

To assess the fidelity of our translation framework, we performed a multivariable logistic regression analysis. It is critical to note that this analysis was designed solely for internal validation—to confirm that the "Interpretation Rulebook" was consistently applied to the raw data—rather than to estimate biological effect sizes.

The results (Supplementary Table S3 and FIGURE 3) confirm that the predefined genetic variants were the statistical drivers of their corresponding phenotype classifications. For instance, the IRS1 rs1801278 variant acted as a deterministic predictor for the insulin resistance classification, yielding an adjusted Odds Ratio (OR) of > 6000 ($P < 0.001$). Similarly, strong statistical signals were observed for CDKAL1 variants regarding T2DM risk and SLCO1B1 variants regarding predicted repaglinide efficacy.

We explicitly caution against misinterpreting these extreme ORs as measures of biological risk. These values are indicative of quasi-complete separation, a statistical phenomenon that occurs because the predictor (genotype) deterministically defines the outcome (predicted phenotype) based on our pre-specified rulebook. Consequently, the magnitude of these ORs reflects the rigidity of our coding definitions rather than the strength of a physiological association. These findings validate the internal consistency of our data processing pipeline, demonstrating that the translation from genotype to clinical phenotype was executed without error.

3.3 T2DM polygenic risk score (PRS) analysis

A weighted PRS was constructed for all 908 patients using T2DM-associated loci (TCF7L2, IRS1, and CDKAL1). Detailed results of this analysis are presented in Supplementary Table S4, with key findings summarized in FIGURE 4.

The mean PRS for the cohort was 2.23 ± 1.58. The PRS was not correlated with patient age (Pearson's r = -0.06, P = 0.073). However, female patients exhibited a significantly higher mean PRS compared to male patients (2.48 vs. 2.08, $P < 0.001$).

The primary validation of the PRS was its strong association with the predicted T2DM_Risk phenotype. Each one-unit increase in the PRS corresponded to a 373.28-fold increase in the odds of being in an increased-risk classification (OR = 373.28; 95% CI, 130.54–1067.40; P < 0.001). Furthermore, the score demonstrated outstanding discriminatory power for this classification, achieving an Area Under the Receiver Operating Characteristic Curve (AUC) of 0.990. An analysis of the components of the score showed that TCF7L2 had the largest relative contribution to the PRS, followed by CDKAL1 and IRS1.

3.4 Exploratory gene-gene interaction analysis

To explore potential epistatic or synergistic effects on the predicted clinical phenotypes, eight biologically-driven gene-gene interaction hypotheses were evaluated. Due to insufficient outcome variation or model convergence issues in certain subgroups, four interaction models were successfully evaluated. The comprehensive results are presented in Supplementary Table S5.

Among the tested pairs, a potential synergistic interaction was observed between the two loci within the CDKAL1 gene (rs7756992 and rs7754840) in modulating the T2DM_Risk phenotype. The inclusion of the interaction term improved the logistic regression model fit compared to a main-effects-only model (Likelihood Ratio Test $P = 0.005$), yielding an interaction Odds Ratio of 8.21. However, these findings must be interpreted with caution as exploratory data.

A joint genotype contingency table (Supplementary Table S6) reveals that sample sizes for specific compound genotypes were limited; for instance, only 2 patients carried the rs7756992 AA / rs7754840 CC combination, and only 7 carried the double wild-type (GG/GG). This sparsity suggests that the inflated effect estimates may be partially driven by quasi-complete separation in these high-risk subgroups. Consequently, while the non-parallel lines in Supplementary Figure S1 suggest a synergistic effect, this interaction requires replication in larger cohorts to confirm its biological stability. No significant interactions were detected for the other evaluated pairs.

4. Discussion

The primary contribution of this study is not the discovery of novel gene-phenotype associations, but rather the development and validation of a methodological framework to address a critical bottleneck in translational pharmacogenomics: the standardization of heterogeneous, real-world genetic data. Our "Interpretation Rulebook" provides a systematic, evidence-based blueprint for converting raw genotypes into uniform, clinically relevant—albeit predicted—phenotypes. The exceptionally strong statistical signals and the high-performance PRS reported herein should be viewed through this lens; they serve as a robust validation of our framework's internal logic and consistency, confirming that the model effectively translates genetic information into risk strata as intended. This foundational work paves the way for future prospective studies to test whether these genotype-derived classifications correlate with actual, measured clinical outcomes.

The development of a high-performing PRS for a Chinese cohort aligns with the broader push towards ancestry-aware genetic risk stratification. Recent multi-ancestry analyses have highlighted that the genetic architecture of T2DM varies across populations, with East Asians showing distinct mechanistic clusters [18, 19]. This contextualizes our strong PRS performance and underscores the importance of using ancestry-matched models, as cross-ancestry PRSs often perform poorly [20–22]. Furthermore, our observation that the effect of the canonical TCF7L2 rs7903146 variant was muted is biologically plausible, as its risk allele is rare in East Asian populations (≈ 2%) compared to European populations (≈

30%) [19, 23]. This finding reinforces the need to prioritize loci with higher allele frequencies and proven effects in East Asian populations when constructing predictive models for this group.

The framework for predicting drug response is firmly grounded in the maturing pharmacogenomic literature, yet its primary value lies in translating these genetic associations into actionable clinical decision support. For instance, the mapping of reduced-function CYP2C9 alleles to higher sulfonylurea-induced hypoglycaemia risk [24] suggests a concrete clinical action: initiating sulfonylureas at a lower dose or prioritizing alternative agents (e.g., metformin or DPP-4 inhibitors) to mitigate risk. Similarly, the identification of SLCO1B1 variants associated with altered repaglinide exposure [25] signals a need for enhanced toxicity monitoring rather than absolute contraindication. The inclusion of a CDKAL1-modulated response to DPP-4 inhibitors, supported by findings from East Asian GWAS [26, 27], further strengthens the rationale for its application in this specific population. However, implementing such predictive tools necessitates a critical assessment of misclassification risks. A genotype-predicted "Normal" phenotype does not guarantee therapeutic safety (false negative), nor does a "Reduced Efficacy" prediction definitively preclude clinical benefit (false positive). Over-reliance on these algorithmic predictions could potentially lead to the premature exclusion of viable therapeutic options or a false sense of security regarding adverse events. Therefore, these genotype-derived phenotypes are best used as probabilistic flags to guide monitoring intensity and dosage titration strategies, complementing, rather than substituting for, standard clinical judgment and measured metabolic outcomes.

Regarding gene-gene interactions, the observation of potential intragenic epistasis within the CDKAL1 gene offers an exploratory insight. While the interaction between rs7756992 and rs7754840 appeared to modulate the predicted T2DM risk, these results must be interpreted with caution. As noted in the contingency analysis (Supplementary Table S6), the sample sizes for specific compound genotypes—such as the double homozygote carriers—were extremely limited. This sparsity can result in unstable effect estimates. Consequently, while this finding aligns with a growing recognition of the importance of intra-gene interactions in complex traits [28–30], it is positioned here as hypothesis-generating, motivating the need for prospective validation with pre-specified interaction tests in larger, independent cohorts to confirm its biological relevance.

This study has several notable strengths, including its novel methodological framework, its validation in a real-world clinical cohort, and its specific focus on the underrepresented Chinese population. However, several limitations must be acknowledged. First and foremost, the retrospective design precludes any inference of causality. The most critical limitation is that our primary outcomes were predicted phenotypes, not measured clinical endpoints. Therefore, while our results validate the internal logic of the classification system, they do not yet establish its clinical utility. Additionally, our PRS was constructed from a limited set of established SNPs and is not a comprehensive genome-wide score. Finally, as a single-center study, our findings may have limited generalizability, and the potential for unmeasured confounding variables cannot be entirely excluded.

5. Conclusion

This study establishes and validates a novel, systematic framework for standardizing real-world pharmacogenomic data within a Chinese Type 2 Diabetes Mellitus cohort. The high performance of the resulting polygenic risk score and the strong internal consistency of the genotype-phenotype classifications confirm that our "Interpretation Rulebook" provides a transparent and robust blueprint for translating raw genetic data into clinically relevant, predicted phenotypes. While the reliance on predicted rather than measured clinical outcomes serves as a primary limitation, this work addresses a critical bottleneck in translational pharmacogenomics. By explicitly defining the logic for clinical decision support and demonstrating the framework's internal fidelity, this study offers a foundational model for implementing pre-emptive, genomics-guided precision therapy. Future prospective trials are now required to validate these genotype-derived predictions against measured patient outcomes to fully establish their clinical utility in personalized diabetes care.

Declarations

Conflict of Interest

The authors declare no conflict of interest.

Ethics Approval and Consent to Participate

This study was approved by the Medical Ethics Committee of Zaozhuang Mining Group Central Hospital (Approval No. 2022-002-001). The requirement for individual informed consent was waived by the Medical Ethics Committee of Zaozhuang Mining Group Central Hospital due to the retrospective nature of the study and the use of de-identified data.

Author Contributions

Ye Liu: Conceptualization, Writing – Original Draft, Writing – Review & Editing. Li Lu: Data Curation, Investigation. Chun-xiao Yu: Supervision, Writing – Review & Editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used Gemini in order to improve readability and language. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Data availability statement

The original genotype and phenotype datasets analyzed during the current study are included in the article and its Supplementary Material. The comprehensive "Interpretation Rulebook" logic is fully detailed in Supplementary Table S1. To facilitate reproducibility, the custom Python source code developed for genotype-to-phenotype mapping, Polygenic Risk Score (PRS) calculation, and statistical interaction analysis has been deposited in the Zenodo repository and is openly accessible at https://doi.org/10.5281/zenodo.17940117. Further inquiries can be directed to the corresponding author.

References

1. Sun H et al. *IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045.* 2022. 183: p. 109119.
2. Florez JC, Pearson ERJD. roadmap achieve Pharmacol precision Med diabetes. 2022;65(11):1830–8.
3. Gloyn AL. D.J.J.T.I.D. Drucker, and endocrinology, *Precision medicine in the management of type 2 diabetes*. 2018. 6(11): pp. 891–900.
4. Grant SF et al. Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. 2006. 38(3): pp. 320–3.
5. Rung J et al. Genetic variant near IRS1 is associated with type 2 diabetes, insulin resistance and hyperinsulinemia. 2009. 41(10): pp. 1110–5.
6. Zhou K et al. Loss-of-function CYP2C9 variants improve therapeutic response to sulfonylureas in type 2 diabetes: a Go-DARTS study. 2010. 87(1): pp. 52–6.
7. Van Leeuwen N et al. A gene variant near ATM is significantly associated with metformin treatment response in type 2 diabetes: a replication and meta-analysis of five cohorts. 2012. 55(7): pp. 1971–7.
8. He J et al. *Effects of SLCO1B1 polymorphisms on the pharmacokinetics and pharmacodynamics of repaglinide in healthy Chinese volunteers.* 2011. 67(7): pp. 701–707.
9. Dujic T et al. Variants in pharmacokinetic transporters and glycemic response to metformin: a metgen meta-analysis. 2017. 101(6): pp. 763–72.
10. Zimdahl H et al. Influence of TCF7L2 gene variants on the therapeutic response to the dipeptidylpeptidase-4 inhibitor linagliptin. 2014. 57(9): pp. 1869–75.
11. Whirl-Carrillo M et al. *Pharmacogenomics knowledge for personalized medicine.* 2012. 92(4): pp. 414–417.
12. Martin AR et al. *Clinical use of current polygenic risk scores may exacerbate health disparities.* 2019. 51(4): pp. 584–591.
13. Khera AV et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. 2018. 50(9): pp. 1219–24.

14. Udler MS et al. Genetic risk scores for diabetes diagnosis and precision medicine. 2019. 40(6): pp. 1500–20.
15. *2. Diagnosis and classification of diabetes: standards of care in diabetes—2025%J Diabetes Care.* 2025. 48(Supplement_1): pp. S27-S49.
16. Wolthuis DF et al. Dutch Pharmacogenetics Working Group (DPWG) guideline for the gene-drug interaction between SLCO1B1 and statins and CYP2C9 and sulfonylureas. 2025. 33(4): pp. 413–20.
17. Whirl-Carrillo M et al. An evidence-based framework for evaluating pharmacogenomics knowledge for personalized medicine. 2021. 110(3): pp. 563–72.
18. Smith K et al. Multi-ancestry polygenic mechanisms of type 2 diabetes. 2024. 30(4): pp. 1065–74.
19. Suzuki K et al. Genetic drivers of heterogeneity in type 2 diabetes pathophysiology. 2024. 627(8003): pp. 347–57.
20. Lennon NJ et al. Selection, optimization and validation of ten chronic disease polygenic risk scores for clinical implementation in diverse US populations. 2024. 30(2): pp. 480–7.
21. Furukawa T et al. *Comparison of polygenic risk scores for type 2 diabetes developed from different ancestry groups.* 2025. 3(1): p. 17.
22. Song M, Kwak SH, Kim JJSR. Risk prediction and interaction analysis using polygenic risk score of type 2 diabetes. Korean Popul. 2024;14(1):6790.
23. Fang C et al. Impaired glucolipid metabolism in gestational diabetes mellitus with T variation of TCF7L2 rs7903146: A case–control study. 2024. 44(1): pp. 182–9.
24. Yee J et al. Association between the CYP2C9 genotype and hypoglycemia among patients with type 2 diabetes receiving sulfonylurea treatment: A meta-analysis. 2021. 43(5): pp. 836–43. e4.
25. Kalliokoski A et al. The effect of SLCO1B1 polymorphism on repaglinide pharmacokinetics persists over a wide dose range. 2008. 66(6): pp. 818–25.
26. Dawed AY et al. CYP2C8 and SLCO1B1 variants and therapeutic response to thiazolidinediones in patients with type 2 diabetes. 2016. 39(11): pp. 1902–8.
27. Liao Y, et al. Past, present and future of Industry 4.0-a systematic literature review and research agenda proposal. Int J Prod Res. 2017;55(12):3609–29.
28. Balvert M et al. *Considerations in the search for epistasis.* 2024. 25(1): p. 296.
29. Tang D, Freudenberg J, A.J.T.A.J.o.H G, Dahl. Factorizing polygenic epistasis improves prediction uncovers Biol pathways complex traits. 2023;110(11):1875–87.
30. Lappalainen T, et al. Genetic Mol Archit complex traits. 2024;187(5):1059–75.
31. Guan Y, et al. Correlation of the TCF7L2 (rs7903146) polymorphism with an enhanced risk of type 2 diabetes mellitus: a meta-analysis. Genet Mol Res. 2016;15(3):10–4238.
32. Zhou WL, Wu TM, Chen JS. Research Progress on the Type 2 Diabetes Mellitus Susceptibility Gene TCF7L2. Med Recapitulate. 2018;24(12):6. (In Chinese).
33. Yousef AA, et al. IRS-1 genetic polymorphism (r. 2963G > A) in type 2 diabetes mellitus patients associated with insulin resistance. The application of clinical genetics; 2018. pp. 99–106.

34. Tuerxunyiming M, et al. Association of rs7754840 G/C polymorphisms in CDKAL1 with type 2 diabetes: a meta-analysis of 70141 subjects. Int J Clin Exp Med. 2015;8(10):17392.

Figures

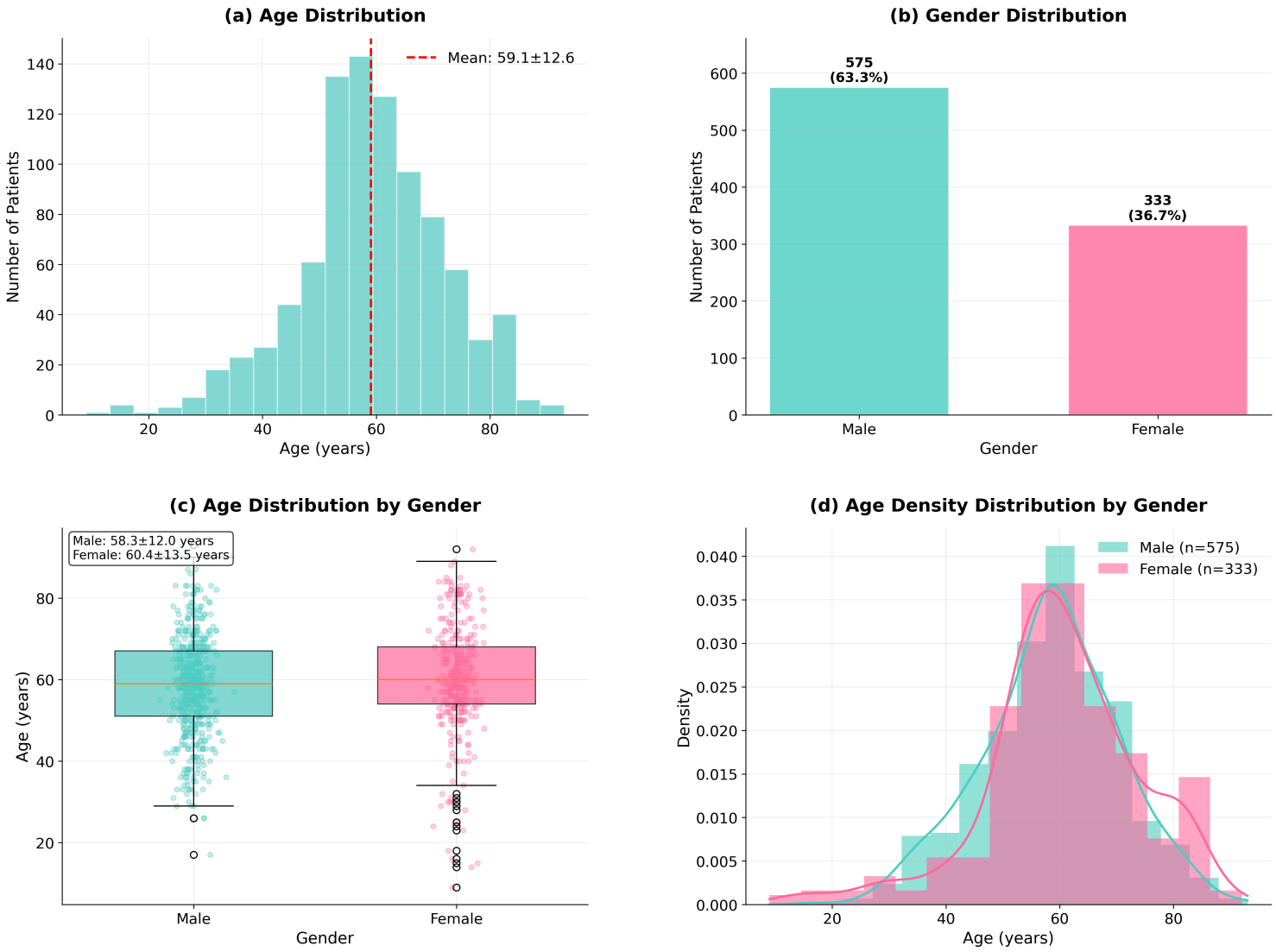


Figure 1

Demographic characteristics of the T2DM study cohort. The visualizations illustrate the age and gender distribution of the 908 Chinese patients included in the retrospective analysis. The cohort exhibits a male predominance (63.3%) and a mean age of 59.1 ± 12.6 years, characterizing the dataset as a representative real-world baseline for the pharmacogenomic evaluation.

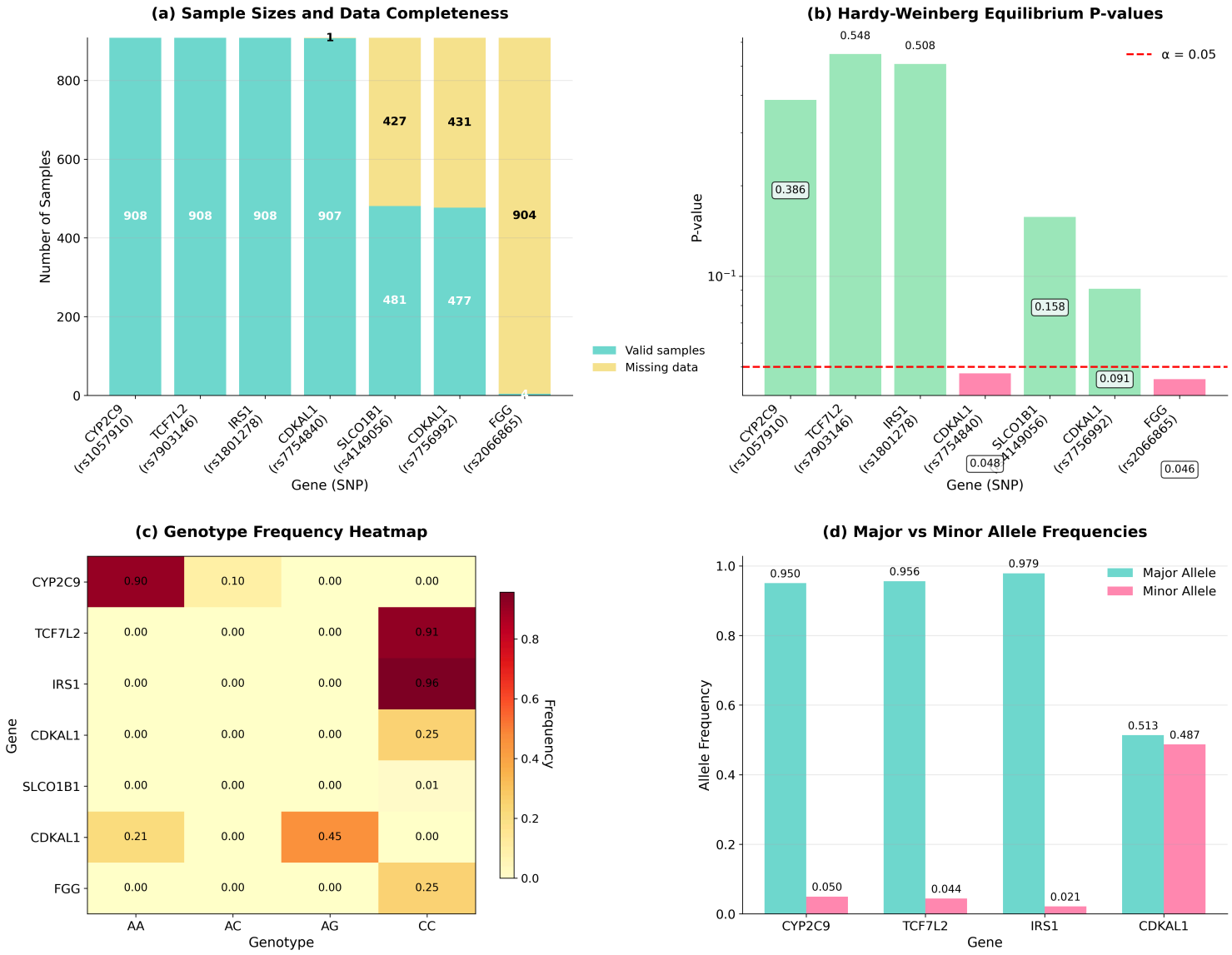


Figure 2

Genetic characteristics of the analyzed polymorphisms. (a) Sample size sufficiency and data completeness for the seven analyzed loci. (b) Hardy-Weinberg Equilibrium (HWE) P-values confirm high-quality genotyping for the majority of variants. (d) Allele frequency analysis highlights the distinct genetic architecture of this cohort, particularly the low minor allele frequency of the canonical TCF7L2 risk variant (~4%) compared to European populations, underscoring the scientific necessity for the ancestry-specific risk modeling approach utilized in this study.

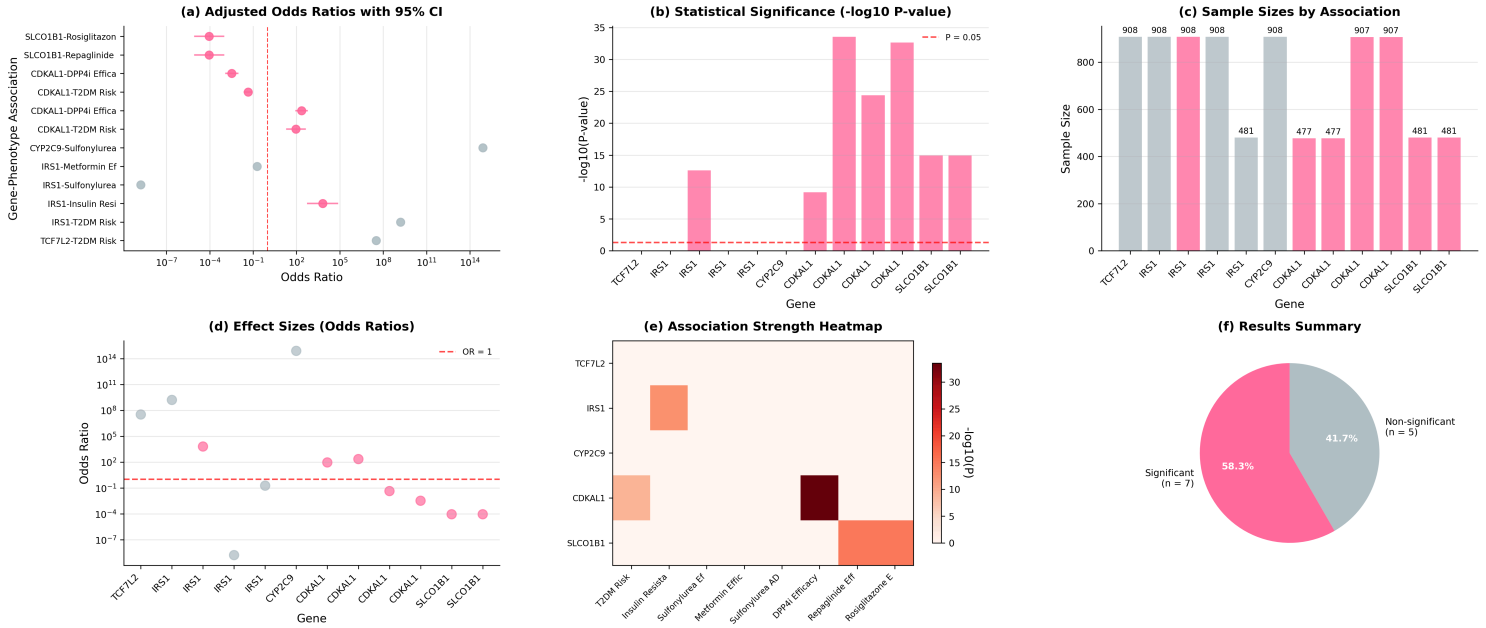


Figure 3

Internal validation of the phenotype mapping logic. The forest plot displays the adjusted Odds Ratios (log scale) for genetic variants predicting their corresponding clinical phenotypes. The extremely high ORs observed (e.g., for IRS1) are expected statistical artifacts that demonstrate the deterministic nature of the "Interpretation Rulebook," confirming that the translation logic was applied consistently and accurately across the cohort.

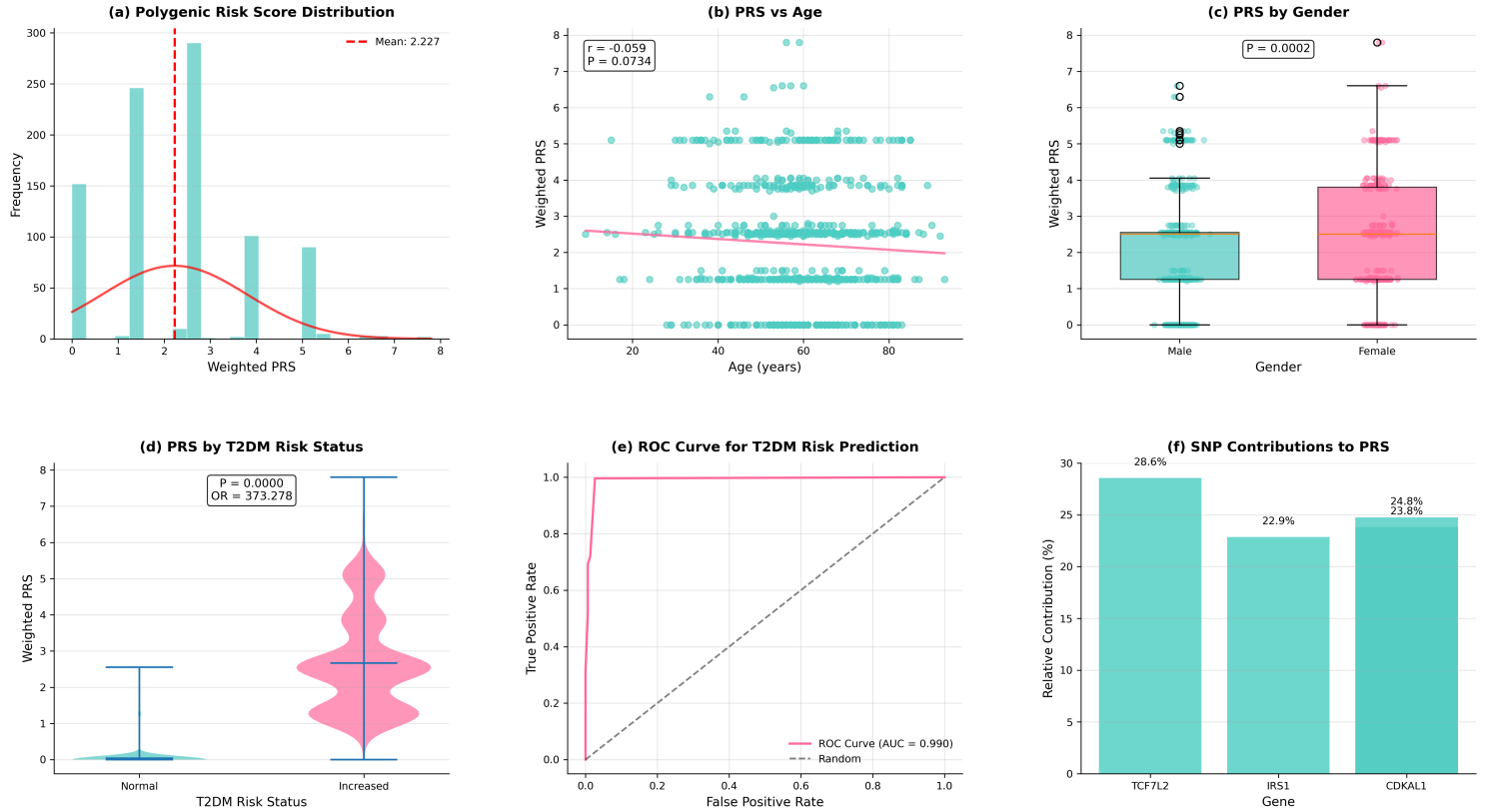


Figure 4

Evaluation of the Polygenic Risk Score (PRS) for T2DM classification. (a) Distribution of weighted PRS in the cohort. (b-d) Associations with age, gender, and risk status. (e) Receiver Operating Characteristic (ROC) curve showing near-perfect discrimination (AUC=0.990). This performance metric confirms that the PRS effectively captures the aggregate risk defined by the study's internal rulebook, validating the cumulative consistency of the risk stratification model.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarytable.docx