Supplemental Methods

GWAS QC procedures

The TWB chip was used to genotype samples from the Taiwan Biobank. This chip is a customized Affymetrix Axiom Genome-Wide Array, encompassing approximately 653,000 single nucleotide polymorphisms (SNPs) [1]. A rigorous quality control (QC) process was conducted at both the SNP and individual levels. SNPs exhibiting call rates less than 95% and possessing Hardy-Weinberg Equilibrium p-values $< 10^{-4}$ were eliminated from the analysis. Similarly, samples demonstrating call rates below 95%, those that were duplicates (PLINK [2] pi_hat statistics exceeding 0.9), or those that failed to pass the PLINK sex check (specifically, inconsistencies between self-reported and actual biological sex) were also excluded. Additional criteria for exclusion included potential sample contamination, which was identified when the median of the PLINK pi_hat statistics of a sample with all other samples > 0.05. Finally, PRIMUS software [3] was employed to identify and retain the largest possible set of unrelated individuals by utilizing a PLINK pi_hat threshold of 0.1, thereby eliminating first-degree relatives.

Supplemental Figures

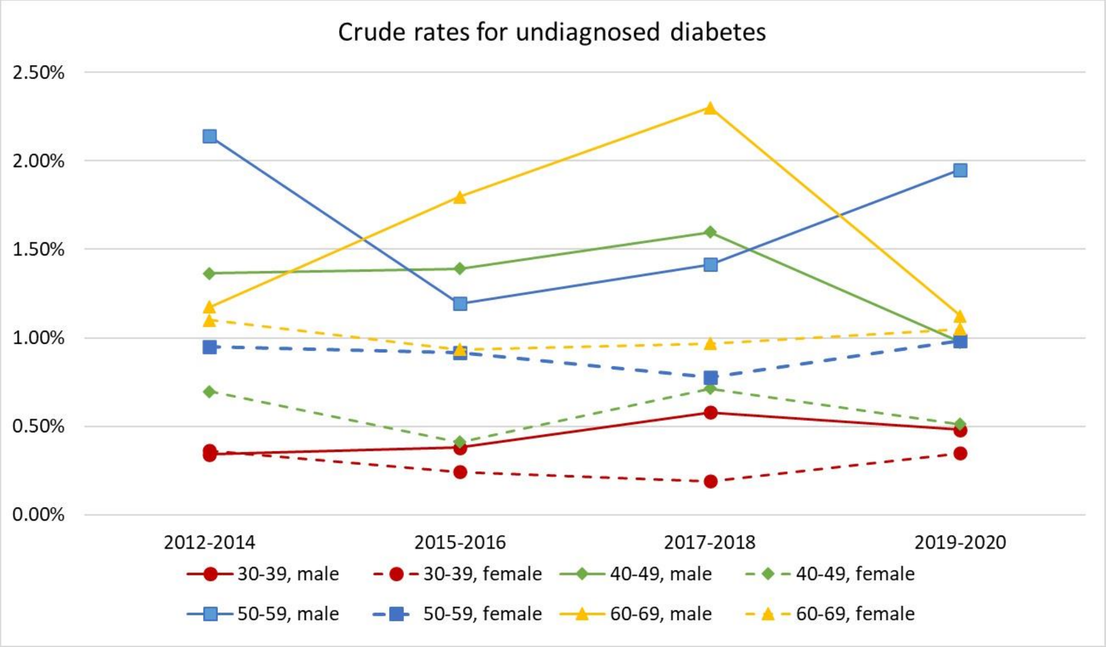


Figure S1. The age- and sex-specific crude rates for undiagnosed diabetes over years

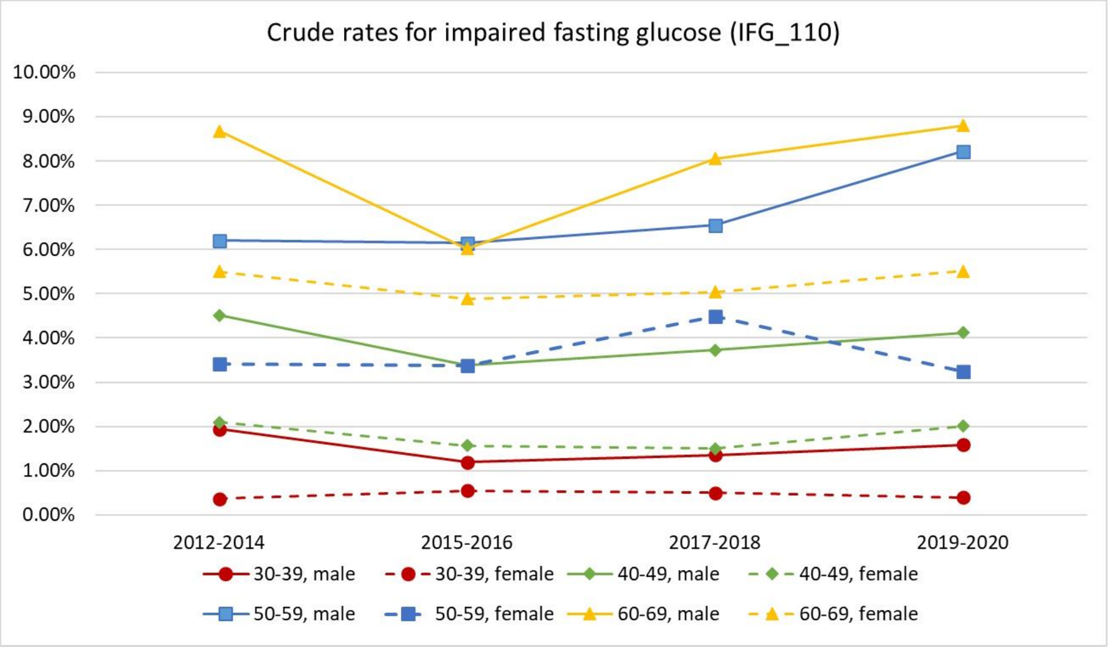


Figure S2. The age- and sex-specific crude rates for IFG_110 (fasting glucose between 110 and 125 mg/dl) over years

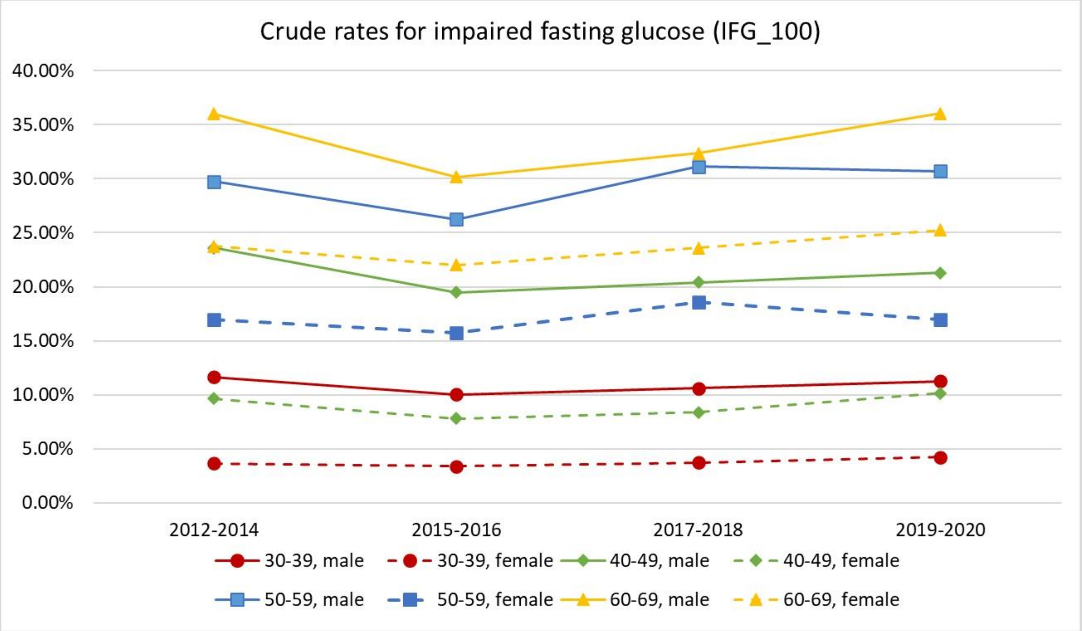
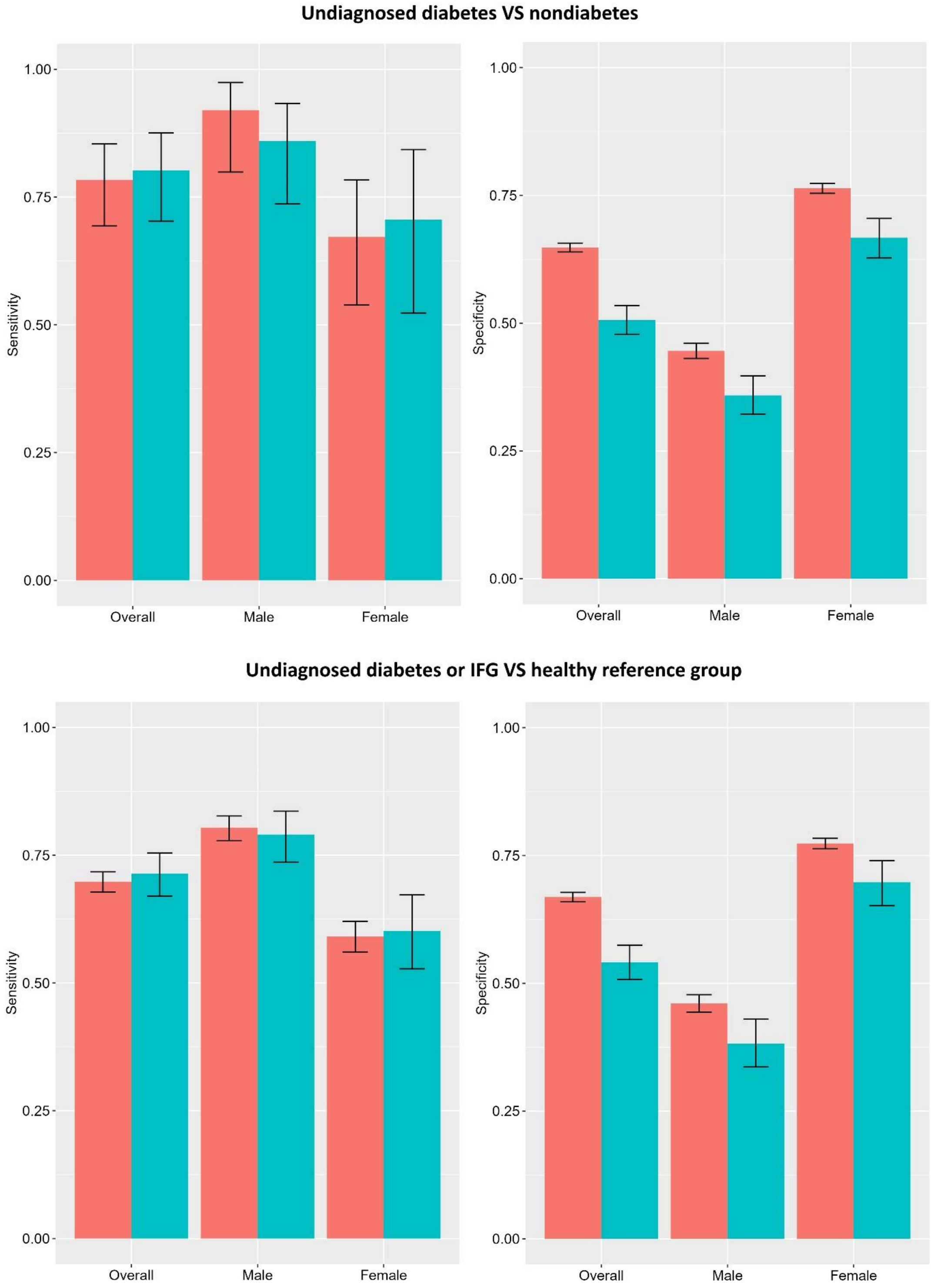


Figure S3. The age- and sex-specific crudes rate for IFG_100 (fasting glucose between 100 and 125 mg/dl) over years



Undiagnosed diabetes VS nondiabetes

Undiagnosed diabetes or IFG VS healthy reference group

Figure S4. Sensitivities and specificities for Model 2. The upper section of the figure illustrates the sensitivities (left) and specificities (right) of the model in predicting undiagnosed diabetes versus nondiabetes. The lower section of the figure shows the

sensitivities (left) and specificities (right) for predicting undiagnosed diabetes or IFG_100 (fasting glucose between 100 and 125 mg/dl) versus healthy reference group. The sensitivities and specificities were calculated in the overall, male, and female samples in the Taiwan Biobank testing dataset (TWB) and the external CVDFACTS validation dataset. The 95% confidence intervals are depicted as error bars in the figure. The results for TWB and CVDFACTS are represented in orange and green bars, respectively.

References

[1] Chen CH, Yang JH, Chiang CWK, et al. (2016) Population structure of Han Chinese in the modern Taiwanese population based on 10,000 participants in the Taiwan Biobank project. Human molecular genetics 25(24): 5321-5331. 10.1093/hmg/ddw346

[2] Purcell S, Neale B, Todd-Brown K, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics 81(3): 559-575. 10.1086/519795

[3] Staples J, Nickerson DA, Below JE (2013) Utilizing graph theory to select the largest set of unrelated individuals for genetic analysis. Genetic epidemiology 37(2): 136-141. 10.1002/gepi.21684