Comparisons of Polyexposure, Polygenic, and Clinical Risk Scores in Risk Prediction of Type 2 Diabetes

Yixuan He,[1,2] Chirag M. Lakhani,[2] Danielle Rasooly,[2,3] Arjun K. Manrai,[2,3] Ioanna Tzoulaki,[4,5] and Chirag J. Patel[2]

Diabetes Care 2021;44:935–943 | https://doi.org/10.2337/dc20-2049

OBJECTIVE

To establish a polyexposure score (PXS) for type 2 diabetes (T2D) incorporating 12 nongenetic exposures and examine whether a PXS and/or a polygenic risk score (PGS) improves diabetes prediction beyond traditional clinical risk factors.

RESEARCH DESIGN AND METHODS

We identified 356,621 unrelated individuals from the UK Biobank of White British ancestry with no prior diagnosis of T2D and normal HbA_{1c} levels. Using self-reported and hospital admission information, we deployed a machine learning procedure to select the most predictive and robust factors out of 111 nongenetically ascertained exposure and lifestyle variables for the PXS in prospective T2D. We computed the clinical risk score (CRS) and PGS by taking a weighted sum of eight established clinical risk factors and >6 million single nucleotide polymorphisms, respectively.

RESULTS

In the study population, 7,513 had incident T2D. The C-statistics for the PGS, PXS, and CRS models were 0.709, 0.762, and 0.839, respectively. Individuals in the top 10% of PGS, PXS, and CRS had 2.00-, 5.90-, and 9.97-fold greater risk, respectively, compared to the remaining population. Addition of PGS and PXS to CRS improved T2D classification accuracy, with a continuous net reclassification index of 15.2% and 30.1% for cases, respectively, and 7.3% and 16.9% for controls, respectively.

CONCLUSIONS

For T2D, the PXS provides modest incremental predictive value over established clinical risk factors. However, the concept of PXS merits further consideration in T2D risk stratification and is likely to be useful in other chronic disease risk prediction models.

Type 2 diabetes (T2D) is a chronic condition that results in impaired well-being for individuals with both genetic and environmental contributors (1). Genome-wide association studies (GWAS) have unveiled important genetic determinants and novel biological pathways in T2D, but their use in prediction is of debate (2). Recent studies have shown that genetic information in the form of the polygenic risk score (PGS) may be useful in identifying individuals with increased genetic risk of developing T2D (3,4). At the same time, T2D is also influenced by nongenetic environmental or lifestyle factors, such as smoking (5), diet (6), and socioeconomic status (7). However, the

[1]Program in Bioinformatics and Integrative Genomics, Harvard Medical School, Boston, MA
[2]Department of Biomedical Informatics, Harvard Medical School, Boston, MA
[3]Computational Health Informatics Program, Boston Children's Hospital, Boston, MA
[4]Department of Epidemiology and Biostatistics, Imperial College London School of Public Health, London, U.K.
[5]Department of Hygiene and Epidemiology, University of Ioannina Medical School, Ioannina, Greece

Corresponding author: Chirag J. Patel, chirag_patel@hms.harvard.edu

Received 18 August 2020 and accepted 13 January 2021

This article contains supplementary material online at https://doi.org/10.2337/figshare.13607828.

© 2021 by the American Diabetes Association. Readers may use this article as long as the work is properly cited, the use is educational and not for profit, and the work is not altered. More information is available at https://www.diabetesjournals.org/content/license.

prospective predictive value of multiple nongenetic factors in comparison with and complementary to PGS has not been thoroughly examined. Studies of nongenetic exposures often only consider a single or handful of factors at a time without much consideration for dense correlation between exposures (8,9,10). Specifically, a polyexposure score (PXS) that combines multiple correlated nongenetic exposure and lifestyle factors has not been evaluated.

In this study, we examined the predictive ability of a PXS for T2D derived from a data-driven machine learning "feature selection" procedure from 111 exposure and lifestyle factors recorded in the UK Biobank (UKB). Our machine learning approach procedure yielded a final model of 12 variables for the T2D PXS. We subsequently tested the added value of the PXS against the PGS and established clinical risk factors for T2D prediction through discrimination and reclassification methodologies.

RESEARCH DESIGN AND METHODS

Study Participants

The UKB examines the role of genetics and environmental exposures in human health (11). The UKB resource comprises 502,655 U.K. participants between 40 and 69 years of age at the time of recruitment between 2006 and 2010. Participants attended 1 of 22 assessment centers across England, Scotland, and Wales where they completed touchscreen and nurse-led questionnaires, had physical measurements taken, and provided biological samples. The study collected extensive data from questionnaires, interviews, health records, physical measures, biological samples, and imaging as well as genome-wide genotype data. Individuals in the UKB underwent genotyping with two similar arrays—UK BiLEVE Axiom Array or UK Biobank Axiom Array—consisting of >800,000 genetic markers scattered across the genome. Additional genotypes were imputed using the Haplotype Reference Consortium resource, the UK10K panel, and the 1000 Genomes panel, increasing the number of testable variants to 96 million (12). UKB also collected information on individual background and lifestyle, cognitive and physical assessments, sociodemographic factors, and medical history. The UKB has ethical approval from the National Health Service National Research Ethics Service (Manchester, U.K.). All participants provided informed consent.

We defined the follow-up time as the time from the first exposure measurement from the first instance until either T2D incidence, competing event (death), or censorship date according to origin of the hospital data (England 31 March 2017, Scotland 30 October 2016, Wales 29 February 2016). Since the existing T2D PGS was derived from individuals with White European ancestry, our analysis was also restricted to White Europeans. The inclusion criteria of our primary study population were unrelated individuals of White European ancestry with complete covariate information and baseline normal HbA_{1c} levels (defined as <6.5%). We divided this population randomly into three roughly equal populations with no overlapping individuals: the training (n = 119,145), validation (n = 118,654), and testing (n = 118,822) sets (Fig. 1, bottom). We used the entire training set of 119,148 individuals (2,511 incident cases) to conduct exposure variable selection and optimized weights of clinical risk factors. The number of individuals with complete exposure data for each exposure association varied (Supplementary Table 1). There were 84,791 individuals (1,586 incident cases) with complete exposure data in the validation set, which we used to optimize weights of the variables in the PXS. In the final testing set for evaluating model performances, there were 68,299 individuals (1,281 incident cases) with complete exposure data for PXS, clinical factors for clinical risk score (CRS), and PGS. In a secondary analysis (a sensitivity analysis for participants with undiagnosed diabetes) (Fig. 1, bottom), we identified 3,658 individuals of unrelated White British ancestry with undiagnosed T2D (HbA_{1c} ≥6.5%) at baseline.

T2D Assessment

UKB contains self-reported information obtained during an interview with a trained nurse in addition to ICD-10 diagnostic codes recorded across all episodes of hospital visit. T2D cases were defined as having an ICD-10 code of E11.X or having self-reported T2D in the interview. We considered only cases in which the individuals did not have T2D during the first assessment visit period (2006–2010) but were subsequently followed up for incident T2D events.

The primary clinical risk factor analysis examined sex, age, family history, BMI, glucose level, systolic blood pressure, HDL, and triglycerides during the first assessment visit period. The binary family history variable was defined as positive if either the mother or the father had T2D and negative if otherwise. We selected these factors on the basis of their precedence in predicting T2D risk (13). Furthermore, Meigs et al. (2) used the same eight clinical factors in their comparison between clinical and genetic factors in T2D. We classified individuals with undiagnosed T2D as those with HbA_{1c} >6.5% [48 mmol/mol] per American Diabetes Association guidelines (14) and no diagnosis of T2D at study baseline (as above).

Deploying the PGS in UKB Participants

To calculate the PGS, we used weights developed by Khera et al. (3) (downloaded from https://cvd.hugeamp.org/downloads.html) on the basis of summary statistics from a meta-analysis of GWAS data from individuals of European ancestry. In summary, Khera et al. generated a set of candidate scores using various single nucleotide polymorphism (SNP) selection methods for White European participants in the UKB and evaluated the scores on the basis of the maximum area under the curve. The study found that the best score consisted of 6,630,149 SNPs and was created using LDPred, an algorithm with a linkage disequilibrium SNP-reweighting approach. With the weights, we calculated the PGS of individuals in our testing set using the built-in allelic scoring procedure of PLINK (–score) (15). PLINK takes the sum of the number of each reference allele multiplied by the weighted coefficient of the allele across all alleles.

Development of the PXS

We classified initial exposure variables as indicators of physiological state, environmental exposure, and self-reported behavior collected during the first assessment visit period (2006–2010). We first extracted all variables in the categories of reception, employment, sociodemographics, lifestyle and environment, estimated nutrients yesterday, early life factors, typical diet yesterday, meal type yesterday, spreads/sauces/cooking oils yesterday, alcoholic beverages yesterday, hot/cold beverages yesterday, cereal

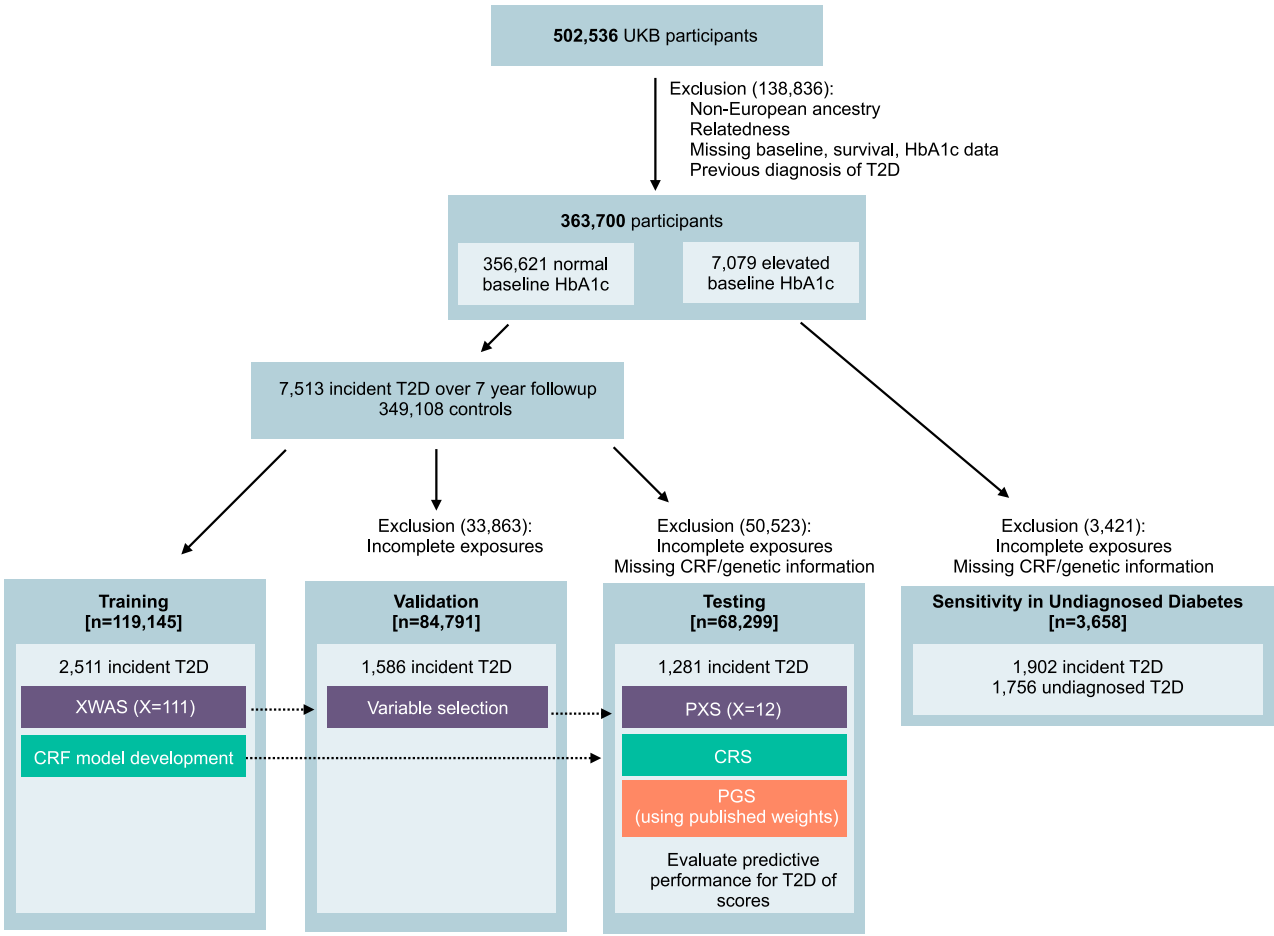

Figure 1—Study design. PXS, CRS, and PGS were calculated and compared for predictive accuracy. PGS was calculated using previously published weights. CRS factors included sex, age, family history, BMI, systolic blood pressure, serum glucose levels, serum HDL-C, and serum triglycerides. PXS factors were selected using a lasso-based method that relied on summary statistics from XWAS. CRF, clinical risk factor.

yesterday, milk/eggs/cheese yesterday, bread/pasta/rice yesterday, soup/snacks/pastries yesterday, meat/fish yesterday, milk/eggs/cheese yesterday, vegetarian alternatives yesterday, fruit/vegetables yesterday, residential air pollution, and residential noise pollution. There were 519 unique variables in total (Supplementary Table 2). From these, we considered only the variables that had data for >90% of the participants as potential correlates (referred to as factors). There were 111 variables that remained (Supplementary Table 1).

We used the PHESANT software, a tool for automated phenome scan analysis, to process the exposure data (16). In summary, PHESANT processes all UKB data and assigns one of four data types: continuous, ordered categorical, unordered categorical, and binary. PHESANT removes responses with negative-encoded values, such as "prefer not to answer" and "do not know." Continuous variables are transformed to a normal distribution using an inverse normal rank transformation. In cases where the variable cannot be transformed because of a large number of participants with the same value (e.g., rank order variables), the variables are encoded as an ordered categorical variable with three levels of equal sample size. For unordered categorical variables, the response with the largest number of participants was selected as the reference group.

Analogous to PGS, the first step in PXS estimation is an exposure-wide association study (XWAS) (17,18). We first conducted XWAS in the training set in which we associated each of 111 nongenetically measured environmental exposures while adjusting for age, assessment center, sex, and the first 40 principal components. Specifically, we modeled time to T2D onset as an independent variable in 111 separate Cox regression models. To maximize sample size, we removed individuals with missing data for each exposure before running its regression model. The sample size for each of the 111 associations can be found in Supplementary Table 1. We deemed Benjamini-Hochberg false discovery rate (FDR)–adjusted $P < 0.05$ as significant. We retained categorical variables that had at least one significant response. Since many exposure variables were assessed in <100,000 individuals in the training set (Supplementary Table 1), we sought to balance the tension between variables with robust XWAS associations but that were also measured in a large number of individuals. Only 10,044 of the 119,145 individuals in the test set had complete and positive-encoded values for all 111 variables. Running an XWAS allowed us to focus on 81 variables for the lasso model, increasing the sample size to 15,261 individuals with complete and positive-encoded values (increase of 5,217 individuals).

We used lasso regression to select variables from those that were FDR significant (19). Lasso is a type of linear regression that uses shrinkage and selection to enhance prediction generalizability of a model. After selection, we then iteratively removed nonsignificant variables on the basis of their association until only independently significant variables (at $P < 0.05$) remained in the model. The coefficients of categorical response represent the difference in effect of each response to the reference response. These coefficients were retained for the PXS calculation. Individuals with missing data in any of the remaining exposure variables were removed from the testing set. We calculated the final PXS by taking the weight sum of the responses of the 12 independent exposure variables.

Statistical Analysis

We fit T2D to each of the risk scores in a Cox regression model while adjusting for sex, age, principal components, and assessment center. We placed individuals into 100 bins by their PGS, PXS, or CRS percentiles and calculated the prevalence of T2D within each bin. We estimated the Pearson correlation coefficient and *P* value between each of the risk scores to measure gross gene environment correlation. Hazard ratios (HRs) for individuals with the highest risk of T2D were calculated by fitting a Cox regression model with binary indicator of being in the top *X* percentile or not.

We assessed discrimination of each model using the Harrell C-statistic, a rank order statistic for predictions against true outcomes. We calculated the continuous net reclassification improvement (NRI) and categorical NRI for each model at a range of thresholds from 0.025 to 0.20. Correct reclassification was when the nested model improved classification, and incorrect reclassification highlighted worse reclassification. Improvements were quantified as the sum of differences in proportion of correct minus incorrect for events and nonevents. NRI is sensitive to the choice of threshold (20); thus, the continuous (category-free) NRI was also used in this analysis.

We graphically assessed calibration of the original models by plotting the observed probability (Kaplan-Meier estimates) against the mean predicted probability within 10ths of the predicted probabilities at $t = 2{,}000$ days (20).

In the secondary analysis, we used HbA_{1c} instead of time to T2D event as the outcome of interest. Individuals were binned according to the percentiles of the validation set in the primary analysis.

We fit Cox regression models with the survival package in R (21). To adjust for multiple tests, we used the p.adjust function of the base statistics R package for Benjamini-Hochberg FDR adjustment (22). We used a modified version of the UpSetR package (23) to visualize overlapping incidence for T2D in the Supplementary Figures.

RESULTS

Our study sample consisted of 356,621 total unrelated individuals (median year of birth 1950, 195,657 females [54.86%]) of White British ancestry with no prior diagnosis of T2D and complete baseline data. Of these, 7,513 participants had incident T2D over a median follow-up time of 3.58 years (interquartile range 2.96 years) (Fig. 1).

We derived the PXS from univariate XWAS summary statistics, analogous to SNP weights estimated from GWAS summary statistics. The XWAS assessed associations between time to T2D onset and 111 total exposure indicators: alcohol ($n = 3$), air pollution ($n = 17$), noise pollution, diet ($n = 25$), disability allowance ($n = 1$), early life factors ($n = 9$), education levels ($n = 1$), employment status ($n = 1$), household information ($n = 7$), noise pollution ($n = 5$), physical activity ($n = 16$), population density ($n = 1$), sexual habits ($n = 4$), sleeping habits ($n = 7$), smoking ($n = 7$), and sun exposure ($n = 7$) (Supplementary Table 1). The XWAS regression analysis had a range of sample sizes ($n = 79{,}024$–$119{,}145$ individuals), with a mean of 116,311 individuals per test (SE 340 individuals) (Supplementary Table 1). There were 81 significantly associated exposure variables with T2D onset after adjusting for multiple hypothesis testing (Benjamini-Hochberg FDR-adjusted $P < 0.05$). Among these, responses in major dietary changes in the past 5 years, current employment status, rent/own accommodation lived in, and alcohol intake frequency were most significantly associated with the disease (Supplementary Table 1).

Next, we selected for variables that were independently associated with T2D onset resulting in 12 exposure variables (RESEARCH DESIGN AND METHODS and Supplementary Table 3). These 12 variables were alcohol intake, comparative body size at age 10, major dietary changes in past 5 years, household income, insomnia, snoring, milk type used (skim, whole, etc.), dietary restriction (eggs, dairy, wheat, etc.), spread type used (butter, etc.), tea intake per day, own or rent accommodations, and past tobacco usage. For example, in the final multivariate model, responses including "Yes, because of illness" and "Yes, because of other reasons" for major dietary changes in past 5 years had HRs of 3.9 and 1.3, respectively, compared with responding no ($P = 5.90 \times 10^{-95}$ and 3.51×10^{-6}, respectively). Responding no for snoring had an HR of 0.75 compared with yes ($P = 6.32 \times 10^{-15}$) (Supplementary Table 3).

We also conducted a sensitivity analysis by removing individuals who had incident T2D within 1 year of follow-up ($n = 2{,}193$) and repeating the PXS derivation, yielding 10 variables: average total household income before tax; morning/evening type of person (chronotype); sleeplessness/insomnia; milk type used; major dietary changes in the past 5 years; alcohol intake frequency; never eat eggs, dairy, wheat, sugar; comparative body size at age 10; cereal intake; and tea intake. The HRs of the 10 variables found in the both analyses were concordant (Pearson $r = 0.997$, $P < 0.001$).

We calculated the PGS for T2D from previously derived SNP weights of >6 million SNPs (3). We estimated the CRS with established risk factors for T2D, including sex, age, family history, BMI, systolic blood pressure, serum glucose levels, serum HDL cholesterol (HDL-C), and serum triglycerides (2). All clinical factors were significantly associated with T2D onset in the multivariate model (Supplementary Table 4). The median fasting time for all participants was 3 h for blood measurements (interquartile range 2 h). Family history and triglycerides had the largest change in response frequency between first and last CRS deciles (Supplementary Fig. 1).

With the addition of PGS, clinical risk factors, or both to the PXS model, all PXS factors remained significant in the multivariate model ($P < 0.05$). The direction

of effect for the PXS factors also remained the same (Supplementary Table 3).

The three risk scores were well calibrated, as seen in the comparison between the observed and predicted cumulative incidences of T2D events across each 10th of predicted risk (Supplementary Fig. 2). To better understand the most important factors in the PXS, we conducted a sensitivity analysis where each individual exposure was iteratively removed from the PXS. Removing major dietary changes in the past 5 years, snoring, or alcohol intake frequency from the model resulted in the largest changes to the C-statistic (Supplementary Table 5). Furthermore, a higher proportion of individuals with low PXS reported no major dietary change in the past 5 years and no snoring compared with individuals with high PXS (Supplementary Fig. 1).

The risk of T2D rose monotonically with all three scores. The CRS and PXS had the steepest changes in T2D risk from lowest to highest percentiles, while PGS changed moderately (Supplementary Fig. 3). In the lowest deciles of PGS, PXS, and CRS, the incidence of T2D was 0.43%, 0.15%, and 0.00%, respectively, while the highest deciles of PGS, PXS, and CRS had T2D incidences of 4.84%, 15.15%, and 22.00%, respectively (Supplementary Fig. 3). HRs associated with risk score values in the top 10% of PGS, PXS, and CRS versus the remaining populations were 2.00 (95% CI 1.73, 2.31), 5.90 (5.28, 6.61), and 9.97 (8.94, 11.13), respectively. The HRs for individuals in the top 1% of PGS, PXS, and CRS versus the remaining populations were 2.64 (1.87, 3.73), 9.74 (7.92, 11.91), and 15.11 (12.74, 17.92), respectively. The HRs for the upper 1–20% versus the remaining distribution are shown in Supplementary Table 7. There were significant differences in the survival times between risk score quartiles (all $P < 0.0001$ from log-rank test) (Supplementary Fig. 4).

All three risk scores were able to discriminate between individuals with and without future T2D in the entire cohort, with C-statistics >0.7. The discrimination for T2D was highest in the CRS model (C-statistic 0.839) and lowest in the PGS model (0.709) (Table 1). Subgroup analysis by sex and age separately showed overall higher discrimination in women and older populations (Table 1). The PGS and PXS model had a C-statistic of 0.776 (95% CI 0.764, 0.788), which additively approaches that of the CRS model (0.839 [0.829, 0.849]). Addition of PGS or PXS to the CRS model increased the C-statistics to 0.844 (0.834, 0.854) or 0.850 (0.840, 0.860), respectively. Use of all three types of scores marginally increased the predictive power (0.855 [0.845, 0.865]).

PGS, PXS, and CRS were significantly, but modestly, correlated with one another (all $P < 0.0001$). The PXS and CRS had the largest correlation ($r = 0.155$ [95% CI 0.147, 0.162]), while PXS and PGS had the weakest correlation (0.0227 [0.0152, 0.0302]). PGS and CRS were modestly correlated (0.0358 [0.0283, 0.0433]). The PXS-PGS interaction term was marginally significant ($P = 0.0457$) but increased predictive value by only 0.001 (C-statistic 0.777 [95% CI 0.765, 0.789]).

The addition of PXS to the CRS model improved T2D classification accuracy, with an overall continuous NRI of 30.1% (95% CI 25.9, 33.6) for cases and 16.9% (14.4, 19.3) for controls. In comparison, addition of PGS increased the NRI to 15.2% (11.5, 19.1) for cases and 7.3% (5.5, 9.2) for controls (Fig. 2). Categorical NRIs are presented in Supplementary Figs. 5–8.

T2D can also be diagnosed by elevated glucose. Thus, we derived a second CRS with serum blood glucose removed to evaluate the importance of glucose for prediction of T2D. Compared with the CRS with the full set of variables, the C-statistics decreased by 0.019 (0.820 [95% CI 0.810, 0.820]); however, the general pattern of predictive and discriminatory abilities of the CRS still held without serum glucose (i.e., addition of PXS and PGS to the CRS resulted in continuous NRI of 0.494 [95% CI 0.440, 0.548] and 0.241 [0.186, 0.296], respectively) (Supplementary Table 6 and Supplementary Figs. 9 and 10).

We defined individuals in the testing cohort with scores in the top 10% as having high risk. There were 255 individuals with high-risk CRS, PXS, and PGS simultaneously (e.g., in the high-risk category for all three scores). Of these, 58 (22.75%) had incident T2D. For the 6,829 individuals with high-risk PGS, PXS, or CRS (nonmutually exclusive), 232 (3.40%), 491 (7.19%), and 570 (8.35%) had incident T2D, respectively (Table 2 and Supplementary Fig. 11). In contrast, we defined individuals in the testing cohort with scores in the bottom 10 percentile as having low risk. For the 6,830 individuals with only low-risk PGS, PXS, or CRS (nonmutually exclusive), 55 (0.81%), 14 (0.20%), and 8 (0.12%) had incident T2D, respectively (Table 2 and Supplementary Fig. 11). Finally, there were 1,281 individuals who had incident T2D in the total testing group population; of these, 232 (18.11%), 491 (38.33%), and 570 (44.50%) had only high-risk PGS, PXS, and CRS, respectively. We found significant positive correlations between HbA_{1c} levels and all three risk scores in the testing set (Supplementary Fig. 12). HbA_{1c} had the strongest correlation to CRS ($r = 0.311$, $P < 0.001$), followed by PXS (0.232, $P < 0.001$), and the weakest correlation to PGS (0.092, $P < 0.001$).

In a secondary analysis, we calculated the PGS, PXS, and CRS of 3,658 unrelated individuals (median year of birth 1948, 1,272 females [34.77%]) with undiagnosed T2D (defined as $HbA_{1c} \geq 6.5\%$ but without a diagnosis of diabetes at baseline) (Fig. 1). The PGS, PXS, and CRS of the undiagnosed participants were in the 70th, 84th, and 99th percentiles, respectively, of the reference group (interquartile differences of 44, 34, and 7 percentiles, respectively). High-risk PXS had the highest sensitivity in undiagnosed participants. Eight hundred seventy-eight (61.14%) of 1,436 and 1,572 (54.79%) of 2,869 individuals with high-risk PXS and CRS, respectively, were later formally diagnosed with T2D (Supplementary Table 8). At baseline, the PXS could discriminate undiagnosed T2D with C-statistics of 0.756 (95% CI 0.748, 0.764), compared with 0.696 (0.688, 0.705) for the PGS.

CONCLUSIONS

Here, we make concrete the concept of PXS by training and evaluating environmental exposure and lifestyle variable scores in association with time to T2D onset. We found that in addition to standard clinical risk factors (e.g., sex, age, family history, BMI, systolic blood pressure, serum glucose levels, serum HDL-C, serum triglycerides), the PXS provides an increase in reclassification. Our holistic approach also demonstrated that the PXS had a greater, but modest improvement in predictive accuracy and

Table 1—C-statistics for evaluating performance of CRS, PGS, and PXS risk models in the full test population and stratified by sex and year of birth

	C-statistic (95% CI)						
	All	Male	Female	Born before 1945	Born between 1945 and 1950	Born between 1950 and 1958	Born after 1958
n	68,299	32,657	35,642	15,032	16,529	18,547	18,191
Events, *n*	1,281	844	437	468	377	291	145
Sex + age	0.670 (0.656, 0.684)	0.629 (0.612, 0.646)	0.637 (0.612, 0.662)	0.594 (0.569, 0.619)	0.606 (0.579, 0.633)	0.624 (0.593, 0.655)	0.592 (0.548, 0.636)
PGS*	0.709 (0.696, 0.722)	0.680 (0.663, 0.697)	0.705 (0.682, 0.728)	0.658 (0.634, 0.682)	0.674 (0.648, 0.700)	0.713 (0.687, 0.739)	0.719 (0.678, 0.76)
PXS*	0.762 (0.749, 0.775)	0.732 (0.716, 0.748)	0.774 (0.753, 0.795)	0.714 (0.690, 0.738)	0.718 (0.692, 0.744)	0.785 (0.759, 0.811)	0.782 (0.743, 0.821)
CRS*	0.839 (0.829, 0.849)	0.817 (0.804, 0.830)	0.855 (0.838, 0.872)	0.800 (0.781, 0.819)	0.817 (0.796, 0.838)	0.854 (0.834, 0.874)	0.857 (0.824, 0.89)
PGS + PXS*	0.776 (0.764, 0.788)	0.749 (0.734, 0.764)	0.786 (0.765, 0.807)	0.729 (0.706, 0.752)	0.730 (0.705, 0.755)	0.803 (0.779, 0.827)	0.802 (0.764, 0.84)
CRS + PGS*	0.844 (0.834, 0.854)	0.821 (0.808, 0.834)	0.859 (0.842, 0.876)	0.805 (0.787, 0.823)	0.820 (0.800, 0.840)	0.861 (0.842, 0.880)	0.865 (0.833, 0.897)
CRS + PXS*	0.850 (0.840, 0.860)	0.829 (0.816, 0.842)	0.866 (0.850, 0.882)	0.811 (0.793, 0.829)	0.823 (0.802, 0.844)	0.873 (0.854, 0.892)	0.866 (0.833, 0.899)
CRS + PXS + PGS*	0.855 (0.845, 0.865)	0.834 (0.821, 0.847)	0.869 (0.853, 0.885)	0.816 (0.798, 0.834)	0.826 (0.806, 0.846)	0.879 (0.861, 0.897)	0.873 (0.841, 0.905)

*Models adjusted for all covariates (sex, age, assessment center, and genetic principal components) in the full test population, all covariates except sex in the sex-stratified analysis, and all covariates except age in year of birth–stratified analysis.

A

CRS Model	CRS+PGS Model		
	# Participants	Continuous NRI	Categorical NRI
Cases	1281	0.152 (0.115 to 0.191)	0.065 (0.021 to 0.118)
Noncases	67018	0.073 (0.055 to 0.092)	-0.005 (-0.009 to -0.002)
Full population	68299	0.225 (0.174 to 0.280)	0.060 (0.020 to 0.109)

B

CRS Model	CRS+PXS Model		
	# Participants	Continuous NRI	Categorical NRI
Cases	1281	0.301 (0.259 to 0.336)	0.091 (0.033 to 0.154)
Noncases	67018	0.169 (0.144 to 0.193)	-0.005 (-0.011 to -0.001)
Full population	68299	0.470 (0.406 to 0.523)	0.085 (0.032 to 0.144)

C

CRS Model	CRS+PGS+PXS Model		
	# Participants	Continuous NRI	Categorical NRI
Cases	1281	0.216 (0.182 to 0.275)	0.144 (0.105 to 0.194)
Noncases	67018	0.215 (0.186 to 0.238)	-0.011 (-0.016 to -0.007)
Full population	68299	0.431 (0.377 to 0.503)	0.132 (0.098 to 0.179)

Figure 2—Reclassification of predicted T2D risk. The reclassified predicted risk with addition of PGS (*A*), PXS (*B*), or PGS + PXS (*C*) to the CRS model in the continuous case or the categorical case with a threshold of 12.5% risk. The overall NRI is the sum of the net reclassifications for cases (P[up|case] − P[down|case]) and noncases (P[down|noncase] − P[up|noncase]). A positive NRI indicates improved reclassification.

reclassification of incident T2D compared with that of PGS. Finally, individuals at the top 10% of PXS had an HR of 5.90 (95% CI 5.28, 6.61) compared with the remaining population.

Previous studies on nongenetic exposures and lifestyle factors focused, for the most part, on a small set of variables at a time without much consideration for the dense correlation that has been documented between exposures (8,9,10,24–26). While there is growing interest for nongenetic scores, such as the exposome risk score (27) or polysocial risk score (28), none have been evaluated and compared against PGSs or baselined against CRSs until recently (29). Abbasi et al. presented a systematic analysis of various T2D risk models, but most have only examined a limited number of nongenetic exposures, each included with limited justification for their variable selection (30). At most, the German Diabetes Risk Score includes physical activity, smoking, and consumption of red meat, whole-grain bread, coffee, and alcohol in addition to clinical risk factors. It is hypothesized that a candidate selection of a priori variables for risk prediction may lead to false positives (31), and a data-driven search for variables may mitigate incorporation of false-positive associations in risk models.

Here, we used a data-driven selection method to build the T2D PXS, ultimately using 12 indicators of alcohol, diet, early life factors, household information, sleep, and smoking. Definition of what lifestyle variables are has been elusive (24,32). We claim that the approach can more precisely define lifestyle through comprehensive inclusion in a data-driven procedure. Specifically, our T2D PXS selection procedure considers not only all variables in the models reviewed by Abbasi et al. (30) but also additional markers with even higher predictive power, arguing for a more comprehensive view of the totality of environmental (nongenetic) exposures to predict T2D. While our approach reidentified associations between exposures and T2D risk that are considered in U.S. clinical guidelines, such as low family income (7) and smoking (5), it is rarely useful to assess the contributions of individual exposures toward disease alone (28). Rather, a PXS captures holistic patient-level nongenetic risk that can inform clinicians about the characteristics of high-risk patients independent of genetic and clinical risk factors. Furthermore, early screening for T2D risk may reduce time between disease onset and clinician diagnosis, allowing for prompt treatment if necessary; however, we emphasize that recalling by a PXS must be tested prospectively.

We compared the predictive accuracy and discriminatory abilities of the PGS, PXS, and CRS for T2D. PGS provided modest, if any, incremental value over established clinical risk factors. A recent study also reported a modest categorical NRI of 0.048, with 33% as the risk cutoff when PGS was added to clinical risk factors for T2D (4). However, PGS for T2D can be both evaluated at birth and used any time during the life course to stratify individuals at highest genetic risk.

Table 2—T2D incidence in test set for individuals with high- and low-risk combinations of CRS, PGS, and/or PXS

Risk score combination	High risk, *n*	Low risk, *n*	High-risk incidence, *n* (%)	Low-risk incidence, *n* (%)
CRS	6,829	6,830	655 (9.59)	8 (0.12)
PXS	6,829	6,830	491 (7.19)	14 (0.20)
PGS	6,829	6,830	232 (3.40)	55 (0.81)
PXS and CRS	2,014	2,687	310 (15.39)	1 (0.04)
PGS and CRS	881	810	124 (14.07)	0 (0.00)
PGS and PXS	735	746	87 (11.84)	2 (0.27)
PGS and PXS and CRS	255	321	58 (22.75)	0 (0.00)

The total number of individuals within each risk score group is indicated as well as the number of T2D incidence cases. For example, there were 6,829 individuals in the top 10 percentiles of PXS; of those, 491 (7.19%) had incident T2D. As another example, there were 2,014 with a high percentile of both PXS and CRS; of those, 310 (15.39%) had incident T2D.

The PXS, on the other hand, allows individuals to appreciate the remediation of disease risk through potential modification in diet and behavior (e.g., alcohol and tobacco usage [33]). The PXS also summarizes the risk of combinations of exposures related to health outcomes. However, we caution that the contents of the PXS, like the PGS, may not be causal. While some factors are arguably modifiable and may causally reduce risk for T2D, others (e.g., socioeconomic status, history of cigarette smoking [34]) may not be.

Abbasi et al. (30) also presented several models that have preestablished weights for clinical factors. We chose to derive our own weights to mitigate effects of population differences between the UKB population for which the previous weights were derived. For the most part, our weights are concordant in direction with those derived by Meigs et al. (2) using the Framingham Offspring Study (Supplementary Table 4). Our refitted model gives the most generous estimate for CRS. We randomly assigned the UKB into training, testing, and validation sets to evaluate the PXS. We emphasize that our findings should be additionally validated in an external cohort; to this end, we have provided all weights for the PXS. One of the most significant challenges in observational exposure studies is the deduction of direction of causality or potential confounding variables. By selecting only individuals with no incident T2D when exposures were measured, we can more confidently report the exposures as conferring risk of T2D. For example, it is possible that the significant association of the response, "Yes, because of illness," to major dietary changes in the past 5 years is explained by another illness (comorbidity). However, because our study focuses on risk prediction, we argue that the issue of confounding variables is not as relevant.

While easy to measure, some of the exposures included self-reported variables, such as diet, which may be prone to measurement error and recall bias (35). If these errors occur at random across all variables considered in the PXS, the association sizes and PXS will be diluted. If, on the other hand, case versus control individuals report their intakes differently, the PXSs will also be directionally biased. It is less clear how PXS will be affected if the types of the errors are different (both random and differential with respect to the exposure or disease) across the variable inputs. We only considered exposure variables if they contained $<10\%$ missing data. Increasing data completeness of variables, or imputing exposure information, would be valuable to the eventual use of machine learning techniques for modeling. Furthermore, exposure responses are highly heterogeneous. Approaches such as the PHESANT pipeline that automate variable codification is but one way to make analyses of heterogeneous data scalable and reproducible (16).

An inherent challenge to environment/nongenetic and genetic studies is that such factors are often examined in isolation. For example, genetic and exposure factors may be correlated, a phenomenon known as gene-environment correlation. To this end, we found that PXS and PGS had a modest, but significant correlation with each other. The PXS-PGS interaction term was significantly associated with T2D onset; however, its added value to discrimination was trivial. It is hypothesized that the gene-environment interaction plays a large role in T2D, but its effect on phenotypic variation is widely debated (36). Gene-environment studies require approaches to prune the vast space of potential gene-environment interactions to test the power of their identification (37). One potential way to increase the power of detection of gene-environment interactions is to examine the PGS and PXS rather than each genetic variant or exposure separately. Furthermore, the clinical risk factors are likely influenced by genetics and environmental exposures. While we found the scores to be significantly correlated to one another, we also demonstrated that PGS, PXS, and CRS provide independent information and are additive in predicting T2D.

Because the UKB consists of primarily individuals with European ancestry, we limited our analysis to only participants of White British ancestry. It is difficult to extrapolate these results to other ethnic populations because European GWAS are often biased when applied to more diverse populations (38). Furthermore, exposure disparities, such as socioeconomic status (39), education attainment (39), and smoking (40), are correlated with ethnicity. Therefore, there is a clear need for more diverse populations in both genetic and environmental exposure studies. A few notable studies exist or will be available in the future, such as the All of Us Project (41) and the Kadoorie Biobank (42). To capture the comprehensive variation of environment and genetics in diseases, and to test the utility of precision medicine, investigations in other populations will be instrumental.

Acknowledgments. The authors thank all the volunteers who participated in this project.

Funding. Our analysis was conducted using the UKB resource through application number 22881. This work was supported by the Bioinformatics and Integrative Genomics training grant from the National Human Genome Research Institute under award number T32-HG-002295 (to Y.H.), the National Institute of Environmental Health Sciences under award numbers R00-ES-23504 and R21-ES-205052, the National Institute of Allergy and Infectious Diseases under award number R01-AI-12725003, the National Science Foundation Graduate Research Fellowship under award number DGE1745303 (to Y.H.), the UKB Early-Career Researcher Award (to Y.H.), and the National Science Foundation under award number 1636870.

Duality of Interest. A.K.M. and C.J.P. are consultants and cofounders of XY Health, Inc. No other potential conflicts of interest relevant to this article were reported.

Author Contributions. Y.H., I.T., and C.J.P. designed the study. Y.H., C.M.L., and D.R. were involved in data processing. Y.H. conducted the statistical analysis and literature search and wrote the first version of the manuscript. C.M.L., D.R., A.K.M., I.T., and C.J.P. revised the manuscript. D.R. provided data. Y.H., C.M.L., and C.J.P. are the guarantors of this work and, as such, had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

Prior Presentation. Parts of this study were presented in poster form at the American Society of Human Genetics 2020 Virtual Meeting, 27–30 October 2020.

References

1. Morgan CL, Currie CJ, Peters JR. Relationship between diabetes and mortality: a population study using record linkage. Diabetes Care 2000; 23:1103–1107

2. Meigs JB, Shrader P, Sullivan LM, et al. Genotype score in addition to common risk factors for prediction of type 2 diabetes. N Engl J Med 2008;359:2208–2219

3. Khera AV, Chaffin M, Aragam KG, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet 2018;50:1219–1224

4. Mars N, Koskela JT, Ripatti P, et al.; FinnGen. Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. Nat Med 2020;26:549–557

5. Pan A, Wang Y, Talaei M, Hu FB, Wu T. Relation of active, passive, and quitting smoking with incident type 2 diabetes: a systematic review and meta-analysis. Lancet Diabetes Endocrinol 2015; 3:958–967

6. Neuenschwander M, Ballon A, Weber KS, et al. Role of diet in type 2 diabetes incidence: umbrella review of meta-analyses of prospective observational studies. BMJ 2019;366:l2368
7. Martinell M, Pingel R, Hallqvist J, et al. Education, immigration and income as risk factors for hemoglobin A1c >70 mmol/mol when diagnosed with type 2 diabetes or latent autoimmune diabetes in adult: a population-based cohort study. BMJ Open Diabetes Res Care 2017; 5:e000346
8. Smith GD, Lawlor DA, Harbord R, Timpson N, Day I, Ebrahim S. Clustered environments and randomized genes: a fundamental distinction between conventional and genetic epidemiology. PLoS Med 2007;4:e352
9. Ioannidis JPA, Loy EY, Poulton R, Chia KS. Researching genetic versus nongenetic determinants of disease: a comparison and proposed unification. Sci Transl Med 2009;1:7ps8
10. van der Meer T, Wolffenbuttel B, Patel CJ. Data-driven assessment, contextualization and implementation of 134 variables in their risk for type 2 diabetes: An analysis of Lifelines, a prospective cohort study in the Netherlands. Diabetologia. In press
11. Sudlow C, Gallacher J, Allen N, et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med 2015;12:e1001779
12. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature 2018;562:203–209
13. Wilson PWF, Meigs JB, Sullivan L, Fox CS, Nathan DM, D'Agostino RB Sr. Prediction of incident diabetes mellitus in middle-aged adults: the Framingham Offspring Study. Arch Intern Med 2007;167:1068–1074
14. American Diabetes Association. *Standards of Medical Care in Diabetes-2018* abridged for primary care providers. Clin Diabetes 2018;36:14–37
15. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 2015;4:7
16. Millard LAC, Davies NM, Gaunt TR, Davey Smith G, Tilling K. Software application profile: PHESANT: a tool for performing automated phenome scans in UK Biobank. Int J Epidemiol 2018; 47:29–35
17. Patel CJ, Bhattacharya J, Butte AJ. An environment-wide association study (EWAS) on type 2 diabetes mellitus. PLoS One 2010;5:e10746
18. Patel CJ, Cullen MR, Ioannidis JP, Butte AJ. Systematic evaluation of environmental factors: persistent pollutants and nutrients correlated with serum lipid levels. Int J Epidemiol 2012;41: 828–843
19. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc B 1996;58:267–288
20. Balding DJ, Moltke I, Marioni J, Eds. *Handbook of Statistical Genomics*. 4th ed. Hoboken, NJ, Wiley, 2019
21. Therneau TM, Grambsch PM. *Modeling Survival Data: Extending the Cox Model*. New York, Springer, 2000
22. Benjamini Y, Hochberg Y. Controlling the false Discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B 1995; 57:289–300
23. Conway JR, Lex A, Gehlenborg N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* 2017; 33:2938–2940
24. Patel CJ, Ioannidis JPA. Studying the elusive environment in large scale. JAMA 2014;311: 2173–2174
25. Patel CJ, Ioannidis JPA. Placing epidemiological results in the context of multiplicity and typical correlations of exposures. J Epidemiol Community Health 2014;68:1096–1100
26. Manrai AK, Patel CJ, Gehlenborg N, Tatonetti NP, Ioannidis JPA, Kohane IS. Methods to enhance the reproducibility of precision medicine. Pac Symp Biocomput 2016;21:180–182
27. Vermeulen R, Schymanski EL, Barabási A-L, Miller GW. The exposome and health: where chemistry meets biology. Science 2020;367:392–396
28. Figueroa JF, Frakt AB, Jha AK. Addressing social determinants of health: time for a polysocial risk score. JAMA 2020;323:1553–1554
29. Elliott J, Bodinier B, Bond TA, et al. Predictive accuracy of a polygenic risk score-enhanced prediction model vs a clinical risk score for coronary artery disease. JAMA 2020;323:636–645
30. Abbasi A, Peelen LM, Corpeleijn E, et al. Prediction models for risk of developing type 2 diabetes: systematic literature search and independent external validation study. BMJ 2012;345:e5900
31. Ioannidis JPA. Why most published research findings are false. PLoS Med 2005;2:e124
32. Tzoulaki I, Elliott P, Kontis V, Ezzati M. Worldwide exposures to cardiovascular risk factors and associated health effects: current knowledge and data gaps. Circulation 2016;133: 2314–2333
33. Johansen MY, MacDonald CS, Hansen KB, et al. Effect of an intensive lifestyle intervention on glycemic control in patients with type 2 diabetes: a randomized clinical trial. JAMA 2017; 318:637–646
34. Borgnakke WS. "Non-modifiable" risk factors for periodontitis and diabetes. Curr Oral Health Rep 2016;3:270–281
35. Ioannidis JPA. The challenge of reforming nutritional epidemiologic research. JAMA 2018; 320:969–970
36. Aschard H, Lutz S, Maus B, et al. Challenges and opportunities in genome-wide environmental interaction (GWEI) studies. Hum Genet 2012; 131:1591–1613
37. Patel CJ, Chen R, Kodama K, Ioannidis JPA, Butte AJ. Systematic identification of interaction effects between genome- and environment-wide associations in type 2 diabetes mellitus. Hum Genet 2013;132:495–508
38. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. Nat Genet 2019;51:584–591
39. Braveman PA, Cubbin C, Egerter S, Williams DR, Pamuk E. Socioeconomic disparities in health in the United States: what the patterns tell us. Am J Public Health 2010;100(Suppl. 1):S186–S196
40. Drope J, Liber AC, Cahn Z, et al. Who's still smoking? Disparities in adult cigarette smoking prevalence in the United States. CA Cancer J Clin 2018;68:106–115
41. Denny JC, Rutter JL, Goldstein DB, et al.; All of Us Research Program Investigators. The "All of Us" research program. N Engl J Med 2019;381: 668–676
42. Chen Z, Chen J, Collins R, et al.; China Kadoorie Biobank (CKB) collaborative group. China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. Int J Epidemiol 2011;40: 1652–1666