Supplementary Materials: Tables S1–S8

Who should have an HbA1c? Development and external validation of a dysglycaemia prediction model using only routinely available data

Table S1

Table S1. Predictors, model terms, fitted coefficients, spline knot locations and intercept of the penalised logistic regression.

Model term	Variable	Form	Coefficient	Spline knots
(intercept)	Intercept		-1.0242	
RIDAGEYR	Age (years)	linear	0.0597	
SEX_MALE	Male sex	linear	0.1637	
DMDEDUC2	Education level	linear	-0.0843	
INDFMPIR	Family income-to-poverty ratio	linear	-0.0138	
BMXBMI	Body mass index (kg/m2)	linear	0.0219	
BMXWAIST	Waist circumference (cm)	linear	-0.0025	
BMXHIP	Hip circumference (cm)	linear	-0.0129	
BMXWT	BMXWT	linear	0.0037	
BMXHT	BMXHT	linear	-0.0005	
BMXARMC	Arm circumference (cm)	linear	0.0139	
WHTR	Waist-to-height ratio	linear	0.1466	
WHR	Waist-to-hip ratio	linear	1.1464	
SBP	Systolic blood pressure (mmHg)	linear	0.0075	
DBP	Diastolic blood pressure (mmHg)	linear	0.0060	
PULSE	Pulse (beats/min)	linear	0.0002	
LBXSATSI	Alanine aminotransferase (U/L)	linear	0.0059	
LBXSASSI	Aspartate aminotransferase (U/L)	linear	0.0055	
LBXSGTSI	Gamma-glutamyl transferase (U/L)	linear	0.0047	
LBXSAPSI	Alkaline phosphatase (U/L)	linear	0.0024	
LBXSTB	Total bilirubin (mg/dL)	linear	-1.0613	
LBXSAL	Albumin (g/dL)	linear	-0.2187	
LBXSTP	Total protein (g/dL)	linear	0.2626	
AST_ALT_RATIO	AST/ALT ratio	linear	-0.6333	

LBXSTR	Triglycerides (mg/dL)	linear	-0.0010	
LBXSCH	Total cholesterol (mg/dL)	linear	0.0028	
LBDHDD	HDL-cholesterol (mg/dL)	linear	-0.0057	
LBXSUA	Uric acid (mg/dL)	linear	0.0524	
TG_HDL_RATIO	TG_HDL_RATIO	linear	0.0198	
LBXSCR	Creatinine (mg/dL)	linear	-0.2724	
LBXSBU	Urea nitrogen (mg/dL)	linear	-0.0018	
EGFR	eGFR (mL/min/1.73m2)	linear	-0.0017	
LBXPLTSI	Platelet count (1000/uL)	linear	0.0005	
LBXWBCSI	White cell count (1000/uL)	linear	0.0614	
LBXHGB	Haemoglobin (g/dL)	linear	-0.0413	
LBXMCVSI	Mean corpuscular volume (fL)	linear	-0.0490	
LBXLYPCT	Lymphocytes (%)	linear	0.0056	
LBXNEPCT	Neutrophils (%)	linear	-0.0137	
LBXHSCRP	High-sensitivity CRP (mg/L)	linear	0.0478	
ETOH_GPD	Alcohol intake (g/day)	linear	-0.0110	
SMOKE_CURRENT	Current smoking	linear	0.3064	
SMOKE_EVER	SMOKE_EVER	linear	-0.1158	
RX_ANTIHTN	Taking antihypertensive medication	linear	0.1379	
RX_LIPID	Taking lipid-lowering medication	linear	0.4006	
HX_CHD	History of coronary heart disease	linear	0.1659	
HX_CHF	HX_CHF	linear	0.2466	
HX_STROKE	History of stroke	linear	-0.3954	
HTN_TOLD	Told has high blood pressure	linear	-0.0785	
RETH_2	Race/Hispanic origin, category 2	indicator	-0.0584	
RETH_3	Race/Hispanic origin, category 3	indicator	-0.2966	
RETH_4	Race/Hispanic origin, category 4	indicator	0.6823	
RETH_6	Race/Hispanic origin, category 6	indicator	0.5010	
RETH_7	Race/Hispanic origin, category 7	indicator	0.1298	
RIDAGEYR_s	Age (years), spline term	restricted cubic spline	-0.0077	25, 48, 74
BMXBMI_s	Body mass index (kg/m2), spline term	restricted cubic spline	0.0015	21.6, 28.30552276, 38.7
BMXWAIST_s	Waist circumference (cm), spline term	restricted cubic spline	-0.0025	78.9, 97.89335975, 120.668
WHTR_s	Waist-to-height ratio, spline term	restricted cubic spline	1.7482	0.4730843, 0.5900907643, 0.7224644632
SBP_s	Systolic blood	restricted cubic spline	-0.0062	102.5993525,

	pressure (mmHg), spline term			121.3253094, 148.8187661
DBP_s	Diastolic blood pressure (mmHg), spline term	restricted cubic spline	-0.0060	61.08013927, 74.02887357, 89.61543656
LBXSATSI_s	Alanine aminotransferase (U/L), spline term	restricted cubic spline	-0.0080	10, 17, 38
LBXSASSI_s	Aspartate aminotransferase (U/L), spline term	restricted cubic spline	-0.0029	14, 19, 31
LBXSGTSI_s	Gamma-glutamyl transferase (U/L), spline term	restricted cubic spline	-0.0073	10.8545686, 21, 54.71708261
LBXSTR_s	Triglycerides (mg/dL), spline term	restricted cubic spline	-0.0000	58.6, 108.9130349, 229.6715019
LBDHDD_s	HDL-cholesterol (mg/dL), spline term	restricted cubic spline	-0.0104	37, 52, 75
LBXPLTSI_s	Platelet count (1000/uL), spline term	restricted cubic spline	-0.0006	173, 240, 328
LBXHSCRP_s	High-sensitivity CRP (mg/L), spline term	restricted cubic spline	-0.0742	0.43, 1.8645, 8.3272
EGFR_s	eGFR (mL/min/1.73m2), spline term	restricted cubic spline	-0.0049	66.52945163, 97.55930491, 121.4857792

Coefficients are on the log-odds scale for the single pooled equation fitted to the development cycle. Continuous predictors marked as restricted cubic splines enter as two terms, the linear term and one nonlinear basis term, with knots at the 10th, 50th and 90th percentiles of the development distribution.

Table S2

Table S2. Missing data by predictor and cycle.

Predictor	Variable	Missing, development (%)	Missing, validation (%)
PULSE	Pulse (beats/min)	17.39	3.21
INDFMPIR	Family income-to-poverty ratio	13.78	12.72
DBP	Diastolic blood pressure (mmHg)	9.63	3.21
SBP	Systolic blood pressure (mmHg)	9.63	3.21
ETOH_GPD	Alcohol intake (g/day)	5.9	13.24
WHTR	Waist-to-height ratio	4.35	4.08
WHR	Waist-to-hip ratio	4.35	4.81
BMXWAIST	Waist circumference (cm)	4.17	3.98
BMXHIP	Hip circumference (cm)	4.08	4.72
BMXARMC	Arm circumference (cm)	3.41	2.32
AST_ALT_RATIO	AST/ALT ratio	2.95	5.76
LBXSASSI	Aspartate aminotransferase (U/L)	2.95	5.76

LBXHSCRP	High-sensitivity CRP (mg/L)	2.72	2.94
TG_HDL_RATIO	TG_HDL_RATIO	2.62	4.91
LBXSAPSI	Alkaline phosphatase (U/L)	2.57	5.47
LBXSUA	Uric acid (mg/dL)	2.57	5.43
LBXSBU	Urea nitrogen (mg/dL)	2.57	5.47
LBXSTP	Total protein (g/dL)	2.57	4.87
LBXSTR	Triglycerides (mg/dL)	2.57	4.91
LBXSGTSI	Gamma-glutamyl transferase (U/L)	2.57	5.47
LBXSATSI	Alanine aminotransferase (U/L)	2.57	5.57
LBXSTB	Total bilirubin (mg/dL)	2.54	5.49
EGFR	eGFR (mL/min/1.73m2)	2.54	5.45
LBXSCR	Creatinine (mg/dL)	2.54	5.45
LBXSCH	Total cholesterol (mg/dL)	2.54	5.45
LBXSAL	Albumin (g/dL)	2.53	4.85
LBDHDD	HDL-cholesterol (mg/dL)	2.08	4.52
BMXBMI	Body mass index (kg/m2)	1.43	1.22
BMXHT	BMXHT	1.35	0.91
BMXWT	BMXWT	1.19	1.04
RX_LIPID	Taking lipid-lowering medication	0.5	0.54
HX_CHF	HX_CHF	0.2	0.06
HX_CHD	History of coronary heart disease	0.2	0.27
DMDEDUC2	Education level	0.15	0.04
HX_STROKE	History of stroke	0.12	0.15
RX_ANTIHTN	Taking antihypertensive medication	0.09	0.06
LBXNEPCT	Neutrophils (%)	0.09	0.54
HTN_TOLD	Told has high blood pressure	0.09	0.06
LBXLYPCT	Lymphocytes (%)	0.09	0.54
LBXMCVSI	Mean corpuscular volume (fL)	0.05	0.37
LBXHGB	Haemoglobin (g/dL)	0.05	0.37
LBXWBCSI	White cell count (1000/uL)	0.05	0.37
LBXPLTSI	Platelet count (1000/uL)	0.05	0.37
SMOKE_CURRENT	Current smoking	0.03	0.12
SMOKE_EVER	SMOKE_EVER	0.03	0.12
SEX_MALE	Male sex	0.0	0.0
RIDRETH3	Race and Hispanic origin	0.0	0.0
RIDAGEYR	Age (years)	0.0	0.0

Table S3

Table S3. Pre-specified sensitivity analyses S1-S10.

Analysis	**Specification**	**Model**	**n**	**Events**	**AUROC or Δ (95% CI)**	**p**
S1	POP B (fasting subsample); outcome HbA1c >=5.7% or FPG >=100 mg/dL	Logistic regression	3090	2007	0.7790 (0.755-0.803)	

S1	POP B (fasting subsample); outcome HbA1c >=5.7% or FPG >=100 mg/dL	LightGBM	3090	2007	0.7770 (0.751-0.801)	
S1	LightGBM minus logistic	delta	3090	2007	-0.0022 (-0.0131 to +0.0081)	0.690
S2	HbA1c threshold 6.0 instead of 5.7	Logistic regression	6649	984	0.8390 (0.818-0.859)	
S2	HbA1c threshold 6.0 instead of 5.7	LightGBM	6649	984	0.8510 (0.834-0.868)	
S2	LightGBM minus logistic	delta	6649	984	0.0124 (+0.0004 to +0.0257)	0.038
S3	age 35 and over (ADA screening age)	Logistic regression	4904	2119	0.7610 (0.741-0.782)	
S3	age 35 and over (ADA screening age)	LightGBM	4904	2119	0.7660 (0.745-0.787)	
S3	LightGBM minus logistic	delta	4904	2119	0.0045 (-0.0048 to +0.0133)	0.336
S4	leisure-time physical activity added	Logistic regression	6649	2346	0.8100 (0.795-0.826)	
S4	leisure-time physical activity added	LightGBM	6649	2346	0.8120 (0.796-0.827)	
S4	LightGBM minus logistic	delta	6649	2346	0.0014 (-0.0042 to +0.0074)	0.612
S5	self-reported prediabetes (DIQ160) re-included	Logistic regression	6649	2346	0.8200 (0.804-0.835)	
S5	self-reported prediabetes (DIQ160) re-included	LightGBM	6649	2346	0.8220 (0.807-0.837)	
S5	LightGBM minus logistic	delta	6649	2346	0.0026 (-0.0032 to +0.0090)	0.411
S6	survey-weighted fitting	Logistic regression	6649	2346	0.8060 (0.790-0.822)	
S6	survey-weighted fitting	LightGBM	6649	2346	0.8050 (0.788-0.821)	
S6	LightGBM minus logistic	delta	6649	2346	-0.0019 (-0.0082 to +0.0044)	0.601
S7	complete cases only	Logistic regression	4349	1503	0.8130 (0.793-0.832)	
S7	complete cases only	LightGBM	4349	1503	0.8070 (0.786-0.827)	

S7	LightGBM minus logistic	delta	4349	1503	-0.0063 (-0.0135 to +0.0014)	0.113
S8	ten strongest predictors	Logistic regression	6649	2346	0.8000 (0.784-0.816)	
S8	ten strongest predictors	LightGBM	6649	2346	0.8040 (0.788-0.819)	
S8	LightGBM minus logistic	delta	6649	2346	0.0043 (-0.0008 to +0.0099)	0.111
S9	undiagnosed diabetes, restricted set	Logistic regression	6649	237	0.8260 (0.776-0.868)	
S9	undiagnosed diabetes, restricted set	LightGBM	6649	237	0.8560 (0.815-0.890)	
S9	LightGBM minus logistic	delta	6649	237	0.0291 (+0.0069 to +0.0523)	0.010
S10	chronic kidney disease excluded	Logistic regression	6236	2126	0.8120 (0.796-0.828)	
S10	chronic kidney disease excluded	LightGBM	6236	2126	0.8100 (0.794-0.826)	
S10	LightGBM minus logistic	delta	6236	2126	-0.0021 (-0.0080 to +0.0040)	0.467

Rows labelled "delta" give the paired difference between LightGBM and the penalised logistic regression within that analysis, with a bootstrap p value. Comparisons are within each analysis, not against the main model.

Table S4

Table S4. SHAP importance ranking with stability statistics.

Predictor	**Label**	**Mean**	**SHAP**	**, LightGBM**	**Rank, LightGBM**	**Mean**	**SHAP**	**, logistic**	**Rank, logistic**	**Rank SD across imputations**	**Rank SD across folds**
RIDAGEYR	Age	0.740	1	1.005	1	0.00	0.00				
RIDRETH3	Race/Hispanic origin	0.358	2	0.523	2	0.00	0.00				
LBXMCVSI	Mean corpuscular volume	0.230	3	0.223	4	0.00	0.00				
LBXSTB	Total bilirubin	0.141	4	0.203	6	0.31	0.97				

ETOH_GPD	Alcohol intake	0.127	5	0.104	14	0.39	0.97				
AST_ALT_RATIO	AST/ALT ratio	0.115	6	0.197	7	0.76	1.60				
LBXHSCRP	High-sensitivity CRP	0.113	7	0.273	3	0.75	1.43				
WHR	Waist-to-hip ratio	0.107	8	0.086	20	1.09	2.10				
LBXSGTSI	Gamma-glutamyl transferase	0.098	9	0.182	8	1.39	1.17				
WHTR	Waist-to-height ratio	0.092	10	0.111	12	0.86	3.97				
LBDHDD	HDL-cholesterol	0.090	11	0.165	9	1.34	2.31				
LBXSUA	Uric acid	0.085	12	0.065	26	0.93	3.12				
LBXHGB	Haemoglobin	0.082	13	0.044	36	0.64	1.40				
SBP	Systolic blood pressure	0.072	14	0.215	5	2.11	5.51				
EGFR	eGFR	0.069	15	0.094	17	1.26	5.95				
RX_LIPID	Lipid-lowering use	0.064	16	0.093	19	1.41	3.14				
LBXLYPCT	Lymphocyte percent	0.063	17	0.041	38	1.32	2.31				
LBXNEPCT	Neutrophil percent	0.063	18	0.104	15	1.73	1.72				
LBXWBCSI	White blood cell count	0.062	19	0.105	13	0.73	2.62				
LBXSTP	Total protein	0.054	20	0.095	16	0.89	3.13				
TG_HDL_RATIO	Triglyceride/ HDL ratio	0.049	21	0.027	43	1.71	5.12				
BMXHIP	Hip circumference	0.045	22	0.124	11	0.83	5.44				
BMXBMI	Body mass index	0.038	23	0.127	10	1.24	4.08				
INDFMPIR	Income-to-poverty ratio	0.034	24	0.028	42	3.57	3.72				
BMXARMC	Arm circumference	0.034	25	0.056	31	1.32	6.75				
LBXSATSI	Alanine aminotransferase	0.032	26	0.073	23	3.05	6.21				

DMDEDUC2	Education	0.031	27	0.080	22	2.13	5.17				
SMOKE_CURRENT	Current smoking	0.027	28	0.094	18	1.98	5.02				
LBXSAPSI	Alkaline phosphatase	0.025	29	0.041	37	2.65	3.92				
BMXWAIST	Waist circumference	0.025	30	0.056	30	2.88	6.63				
LBXSCR	Creatinine	0.025	31	0.039	39	2.04	5.25				
LBXSTR	Triglycerides	0.024	32	0.060	29	2.68	6.40				
LBXSCH	Total cholesterol	0.023	33	0.082	21	2.52	2.41				
LBXSASSI	Aspartate aminotransferase	0.020	34	0.046	35	3.72	4.31				
DBP	Diastolic blood pressure	0.019	35	0.072	24	3.17	4.39				
LBXSAL	Albumin	0.019	36	0.062	28	2.72	5.46				
RX_ANTIHTN	Antihypertensive use	0.018	37	0.053	33	1.70	6.77				
BMXWT	Weight	0.017	38	0.065	27	2.18	4.69				
PULSE	Pulse	0.016	39	0.015	44	4.82	6.28				
LBXPLTSI	Platelet count	0.015	40	0.053	34	1.39	4.15				
BMXHT	Height	0.011	41	0.012	45	0.75	4.37				
SEX_MALE	Male sex	0.009	42	0.069	25	1.26	5.31				
LBXSBU	Blood urea nitrogen	0.009	43	0.000	48	1.14	5.06				
HTN_TOLD	HTN_TOLD	0.005	44	0.029	41	0.97	1.97				
HX_STROKE	Stroke	0.004	45	0.031	40	0.41	1.17				
HX_CHD	Coronary heart disease	0.002	46	0.011	47	0.55	1.34				
HX_CHF	Congestive heart failure	0.001	47	0.011	46	0.44	1.40				
SMOKE_EVER	Ever smoked	0.000	48	0.055	32	0.00	1.51				

Importance is the mean absolute SHAP value across the development cycle. Rank standard deviations are computed across the twenty imputations and across the ten outer cross-validation folds. ## Table S5

Table S5. Discrimination of the penalised logistic regression within subgroups defined by sex, age band and race and Hispanic origin, both cycles.

Cycle	Subgroup	n	Events	AUROC	LCI	UCI
Development	Overall	6649	2346	0.81	0.795	0.827
Development	Men	3172	1154	0.805	0.782	0.827
Development	Women	3477	1192	0.816	0.797	0.836
Development	Age 20-39	2290	372	0.838	0.806	0.864
Development	Age 40-59	2260	878	0.759	0.725	0.79
Development	Age 60 or older	2099	1096	0.712	0.679	0.743
Development	Mexican American	768	238	0.801	0.761	0.838
Development	Other Hispanic	695	228	0.834	0.797	0.867
Development	Non-Hispanic White	2380	675	0.808	0.783	0.831
Development	Non-Hispanic Black	1686	806	0.79	0.763	0.815
Development	Non-Hispanic Asian	810	308	0.823	0.791	0.851
Development	Other or multiracial	310	91	0.724	0.638	0.809
Validation	Overall	4826	1474	0.818	0.802	0.833
Validation	Men	2147	676	0.807	0.78	0.83
Validation	Women	2679	798	0.829	0.809	0.847
Validation	Age 20-39	1369	161	0.797	0.754	0.841
Validation	Age 40-59	1418	438	0.795	0.765	0.822
Validation	Age 60 or older	2039	875	0.724	0.699	0.749
Validation	Mexican American	316	120	0.835	0.782	0.88
Validation	Other Hispanic	482	148	0.8	0.746	0.848
Validation	Non-Hispanic White	2938	784	0.832	0.81	0.852
Validation	Non-Hispanic Black	529	240	0.727	0.675	0.776
Validation	Non-Hispanic Asian	269	87	0.852	0.791	0.902
Validation	Other or multiracial	292	95	0.755	0.667	0.828
Weighted area under the receiver operating characteristic curve with 95% confidence intervals from 1,000 bootstrap						

replicates resampled within outcome strata. Subgroups with fewer than 50 participants or fewer than 10 events are not shown.						

Table S6. Development cycle: differences in net benefit and in operating characteristics between the penalised logistic regression and the ADA risk test applied at its action threshold of 5, with 95% confidence intervals from 2000 Rao–Wu bootstrap replicates resampling primary sampling units within strata.

Threshold probability	Net benefit, model minus ADA rule	95% CI	Net benefit, model minus test-everyone	95% CI
0.050	0.0624	0.0514 to 0.0746	0.0047	0.0022 to 0.0071
0.100	0.0490	0.0373 to 0.0618	0.0196	0.0135 to 0.0253
0.150	0.0384	0.0266 to 0.0511	0.0408	0.0316 to 0.0489
0.200	0.0308	0.0190 to 0.0430	0.0689	0.0575 to 0.0791
0.250	0.0293	0.0176 to 0.0412	0.1077	0.0951 to 0.1191
0.265	0.0268	0.0149 to 0.0399	0.1185	0.1044 to 0.1306
0.300	0.0298	0.0160 to 0.0445	0.1545	0.1368 to 0.1706
0.350	0.0392	0.0239 to 0.0555	0.2171	0.1946 to 0.2372
0.400	0.0517	0.0307 to 0.0737	0.2919	0.2639 to 0.3178
0.500	0.1048	0.0806 to 0.1304	0.5066	0.4712 to 0.5383

0.600	0.2071	0.1747 to 0.2417	0.8513	0.8076 to 0.8895

Out-of-fold predictions; mobile-examination-centre weights.

Table S6 (continued). Operating-point differences.

Quantity	Difference	95% CI	p
Sensitivity at the questionnaire's specificity, model minus questionnaire (percentage points)	10.6	5.9 to 15.0	<0.001
Proportion tested at threshold 0.30, model minus questionnaire (percentage points)	-3.6	-5.4 to -1.7	<0.001
Sensitivity at threshold 0.30, model minus questionnaire (percentage points)	3.8	-0.4 to 8.1	0.08

Table S7. Model versus ADA risk test under two treatments of unobserved questionnaire items, and paired sensitivity comparisons of the penalised logistic regression against the main model.

Population	ADA risk test	Penalised	Paired difference	p

	AUROC (95% CI)	logistic AUROC (95% CI)	(95% CI)	
All development-cycle participants (n = 6649); unobserved questionnaire items scored as zero per the published rule	0.741 (0.721 to 0.758)	0.810 (0.788 to 0.830)	+0.0694 (+0.0493 to +0.0906)	<0.001
Participants with every questionnaire item observed (n = 5763)	0.730	0.802	+0.0720 (+0.0483 to +0.0968)	<0.001

Table S7 (continued). Paired comparisons against the main penalised logistic regression (AUROC 0.8104), same participants and folds, 20 imputations, fixed penalty.

Analysis	AUROC	Difference vs main model (95% CI)	p

Physical activity added (S4)	0.8102	-0.0002 (-0.0007 to +0.0001)	0.23
Self-reported high blood glucose re-included (S5)	0.8199	+0.0095 (+0.0035 to +0.0156)	<0.001
Non-fasting serum glucose added (S11, added after the analysis plan)	0.8344	+0.0240 (+0.0184 to +0.0295)	<0.001
Imputed values constrained to the observed range of each predictor	0.8105	+0.0001 (-0.0002 to +0.0005)	0.56
Cross-validation folds formed at the primary-sampling-unit level	0.8117	not paired (different fold structure)	

Table S8. Imputation diagnostics, development cycle, first of 20 imputations: observed versus imputed distributions for predictors with more than 5% missingness.

Predictor	Missing, %	n imputed	Observed mean (SD)	Imputed mean (SD)	Observed 5th–95th percentile	Imputed range	Imputed values outside

							observed range, %
Pulse rate, beats/min	17.39	1156	69.35 (11.23)	69.02 (11.71)	52.67 to 88.67	31.29 to 103.78	0.4
Family income to poverty ratio	13.78	916	2.64 (1.63)	2.52 (1.62)	0.40 to 5.00	-3.06 to 7.28	13.0
Diastolic blood pressure, mmHg	9.63	640	74.97 (11.48)	74.14 (11.95)	57.67 to 95.00	38.97 to 113.99	0.3
Systolic blood pressure, mmHg	9.63	640	123.75 (18.65)	123.55 (19.83)	98.00 to 159.67	69.82 to 178.52	0.8
Alcohol intake, g/day	5.90	392	6.49 (16.11)	4.83 (17.46)	0.00 to 32.51	-42.94 to 47.59	38.3

Imputation by chained equations with Bayesian ridge regression and posterior sampling, five iterations, fitted on the 48 predictors without the outcome or design variables. Posterior sampling

is unbounded, so a proportion of imputed values lies outside the observed range; a sensitivity analysis constraining imputed values to the observed range is reported in Table S7.