Detection of Type 2 Diabetes from 20-second Speech Recordings: A Large-Scale Validation Study

Elisa Brann[1,2†], Roseline Polle[1†], Giedrė Čepukaitytė[1], Alexandra L. Georgescu[1], Owen Parsons[1], Emilia Molimpakis[1], Stefano Goria[1*]

[1]thymia Ltd., 31 Finsbury Circus, London, EC2M 5SQ, UK.
[2]Institute of Psychiatry, Psychology & Neuroscience (IoPPN), King's College London, 16 De Crespigny Park, London, SE5 8AB, UK.

*Corresponding author(s). E-mail(s): stefano@thymia.ai;
Contributing authors: elisa@thymia.ai; roseline@thymia.ai; giedre@thymia.ai; alexandra@thymia.ai; owen@thymia.ai; emilia@thymia.ai;
†These authors contributed equally to this work.

Abstract

Accessible screening for type 2 diabetes (T2D) is critical, with millions of cases remaining undiagnosed globally. Here, we present the largest known real-world validation study for a speech-based T2D prediction model, trained on speech data from over 21,000 individuals, that works on features extracted from 20-second speech recordings. The model was evaluated in two stages: 1) Against self-reported diagnoses in 7,319 English-speaking participants using AUC, and 2) Against HbA1c blood tests in a subset of 801 participants drawn from the full cohort. Performance was also compared against QDiabetes and in the presence of key confounding variables. The model demonstrated clinically useful predictive capacity on self-reported data (AUC = 0.80 ± 0.03), approaching QDiabetes (AUC = 0.86 ± 0.03). It was robust to most demographic confounds (e.g., age and sex) and medication use, with reduced performance in the presence of comorbidities (e.g., cardiovascular disease and hypertension). At diabetes threshold of HbA1c ≥48 mmol/-mol, the model achieved an AUC of 0.75 (±0.07). This biomarker-validated speech-based tool demonstrates potential to complement existing methods through accessible, scalable screening requiring only a 20-second speech sample.

Keywords: Type 2 diabetes, Voice biomarkers, Machine learning, Digital Health, Screening, Speech analysis

1 Introduction

Type 2 diabetes (T2D) affects an estimated 7% of UK adults, with approximately 30% of cases remaining undiagnosed, equating to roughly one million adults living with undetected diabetes[1]. While early detection is crucial for preventing complications and reducing healthcare costs, current screening methods are opportunistic and/or face significant barriers. For example, the NHS Health Check offers UK adults aged 40-74 a 20-30-minute in-person assessment of lifestyle, family history, and clinical measurements (e.g., blood pressure, cholesterol and BMI) to calculate their risk of developing T2D and other chronic conditions. Although the programme has improved disease detection[2], uptake remains limited, with only 40.4% of eligible adults attending as of 2025[3]. Some of the reasons behind low uptake include aversion to preventive medicine, competing priorities and difficulty accessing GP services[2]. Both screening access and T2D prevalence also vary substantially across demographic groups[1]. Remote, non-invasive screening tools that require minimal time commitment from both patients and GPs may enhance detection rates while also reaching populations currently underserved by traditional healthcare pathways.

Speech-based tools offer particular promise as an accessible, scalable and low-burden option for early-stage screening. Speech contains rich linguistic and paralinguistic signal reflecting both physical and mental health[4]. Everyday digital devices, such as smartphones, can be used to capture audio recordings containing this signal on a large scale[5–7]. Patterns in speech signal, including subtle characteristics that may be imperceptible to conventional assessment, can in turn be extracted using machine learning approaches, enabling automated screening from speech recordings.

However, large-scale studies predicting disease from speech and voice data remain scarce, particularly those capturing the demographic diversity and medical complexity of real-world clinical populations. One notable example, focused on depression and anxiety, gathered data from over 30,000 participants [7]. Models evaluated on an unseen dataset from 2,431 participants predicted symptoms with an area under the curve (AUC) of 0.84 [7]. Given the accessibility of speech-based assessments, a model of equivalent accuracy deployed at population scale could transform existing diabetes screening, reaching individuals who would otherwise miss out on conventional services and increasing uptake of the more time- and resource-intensive tests by those who actually need it.

T2D affects voice through multiple physiological pathways, with severity of vocal changes generally corresponding to disease progression. Individuals with diabetes show a higher incidence of noticeable voice problems such as vocal straining, hoarseness, and significantly shorter maximum phonation time compared to healthy controls[8–10]. These more pronounced, perceptible symptoms are positively correlated with markers of advanced disease, specifically, poor glycemic control and the presence of neuropathy. However, research has historically focused on long-term cumulative damage rather than tracking voice changes across disease stages or investigating the subtler vocal alterations that may occur earlier in disease development[11].

Initial studies demonstrate the potential of speech-based approaches for T2D detection[12–17]. For example, in a large study that utilised data from 3129 participants, Guo and colleagues[12] found that models trained on paralinguistic features outperformed those that relied on clinical features, achieving AUCs of >80%, with jitter and shimmer as important paralinguistic predictors. However, given the tightly controlled conditions under which data for this study were collected, including the quiet setting and professional recording equipment[12], these findings may not generalise to real-world settings. Studies by Kaufman and colleagues [13] and Elbéji and colleagues [14] achieved detection accuracies of 70-75% using smartphone recordings. They identified significant vocal changes in T2D patients related to pitch, intensity, and vocal perturbations[13, 14]. However, these studies were constrained by modest sample sizes (n=267 and 607, respectively) and while both attempted to incorporate diversity into recruitment strategies[13, 14], replication in larger, population-representative cohorts is required before clinical translation to avoid amplifying preexisting disparities in early screening programmes [1, 18, 19].

The present study addresses this critical gap by validating a speech-based T2D detection model, trained on more than 21,000 participants—the largest known training dataset for this task—against the largest and most comprehensive real-world validation dataset to date, comprising over 7000 unique users (more than 12 times larger than previous real-world studies), enabling robust assessment of performance across diverse demographic and clinical subgroups. Beyond sample size, this study has collected extensive personal and health information, allowing systematic evaluation of how the model performs in the presence of key confounding factors and against established screening methods. In line with previous studies[12–15], but on a larger scale, we also incorporated clinical validation through

haemoglobin A1c (HbA1c) biomarker testing. This allowed us to quantify changes in predictive performance from mild (prediabetes) to severe (diabetes) disease stages—a step not undertaken in previous voice-based T2D studies.

Our investigation pursued three primary objectives. First, to evaluate the speech-based model's ability to predict self-reported diabetes status. The rich demographic and health data collected enabled stratified analysis by typical T2D risk factors, such as age, sex, ethnicity, diabetes medication use, namely, no medication or any medication, and comorbid conditions frequently associated with T2D, including cardiovascular disease, hypertension, obesity and chronic kidney disease. This comprehensive approach allowed us to identify where voice-based screening performs optimally and where additional refinement may be needed.

Second, to validate model predictions beyond self-reported diagnoses, we distributed HbA1c blood test kits to over 800 participants, making this the only known large, fully remote speech-based diabetes study with concurrent blood sampling, as previous studies utilised health records or self-report to obtain HbA1c levels[12–15]. Participants in this sample were drawn from the initial cohort stratified by predicted diabetes risk. By linking voice-based risk predictions to objective biomarker measurements carried out within three months of the speech samples, we evaluated model's ability to predict T2D across prediabetes and diabetes stages, providing validation against a clinical gold standard independent of participant self-report.

Third, we compared our speech model against QDiabetes, the National Institute for Health and Care Excellence (NICE)-recommended tool that uses demographic and clinical variables to estimate 10-year T2D risk[20, 21]. In practice, no systematic non-invasive screening pathway exists for T2D in the UK: current detection is largely opportunistic, relying on blood tests taken incidentally during GP appointments. While QDiabetes is recommended for risk identification[21], it is underutilised in practice[2, 3], and, critically, predicts future risk rather than the current disease status[20]. In the absence of a true non-invasive screening comparator, QDiabetes represents the most relevant available benchmark. By directly comparing the performance, we assessed whether analysis of 20-second speech segments offers comparable detection capabilities to existing but under-exploited screening tools based on risk scores computed from demographic and clinical variables and requiring greater time commitment.

2 Methods

2.1 Study Design

This study employed a two-stage approach to validate a voice-based T2D screening tool. Stage 1 evaluated the speech-model performance on a prospectively collected cohort of 7,319 UK adults who completed voice recordings and health questionnaires, with predictions compared against self-reported T2D diagnoses. This cohort was distinct from the model's training dataset (n = 21,129 participants; see Section 2.4). Stage 2 validated model predictions against HbA1c blood biomarkers in a stratified subset of 801 UK adults, a subset of the Stage 1 participant cohort, selected to ensure representation across model-predicted levels of risk of T2D. This design, represented in Figure 1, enabled evaluation of model discrimination against both patient-reported diagnoses and objective clinical biomarkers. This study was retrospectively registered on ClinicalTrials.gov (NCT07421921) on 11/27/2025.

2.2 Participants

Participants were recruited using an online research participation platform (Prolific) and were required to be 18 years or older, speak English as a first language, have no language difficulties, have normal or corrected to normal eye-sight, no hearing impairments, and be a current resident in the UK.

2.3 Ethical Considerations

The ethical review process for this validation study was led by the University of Portsmouth Research Ethics Committee (25/ETHICS/011), with a favourable opinion granted on August 19, 2025. All participants gave written informed consent and were compensated for their time. Individuals who completed the HbA1c test were also provided with access to a written personalised doctor's report

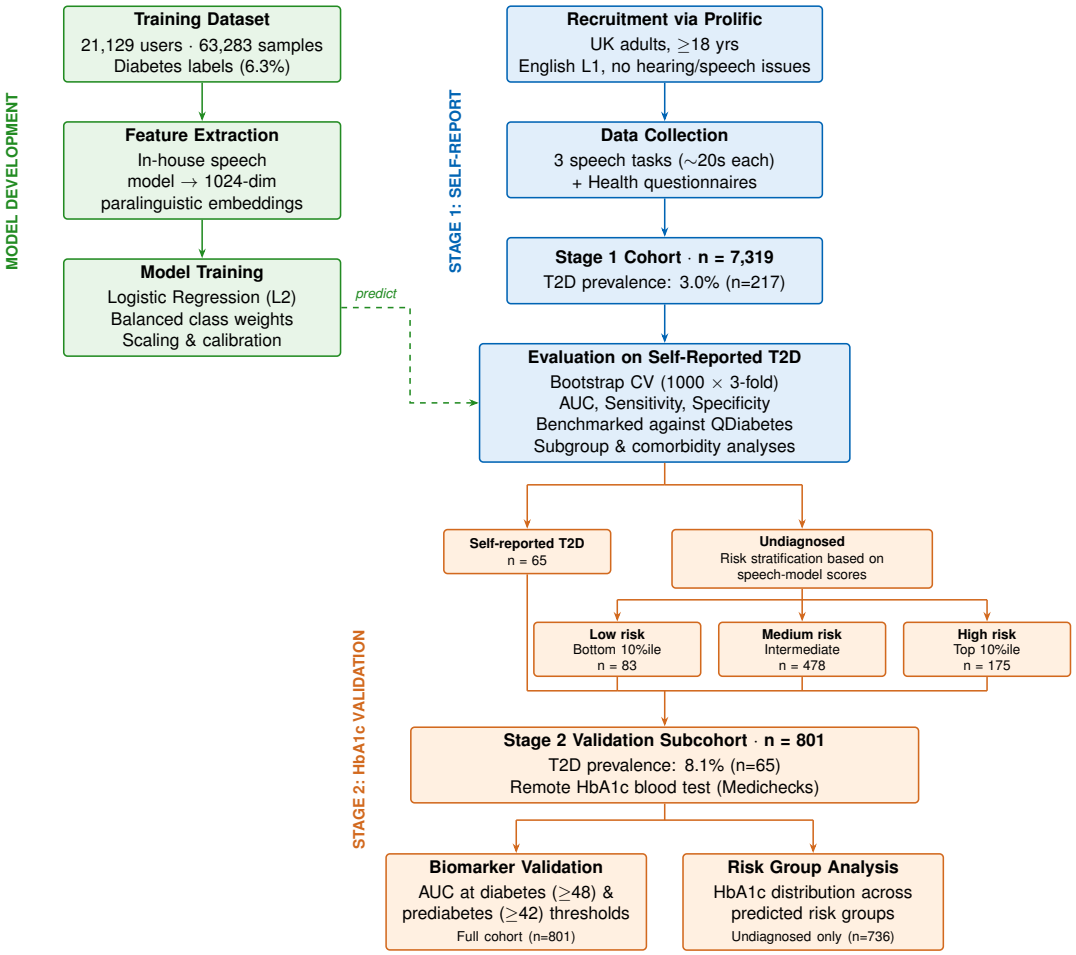

Fig. 1 Study design. The speech model was developed using a proprietary dataset of 63,283 voice samples from 21,129 unique English-speaking participants residing in the UK or US. The model was evaluated in two stages: Stage 1 included a separate cohort of 7,319 English-speaking participants from the UK who provided self-report data; Stage 2 involved a sub-cohort of Stage 1 who completed home HbA1c tests for validation against gold-standard. This subset of participants included both individuals with self-reported type 2 diabetes (T2D) and without self-reported diabetes diagnosis (Undiagnosed). Undiagnosed participants were selected for Stage 2 based on model-predicted diabetes risk.

detailing the results of their test and follow-on care guidance. All study procedures were performed in accordance with the Declaration of Helsinki.

2.4 Model Development and Training

Training data. The training data processing and model development pipeline is illustrated in Figure 1 (top left). The model was developed using a proprietary dataset of 63,283 voice samples from 21,129 unique English-speaking participants whose country of residence was either the United Kingdom or the United States, collected via the thymia research platform [22]. Recordings comprised a reading-aloud task and two free-speech tasks. Alongside speech data, participants completed a general health questionnaire that asked whether they had been diagnosed with diabetes of *any* type. As the original data collection was not designed around Type 2 diabetes specifically, the questionnaire did not distinguish between diabetes types; we refer to this as the *noisy* or *mixed-type* diabetes label, with an estimated prevalence of 6.3% (1,318 participants). Of these, an estimated 80-90% are type 2, consistent with UK and US population estimates [23, 24]. The model was subsequently evaluated in two stages using *clean* Type 2 diabetes labels: first against self-reported T2D status from a targeted health questionnaire (Stage 1; Section 2.5.1), and then against HbA1c-derived labels from remote blood testing (Stage 2; Section 2.5.2).

Feature extraction. Voice recordings were preprocessed through an in-house pipeline involving resampling to 16kHz and trimming silences to produce segments of 10-20s. Paralinguistic 1024-dimensional features were extracted from trimmed audio segments using an in-house speech model based on TRILLsson5[25]. To balance data quality with real-world generalisability, minimal pre-processing was applied. Samples were excluded only if insufficient speech was present to reach the 10-second threshold or recording quality was inadequate for automated transcription. Here we use an automated-transcription tool (Deepgram base model) only as a quality control tool, as well as to trim silences. These criteria excluded approximately 5% of collected samples. We additionally excluded samples missing questionnaire data required for QDiabetes comparison (see Section 2.6.2) calculation (e.g., weight), corresponding to 7% of collected samples. Due to overlap between these exclusion criteria, 10% of samples were dropped in total. All reported numbers and results use this filtered dataset.

Prediction model. A logistic regression classifier with L2 regularisation ($C = 0.001$) was trained on the 1024-dimensional embeddings using the `lbfgs` solver and balanced class weights to account for the low prevalence of diabetes in the training population. A standard scaler was applied to normalise the input features. The model was calibrated using Platt scaling (sigmoid method) via 3-fold cross-validation on the training set, so that among individuals assigned a predicted probability of X%, approximately X% are expected to have the condition.

2.5 Data Collection

2.5.1 Stage 1: Voice and Self Reported Health Information

Individuals completed the same speech activities as described in Section 2.4. The reading out loud task involved a standard text commonly used as a speech elicitation task due to its phonetic range (the Aesop fable "The North Wind and the Sun"[26]), whilst the free-speech tasks required participants to speak at their usual volume and pace in response to two different questions ("What did you do last weekend?"; "What has your mood been like over the past couple of weeks?"), similar to the protocols used in [7, 27–29]. Self-report measures included three questionnaires covering demographics, current state (including questions relevant to voice quality at the time of recording, such as current physical and respiratory health, smoking/substance use, sleep, and device used), and health information (covering in depth questions relevant to chronic health conditions including diabetes, cardiovascular health, kidney disease, respiratory health, medication use, neurological/psychiatric health, and screening questions relevant to HbA1c testing).

2.5.2 Stage 2: Remote HbA1c testing

HbA1c measures glycated hemoglobin, reflecting average blood glucose levels over the preceding three months, and is the gold-standard diagnostic test for T2D in the NHS. The test involved finger-prick blood collection (5ml) using a lancet, with results categorised as: <42 mmol/mol (6.0%): normal; 42-47 mmol/mol (6.0-6.4%): prediabetes; ≥48 mmol/mol (≥6.5%): diabetes. Tests were sourced and processed via Medichecks Ltd., a United Kingdom Accreditation Service (UKAS)-accredited direct-to-consumer blood testing company.

To administer the HbA1c tests, participants were first divided into those with self-reported T2D and those without. Low, medium and high risk groups among participants without self-reported T2D were then defined based on their predicted speech model scores, with individuals from across these groups recruited via Prolific to complete at-home HbA1c blood tests within three months of providing their voice data. High-risk and low-risk groups comprised those in the highest and lowest 10th percentile, respectively, while medium-risk participants were randomly selected from the intermediate score range. This sampling strategy resulted in a skewed distribution for HbA1c testing, with extreme risk scores overrepresented compared to the broader cohort that completed the self-report questionnaire. This stratified sampling strategy both ensured adequate representation across the predicted risk spectrum to assess discrimination between risk groups, and enriched for potential diabetes cases relative to random sampling, improving statistical power to evaluate model performance given the limited HbA1c testing resources.

HbA1c results provided objective, current diabetes status independent of self-reported diagnoses, accounting for individuals in remission or with undiagnosed diabetes. The continuous HbA1c values also enabled analysis across the spectrum of abnormal glycemic control, from prediabetes to diabetes. For model validation, self-reported diabetes diagnoses from the health questionnaire were

disregarded in favour of HbA1c-confirmed status, enabling assessment of the model's ability to detect physiologically-confirmed diabetes.

2.6 Statistical Analysis and Performance Evaluation

2.6.1 Evaluation strategy

All performance evaluations were conducted using the reading out loud task only, with a single recording of 10 to 20s per participant. To obtain robust performance estimates with optimised decision thresholds, we employed a bootstrap cross-validation (CV) procedure. We created 1000 bootstrap resamples of the test set (N=7,319 participants), and for each bootstrap sample performed 3-fold cross-validation. Within each CV fold, the decision threshold was optimised on 2/3rd of the data (validation split) to maximize balanced accuracy, then performance metrics were calculated on the remaining 1/3rd (evaluation split). This yielded 3000 replicate performance measurements (1000 bootstrap iterations $\times$ 3 CV folds), allowing us to report mean $\pm$ standard deviation for all metrics. The primary outcomes of the speech-model included area under the curve (AUC), sensitivity (recall) and false positive rate (1 - specificity). We also evaluated Expected Calibration Error (ECE) using 10 equal-width bins to assess the alignment between predicted probabilities and observed accuracy across confidence levels.

2.6.2 Comparison with QDiabetes

Model performance was compared against QDiabetes-2018, the NICE-recommended demographic risk assessment tool for diabetes screening in the UK[21]. QDiabetes calculates 10-year diabetes development risk based on age, BMI, ethnicity, family history, and clinical factors. While QDiabetes and our voice-based model address different clinical questions (future risk vs. current detection), QDiabetes functions as a case-finding tool in practice, and in the absence of any systematic non-invasive screening alternative, represents the closest available clinical benchmark. This comparison provides context for performance in population screening scenarios.

When comparing model performance, we use bootstrap significance testing[30, 31]. For each bootstrap sample, we compute the difference in the metric of interest (e.g., AUC) between models (speech-model minus QDiabetes), such that positive values indicate higher speech-model performance. This yields a distribution of differences; if the 95% confidence interval excludes 0, the difference is statistically significant. We also report a two-tailed p-value, calculated as twice the smaller proportion of bootstrap differences falling on either side of zero.

2.6.3 Subgroup Analyses

To assess model generalisability and identify potential confounds, we evaluated performance across demographic subgroups (age, sex, ethnicity, BMI), clinical characteristics (comorbid conditions and diabetes medication use). Unlike the full evaluation strategy (Section 2.6.1), subgroup analyses used AUC as the sole metric. Since AUC is threshold-independent, no threshold optimisation or cross-validation was required; bootstrap resampling (1000 iterations) was retained to estimate confidence intervals.

2.6.4 Sensitivity to training label noise

To assess the impact of mixed-type diabetes training labels on model performance (i.e. type 1, type 2, gestational), we compared two speech models using 5-fold stratified cross-validation within the Stage 1 cohort (n=7,319). The first model was trained on verified T2D labels collected in the present study. The second model replicated the model's training conditions, using mixed-type diabetes labels and speech embeddings from an earlier data collection wave involving the same participants. Both models were evaluated against verified T2D status using the Stage 1 embeddings. Performance was compared using bootstrap resampling (1,000 iterations per fold, 5,000 total), with a two-tailed p-value computed as twice the minimum of the proportions of bootstrap samples where the AUC difference was less than or equal to zero or greater than or equal to zero.

3 Results

3.1 Participant Characteristics

Demographics and participant characteristics across the different stages of the study are outlined in Table 1.

Table 1 Participant Demographics and Characteristics Across Study Datasets. Demographic and clinical characteristics of participants in the training dataset (used for model development), self-report health dataset (Stage 1), and HbA1c dataset (Stage 2). To maximise data availability for model training, we utilised an existing dataset that employed broader, less granular labels for diabetes status (combining Type 1, Type 2, gestational, and other forms without distinction). Consequently, detailed demographic and clinical information collected in validation stages (such as specific ethnicity categories, BMI, and comorbid conditions) are not available for the training cohort.

Datasets	**Training data**	**Self-report health data (Phase 1)**	**HbA1c data (Phase 2)**
N	21,129	7,319	801
Diabetes prevalence			
Any	1,318 (6.3%)	467 (6.4%)	35 (4.4%)
T2D	-	217 (3.0%)	35 (4.4%)
Confirmed	-	-	29 (3.6%)
Undiagnosed	-	-	6 (0.7%)
T1D	-	58 (0.8%)	-
Gestational	-	160 (2.2%)	-
Other (induced)	-	32 (0.4%)	-
# speech samples	63,283	7,319	801
Birth Sex			
Female	13,341 (63.1%)	4,917 (67.2%)	496 (61.0%)
Male	7,715 (36.5%)	2,402 (32.8%)	305 (38.1%)
Other	73 (0.3%)	-	-
Ethnicity			
White	15,368 (72.7%)	6,003 (82.0%)	685 (85.5%)
Black	3,079 (14.6%)	547 (7.5%)	45 (5.6%)
Asian & South Asian	1,109 (5.2%)	476 (6.5%)	46 (5.7%)
Asian	-	158 (2.2%)	17 (2.1%)
South Asian	-	318 (4.3%)	29 (3.6%)
Mixed	1,107 (5.2%)	244 (3.3%)	20 (2.5%)
Other	462 (2.2%)	49 (0.7%)	5 (0.6%)
BMI			
<18.5	-	119 (1.6%)	8 (1.0%)
18.5–24.9	-	2,855 (39.0%)	272 (34.0%)
25.0–29.9	-	2,403 (32.8%)	229 (28.6%)
>30.0	-	1,942 (26.5%)	292 (36.5%)
Age (years)			
<40	12,162 (57.6%)	3,970 (54.2%)	312 (39.0%)
>40	8,967 (42.4%)	3,349 (45.8%)	489 (61.0%)
Co-morbid chronic health condition			
CVD	-	185 (2.5%)	33 (4.1%)
Hypertension	-	1,089 (14.9%)	202 (25.1%)
CKD	-	190 (2.6%)	8 (1.0%)
Medications			
No medication	-	7,090 (96.9%)	745 (93%)
Any diabetes medication	-	229 (3.1%)	56 (7.0%)

Abbreviations: T2D = Type 2 diabetes; T1D = Type 1 diabetes; CVD = cardiovascular disease; CKD = chronic kidney disease; BMI = body mass index; Confirmed = self-reported cases of T2D; Undiagnosed = newly diagnosed cases of T2D in individuals who did not self-report as having T2D during Stage 1 data collection.

3.2 Model Performance

3.2.1 Predicting T2D from self-reported health information

Discrimination of T2D detection based on self-reported data are described in Table 2. Assuming prediction models with AUC values above 0.80 can be considered clinically-useful[32, 33], we observed the speech-model was effective at distinguishing between people who self-reported a diagnosis of T2D and those who did not. The speech model also achieved an ECE of 0.019, usually considered to represent good calibration.

Table 2 Discrimination of T2D detection based on self report across speech and QDiabetes models. Values are shown as mean ± standard deviation across 3000 iterations.

Metric	Speech model	QD
AUC	0.80 ± 0.03	0.86 ± 0.03
Recall (Sensitivity)	0.76 ± 0.10	0.82 ± 0.07
False Positive Rate (1 - Specificity)	0.31 ± 0.06	0.24 ± 0.03
ECE	0.019 ± 0.004	0.017 ± 0.004

Abbreviations: AUC = Area Under Receiver Operating Characteristic Curve; ECE = Expected Calibration Error; QD = QDiabetes.

3.2.2 Speech model vs QDiabetes for predicting self-reported diabetes

For overall prediction of T2D, the speech model showed modestly lower overall discrimination than QDiabetes (ΔAUC = -0.06, 95% CI [-0.08, -0.03], $p < 0.001$; Table 2, Figure 2). Differences in sensitivity ($p = 0.29$) and specificity ($p = 0.11$) were not statistically significant. Calibration was good for both models (ECE of 0.019 and 0.017), with no statistical difference (p=0.15). Similarly, both models discriminated well between self-reported T2D and type 1 diabetes, with no statistically significant difference in AUC (ΔAUC = -0.02, 95% CI [-0.09, 0.05], $p = 0.54$; Table 3). However, when extending to all past diabetes diagnoses (including gestational and other induced forms), the speech model significantly outperformed QDiabetes (ΔAUC = 0.17, 95% CI [0.12, 0.22], $p < 0.001$), with QDiabetes discrimination dropping substantially (AUC = 0.62) while the speech model remained stable (AUC = 0.79).

Table 3 Performance of the speech model and QDiabetes when discriminating T2D from self-reported type 1 diabetes and any present or past diabetes diagnosis (including type 1, gestational and other induced forms of diabetes).

Subgroup/Condition	Sample size (T2D/ subgroup)	Prevalence	Speech model AUC	QD AUC
Overall	217/7319	3.0%	0.80	0.86
On population with:				
Type 1 + Type 2 diabetes only	217/275	79.0%	0.81	0.83
Any diabetes diagnosis	217/467	46.5%	0.79	0.62

Abbreviations: AUC = Area Under Receiver Operating Characteristic Curve; T2D = Type 2 Diabetes; QD = QDiabetes.

3.2.3 Speech model's performance across demographic subgroups

Figure 3 shows the AUCs over different demographic and comorbidity subgroups.

T2D prevalence varies substantially across demographic groups and frequently co-occurs with conditions that may themselves affect voice characteristics. To evaluate model robustness and identify potential sources of bias or confounding, we assessed discrimination across demographic subgroups (birth sex, age and ethnicity), chronic comorbid conditions and medication status (Table 4).

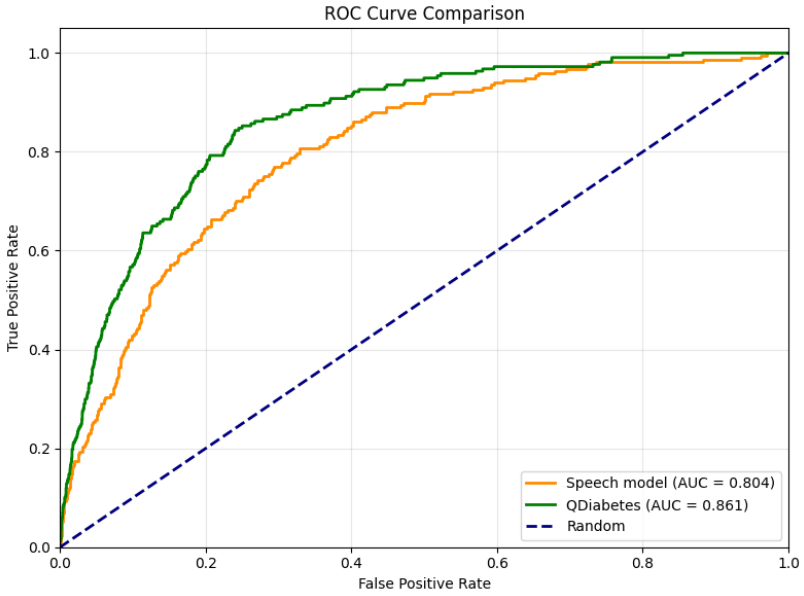

Fig. 2 ROC curves comparing the speech model and QDiabetes for T2D prediction on self-reported health data (Stage 1).

The model demonstrated robust performance across most demographic groups. Sex-specific performance was comparable (male AUC 0.79; female AUC 0.81), as was performance across most age categories (AUC $\geq$ 0.75). Ethnicity-specific performance remained strong for most subgroups (AUC $\geq$ 0.8), with lower performance observed in the Black (AUC 0.69) and Asian (AUC 0.65) populations. However, these subgroups have limited representation of T2D cases ($n = 18$ and $n = 5$, respectively), despite reasonable overall sample sizes ($n = 547$ and $n = 159$), which limits the reliability of these findings.

3.2.4 Speech model's performance across comorbidities and medication status

Type 2 diabetes is often co-morbid with multiple other chronic health conditions, including cardiovascular disease (CVD), hypertension, BMI > 30, indicating obesity, and chronic kidney disease (CKD). We observed lower performance predicting T2D in groups with most of the conditions that commonly co-occur with T2D (CVD AUC 0.69; hypertension AUC 0.65; BMI > 30 AUC 0.73) but comparable performance for CKD (AUC 0.82) and other groups without these respective conditions (AUC $\geq$ 0.8). A similar pattern was observed for QDiabetes, which also showed reduced performance in comorbid groups while maintaining strong discrimination in non-comorbid populations (Table 4).

We also examined performance in individuals with long COVID, a condition increasingly linked to T2D[34]. Here, the speech model retained reasonable discrimination (AUC = 0.74) whereas QDiabetes performance dropped substantially (AUC = 0.59), though the difference did not reach statistical significance (ΔAUC = 0.15, 95% CI $[-0.005, 0.31]$, $p = 0.054$; Table 4).

Finally, we examined how diabetes medication use affects model performance. We found that both models maintain an AUC of 0.81 within people taking at least one type of diabetes medication ($N = 229$ of which 66.8% have T2D), and drops slightly for people taking no diabetes medication ($N = 7,090$ of which 0.9% have T2D) for the speech model (AUC=0.74) but stayed constant for QDiabetes (AUC=0.80, ΔAUC $=-0.06$, 95% CI $[-0.119, -0.003]$, $p = 0.034$).

3.2.5 Speech model validation using HbA1c across diabetes severity thresholds

At the diabetes threshold (HbA1c $\geq$ 48 mmol/mol), the speech model achieved an AUC of 0.75 ($\pm$0.07), correctly identifying 82% ($\pm$23%) of diabetes cases, though 47% ($\pm$11%) of non-diabetic individuals were classified as high-risk. By comparison, QDiabetes demonstrated comparable performance in the same cohort (AUC 0.77 ± 0.08) with a 72% ($\pm$22%) sensitivity and 32% ($\pm$12%) false positives. No differences reached statistical significance for any metric (ΔAUC $= -0.02$, 95% CI $[-0.11, 0.05]$, $p = 0.60$; sensitivity $p = 0.51$; specificity $p = 0.15$).

To check if the lower AUCs as compared to self-reported diabetes were due to the skewed population distribution in the reduced HbA1c evaluation set (801 participants vs 7,319 for self-reported

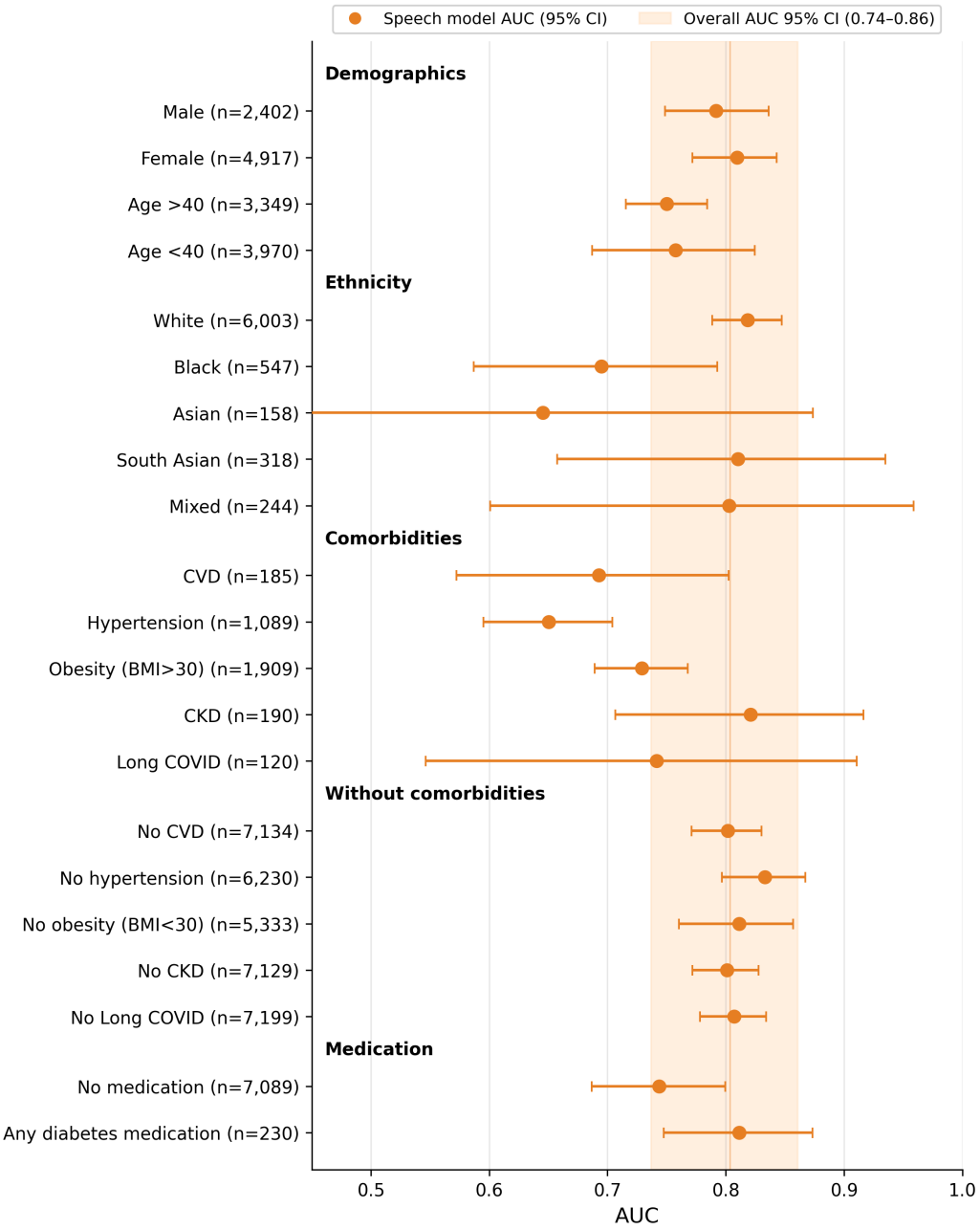

Fig. 3 Area Under the Curve (AUC) and 95% Confidence Interval (CI) when predicting Type 2 Diabetes (T2D) in different subgroups . CVD = cardiovascular disease, CKD = chronic kidney disease, BMI = body mass index.

data), which was selected based on predicted risk groups, as described in 2.5.2., we also computed AUC for the speech model and QDiabetes for this reduced population using the self-reported label. We found AUCs of 0.72 ± 0.06 and 0.76 ± 0.06, respectively, comparable to the results obtained above, and with no significant differences for any metric (ΔAUC $= -0.04$, 95% CI $[-0.09, 0.02]$, $p = 0.20$). This confirms that the lower AUCs observed for HbA1c-based diagnosis are attributable to the skewed population distribution in the reduced HbA1c evaluation set, rather than to a degradation in model discriminative performance.

At the prediabetes threshold (HbA1c $\geq$ 42 mmol/mol), the model achieved AUC 0.73 ($\pm$0.06), detecting 79% ($\pm$16%) of cases with 43% ($\pm$9%) false positives, while QDiabetes demonstrated stronger discrimination at this threshold (AUC 0.80 ± 0.06, 79 ± 13% sensitivity, 30 ± 7% false positives). Again, the difference in AUC was not statistically significant (ΔAUC $= -0.06$, 95% CI $[-0.13, 0.00]$, $p = 0.054$), and the recall and specificity differences were also non significant (p=0.97 and p=0.09 respectively).

Table 4 Overview of type 2 diabetes (T2D) in overall sample and common co-morbid health condition and medication subgroups, speech-model and QDiabetes predictions also listed. ΔAUC = speech-model AUC − QDiabetes AUC. Positive values indicate higher speech model performance.

	Sample size (T2D /sub-group size)	**T2D prevalence (%)**	**Speech model AUC**	**QD AUC**	**ΔAUC**(95% CI)	***p***
Overall	217/7319	3.0%	0.80	0.86	−0.06 (−0.09,−0.03)	<0.001
By condition:						
CVD	21/185	11.4%	0.69	0.77	−0.08 (−0.22,0.05)	0.204
Hypertension	114/1090	10.4%	0.65	0.72	−0.07 (−0.13,−0.01)	0.018
Obesity (BMI>30)	125/1910	6.6%	0.73	0.80	−0.07 (−0.12,−0.03)	<0.001
CKD	18/190	9.4%	0.82	0.83	−0.01 (−0.08, 0.07)	0.908
Long COVID	10/121	8.3%	0.74	0.59	0.15 (−0.01,0.31)	0.054
Without:						
No CVD	196/7134	2.8%	0.80	0.86	−0.06 (−0.09,−0.03)	<0.001
No Hypertension	103/6230	1.6%	0.83	0.86	−0.03 (−0.06,0.01)	0.12
No Obesity (BMI>30)	91/5333	1.7%	0.81	0.86	−0.05 (−0.09,−0.003)	0.04
No CKD	199/7129	2.8%	0.80	0.86	−0.06 (−0.09, −0.03)	<0.001
No Long COVID	207/7199	2.9%	0.81	0.87	-0.06 (−0.09,−0.04)	<0.001
By medication use:						
No medication	64/7090	0.9%	0.74	0.80	−0.06 (−0.12,0.003)	0.034
Any medication	153/229	66.8%	0.81	0.81	0.00 (−0.06,0.07)	0.866

Abbreviations: AUC = Area Under Receiver Operating Characteristic Curve, CKD = Chronic Kidney Disease; QD = QDiabetes.

3.2.6 HbA1c discrimination between predicted risk groups in undiagnosed population

The model effectively distinguished between individuals at different levels of diabetes risk, even within the undiagnosed (no self-reported T2D) population (Figure 4). Participants predicted to be at high risk of undiagnosed diabetes (top 10th percentile; $n = 175$) had significantly higher mean HbA1c levels than both the medium-risk group (intermediate score range; $n = 478$; Δmean = +3.3 mmol/mol, $p < 0.001$; Mann-Whitney U tests with Bonferroni correction for three pairwise comparisons) and the low-risk group (bottom 10th percentile; $n = 83$; Δmean = +5.3 mmol/mol, $p < 0.001$). Conversely, low-risk participants had significantly lower HbA1c levels than the medium-risk group (Δmean = −2.1 mmol/mol, $p < 0.001$). Importantly, none of the participants in the low-risk group had HbA1c levels in the prediabetic ($\geq$ 42 mmol/mol) or diabetic ($\geq$ 48 mmol/mol) range, confirming the model's ability to accurately identify individuals unlikely to have elevated blood glucose.

3.2.7 Sensitivity to training label noise

To assess sensitivity to the mixed-type diabetes labels used for model training, we retrained the model using 5-fold cross-validation within the Stage 1 cohort with clean T2D labels ($n = 7,319$; 217 T2D cases), as well as again with the mixed-type label used in training, as described in Section 2.6.4. The clean-label model achieved AUC 0.775, compared to 0.740 for the mixed-label model (ΔAUC = 0.035, 95% CI [−0.038, 0.113], $p = 0.34$). The difference was not statistically significant, suggesting the mixed-type training labels did not substantially compromise T2D discrimination. Nevertheless, the trend toward higher AUC with clean T2D labels suggests that training on a larger dataset with verified T2D diagnoses could yield meaningful improvements, as the models are trained here with only ∼5,800 speakers per fold.

4 Discussion

This study addressed three critical issues in speech-based diabetes detection: the need for large-scale validation, clinical biomarker validation and comparison to current recommended screening methods. By conducting a two-stage validation study in 7,319 UK adults—a much larger and more

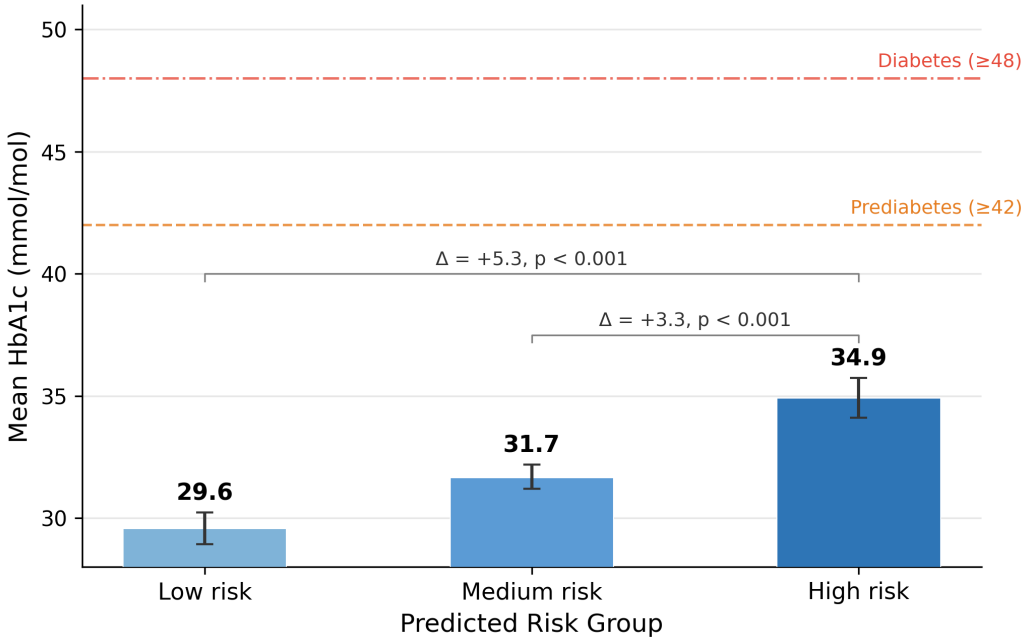

Fig. 4 Mean HbA1c levels for different model-predicted risk groups (low, medium and high). Undiagnosed population only (no self-reported T2D, n=736)

diverse sample than prior validation studies[13–15]—we demonstrated that machine learning analysis of 20-second speech recordings can reliably distinguish individuals with T2D from those without.

4.1 Voice-based screening demonstrates clinically useful detection of T2D

The speech-based model, trained on the largest real-world sample to date, achieved an AUC of 0.80 for distinguishing individuals with and without T2D. This is on par with the 0.80 threshold considered clinically useful[35], though we acknowledge such thresholds lack standardised justification[32]. The model was well calibrated (ECE = 0.019), ensuring that prediction scores are meaningful — i.e., that the predicted probability actually corresponds to approximately the same percentage of individuals having the condition. Performance approached that of QDiabetes (AUC 0.86), the risk assessment tool recommended by NICE for the NHS T2D risk identification programme[21]. In the absence of a system-wide non-invasive screening pathway—T2D detection in the UK being largely opportunistic and reliant on blood testing—QDiabetes represents the most relevant available benchmark. Risk prediction algorithms like QDiabetes, while nominally predicting future T2D incidence (in this case, 10-year risk), often function as case-finding tools in practice. We therefore evaluated both tools on the same task, identifying current T2D status, providing a direct performance comparison[20]. At optimised decision thresholds, our model correctly identified three-quarters of T2D cases (sensitivity 0.76), with a false positive rate of approximately 30%. No significant differences were found for sensitivity ($p = 0.29$) or specificity ($p = 0.11$) between the speech model and QDiabetes, indicating that despite lower overall discrimination, the two methods performed comparably at their respective operating points. This suggests voice-based screening could serve a role in diabetes detection pathways, offering a non-invasive, fast and scalable first-line option for risk prediction that can be followed up with more time- and resource-intensive but objective blood tests if risk is identified.

4.2 Voice-based screening enables greater nuance when distinguishing diabetes type

An unexpected yet valuable finding emerged when evaluating broader diabetes categories: We examined a subset of participants who reported any diabetes diagnosis (past or present, including type 1, gestational, or drug-induced forms) and evaluated how well the model distinguished T2D from these other diabetes types. Voice-based predictions remained consistent while QDiabetes performance dropped substantially. This may be because other forms of diabetes (often temporary, e.g., gestational diabetes), share risk factor profiles with T2D, for example, the demographics and clinical

history of affected individuals[36], and QDiabetes cannot discriminate well between these overlapping characteristics. In contrast, the speech model captures paralinguistic features that may reflect not only the demographic characteristics embedded in voice but also recent physiology, potentially enabling better discrimination between current and past cases.

4.3 Voice-based screening is robust to emerging diseases

Another interesting exploratory finding was that, in contrast to QDiabetes (AUC 0.59), the speech model retained reasonable performance in individuals with long COVID (AUC 0.74). COVID-19 has been associated with a disproportionate increase in T2D cases[37]. In turn, T2D is itself a risk factor for developing long COVID, with the two conditions also sharing several important comorbidities, including cardiovascular and kidney disease[34]. QDiabetes' reliance on predefined risk factors (established through historical population statistics and updated infrequently) may limit its adaptability to emerging conditions like long COVID that alter diabetes risk profiles. In contrast, data-driven models, such as the speech model used in this study, can be readily implemented to discover patterns directly from data without requiring pre-specification of risk factors. This creates potential advantages when new health conditions emerge that affect diabetes risk, or population risk profiles shift over time. Nevertheless, given the modest sample size of long COVID cases (n=121), these findings remain preliminary and warrant validation in larger, dedicated studies.

4.4 Voice-based screening is robust to most confounding (demographic) factors

Stratified analyses across demographic and clinical subgroups revealed generally consistent performance. Sex-specific discrimination was comparable (male AUC 0.79; female AUC 0.81) and also exceeded the gender-specific results reported by an earlier voice-biomarker diabetes screening model who achieved AUC values of 0.75 and 0.71, respectively[14]. Performance also remained robust across most age categories (AUC ≥ 0.75) and ethnic groups (AUC > 0.80). However, discrimination declined among Black and Asian participants (AUC ≤ 0.70). This warrants further investigation using samples with higher representation of T2D cases ($n = 18$ Black and $n = 5$ Asian cases in the current samples of $n = 547$ and $n = 159$ individuals, respectively), both because of the higher diabetes burden in these ethnic groups and the purported differences in T2D-related voice characteristics across ethnicities noted in the literature, although not without skepticism[1, 18, 38].

4.5 Voice-based T2D screening declines in several high-risk comorbid populations and unmedicated participants

The speech model performed worse among samples with CVD (AUC 0.69), hypertension (AUC 0.65) and obesity (BMI > 30, AUC 0.73). The reduced discrimination in individuals with these conditions likely reflects substantial physiological overlap. For example, both obesity and T2D share common pathophysiological mechanisms—insulin resistance, chronic inflammation, and metabolic dysfunction[39]—that may produce similar vocal changes. Additionally, obesity directly affects voice, such as through increased adipose tissue in the neck and chest, altering respiratory dynamics and vocal tract resonance[40]. Similarly, vocal changes have been observed in individuals with hypertension, which, if untreated, ultimately leads to cardiovascular disease, although the underlying mechanisms in this case are less clear[41, 42]. From a screening perspective, this limitation is not necessarily problematic. CVD, hypertension and obesity are strongly associated with undiagnosed T2D prevalence and represent priority populations for screening[43]. A tool that flags these individuals for confirmatory blood glucose testing—even if the vocal signatures partially reflect the comorbidities themselves rather than T2D-specific pathophysiology—still achieves the clinical objective of identifying individuals who warrant further evaluation for potential undiagnosed disease.

Surprisingly, we did not observe a decline in the speech model's performance in individuals with CKD (AUC 0.82), another chronic condition associated with T2D. CKD often develops over time in poorly controlled T2D and is exacerbated by hypertension. Given that the speech models ability to predict T2D in this group is preserved or even slightly better, it may be that vocal changes associated with CKD are distinct from those seen in T2D. Alternatively, CKD as a result of current T2D may result in vocal changes that are informative for the model.

The performance pattern observed in comorbid populations suggests the speech model is not simply exploiting comorbidities as predictive shortcuts. If the model relied primarily on detecting

comorbidities to predict diabetes, we would expect substantially worse performance in their absence. However, the model maintains strong discrimination ($AUC \geq 0.80$) in non-comorbid populations, reinforcing that it is not simply exploiting these factors. Moreover, even in comorbid groups, AUC never drops below 0.65, indicating the model retains meaningful discriminative ability above chance across all subgroups.

It is not well understood whether diabetes medication can independently affect voice, inadvertently helping the speech model identify individuals with T2D. We therefore also looked into how well the model performed in individuals on any diabetes medication and unmedicated participants. We found that prediction performance in the medicated group (n = 229) was similar to that in the overall sample (AUC 0.81). In contrast, in the group taking no diabetes medication (n = 7090), performance dropped to an AUC of 0.74. The complexity of diabetes therapy correlates with disease severity[44]. This suggests that the reduction in performance may reflect a harder discrimination problem, with T2D sufferers in this group potentially having a mild form of the disease. Similarly, the medicated group likely consists of individuals with active and more severe forms of any type of diabetes, suggesting that, as discussed in Section 4.2, the speech model can pick up on differences among these subgroups.

4.6 HbA1c validation: evidence for sustained performance against physiological testing

For Stage 2 HbA1c validation, we tested a subset of 801 participants drawn from the original sample of 7,319 using home blood test kits supplied by a UK-accredited direct-to-consumer blood test company. To our knowledge, no other large scale study performed remote blood testing, making this the only known study with concurrent (within three months) HbA1c sampling. The use of HbA1c tests has confirmed that the speech model can detect changes associated with diabetes across biomarker thresholds, with consistent performance in both diabetes (HbA1c $\geq$ 48 mmol/mol: AUC 0.75) and prediabetes (HbA1c $\geq$ 42 mmol/mol: AUC 0.73). While performance was lower than on self-reported data, a similar drop was observed for QDiabetes (AUC 0.77 and 0.80, respectively), and is likely attributable to the skewed population distribution in the reduced HbA1c evaluation set rather than a fundamental limitation of either model.

The relatively similar performance of the speech model across both diabetes thresholds (diabetic and prediabetic) is notable given that vocal changes are expected to become more pronounced as the disease progresses. HbA1c reflects average blood glucose levels over the preceding three months and may not reflect levels at the time of the speech recording. However, it is possible that paralinguistic aspects of speech picked up by the model vary with the level of blood glucose, irrespective of the disease stage[45]. Several studies have also proposed that it is the variability in blood glucose levels, rather than elevated blood glucose itself, that causes detrimental effects in diabetes, including oxidative stress, neuropathy and retinopathy, more common in advanced disease stages[46–48]. Future studies that measure blood glucose variability continuously could help further characterize how the speech model's paralinguistic features contribute to its predictions. Although the confound analyses suggest the model is largely robust, it remains possible that some demographic signal contributes in part; even so, this would not diminish the model's screening utility—demographic characteristics are well-established predictors of diabetes risk, and the ability to infer them passively from a voice sample, without requiring explicit self-report, represents a practical advantage, especially when competing priorities and limited access to GP services are cited as potential reasons behind the low update for preventative screening[2].

Finally, the model also showed good performance in distinguishing between individuals at different levels of diabetes risk, even within the undiagnosed (by self-report) sample. Participants predicted to be at high risk of T2D had significantly higher mean HbA1c levels as compared to medium- or low-risk groups. Furthermore, none of the low-risk participants, as predicted by the model, had concerning levels of HbA1c. This underscores the potential of speech based tools for use in population-wide screening programmes.

4.7 Clinical implications

The findings of this study suggest that voice-based tools could be used for diabetes screening. We present a model that was trained on the largest real-world dataset to date and which achieved clinically useful discrimination on self-reported data (AUC = 0.80). Crucially, it maintained strong

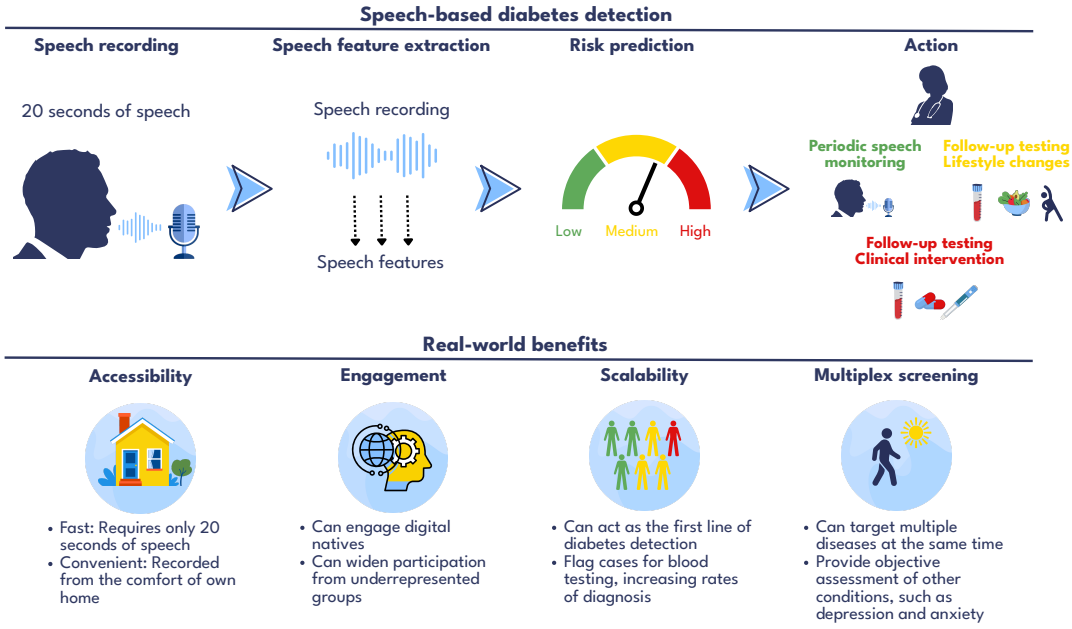

Fig. 5 Summary of what speech-based diabetes detection could look like and real-world benefits of this approach.

performance when validated against HbA1c biomarkers (AUC = 0.75, AUC = 0.73), where both diabetic and non-diabetic labels were objectively confirmed through blood testing. This demonstrates the model's ability to detect diabetes beyond known, self-reported cases—addressing a critical gap given that approximately 1 million individuals in the UK remain undiagnosed.

The practical advantages of voice-based screening extend beyond predictive performance (see Figure 5). Current screening pathways in the UK rely either on opportunistic blood tests taken during regular GP appointments, or time-intensive screening appointments, such as the NHS Health Check, which require 20-30 minutes and include blood draws for glucose and lipids, with approximately 40% of eligible individuals not attending[3]. In contrast, voice-based screening requires only a 20-second recording obtainable remotely without needles, clinical visits, questionnaire completion, or recall of detailed medical history and may also be capable of flagging the risk of other diseases at the same time (e.g., hypertension[42]). This could reduce barriers to participation for individuals who face challenges accessing conventional healthcare—whether due to aversion to blood tests, time constraints, mobility limitations, or geographic distance from clinical facilities[2]. Only those individuals that are flagged to be at high risk could be asked to complete blood testing to confirm diagnosis and administer lifestyle or clinical interventions, optimising diabetes screening and saving patient and clinician time. The technology's minimal friction may particularly appeal to younger, digitally-native populations who are often underrepresented in preventive screening but will age into higher-risk categories. In addition, if done correctly, for example ensuring representation in both model training and tool design [49, 50], the implementation of such technology has the potential to close the gap in health inequities, such as the documented under-diagnosis of Black and Asian ethnic groups with T2D [1]. Therefore, we suggest that future versions of our model may meaningfully complement existing diabetes detection strategies.

Voice biomarkers also offer inherent scalability advantages. Unlike fixed questionnaire-based tools, machine learning models can be iteratively updated to incorporate emerging risk factors—as demonstrated during the COVID-19 pandemic when new metabolic risk profiles emerged. As mentioned previously, voice analysis is not disease-specific: the same recording protocol used here for diabetes screening has demonstrated utility for detecting depression and anxiety[7]. Voice-based tools have also been experimentally deployed for detecting and monitoring respiratory conditions [51, 52]. This creates potential for integrated multi-condition (multiplex) screening from a single voice sample, maximising clinical yield while maintaining low participant burden. Such approaches could enable opportunistic screening across multiple conditions simultaneously, fundamentally expanding the scope of preventive healthcare accessible through digital platforms.

In practice, voice-based tools may be implemented as the first line of T2D screening (see Figure 5). Speech recordings can be made before (e.g., as a stand-alone screening) or during primary care appointments, either on the telephone or in person. Features automatically extracted from these recordings can be fed into the risk prediction model that will generate a risk score. The latter can then be used by the primary care practitioner to triage patients for either follow-up speech-based monitoring that could involve an AI agent, or, in case of medium or high risk scores, blood workup to confirm diagnosis, followed by, depending on the risk level, either lifestyle or clinical interventions. This would reduce the need for time-consuming demographic or invasive blood tests when they are not required, saving both money and clinician time, which can then be allocated to otherwise improving patient care.

5 Conclusions

In summary, this study presents a large-scale, biomarker-validated speech-based screening tool for T2D. Using a cohort twelve times larger than previous studies, we demonstrated that 20-second voice recordings can detect diabetes with discrimination approaching QDiabetes (AUC = 0.80 vs 0.86), the NICE-recommended but underutilised screening tool for UK primary care [21]—and the most relevant available benchmark in the absence of any systematic non-invasive screening alternative. Our speech model was robust to most demographic confounds. Furthermore, validation against HbA1c biomarkers provided preliminary evidence that the model can identify diabetes cases beyond self-reported diagnoses, including individuals with undiagnosed disease. The minimal burden of voice-based screening—requiring no clinical appointments, specialised personnel, or invasive procedures—positions it as a scalable complement to existing screening pathways. Combined with its potential for multi-condition detection from a single voice sample, and future work on improving performance in ethnic minorities and in the presence of comorbidities, this approach could meaningfully expand access to preventive healthcare, particularly for populations facing barriers to conventional screening.

Acknowledgements. We thank all participants who contributed to this study. We are also grateful to the Medichecks team for their support. We acknowledge the contributions of consultant physicians Dr Joanna Bilak (MBBS MRCP) and Dr Andrew Solomon (BM BCh MA DM FRCP), and Paula Mason (RD, CDCES; registered dietitian and diabetes educator) for their consultation on the health questionnaire and experimental design.

Author contributions. S.G., E.M. and E.B conceived the study, E.B. and A.L.G. conducted the study, R.P., E.B. and O.P. analysed the results. R.P., E.B., O.P., A.L.G, G.Č., S.G. and E.M. contributed to the writing and editing of the manuscript.

Funding. This research did not receive any external funding.

Competing interests. E.M. and S.G. are co-founders of thymia Ltd. E.B., R.P., G.Č., O.P., and A.L.G. are employees of thymia Ltd. E.M., S.G., O.P., and A.L.G. hold equity in the company, which may benefit from commercialisation of technologies similar to those described in this paper.

Data availability. The datasets generated and analysed during the current study are not publicly available due to lack of consent for public sharing of raw data, which due to the nature of speech data would compromise participant privacy. They may be made available on reasonable request to the corresponding author for non-commercial research purposes related to detecting or monitoring diabetes (the purposes for which consent for research data sharing was obtained), subject to completion of a Data Sharing Agreement with thymia Ltd.

Code availability.

References

[1] Office for National Statistics (ONS): Risk factors for pre-diabetes and undiagnosed type 2 diabetes in England: 2013 to 2019. statistical bulletin (2024)

[2] Tanner, L., Kenny, R., Still, M., Ling, J., Pearson, F., Thompson, K., Bhardwaj-Gosling, R.: NHS Health Check programme: a rapid review update. BMJ Open **12**(2), 052832 (2022) https://doi.org/10.1136/bmjopen-2021-052832 . Accessed 2026-02-10

[3] Public Health England: NHS Health Check (2025)

[4] Sara, J.D.S., Orbelo, D., Maor, E., Lerman, L.O., Lerman, A.: Guess What We Can Hear—Novel Voice Biomarkers for the Remote Detection of Disease. Mayo Clinic Proceedings **98**(9), 1353–1375 (2023) https://doi.org/10.1016/j.mayocp.2023.03.007 . Accessed 2026-03-09

[5] Budd, J., Baker, K., Karoune, E., Coppock, H., Patel, S., Payne, R., Tendero Cañadas, A., Titcomb, A., Hurley, D., Egglestone, S., Butler, L., Mellor, J., Nicholson, G., Kiskin, I., Koutra, V., Jersakova, R., McKendry, R.A., Diggle, P., Richardson, S., Schuller, B.W., Gilmour, S., Pigoli, D., Roberts, S., Packham, J., Thornley, T., Holmes, C.: A large-scale and PCR-referenced vocal audio dataset for COVID-19. Scientific Data **11**(1), 700 (2024) https://doi.org/10.1038/s41597-024-03492-w . Accessed 2026-03-09

[6] Bot, B.M., Suver, C., Neto, E.C., Kellen, M., Klein, A., Bare, C., Doerr, M., Pratap, A., Wilbanks, J., Dorsey, E.R., Friend, S.H., Trister, A.D.: The mPower study, Parkinson disease mobile data collected using ResearchKit. Scientific Data **3**(1), 160011 (2016) https://doi.org/10.1038/sdata.2016.11 . Accessed 2026-03-09

[7] Norbury, A., Fairs, G., Georgescu, A.L., Nour, M.M., Molimpakis, E., Goria, S.: A multimodal Bayesian network for symptom-level depression and anxiety prediction from voice and speech data. Scientific Reports **16**(1), 5397 (2026) https://doi.org/10.1038/s41598-025-33331-w . Accessed 2026-02-09

[8] Hamdan, A.-l., Jabbour, J., Nassar, J., Dahouk, I., Azar, S.T.: Vocal characteristics in patients with type 2 diabetes mellitus. European Archives of Oto-Rhino-Laryngology **269**(5), 1489–1495 (2012) https://doi.org/10.1007/s00405-012-1933-7 . Accessed 2026-02-10

[9] Hamdan, A.-L., Kurban, Z., Azar, S.T.: Prevalence of phonatory symptoms in patients with type 2 diabetes mellitus. Acta Diabetologica **50**(5), 731–736 (2013) https://doi.org/10.1007/s00592-012-0392-3 . Accessed 2026-02-10

[10] Gölaç, H., Atalik, G., Türkcan, A.K., Yilmaz, M.: Disease related changes in vocal parameters of patients with type 2 diabetes mellitus. Logopedics Phoniatrics Vocology **47**(3), 202–208 (2022) https://doi.org/10.1080/14015439.2021.1917653 . Accessed 2026-02-10

[11] Saghiri, M.A., Vakhnovetsky, A., Vakhnovetsky, J.: Scoping review of the relationship between diabetes and voice quality. Diabetes Research and Clinical Practice **185**, 109782 (2022) https://doi.org/10.1016/j.diabres.2022.109782 . Accessed 2026-02-10

[12] Guo, J., Peng, W., Hu, S., Lu, D., Chen, S.: A Novel Machine Learning-Driven Voice and Clinical Biomarkers Framework for Robust Prediction of Type 2 Diabetes Mellitus. Journal of Voice, 089219972500400 (2025) https://doi.org/10.1016/j.jvoice.2025.09.033 . Accessed 2026-03-10

[13] Kaufman, J.M., Thommandram, A., Fossat, Y.: Acoustic Analysis and Prediction of Type 2 Diabetes Mellitus Using Smartphone-Recorded Voice Segments. Mayo Clinic Proceedings: Digital Health **1**(4), 534–544 (2023) https://doi.org/10.1016/j.mcpdig.2023.08.005 . Accessed 2026-02-10

[14] Elbéji, A., Pizzimenti, M., Aguayo, G., Fischer, A., Ayadi, H., Mauvais-Jarvis, F., Riveline, J.-P., Despotovic, V., Fagherazzi, G.: A voice-based algorithm can predict type 2 diabetes status in USA adults: Findings from the Colive Voice study. PLOS Digital Health **3**(12), 0000679 (2024) https://doi.org/10.1371/journal.pdig.0000679 . Accessed 2026-02-06

[15] Oreskovic, J., Fazli, G., Varma, V., Malik, K., Kaufman, J., Fossat, Y.: Voice-based prediction of prediabetes using classical machine learning models. Frontiers in Clinical Diabetes and Healthcare **6**, 1697769 (2025) https://doi.org/10.3389/fcdhc.2025.1697769 . Accessed 2026-03-11

[16] Jeon, J., Palanica, A., Sarabadani, S., Lieberman, M., Fossat, Y.: Biomarker potential of real-world voice signals to predict abnormal blood glucose levels. Systems Biology (2020). https://doi.org/10.1101/2020.09.25.314096 . http://biorxiv.org/lookup/doi/10.1101/2020.09.25.314096

Accessed 2026-03-11

[17] Summoogum, K., Das, D., Kumaran, S., Bhagra, S.: A Voice-based triage for Type 2 Diabetes using a Conversational Virtual Assistant in the Home Environment: Voice-based triage for Type 2 diabetes using seven non-identifiable acoustic biomarkers collected via a home virtual assistant. In: Proceedings of the 2025 9th International Conference on Medical and Health Informatics, pp. 260–269. ACM, Kyoto Japan (2025). https://doi.org/10.1145/3761712.3761773 . https://dl.acm.org/doi/10.1145/3761712.3761773 Accessed 2026-03-11

[18] Nagai, K., Chung, H.-F., Hayashi, K., Dobson, A.J., Ideno, Y., Sandin, S., Van Der Schouw, Y.T., Hardy, R., Anderson, D.J., Demakakos, P., Brunner, E.J., Mitchell, E.S., Woods, N.F., Eastwood, S.V., El Khoudary, S.R., Hedderson, M.M., Weiderpass, E., Mishra, G.D.: The Association Between Race/Ethnicity and Risk of Type 2 Diabetes in Women Varies by BMI: A Pooled Analysis of Individual Data From 15 Cohort Studies. Diabetes Care **49**(2), 247–256 (2026) https://doi.org/10.2337/dc25-1478 . Accessed 2026-02-11

[19] Whyte, M.B., Hinton, W., McGovern, A., Van Vlymen, J., Ferreira, F., Calderara, S., Mount, J., Munro, N., De Lusignan, S.: Disparities in glycaemic control, monitoring, and treatment of type 2 diabetes in England: A retrospective cohort analysis. PLOS Medicine **16**(10), 1002942 (2019) https://doi.org/10.1371/journal.pmed.1002942 . Accessed 2026-02-20

[20] Hippisley-Cox, J., Coupland, C.: Development and validation of QDiabetes-2018 risk prediction algorithm to estimate future risk of type 2 diabetes: cohort study. BMJ, 5019 (2017) https://doi.org/10.1136/bmj.j5019 . Accessed 2026-02-10

[21] National Institute of Health and Care Excellence (NICE): Type 2 diabetes: prevention in people at high risk. Public health guideline **PH38** (2012)

[22] Fara, S., Goria, S., Molimpakis, E., Cummins, N.: Speech and the n-Back task as a lens into depression. How combining both may allow us to isolate different core symptoms of depression. In: Interspeech 2022. ISCA. https://doi.org/10.21437/Interspeech.2022-10393 . https://www.isca-archive.org/interspeech_2022/fara22_interspeech.html Accessed 2026-02-10

[23] Diabetes UK: How many people in the UK have diabetes? Accessed [date] (2025). https://www.diabetes.org.uk/about-us/about-the-charity/our-strategy/statistics

[24] Centers for Disease Control and Prevention: Type 2 Diabetes. Accessed [date] (2026). https://www.cdc.gov/diabetes/about/about-type-2-diabetes.html

[25] Shor, J., Venugopalan, S.: TRILLsson: Distilled Universal Paralinguistic Speech Representations. In: Interspeech 2022. ISCA. https://doi.org/10.21437/Interspeech.2022-118 . https://www.isca-archive.org/interspeech_2022/shor22_interspeech.html Accessed 2026-02-10

[26] Deterding, D.: The North Wind versus a Wolf: short texts for the description and measurement of English pronunciation. Journal of the International Phonetic Association **36**(2), 187–196 (2006) https://doi.org/10.1017/S0025100306002544 . Accessed 2026-02-10

[27] Mazur, A., Costantino, H., Tom, P., Wilson, M.P., Thompson, R.G.: Evaluation of an ai-based voice biomarker tool to detect signals consistent with moderate to severe depression. The Annals of Family Medicine **23**(1), 60–65 (2025)

[28] Schwoebel, J.W., Schwartz, J., Warrenburg, L.A., Brown, R., Awasthi, A., New, A., Butler, M., Moss, M., Pissadaki, E.K.: A longitudinal normative dataset and protocol for speech and language biomarker research. medrxiv, 2021–08 (2021)

[29] Larsen, E., Murton, O., Song, X., Joachim, D., Watts, D., Kapczinski, F., Venesky, L., Hurowitz, G.: Validating the efficacy and value proposition of mental fitness vocal biomarkers in a psychiatric population: prospective cohort study. Frontiers in Psychiatry **15**, 1342835 (2024)

[30] Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap, 0 edn. Chapman and Hall/CRC, ??? (1994). https://doi.org/10.1201/9780429246593 . https://www.taylorfrancis.com/books/9781000064988 Accessed 2026-02-11

[31] DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L.: Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics **44**(3), 837–845 (1988)

[32] De Hond, A.A.H., Steyerberg, E.W., Van Calster, B.: Interpreting area under the receiver operating characteristic curve. The Lancet Digital Health **4**(12), 853–855 (2022) https://doi.org/10.1016/S2589-7500(22)00188-1 . Accessed 2026-02-10

[33] Van Calster, B., Collins, G.S., Vickers, A.J., Wynants, L., Kerr, K.F., Barreñada, L., Varoquaux, G., Singh, K., Moons, K.G.M., Hernandez-boussard, T., Timmerman, D., Mclernon, D.J., Van Smeden, M., Steyerberg, E.W.: Performance evaluation of predictive AI models to support medical decisions: Overview and guidance. arXiv. Version Number: 1 (2024). https://doi.org/10.48550/ARXIV.2412.10288 . https://arxiv.org/abs/2412.10288 Accessed 2026-02-10

[34] Berends, M.S., Homburg, M., Kupers, T., Meijer, E.N., Bos, I., Verheij, R., Kuiper, J., Berger, M.Y., Peters, L.L.: Impact of pre-existing comorbidities and multimorbidities, demography and viral variants on post-acute sequelae of COVID-19 ('Long COVID') in Dutch primary care: A retrospective cohort study. International Journal of Infectious Diseases **156**, 107912 (2025) https://doi.org/10.1016/j.ijid.2025.107912 . Accessed 2026-02-06

[35] Çorbacıoğlu, K., Aksel, G.: Receiver operating characteristic curve analysis in diagnostic accuracy studies: A guide to interpreting the area under the curve value. Turkish Journal of Emergency Medicine **23**(4), 195–198 (2023) https://doi.org/10.4103/tjem.tjem_182_23 . Accessed 2026-02-24

[36] Dennison, R.A., Chen, E.S., Green, M.E., Legard, C., Kotecha, D., Farmer, G., Sharp, S.J., Ward, R.J., Usher-Smith, J.A., Griffin, S.J.: The absolute and relative risk of type 2 diabetes after gestational diabetes: A systematic review and meta-analysis of 129 studies. Diabetes Research and Clinical Practice **171**, 108625 (2021) https://doi.org/10.1016/j.diabres.2020.108625 . Accessed 2026-02-06

[37] Izzo, R., Pacella, D., Trimarco, V., Manzi, M.V., Lombardi, A., Piccinocchi, R., Gallo, P., Esposito, G., Lembo, M., Piccinocchi, G., Morisco, C., Santulli, G., Trimarco, B.: Incidence of type 2 diabetes before and during the COVID-19 pandemic in Naples, Italy: a longitudinal cohort study. eClinicalMedicine **66**, 102345 (2023) https://doi.org/10.1016/j.eclinm.2023.102345 . Accessed 2026-02-06

[38] Sidorova, J., Anisimova, M.: Impact of Diabetes Mellitus on Voice: A Methodological Commentary. Journal of Voice **36**(2), 294–129412 (2022) https://doi.org/10.1016/j.jvoice.2020.05.015 . Accessed 2026-02-10

[39] Wu, H., Ballantyne, C.M.: Metabolic Inflammation and Insulin Resistance in Obesity. Circulation Research **126**(11), 1549–1564 (2020) https://doi.org/10.1161/CIRCRESAHA.119.315896 . Accessed 2026-02-10

[40] Bosso, J.R., Martins, R.H.G., Pessin, A.B.B., Tavares, E.L.M., Leite, C.V., Naresse, L.E.: Vocal Characteristics of Patients With Morbid Obesity. Journal of Voice **35**(2), 329–732911 (2021) https://doi.org/10.1016/j.jvoice.2019.09.012 . Accessed 2026-02-10

[41] Taghibeyglou, B., Kaufman, J.M., Fossat, Y.: Machine Learning-Enabled Hypertension Screening Through Acoustical Speech Analysis: Model Development and Validation. IEEE Access **12**, 123621–123629 (2024) https://doi.org/10.1109/ACCESS.2024.3443688 . Accessed 2026-02-09

[42] Sara, J.D.S., Maor, E., Borlaug, B., Lewis, B.R., Orbelo, D., Lerman, L.O., Lerman, A.: Noninvasive vocal biomarker is associated with pulmonary hypertension. PLOS ONE **15**(4), 0231441 (2020) https://doi.org/10.1371/journal.pone.0231441 . Accessed 2026-02-09

[43] Klein Woolthuis, E.P., De Grauw, W.J.C., Van Gerwen, W.H.E.M., Van Den Hoogen, H.J.M., Van De Lisdonk, E.H., Metsemakers, J.F.M., Van Weel, C.: Yield of Opportunistic Targeted Screening for Type 2 Diabetes in Primary Care: The Diabscreen Study. The Annals of Family Medicine **7**(5), 422–430 (2009) https://doi.org/10.1370/afm.997 . Accessed 2026-02-24

[44] Luzuriaga, M., Leite, R., Ahmed, H., Saab, P.G., Garg, R.: Complexity of antidiabetic medication regimen is associated with increased diabetes-related distress in persons with type 2 diabetes mellitus. BMJ Open Diabetes Research & Care **9**(1), 002348 (2021) https://doi.org/10.1136/bmjdrc-2021-002348 . Accessed 2026-03-12

[45] Kaufman, J., Jeon, J., Oreskovic, J., Fossat, Y.: Linear effects of glucose levels on voice fundamental frequency in type 2 diabetes and individuals with normoglycemia. Scientific Reports **14**(1), 19012 (2024) https://doi.org/10.1038/s41598-024-69620-z . Accessed 2026-03-12

[46] Ceriello, A., Esposito, K., Piconi, L., Ihnat, M.A., Thorpe, J.E., Testa, R., Boemi, M., Giugliano, D.: Oscillating Glucose Is More Deleterious to Endothelial Function and Oxidative Stress Than Mean Glucose in Normal and Type 2 Diabetic Patients. Diabetes **57**(5), 1349–1354 (2008) https://doi.org/10.2337/db08-0063 . Accessed 2026-02-11

[47] Jia, Y., Long, D., Yang, Y., Wang, Q., Wu, Q., Zhang, Q.: Diabetic peripheral neuropathy and glycemic variability assessed by continuous glucose monitoring: A systematic review and meta-analysis. Diabetes Research and Clinical Practice **213**, 111757 (2024) https://doi.org/10.1016/j.diabres.2024.111757 . Accessed 2026-02-11

[48] Kim, H.U., Park, S.P., Kim, Y.-K.: Long-term HbA1c variability and the development and progression of diabetic retinopathy in subjects with type 2 diabetes. Scientific Reports **11**(1), 4731 (2021) https://doi.org/10.1038/s41598-021-84150-8 . Accessed 2026-02-11

[49] Osonuga, A., Osonuga, A.A., Fidelis, S.C., Osonuga, G.C., Juckes, J., Olawade, D.B.: Bridging the digital divide: artificial intelligence as a catalyst for health equity in primary care settings. International Journal of Medical Informatics **204**, 106051 (2025) https://doi.org/10.1016/j.ijmedinf.2025.106051 . Accessed 2026-02-24

[50] Wilson, S., Tolley, C., Mc Ardle, R., Lawson, L., Beswick, E., Hassan, N., Slight, R., Slight, S.: Recommendations to advance digital health equity: a systematic review of qualitative studies. npj Digital Medicine **7**(1), 173 (2024) https://doi.org/10.1038/s41746-024-01177-7 . Accessed 2026-02-24

[51] Weglarz, K., Szczygieł, E., Masłoń, A., Blaut, J.: Assessment of breathing patterns and voice of patients with COPD and dysphonia. Respiratory Medicine **240**, 108012 (2025) https://doi.org/10.1016/j.rmed.2025.108012 . Accessed 2026-02-24

[52] Rizos, G., Calvo, R.A., Schuller, B.W.: Positive-Pair Redundancy Reduction Regularisation for Speech-Based Asthma Diagnosis Prediction. In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5. IEEE, Rhodes Island, Greece (2023). https://doi.org/10.1109/ICASSP49357.2023.10097087 . https://ieeexplore.ieee.org/document/10097087/ Accessed 2026-02-24