nature medicine

LETTERS

https://doi.org/10.1038/s41591-020-1010-5

Check for updates

A digital biomarker of diabetes from smartphone-based vascular signals

Robert Avram[1], Jeffrey E. Olgin[1], Peter Kuhar[2], J. Weston Hughes[3], Gregory M. Marcus[1], Mark J. Pletcher[4], Kirstin Aschbacher[1,5,6,7] and Geoffrey H. Tison[1,5,7] ✉

The global burden of diabetes is rapidly increasing, from 451 million people in 2019 to 693 million by 2045[1]. The insidious onset of type 2 diabetes delays diagnosis and increases morbidity[2]. Given the multifactorial vascular effects of diabetes, we hypothesized that smartphone-based photoplethysmography could provide a widely accessible digital biomarker for diabetes. Here we developed a deep neural network (DNN) to detect prevalent diabetes using smartphone-based photoplethysmography from an initial cohort of 53,870 individuals (the 'primary cohort'), which we then validated in a separate cohort of 7,806 individuals (the 'contemporary cohort') and a cohort of 181 prospectively enrolled individuals from three clinics (the 'clinic cohort'). The DNN achieved an area under the curve for prevalent diabetes of 0.766 in the primary cohort (95% confidence interval: 0.750–0.782; sensitivity 75%, specificity 65%) and 0.740 in the contemporary cohort (95% confidence interval: 0.723–0.758; sensitivity 81%, specificity 54%). When the output of the DNN, called the DNN score, was included in a regression analysis alongside age, gender, race/ethnicity and body mass index, the area under the curve was 0.830 and the DNN score remained independently predictive of diabetes. The performance of the DNN in the clinic cohort was similar to that in other validation datasets. There was a significant and positive association between the continuous DNN score and hemoglobin A1c (*P* ≤ 0.001) among those with hemoglobin A1c data. These findings demonstrate that smartphone-based photoplethysmography provides a readily attainable, non-invasive digital biomarker of prevalent diabetes.

Globally, half of all people living with diabetes are undiagnosed (~224 million), and 79% live in low- and middle-income countries[1]. Diabetes causes both macrovascular and microvascular multi-organ disease, including coronary heart disease, stroke, neuropathy and kidney disease, among others[3]. A readily attainable, non-invasive digital biomarker of diabetes could facilitate disease detection by making it easier to identify at-risk individuals who would benefit from confirmatory diagnostic testing using hemoglobin A1c (HbA1c) data. Such a tool would have particular impact in underserved populations and those out of reach of traditional medical care.

Photoplethysmography (PPG) is a non-invasive optical technique that detects blood flow changes through a vascular bed[4]. It involves shining light into tissue, such as the fingertip or wrist, and quantifying the backscattered light that corresponds with changes in blood volume[4]. PPG has long been used clinically to measure heart rate (HR) and peripheral blood oxygen saturation[4], and research applications have ranged from detection of hypertension[5] to detection of various cardiovascular abnormalities[6,7]. Until recently, PPG recording required specialized equipment; however, technological developments have enabled PPG measurement from sensors on smart devices, such as smartphones and fitness trackers. The rapid worldwide adoption of smart devices over the past decade[8] provides an opportunity to develop non-invasive, widely scalable digital biomarkers for diseases such as diabetes[9].

PPG is uniquely positioned to capture the multifactorial sequelae of diabetes resulting from a variety of pathophysiologic mechanisms. PPG readily captures sequential heartbeats, enabling not only its long-standing use for HR measurement, but also analysis of HR variability (HRV), which is impacted by diabetic autonomic and neural regulatory effects[10–12]. Recently, a shared genetic etiology between resting HR and diabetes was identified, implicating mechanisms ranging from metabolism to endothelial aging[13]. Indeed, endothelial dysfunction is an early hallmark of diabetic vascular disease, and is readily detectable in the PPG waveform[14]. Similarly, diabetes-related microvascular arteriosclerosis[6,15] and neuropathy can affect PPG[16]. Given the multitude of mechanisms by which diabetes impacts PPG, algorithmic analysis of PPG should ideally leverage the complete PPG recording and all of the morphologic and temporal information contained therein. DNNs are a class of algorithms[17] that have successfully achieved complex pattern recognition for various medical tasks[18–20]. DNNs provide the advantage of being agnostic to specific sets of predetermined PPG features suspected to predict diabetes, and instead detect patterns using the full PPG record. We therefore hypothesized that PPG obtained from commercially available smartphones and analyzed using a DNN could identify individuals with and without diabetes.

In this study, we first developed and validated a DNN to detect prevalent diabetes in a 'primary cohort', composed of 53,870 Health eHeart study[21] participants who contributed 2,589,448 PPG recordings between 1 April 2014 and 30 April 2018 (Fig. 1a). Participants self-reported diabetes status and measured PPG by placing an index fingertip on the smartphone camera using the Azumio Instant Heart Rate iOS application (Azumio, Inc; Fig. 1b). The primary cohort was randomly split into training (70%, $n = 37{,}709$) and development (10%, $n = 4{,}848$) datasets—used to train and tune the DNN, respectively—and a test dataset (20%, $n = 11{,}313$), used for DNN validation. The DNN outputs a 'DNN score' between 0 and 1, with higher

[1]Division of Cardiology, Department of Medicine and Cardiovascular Research Institute, University of California, San Francisco, San Francisco, CA, USA. [2]Azumio, Inc, Redwood City, CA, USA. [3]Department of Computer Science, University of California, Berkeley, Berkeley, CA, USA. [4]Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, CA, USA. [5]Bakar Computational Health Sciences Institute, University of California, San Francisco, San Francisco, CA, USA. [6]Department of Psychiatry, Weill Institute for Neurosciences, University of California, San Francisco, San Francisco, CA, USA. [7]These authors contributed equally: Kirstin Aschbacher, Geoffrey H. Tison. ✉e-mail: geoff.tison@ucsf.edu

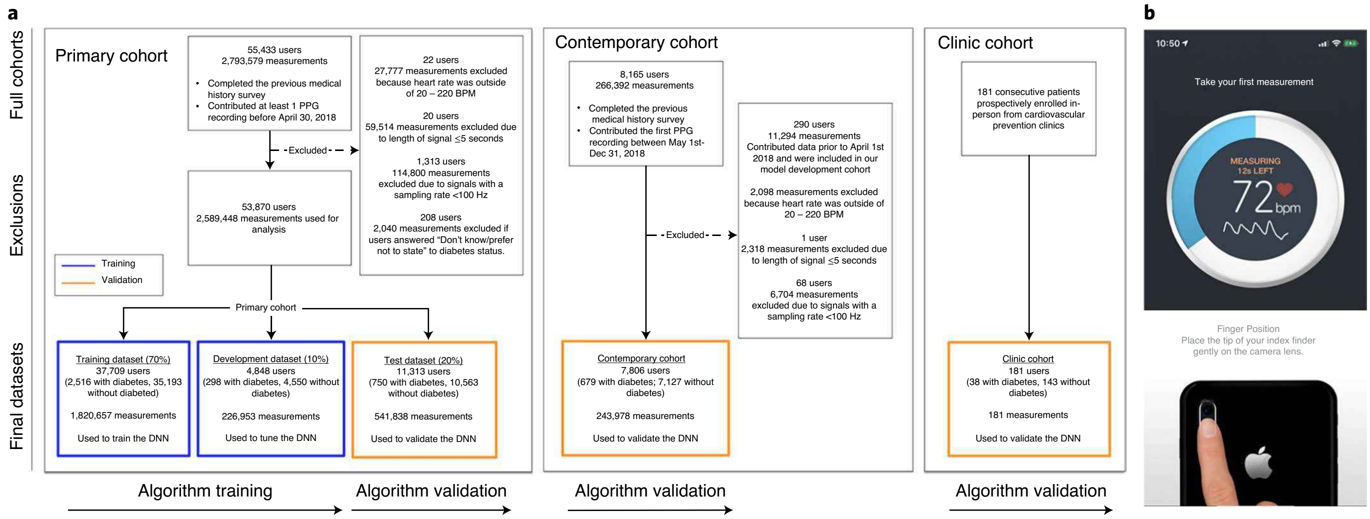

a
Full cohorts
Exclusions
Final datasets
Primary cohort
55,433 users
2,793,579 measurements
Completed the previous medical history survey
Contributed at least 1 PPG recording before April 30, 2018
Excluded
22 users
27,777 measurements excluded because heart rate was outside of 20 – 220 BPM
20 users
59,514 measurements excluded due to length of signal ≤5 seconds
1,313 users
114,800 measurements excluded due to signals with a sampling rate <100 Hz
208 users
2,040 measurements excluded if users answered “Don’t know/prefer not to state” to diabetes status.
53,870 users
2,589,448 measurements used for analysis
Training
Validation
Primary cohort
Training dataset (70%)
37,709 users
(2,516 with diabetes, 35,193 without diabeted)
1,820,657 measurements
Used to train the DNN
Development dataset (10%)
4,848 users
(298 with diabetes, 4,550 without diabetes)
226,953 measurements
Used to tune the DNN
Test dataset (20%)
11,313 users
(750 with diabetes, 10,563 without diabetes)
541,838 measurements
Used to validate the DNN
Algorithm training
Algorithm validation
Contemporary cohort
8,165 users
266,392 measurements
Completed the previous medical history survey
Contributed the first PPG recording between May 1st-Dec 31, 2018
Excluded
290 users
11,294 measurements
Contributed data prior to April 1st 2018 and were included in our model development cohort
2,098 measurements excluded because heart rate was outside of 20 – 220 BPM
1 user
2,318 measurements excluded due to length of signal ≤5 seconds
68 users
6,704 measurements excluded due to signals with a sampling rate <100 Hz
Contemporary cohort
7,806 users
(679 with diabetes; 7,127 without diabetes)
243,978 measurements
Used to validate the DNN
Algorithm validation
Clinic cohort
181 consecutive patients prospectively enrolled in-person from cardiovascular prevention clinics
Clinic cohort
181 users
(38 with diabetes, 143 without diabetes)
181 measurements
Used to validate the DNN
Algorithm validation
b
10:50
Take your first measurement
MEASURING
12s LEFT
72 bpm
Finger Position
Place the tip of your index finder gently on the camera lens.

Fig. 1 | Consort diagram describing the study cohorts and screenshots from the smartphone app used for PPG acquisition. a, Description of the datasets used for algorithm development and validation. The DNN was trained using the training and development dataset of the primary cohort (left), and validated using the test dataset of the primary cohort. We additionally validated the DNN in the temporally distinct contemporary cohort (middle) and the prospectively enrolled, in-person clinic cohort (right). The blue outlines indicate datasets used for model development and training. The yellow outlines indicate datasets used for model validation. All datasets are completely separate and do not contain overlapping participants. **b**, Screenshots from the smartphone app used to acquire user-measured PPG recordings using a smartphone app and camera.

scores suggesting greater likelihood of diabetes (see Methods). Since many participants contributed >1 recording, we reported DNN performance using the area under the receiver operating characteristic curve (AUC)[22] at both the ‘recording level’, which treats each recording independently, and the ‘user level’, which averages the DNN score for all recordings provided by a user; user-level assessment was preferred when possible since clinical application calls for classifying a user as having diabetes or not.

In the primary cohort, 3,564 participants (6.6%) had self-reported diabetes and 50,306 (93.3%) did not (Extended Data Figs. 1 and 2). Compared to those without diabetes, those with diabetes were older, male, had a higher HR and body mass index (BMI), and were less likely to be non-Hispanic white. In the hold-out test dataset, the DNN’s AUC to detect diabetes was 0.766 at the user level (95% confidence interval (CI): 0.750–0.782; recording-level AUC = 0.680, 95% CI: 0.678–0.683; Table 1 and Fig. 2a). At the chosen cutoff threshold (DNN score = 0.427), user-level sensitivity was 75% and specificity was 65%. Owing in part to the low prevalence of diabetes in our cohort (6.6%), the positive predictive value (PPV) of the DNN score at the user level and recording level was 13% and 10%, while the negative predictive value (NPV) was 97% and 96%, respectively (Table 1). DNN performance in the development dataset was not significantly different from the test dataset (user-level AUC = 0.766, 95% CI: 0.740–0.792; recording-level AUC = 0.694, 95% CI: 0.691–0.698).

In addition to validating DNN performance in the primary cohort test dataset, we used two additional validation cohorts (Fig. 1a), providing three total examples of algorithm generalizability to datasets distinct from the training dataset[23]. The first was the ‘contemporary cohort’, composed of PPG recordings from 7,806 participants newly enrolled into Health eHeart from 1 May to 31 December 2018. This temporally distinct validation cohort exhibits the DNN’s robustness to secular changes, such as new smartphone models and cameras, that could affect PPG recording. Then, to validate our approach in a real-world clinical setting, we prospectively enrolled an in-person ‘clinic cohort’ composed of 181 consecutive patients referred to three cardiovascular prevention clinics (two in San Francisco, one in Montreal) between 1 November 2018 and 30 July 2019 (Fig. 1a and Extended Data Fig. 3). The DNN’s user-level AUC to detect diabetes in the contemporary cohort was similar to that in the primary cohort: 0.740 (95% CI: 0.723–0.758; recording-level AUC = 0.661, 95% CI: 0.664–0.667); the DNN had higher sensitivity, but lower specificity, versus the primary cohort (Table 1).

In the prospectively enrolled in-person clinic cohort, 38 patients (21.0%) had medical-record-confirmed diabetes (Extended Data Fig. 3). Compared with the primary cohort, the clinic cohort was substantially older, more male and had more comorbidity. The clinic cohort recording-level AUC (0.682, 95% CI: 0.605–0.755) was similar to the recording-level AUC in the test dataset and contemporary cohorts (Table 1). Compared with the test dataset, there was higher sensitivity and PPV, but lower specificity and NPV. When clinic cohort patients with a prior diabetes diagnosis were excluded (n = 17), 21 patients remained who were newly diagnosed by HbA1c during the clinic visit. In this subset of patients with newly diagnosed diabetes, the DNN AUC was 0.644 (95% CI: 0.546–0.744; Table 1); the DNN correctly identified 16 out of 21 patients with newly diagnosed diabetes (Extended Data Fig. 4f).

To investigate whether PPG was predictive of diabetes independently of other predictors and comorbidities, we built nested logistic regression (LogReg) models in the test dataset with and without the inclusion of the DNN score (Table 2). After adjustment for age, gender, race and BMI, the DNN score remained independently and significantly predictive of prevalent diabetes (Table 2 and Supplementary Table 1); the AUC for this prediction model was 0.830 (95% CI: 0.787–0.873; Fig. 2a). The DNN score was also strongly predictive of diabetes independently of all examined comorbidities, including hypertension, hypercholesterolemia and coronary artery disease, among others (Table 2; LogReg model 5); the AUC for this prediction model was 0.830 (95% CI: 0.815–0.844; Fig. 2a). In all models, the DNN score was a strong diabetes predictor and was only slightly attenuated after adjustment (Table 2 and Supplementary Table 1). HRV was no longer a significant predictor of diabetes after the DNN score was added, while HR was attenuated (Table 2; LogReg model 4). Compared to participants with a DNN score below the cutoff (<0.427), those with a DNN score

Table 1 | Performance of the DNN to detect diabetes using PPG in three validation datasets

	AUC (95% CI)	Sensitivity[a]	Specificity[a]	PPV[a]	NPV[a]
Test dataset, $n=11,313$					
User level	0.766 (0.750–0.782)	75.0% (72.0–77.8%)	65.4% (64.6–66.3%)	13.3% (12.3–14.3%)	97.4% (97.0–97.7%)
Recording level	0.680 (0.678–0.683)	66.2% (65.8–66.7%)	60.2% (60.1–60.3%)	10.2% (10.0–10.3%)	96.3% (96.3–96.4%)
Contemporary cohort, $n=7,806$					
User level	0.740 (0.722–0.756)	80.7% (77.7–83.6%)	54.4% (53.2–55.5%)	14.5% (13.3–15.5%)	96.7% (96.2–97.2%)
Recording level	0.664 (0.661–0.667)	72.8% (72.2–73.3%)	51.6% (51.4–51.8%)	14.6% (14.5–14.8%)	94.3% (94.2–94.4%)
Clinic cohort, $n=181$					
Recording level	0.682 (0.605–0.755)	81.7% (69.2–93.1%)	53.4% (45.8–61.1%)	31.9% (22.9–40.7%)	91.6% (85.7–97.0%)
Newly diagnosed diabetes, recording level ($n=164$)	0.644 (0.546–0.744)	75.9% (56.3–92.9%)	53.0% (45.2–61.2%)	19.1% (11.2–28.3%)	93.8% (88.2–98.4%)

Sample sizes shown indicate numbers of individual people. User-level performance metrics are reported based on the average DNN score for all recordings from an individual user. Recording-level performance metrics are calculated treating each recording independently. Since clinic cohort participants received only one measurement, only the recording-level metric is reported for this cohort.

[a] Metrics are reported at a threshold of DNN score = 0.427; this threshold can be altered to optimize DNN performance on specific metrics as suitable for future applications.

above the cutoff differed demographically and were nearly twice as likely to have any medical condition (69.4% versus 37.3%; $P<0.001$; Supplementary Table 2).

We performed several sensitivity analyses for hypertension specifically, since it is comorbid with diabetes and may directly cause PPG-measurable vascular changes. A subset of test dataset participants provided Bluetooth-linked, home-measured blood pressures within 3 months of a PPG recording, totaling 13,007 PPG-blood pressure recording pairs (55 patients with diabetes, 527 patients without diabetes). Although the systolic (but not diastolic) value was a significant univariate predictor of diabetes, after the DNN score and other (non-hypertension) comorbidities were added into a multivariable model, systolic blood pressure was no longer a significant diabetes predictor; the DNN score, however, remained strongly independent (odds ratio: 3.53, 95% CI: 2.20–5.67; $P<0.001$). Furthermore, after excluding those with self-reported hypertension from the test dataset, DNN performance remained similar to that in the full test dataset at both user and recording levels.

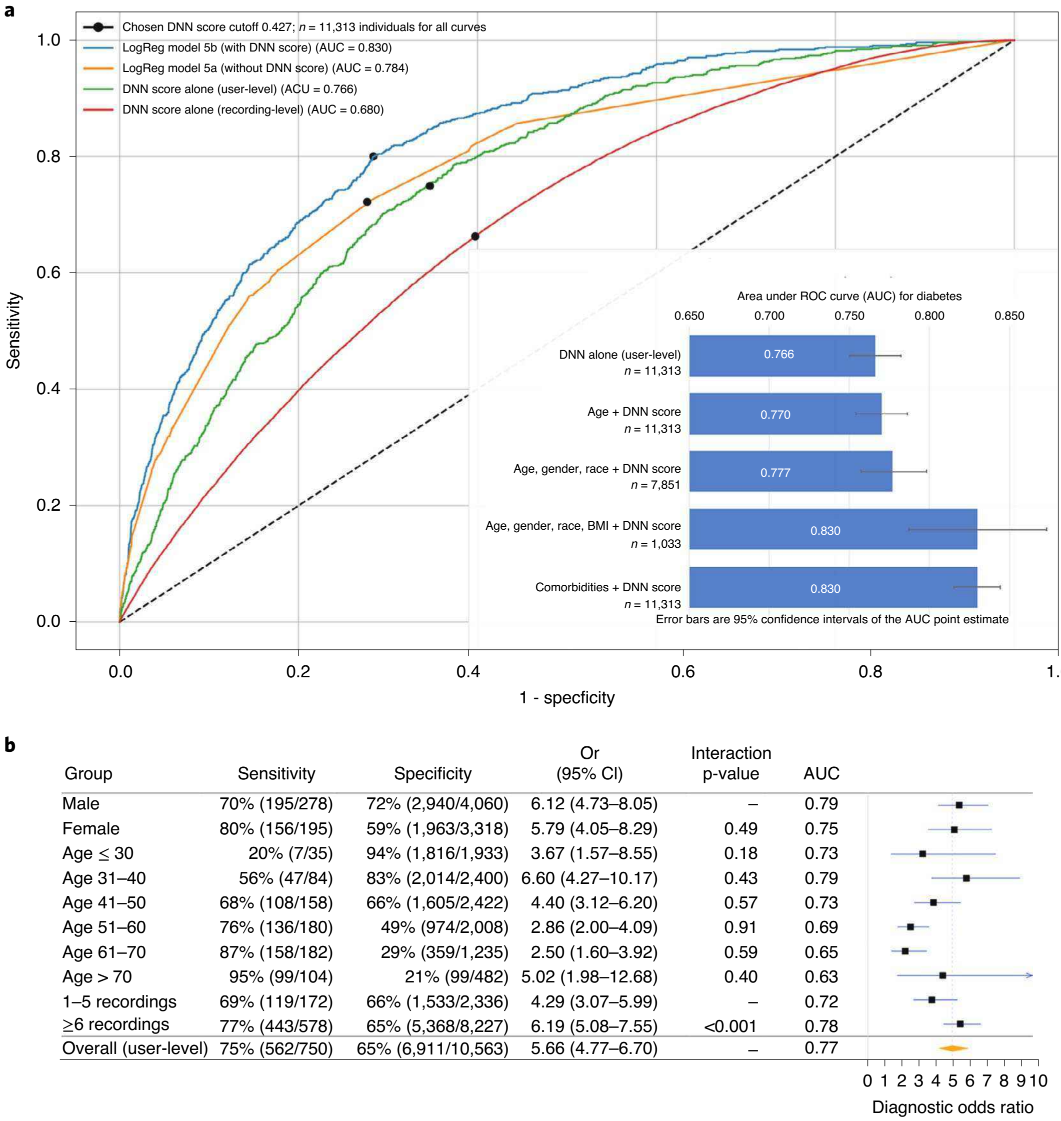
Owing to the limitations of relying on self-reported diabetes in our primary analysis, we performed additional sensitivity analyses aimed at addressing this. We identified Health eHeart participants who had laboratory-confirmed diabetes based on fasting glucose or HbA1c drawn within 180 days of diabetes self-report ($n=12,073$). In this subset, the PPV of self-reported diabetes was 81.8% (1,816/2,220) and the NPV was 88.9% (8,767/9,853). We additionally examined the performance of the DNN amongst participants who had laboratory-confirmed diabetes within 180 days of a PPG measurement in the test dataset ($n=152$ users; 9,327 measurements) and contemporary cohort ($n=94$ users; 3,659 measurements). Sampling up to five measurements per participant, the DNN's recording-level AUCs were similar when using laboratory-confirmed diabetes or self-reported diabetes in both the test dataset (0.670, 95% CI: 0.629–0.710; versus 0.650, 95% CI: 0.606–0.694) and the contemporary cohort (0.669, 95% CI: 0.618–0.719; versus 0.705, 95% CI: 0.657–0.754).

In these laboratory-confirmed diabetes subsets ($n=246$), there was also evidence for a significant linear association between the continuous DNN score and both HbA1c and fasting glucose: a 1-s.d. increase in DNN score was associated with 0.32% increase in HbA1c (beta coefficient = 2.28, 95% CI: 1.27–3.29; $P\leq0.001$) and 0.11 mmol l^{-1} increase in fasting glucose (beta coefficient = 0.82, 95% CI: 0.30–1.34; $P\leq0.001$). Similarly, among clinic cohort patients with an HbA1c measured within seven days of the visit ($n=93$), there was a positive, borderline association between the DNN score and HbA1c values (beta coefficient = 1.58, 95% CI: −0.021 to 3.187; $P=0.053$). Since long-standing poor glycemic control can adversely affect the vasculature and therefore PPG, we also performed a sensitivity analysis comparing DNN performance between HbA1c strata. Among test dataset participants with an HbA1c 7.0–8.0% within 6 months of a PPG measurement, we observed similar recording-level AUC = 0.636 (95% CI: 0.587–0.686) to that in those with an HbA1c > 8.0%, AUC = 0.632 (95% CI: 0.585–0.679), suggesting similar DNN performance regardless of glycemic control. We also examined the diagnostic odds ratio for a positive DNN prediction across different test dataset strata of gender, age, time of day, recording length and HR (Fig. 2b and Extended Data Fig. 5). DNN performance was the highest in those with >6 recordings and HR < 100 beats per minute (b.p.m.).

Finally, we performed several analyses to help illuminate the mechanisms by which PPG may capture diabetes-related information. We plotted activation maps from inner DNN layers that illustrate how it encodes input PPG recordings, and its behavior in the presence of artifacts (Extended Data Figs. 6 and 7). To investigate the role of PPG morphology to predict diabetes in isolation, we trained a separate DNN using a single-cardiac-cycle PPG waveform as the sole input; user-level AUC = 0.691 (95% CI: 0.680–0.700) and recording-level AUC = 0.605 (95% CI: 0.600–0.610). To investigate the role of HR and its derivatives in isolation, we trained a separate DNN using only peak-to-peak PPG intervals as the sole input (which removes all PPG morphology information); user-level AUC = 0.721 (95% CI: 0.703–0.740) and recording-level AUC = 0.645 (95% CI: 0.642–0.647).

Discussion

In this large-scale study and validation across three distinct cohorts, we show that smartphone-measured PPG, analyzed with deep learning, can serve as an independent, non-invasive digital biomarker of prevalent diabetes. Importantly, the ability of this PPG biomarker to predict diabetes was independent of standard risk factors and comorbidities, and discrimination further improved when adding easily obtainable covariates such as age, gender, race/ethnicity and BMI. Our validation of this digital biomarker in three cohorts



Group	Sensitivity	Specificity	Or (95% CI)	Interaction p-value	AUC
Male	70% (195/278)	72% (2,940/4,060)	6.12 (4.73–8.05)	–	0.79
Female	80% (156/195)	59% (1,963/3,318)	5.79 (4.05–8.29)	0.49	0.75
Age ≤ 30	20% (7/35)	94% (1,816/1,933)	3.67 (1.57–8.55)	0.18	0.73
Age 31–40	56% (47/84)	83% (2,014/2,400)	6.60 (4.27–10.17)	0.43	0.79
Age 41–50	68% (108/158)	66% (1,605/2,422)	4.40 (3.12–6.20)	0.57	0.73
Age 51–60	76% (136/180)	49% (974/2,008)	2.86 (2.00–4.09)	0.91	0.69
Age 61–70	87% (158/182)	29% (359/1,235)	2.50 (1.60–3.92)	0.59	0.65
Age > 70	95% (99/104)	21% (99/482)	5.02 (1.98–12.68)	0.40	0.63
1–5 recordings	69% (119/172)	66% (1,533/2,336)	4.29 (3.07–5.99)	–	0.72
≥6 recordings	77% (443/578)	65% (5,368/8,227)	6.19 (5.08–7.55)	<0.001	0.78
Overall (user-level)	75% (562/750)	65% (6,911/10,563)	5.66 (4.77–6.70)	–	0.77

Fig. 2 | Comparison of model performance to detect diabetes in the test dataset. a, Receiver operating characteristic curves for detection of diabetes, as assessed for the DNN score alone or for the output of LogReg model 5, which includes comorbidities, with and without the DNN score. DNN performance is calculated at either the recording level, which treats each recording independently, or at the user level, which is averaged across all recordings of an individual user. The DNN score cutoff used (0.427) is indicated by a black dot on each curve. Inset: bar chart showing the AUC point estimate values for diabetes in the test dataset by the indicated models; 95% CIs are shown as error bars. **b**, DNN sensitivity, specificity, diagnostic odds ratio and AUC to detect prevalent diabetes in the test dataset, as reported across ranges of age, gender and number of recordings. The test dataset sample size is 11,313 individuals. Counts are provided in parentheses for all subgroup metrics. The diagnostic odds ratio was quantified as the ratio of the positive likelihood ratio (sensitivity/(1 – specificity)) to the negative likelihood ratio ((1 – sensitivity)/specificity), with the associated 95% CI. The diagnostic odds ratio is presented at the user level for strata of age, gender and number of recordings. The interaction *P* values were calculated using two-sided Wald tests between the DNN score and the respective covariates for diabetes.

demonstrated that the DNN generalizes to prospectively enrolled and real-world clinical populations. This digital biomarker of diabetes could serve as a readily attainable complement to other established tools, providing novel information about vascular and autonomic sequelae of diabetes for clinical applications ranging from screening to therapeutic monitoring. However, additional research will be needed to determine its utility in these scenarios.

Our work effectively helps to expand the clinical utility of the PPG modality, since physicians do not currently interpret PPG in the context of diabetes. Prior work has reported associations between individually derived PPG features and diabetes-related physiologic changes, mostly using clinic-based pulse oximeters. The physiologic changes most commonly invoked include HRV[24], endothelial dysfunction[14], arterial stiffness[15] and combinations thereof[24,25], providing important early indications that aspects of the PPG waveform contain diabetes-related information. Our study extends these findings, demonstrating that it is not necessary to derive (and be limited to) particular predefined PPG features; rather, the complete PPG recording—containing all of the physiologic information—can be analyzed using a DNN to detect diabetes with strong predictive performance. This PPG-derived DNN biomarker is independent of comorbidities and can be augmented with clinical data, when available, to further improve performance. One of the real-world challenges of using remote-sensor data to identify disease biomarkers in ambulatory patients is the multiple potential sources of environmental noise, user error and demographic heterogeneity. Our study makes this crucial

Table 2 | Performance of LogReg models for prediction of prevalent diabetes with and without the DNN score in the test dataset

Predictor	Multivariable-adjusted OR without DNN score (95% CI)	*P* value[a]	Multivariable-adjusted OR with DNN score (95% CI)	*P* value[a]
	LogReg model 1a: age: AUC = 0.691 (95% CI: 0.672–0.710) (*n* = 11,313)		**LogReg model 1b: 1a + DNN score: AUC = 0.770 (95% CI: 0.754–0.786)** (*n* = 11,313)	
Age, years	1.04 (1.04–1.05)	**<0.001**	1.01 (1.01–1.02)	**<0.001**
DNN score, per s.d.	–	–	2.69 (2.41–2.99)	**<0.001**
	LogReg model 2a: age, gender and race: AUC = 0.698 (95% CI: 0.674–0.722) (*n* = 7,851)		**LogReg model 2b: 2a + DNN score: AUC = 0.777 (0.757–0.798)** (*n* = 7,851)	
Age, years	1.04 (1.04–1.05)	**<0.001**	1.01 (1.00–1.02)	**0.013**
Gender				
Males	Ref.	–	Ref.	–
Females	0.99 (0.82–1.21)	0.996	0.65 (0.53–0.79)	**<0.001**
Race/ethnicity		**0.003**		0.17
Non-Hispanic white, *n* (%)	Ref.	–	Ref.	–
Black or African American, *n* (%)	1.87 (1.11–3.15)	**0.001**	1.40 (0.82–2.38)	0.213
Hispanic, Latino or Spanish origin/ancestry, *n* (%)	0.73 (0.49–1.07)	0.106	0.69 (0.46–1.01)	0.058
Asian, *n* (%)	1.86 (1.30–2.67)	**0.001**	1.46 (1.01–2.12)	**0.047**
Multi-ethnic, *n* (%)	1.27 (0.78–2.07)	0.344	1.29 (0.78–2.13)	0.314
Other, *n* (%)	0.97 (0.49–1.95)	0.941	0.86 (0.42–1.74)	0.674
DNN score, per s.d.	–	–	2.88 (2.51–3.31)	<0.001
	LogReg model 3a: age, gender, race and BMI: AUC = 0.801 (95% CI: 0.752–0.850) (*n* = 1,033)		**LogReg model 3b: 3a + DNN score: AUC = 0.830 (95% CI: 0.787–0.873)** (*n* = 1,033)	
Age, years	1.04 (1.02–1.06)	**<0.001**	1.01 (0.99–1.04)	0.189
Gender				
Males	Ref.	–	Ref.	–
Females	0.67 (0.39–1.13)	0.130	0.51 (0.30–0.88)	**0.015**
Race/ethnicity		0.232		0.415
Non-Hispanic white	Ref.	–	Ref.	–
Black or African American	0.33 (0.04–2.63)	0.294	0.32 (0.04–2.63)	0.291
Hispanic, Latino or Spanish origin/ancestry	1.22 (0.45–3.35)	0.696	1.08 (0.38–3.05)	0.884
Asian or Pacific Islander	2.82 (0.97–8.22)	0.058	2.36 (0.77–7.24)	0.135
Multi-ethnic	0.42 (0.12–1.45)	0.168	0.46 (0.13–1.59)	0.218
Other/prefer not to disclose	0 (0)	0.999	0 (0)	0.999
BMI	1.15 (1.11–1.19)	**<0.001**	1.08 (1.04–1.12)	**<0.001**
DNN score, per s.d.	–	–	2.12 (1.53–2.94)	**<0.001**
	LogReg model 4a: HR and HRV: AUC = 0.586 (95% CI: 0.565–0.606) (*n* = 11,313)		**LogReg model 4b: 4a + DNN score: AUC = 0.765 (0.748–0.782)** (*n* = 11,313)	
HR, b.p.m.	1.02 (1.01–1.02)	**<0.001**	1.01 (1.00–1.01)	**0.024**
HRV (RMSSD), per 10 ms	0.97 (0.94–0.99)	**0.027**	1.02 (1.00–1.05)	0.068
DNN score, per s.d.	–	–	2.92 (2.65–3.21)	<0.001
	LogReg model 5a: comorbidities: AUC = 0.784 (0.766–0.802) (*n* = 11,313)		**LogReg model 5b: 5a + DNN score: AUC = 0.830 (0.815–0.844)** (*n* = 11,313)	
Hypertension, *n* (%)	3.49 (2.93–4.16)	**<0.001**	2.57 (2.15–3.07)	**<0.001**
Hypercholesterolemia, *n* (%)	2.44 (2.05–2.89)	**<0.001**	1.97 (1.66–2.34)	**<0.001**
Coronary artery disease, *n* (%)	1.35 (1.04–1.76)	**0.024**	1.22 (0.94–1.59)	0.144
Prior MI, *n* (%)	1.04 (0.74–1.48)	0.815	1.06 (0.74–1.50)	0.765
CHF, *n* (%)	2.39 (1.67–3.42)	**<0.001**	2.09 (1.46–2.98)	**<0.001**
PVD, *n* (%)	1.49 (1.00–2.21)	0.051	1.43 (0.97–2.11)	0.075
Prior stroke, *n* (%)	1.91 (1.39–2.61)	**<0.001**	1.74 (1.27–2.38)	**0.001**
Sleep apnea, *n* (%)	2.06 (1.72–2.46)	**<0.001**	1.85 (1.54–2.22)	**<0.001**
DNN score, per s.d.	–	–	2.22 (2.00–2.46)	**<0.001**

All models are shown without ('a') and with ('b') inclusion of the DNN score as a predictor. Models 1–3 are nested models, containing incrementally more demographic predictors and BMI. Model 4 adjusts for HR and HRV. Model 5 adjusts for common cardiovascular comorbidities. Sample sizes shown indicate numbers of individual people. OR, diagnostic odds ratio; RMSSD, root mean square of successive peak-to-peak interval differences; MI, myocardial infarction; CHF, congestive heart failure; PVD, peripheral vascular disease. Independent variables were standardized using the *Z* score. Ref. denotes the reference category used for categorical predictors. Bold indicates $P < 0.05$. [a]The *P* value was calculated using the Wald test for the multivariable-adjusted odds ratio (two-sided).

translational step by using remotely measured PPG signals from commercially available smartphones in a free-living population.

There are various potential applications for a PPG-based digital biomarker of diabetes. Diabetes has numerous characteristics that make it an ideal candidate for screening, such as a prolonged asymptomatic period and the availability of disease-modifying therapy. However, since population-wide screening is not currently recommended, a widely accessible smart-device-based tool could be used to identify and encourage individuals at higher risk of having prevalent diabetes to seek medical care and obtain a low-cost confirmatory diagnostic test such as HbA1c[26–28]. Leveraging smart devices to perform diabetes risk prediction without requiring clinic visits would substantially lower barriers to access given the widespread ownership of smartphones, facilitating measurement amongst many of the 224 million people living globally with undiagnosed diabetes[1]. The discriminative performance of our PPG biomarker is comparable to that of other commonly used tests such as mammography for breast cancer (AUC range 0.67–0.74)[29] or cervical cytology for cervical cancer (AUC range 0.81–0.86)[30]. It compares favorably to existing diabetes-specific risk scores that have AUCs between 0.74 and 0.85, some of which require serum glucose measurement and none of which is in common clinical use[23]. Reported AUCs of serum-based diagnostic tests such as HbA1c or fasting plasma glucose depend on the gold-standard comparator used, but for prevalent microvascular complications range from 0.82 to 0.96 (ref. [31]). Comparatively, the ease and non-invasiveness of PPG make it widely scalable, and its painlessness makes it attractive for repeated testing. Furthermore, since the PPG biomarker is predictive independently of the demographic and comorbidity components comprising most risk scores, it could also be used to supplement existing scores by capturing complementary vascular and autonomic information.

Of the various mechanisms by which PPG may detect diabetes, PPG likely captures the majority of the HR and HRV information as relates to diabetes[10–13,32]. Both predictors were attenuated in the presence of the DNN score, and peak-to-peak PPG interbeat intervals had only modestly lower AUC (0.721) than the full PPG record (0.766). While interbeat intervals likely contain the predominant predictive information for diabetes, waveform morphologies likely additionally capture information on diabetic vascular changes ranging from endothelial dysfunction[14] to arterial stiffening[15].

Our study has several limitations. Participants elected to download the iOS smartphone app and therefore may have higher socioeconomic status, technological competence or health literacy relative to the general population. Our reliance on self-reported diabetes is another limitation. However, our results generalized to the unselected clinic cohort, which had medical-record-confirmed diabetes, and sensitivity analysis suggested high PPV/NPV against laboratory-confirmed diabetes. Also, misclassification due to self-report at the algorithm training stage would be expected to bias DNN performance toward the null during validation. In analyses that used laboratory or blood pressure measurements, the time windows we used were large and mainly informative as sensitivity analyses. Future studies are needed to confirm this, and whether PPG signals from other sources, such as smartwatches, or obtained from anatomic locations such as the toe or ear would perform similarly. Given the lower overall prevalence of diabetes, the PPV of our PPG biomarker ranged from 10% to 32%, which is similar to existing diabetes risk scores whose PPVs mostly range between 10% and 25% depending on the population and threshold used[23,33]. While false positives are a concern, confirmatory HbA1c testing is relatively cost-effective; and since individuals with positive DNN predictions were also more likely to have cardiometabolic conditions, they would likely benefit from medical contact. Depending on the intended use of the biomarker, the DNN score threshold can also be altered to maximize sensitivity or specificity for the intended application. The cross-sectional nature of our study design limited direct investigation of PPG as a diabetes screening tool, or prediction of incident diabetes. We also did not have sufficient data in the primary cohort on the type, severity or medication use for diabetes. The DNN score did perform similarly, however, in clinic cohort subsets with newly diagnosed diabetes and between HbA1c strata. Finally, we were limited in our attempts to compare our approach against standard diabetes risk scores owing to the lack of necessary variables in our cohort. These data availability limitations, however, serve to illuminate the difficulty providers also encounter when deploying existing questionnaire-based prediction models, underscoring a strength of non-invasive, objective PPG-based diabetes detection.

In summary, we demonstrate that PPG recorded using consumer-owned smartphones can provide a readily attainable digital biomarker of prevalent diabetes that is independent of standard risk factors and comorbidities. Remote capture of diabetes-predictive PPG information from ambulatory users is feasible and provides an easily scalable, non-invasive complement to diabetes risk prediction. The linear association of the DNN score with HbA1c suggests that PPG may additionally capture information about diabetes severity and control, but this requires further investigation. Although this study leverages a large dataset, additional research is needed ideally in targeted intended-use populations to determine how to best incorporate this digital biomarker into existing practice recommendations for diabetes screening and care—particularly in light of the potential for its wide deployment using existing smart devices outside the purview of traditional medical care.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-020-1010-5.

Received: 11 July 2019; Accepted: 6 July 2020;
Published online: 17 August 2020

References

1. Cho, N. H. et al. IDF Diabetes Atlas: global estimates of diabetes prevalence for 2017 and projections for 2045. *Diabetes Res. Clin. Pract.* **138**, 271–281 (2018).
2. Harris, M. I., Klein, R., Welborn, T. A. & Knuiman, M. W. Onset of NIDDM occurs at least 4–7 yr before clinical diagnosis. *Diabetes Care* **15**, 815–819 (1992).
3. Bertoni, A. G., Krop, J. S., Anderson, G. F. & Brancati, F. L. Diabetes-related morbidity and mortality in a national sample of U.S. elders. *Diabetes Care* **25**, 471–475 (2002).
4. Allen, J. Photoplethysmography and its application in clinical physiological measurement. *Physiol. Meas.* **28**, R1–R39 (2007).
5. Elgendi, M. et al. The use of photoplethysmography for assessing hypertension. *npj Digit. Med.* **2**, 60 (2019).
6. Alty, S. R., Angarita-Jaimes, N., Millasseau, S. C. & Chowienczyk, P. J. Predicting arterial stiffness from the digital volume pulse waveform. *IEEE Trans. Biomed. Eng.* **54**, 2268–2275 (2007).
7. Otsuka, T., Kawada, T., Katsumata, M. & Ibuki, C. Utility of second derivative of the finger photoplethysmogram for the estimation of the risk of coronary heart disease in the general population. *Circ. J.* **70**, 304–310 (2006).
8. *Smartphone Ownership Is Growing Rapidly around the World, but Not Always Equally* (Pew Research Center, 2019).
9. Coravos, A., Khozin, S. & Mandl, K. D. Developing and adopting safe and effective digital biomarkers to improve patient outcomes. *npj Digit. Med.* **2**, 14 (2019).
10. Singh, J. P. et al. Association of hyperglycemia with reduced heart rate variability (The Framingham Heart Study). *Am. J. Cardiol.* **86**, 309–312 (2000).
11. Avram, R. et al. Real-world heart rate norms in the Health eHeart study. *npj Dig. Med.* **2**, 58 (2019).
12. Carnethon, M. R., Golden, S. H., Folsom, A. R., Haskell, W. & Liao, D. Prospective investigation of autonomic nervous system function and the development of type 2 diabetes. *Circulation* **107**, 2190–2195 (2003).

13. Guo, Y. et al. Genome-wide assessment for resting heart rate and shared genetics with cardiometabolic traits and type 2 diabetes. *J. Am. Coll. Cardiol.* **74**, 2162–2174 (2019).
14. Lilia, C.-M. et al. Endothelial dysfunction evaluated using photoplethysmography in patients with type 2 diabetes. *J. Cardiovasc. Dis. Diagn.* **3**, 219 (2015).
15. Pilt, K., Meigas, K., Ferenets, R., Temitski, K. & Viigimaa, M. Photoplethysmographic signal waveform index for detection of increased arterial stiffness. *Physiol. Meas.* **35**, 2027–2036 (2014).
16. Schönauer, M. et al. Cardiac autonomic diabetic neuropathy. *Diabetes Vasc. Dis. Res.* **5**, 336–344 (2008).
17. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
18. Gulshan, V. et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *J. Am. Med. Assoc.* **316**, 2402–2410 (2016).
19. Zhang, H. et al. Comparison of physician visual assessment with quantitative coronary angiography in assessment of stenosis severity in China. *JAMA Intern. Med.* **178**, 239–247 (2018).
20. Hannun, A. Y. et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat. Med.* **25**, 1–11 (2019).
21. Dixit, S. et al. Secondhand smoke and atrial fibrillation: data from the Health eHeart Study. *Heart Rhythm* **13**, 3–9 (2016).
22. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **27**, 861–874 (2006).
23. Noble, D., Mathur, R., Dent, T., Meads, C. & Greenhalgh, T. Risk models and scores for type 2 diabetes: systematic review. *Br. Med. J.* **343**, d7163–d7163 (2011).
24. Moreno, E. M. et al. Type 2 diabetes screening test by means of a pulse oximeter. *IEEE Trans. Biomed. Eng.* **64**, 341–351 (2017).
25. Nirala, N., Periyasamy, R., Singh, B. K. & Kumar, A. Detection of type-2 diabetes using characteristics of toe photoplethysmogram by applying support vector machine. *Biocybern. Biomed. Eng.* **39**, 38–51 (2019).
26. Selvin, E., Steffes, M. W., Gregg, E., Brancati, F. L. & Coresh, J. Performance of A1C for the classification and prediction of diabetes. *Diabetes Care* **34**, 84–89 (2010).
27. Camacho, J. E., Shah, V. O., Schrader, R., Wong, C. S. & Burge, M. R. Performance of A1C versus OGTT for the diagnosis of prediabetes in a community-based screening. *Endocr. Pract.* **22**, 1288–1295 (2016).
28. Karakaya, J., Akin, S., Karagaoglu, E. & Gurlek, A. The performance of hemoglobin A1c against fasting plasma glucose and oral glucose tolerance test in detecting prediabetes and diabetes. *J. Res. Med. Sci.* **19**, 1051–1057 (2014).
29. Pisano, E. D. et al. Diagnostic performance of digital versus film mammography for breast-cancer screening. *N. Engl. J. Med.* **353**, 1773–1783 (2005).
30. Mathews, W. C., Agmas, W. & Cachay, E. Comparative accuracy of anal and cervical cytology in screening for moderate to severe dysplasia by magnification guided punch biopsy: a meta-analysis. *PLoS ONE* **6**, e24946 (2011).
31. *Use of Glycated Haemoglobin (HbA1c) in the Diagnosis of Diabetes Mellitus: Abbreviated Report of a WHO Consultation* (World Health Organization, 2011).
32. Kim, D.-I. et al. The association between resting heart rate and type 2 diabetes and hypertension in Korean adults. *Heart* **102**, 1757–1762 (2016).
33. Lindström, J. & Tuomilehto, J. The diabetes risk score: a practical tool to predict type 2 diabetes risk. *Diabetes Care* **26**, 725–731 (2003).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Data sources and study population. *The primary cohort.* The primary cohort was derived from 55,433 Health eHeart participants aged ≥18 years, who self-reported a diabetes diagnosis by a healthcare provider and made at least one PPG recording between 1 April 2014 and 30 April 2018 (Fig. 1a). Health eHeart is a worldwide, Internet-based, longitudinal electronic cohort of English-speaking adults[21]. PPG waveforms were obtained by placing an index fingertip on the smartphone camera (Fig. 1b and Extended Data Fig. 8). To assess self-reported diabetes status, participants were asked, "Have you ever been told by a doctor, nurse or other healthcare provider that you have diabetes?" and provided the answer options of "Yes", "No" or "Don't know/prefer not to state". Participants who answered "Don't know/prefer not to state" were excluded from our analysis. Participants completed additional surveys regarding demographics, anthropometrics and medical history to varying degrees. We have demonstrated previously[21] that self-reported past medical history in Health eHeart is strongly correlated with the medical record.

The primary cohort was randomly split into training (70%, $n=37{,}709$), development (10%, $n=4{,}848$) and test (20%, $n=11{,}313$) datasets (Extended Data Fig. 2 and Life Sciences Reporting Summary). The training dataset was used for DNN development and training, and DNN hyperparameters were tuned in the development dataset. Final model performance is reported in the test dataset, which was kept completely separate until the final evaluation step.

Two additional validation cohorts. In addition to validating the performance of the DNN algorithm in the primary cohort test dataset, we additionally reported DNN performance in two validation cohorts (Fig. 1a), providing three examples of validation in datasets separate from training data[23]. The first was the 'contemporary cohort', which was composed of PPG recordings from 7,806 participants newly enrolled into Health eHeart between 1 May 2018 and 31 December 2018 (Extended Data Fig. 3 and Supplementary Table 3). This temporally distinct validation cohort helps to account for secular changes, such as changes in smartphone models and cameras, that could affect PPG recording. Then, to test the validity of our approach in a real-world clinical setting, we prospectively enrolled an in-person 'clinic cohort' composed of 181 consecutive patients referred to 3 cardiovascular prevention clinics (2 in San Francisco, 1 in Montreal) between 1 November 2018 and 30 July 2019 (Extended Data Fig. 3 and Supplementary Table 4). Clinic cohort participants were consented, and assessed for height, weight and BMI, and a trained coordinator obtained at least 15 s of a single PPG recording using an iPhone and determined diabetes status by medical chart review. For the subset of clinic cohort patients who also had fasting glucose and HbA1c obtained within seven days of the in-clinic visit, we used the American Diabetes Association diagnosis criteria to classify participants as having/not having diabetes[34].

The University of California, San Francisco Institutional Review Board approved the study and all participants gave informed consent.

PPG waveform acquisition and preprocessing. PPG waveforms were obtained by placing the index fingertip[35] on the smartphone camera using the Azumio Instant Heart Rate iOS smartphone application. Although the app is available for Android and iOS operating systems, data were limited to iOS app versions in this study owing to data availability in Health eHeart. Changes in reflected light intensity recorded by the smartphone camera are interpreted as pulsatile blood volume change. The waveforms were pre-processed by the Azumio algorithm for camera artifact removal, utilizing standard detrending and low-pass-filter techniques (Fig. 1). A low-pass ~0.4-Hz, second-order, zero-phase-shift infinite impulse response (IIR) filter is used to find the trend; the trend is subtracted to get the detrended signal. Another low-pass ~10-Hz, second-order, zero-phase-shift IIR filter is used to remove high-frequency noise. Individual beats corresponding to cardiac cycles were identified using the rising edge of the PPG signal. If the recording did not have at least 5 s of continuous discernible peak-to-peak intervals, it was removed. Waveforms with a length under 5 s or with an amplitude of 0, indicating a null signal, were also removed. We excluded outlier PPG measurements defined as HR values of outside the biologically plausible range of 20–220 b.p.m. We limited waveforms in our dataset to those collected at either 100 Hz or 120 Hz, and upsampled recordings of 100 Hz to 120 Hz using the standard polyphase method[36] to minimize variance due to sampling frequency. We derived the onset of each cardiac cycle by identifying the rising edge of the waveform, used to determine HR and HRV (using the RMSSDs).

DNN development and performance. We built a 39-layer convolutional DNN to detect prevalent diabetes (Extended Data Fig. 9). The DNN takes the PPG waveform as the sole input, which consists of 2,560 samples equivalent to ~21.3 s (approximately the mean signal duration), and outputs a DNN score between 0 and 1 per signal; higher scores suggest greater likelihood of diabetes. Shorter signals were zero-padded up to the fixed length and longer examples were cropped. All PPG waveforms were standardized using the mean and s.d. values of the entire training dataset. The network architecture had 39 layers organized in a block structure, consisting of convolutional layers with an initial filter size of 15 and filter number of 16. The size of the filters decreased, and the number of filters increased, as network depth increased. After each convolutional layer, we applied batch normalization[37], rectified linear activation[38] and dropout[39] with a probability of 0.2. The final flattened and fully connected softmax layer produced a distribution across the classes of diabetes/no diabetes[40]. Weights were initialized randomly as described by He et al.[41].

We used grid-search to tune the network hyperparameters by searching over the best optimizer, the best initializer, the number of convolutional layers, the stride size, the filter length, the number of filters, the class weight, the learning rate, the input length of the signal, the batch size, the dropout, the early stopping criteria and the amount of cropping of the start/end of the signal, based on the recording-level development dataset performance. The best performance was achieved by cropping two beats from the beginning and one beat from the end of the signal; this was applied to all PPG records. For all of the models presented, we used the Rectified ADAM optimizer with the default parameters[42], and a mini-batch size of 512. The learning rate was initialized at 1×10^{-3} and was adjusted on the basis of the effects of variance and momentum during training[42]. We halted training after an absence of improvement in the loss within the development set for eight consecutive epochs. A class weight of 10:1 for diabetes to non-diabetes recordings was applied to our loss function. The best performing model was chosen on the basis of the development dataset recording-level AUC performance and was then applied to all validation sets. We explored different architectures involving recurrent layers, such as long-short-term memory cells and residual blocks (ResNet), and with age or hour of the day added as additional inputs to the DNN, but found no improvement in AUC despite substantial increases in model complexity and runtime. The DNN was trained for 18 epochs.

Grid-search of hyperparameters. We performed a systematic search of hyperparameters among these values:

- Model architecture: convolutional neural network, ResNet, LSTM
- Number of convolutional layers: 7, 15, 19, 25, 29, 35, 39
- Filter length: 5, 7, 9, 11, 13, 15
- Number of filters to start: 8, 16, 32, 64
- Optimizer: Adam, Rectified Adam
- Class weight for 'diabetes': 5, 10, 15, 20
- Initializer: Glorot, He
- Learning rate: 10×10^{-1}, 10×10^{-2}, 10×10^{-3}, 10×10^{-4}, 10×10^{-5}
- Input shape [2,560, 1]; [2,048, 1]
- Batch size: 64, 128, 256, 512
- Dropout: 0.2, 0.4, 0.6
- Early stopping criteria: 6, 8, 12, 20
- (Preprocessing) Number of beats cropped at the start of the signal: 0, 1, 2, 3
- (Preprocessing) Number of beats cropped at the end of the signal: 0, 1, 2, 3

We reported DNN performance using the AUC[22] in three separate test datasets: the primary cohort test dataset; the contemporary cohort; and the in-person clinic cohort. Since many participants contributed >1 recording, we assessed model performance both at the 'recording level', which treats each recording independently, and at the 'user level', which averages the DNN score for all recordings provided by an individual user. Our primary aim was to evaluate the user-level DNN score, since the clinical goal would be to classify a patient as having diabetes or not. Clinic cohort patients have only recording-level performance since only a single recording was obtained per patient during their visit. We also plotted the activation maps of several hidden convolutional layers of the trained DNN[43] from an example PPG record to help illuminate some of the higher-level PPG features derived by the DNN (Extended Data Figs. 8 and 9).

Sensitivity analyses. To better ascertain the reliability of self-reported diabetes in the primary cohort, we described the PPV and NPV of self-reported diabetes in the larger Health eHeart study[21] using fasting glucose or HbA1c drawn within 180 days of self-reported diabetes; if >1 laboratory value was available, the value closest in time to self-report was used. Laboratory-confirmed diabetes was defined according to the American Diabetes Association guidelines: HbA1c ≥ 7.0%, fasting glucose (fasting glucose ≥ 126 mg dl^{-1} or 7.0 mmol l^{-1})[34] or non-diabetic range of HbA1c/fasting glucose but self-report of taking diabetes medications. We also examined DNN performance among the subset of test dataset and contemporary cohort participants who had laboratory-confirmed diabetes using laboratory values drawn within 180 days of a PPG measurement. For those with multiple measurements, we randomly sampled up to five measurements. To understand the performance of the DNN according to glycemic control in the laboratory-confirmed diabetes cohort, we examined DNN performance in strata of HbA1c above and below 8.0%. Additionally, in the clinic cohort, we examined DNN performance after excluding those with a prior diagnosis of diabetes. Linear regression models were fitted with the DNN score as the predictor and either HbA1c or glucose value as the dependent variable in the test dataset and contemporary cohort. To investigate the role of HR in isolation, we trained a separate DNN to detect diabetes using only peak-to-peak intervals as input and the same architecture and training data as the primary DNN. To investigate the role of the PPG waveform in isolation, we trained a separate DNN using the PPG waveform from a single cardiac cycle, removing the time-domain contribution from consecutive cardiac cycles. In the clinic cohort, we also modeled the DNN score against HbA1c as the dependent variable with linear regression.

Statistical analysis. Basic demographics and previous medical conditions are presented for each dataset, and continuous data are presented as mean ± s.d. The 'DNN score' is the final layer of the DNN, which is an output distribution for diabetes based on the PPG input. We identified a discrimination threshold for the DNN score that maximized the macro average sensitivity between the 'diabetes' and 'no diabetes' classes in the training dataset[44]; this threshold is applied to all relevant performance metrics. We present sensitivity, specificity, PPV and NPV for each of our test datasets[22]. CIs for these metrics were derived by bootstrapping 80% of the test data over 1,000 iterations to obtain the 5th and 95th percentile values. The diagnostic odds ratio (odds ratio) is a measure of the effectiveness of a diagnostic test and is defined as the ratio of the odds of the DNN score being positive for diabetes if the subject has diabetes, relative to the odds of the DNN score being positive if the subject does not have diabetes[45]. Odds ratios, two-sided P values for interaction (between the DNN score, the covariates and diabetes, calculated by the Wald test), sensitivity and specificity were calculated separately between different strata of age, gender and PPG recording characteristics.

To understand the incremental contribution of PPG-based predictions alongside commonly available demographic and clinical predictors of diabetes, we built nested LogReg models for prevalent diabetes both with and without the inclusion of the standardized DNN score. LogReg model 1 included age as a covariate; LogReg model 2 additionally included gender and race/ethnicity; LogReg model 3 additionally included BMI. Since HR[12,13] and HRV[46] are known independent predictors of diabetes and can be derived from the PPG signal, we examined the specific role of HR in the PPG-based prediction of diabetes by including the per-record average HR and HRV (calculated using the RMSSDs) as covariates in LogReg model 4. Finally, in LogReg model 5 we included clinical comorbidities commonly known to co-occur with diabetes, to ascertain the independent value of the PPG DNN score for identifying diabetes. All continuous logistic regression variables were standardized using the Z score to allow comparison between odds ratios and we used complete-case analysis, excluding individuals with missing covariates. Tests for normality were performed and met by all continuous predictors, and there were no adjustments made for multiple comparisons.

A two-sided P value <0.05 was considered significant. The convolutional neural network was built in Python 2.7 using Keras (version 2.0.3) and TensorFlow (version 1.13.2). The LogReg models and AUC were derived in SPSS v24.0 (IBM).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data that support the findings of this study are available from the authors and Azumio, but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of Azumio.

Code availability

The code that supports this work is copyright of the Regents of the University of California and can be made available through license.

References

34. American Diabetes Association Classification and diagnosis of diabetes: standards of medical care in diabetes—2018. *Diabetes Care* **41**, S13–S27 (2018).
35. Elgendi, M. On the analysis of fingertip photoplethysmogram signals. *Curr. Cardiol. Rev.* **8**, 14–25 (2012).
36. Emami, S. New methods for computing interpolation and decimation using polyphase decomposition. *IEEE Trans. Educ.* **42**, 311–314 (1999).
37. Ioffe, S. & Szegedy, C. Batch normalization: accelerating deep network training by reducing internal covariate shift. in *Proc. 32nd International Conference on Machine Learning* 448–456 (2015).
38. Nair, V. & Hinton, G. E. Rectified linear units improve restricted Boltzmann machines. in *Proc. 27th International Conference on Machine Learning* 807–814 (2010).
39. Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
40. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* (MIT Press, 2016).
41. He, K., Zhang, X., Ren, S. & Sun, J. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. in *Proc. IEEE International Conference on Computer Vision* **2015**, 1026–1034 (2015).
42. Liu, L. et al. On the variance of the adaptive learning rate and beyond. Preprint at https://arxiv.org/abs/1908.03265 (2019).
43. Yosinski, J., Clune, J., Nguyen, A., Fuchs, T. & Lipson, H. Understanding neural networks through deep visualization. Preprint at https://arxiv.org/abs/1506.06579 (2015).
44. Ferri, C., Hernández-Orallo, J. & Modroiu, R. An experimental comparison of performance measures for classification. *Pattern Recognit. Lett.* **30**, 27–38 (2009).
45. Glas, A. S., Lijmer, J. G., Prins, M. H., Bonsel, G. J. & Bossuyt, P. M. M. The diagnostic odds ratio: a single indicator of test performance. *J. Clin. Epidemiol.* **56**, 1129–1135 (2003).
46. Benichou, T. et al. Heart rate variability in type 2 diabetes mellitus: a systematic review and meta-analysis. *PLoS ONE* **13**, e0195166 (2018).

Acknowledgements

We acknowledge A. Markowitz for editorial support made possible by CTSI grant KL2 TR001870. R.A. received support from the "Fonds de la recherche en santé du Québec" (grant 274831). G.H.T. received support from the National Institutes of Health NHLBI K23HL135274. J.E.O., M.J.P. and G.M.M. received support from the National Institutes of Health (U2CEB021881). M.J.P. is partially supported by a PCORI contract supporting the Health eHeart Alliance (PPRN-1306-04709). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

J.E.O., R.A., G.H.T. and K.A. contributed to the study design. P.K., J.E.O., R.A., K.A. and G.H.T. contributed to data collection. R.A. and G.H.T. performed data cleaning and analysis, ran experiments and created tables and figures. R.A., J.E.O., P.K., J.W.H., G.M.M., M.J.P., K.A. and G.H.T. contributed to data interpretation and writing. G.H.T., J.E.O. and K.A. supervised. G.H.T. and K.A. contributed equally as co-senior authors. All authors read and approved the submitted manuscript.

Competing interests

J.E.O. has received research funding from Samsung and iBeat. G.M.M. has received research funding from Medtronic, Jawbone and Eight. K.A. received funding from Jawbone Health Hub. P.K. is an employee of Azumio. G.H.T has received research grants from Janssen Pharmaceuticals and Myokardia and is an advisor to Cardiogram, Inc. None of the remaining authors have potential conflicts of interest. Azumio provided no financial support for this study and provided only access to the data. Data analysis, interpretation and decision to submit the manuscript were performed independently from Azumio.

Additional information

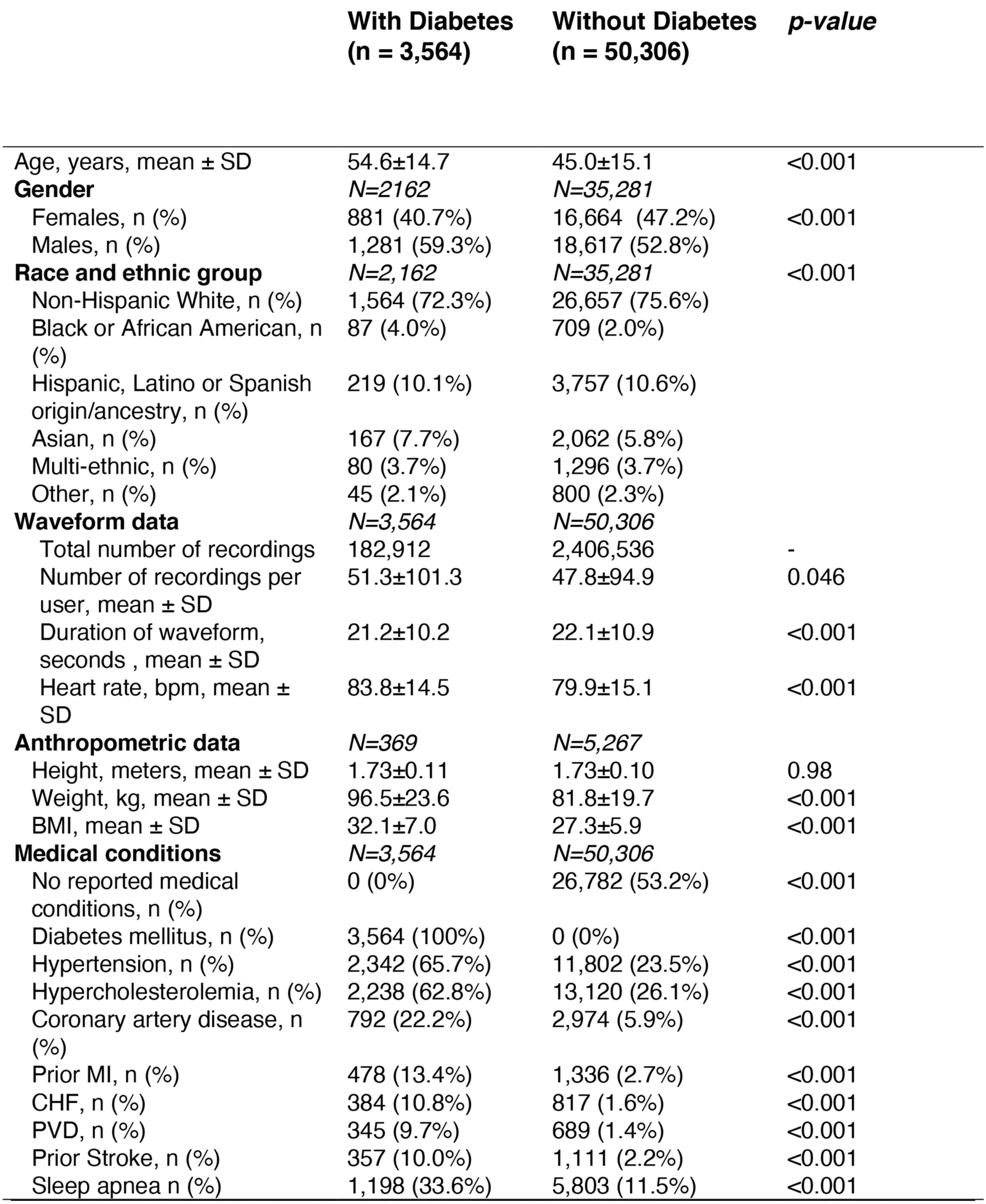
Extended data is available for this paper at https://doi.org/10.1038/s41591-020-1010-5.

Supplementary information is available for this paper at https://doi.org/10.1038/s41591-020-1010-5.

Correspondence and requests for materials should be addressed to G.H.T.

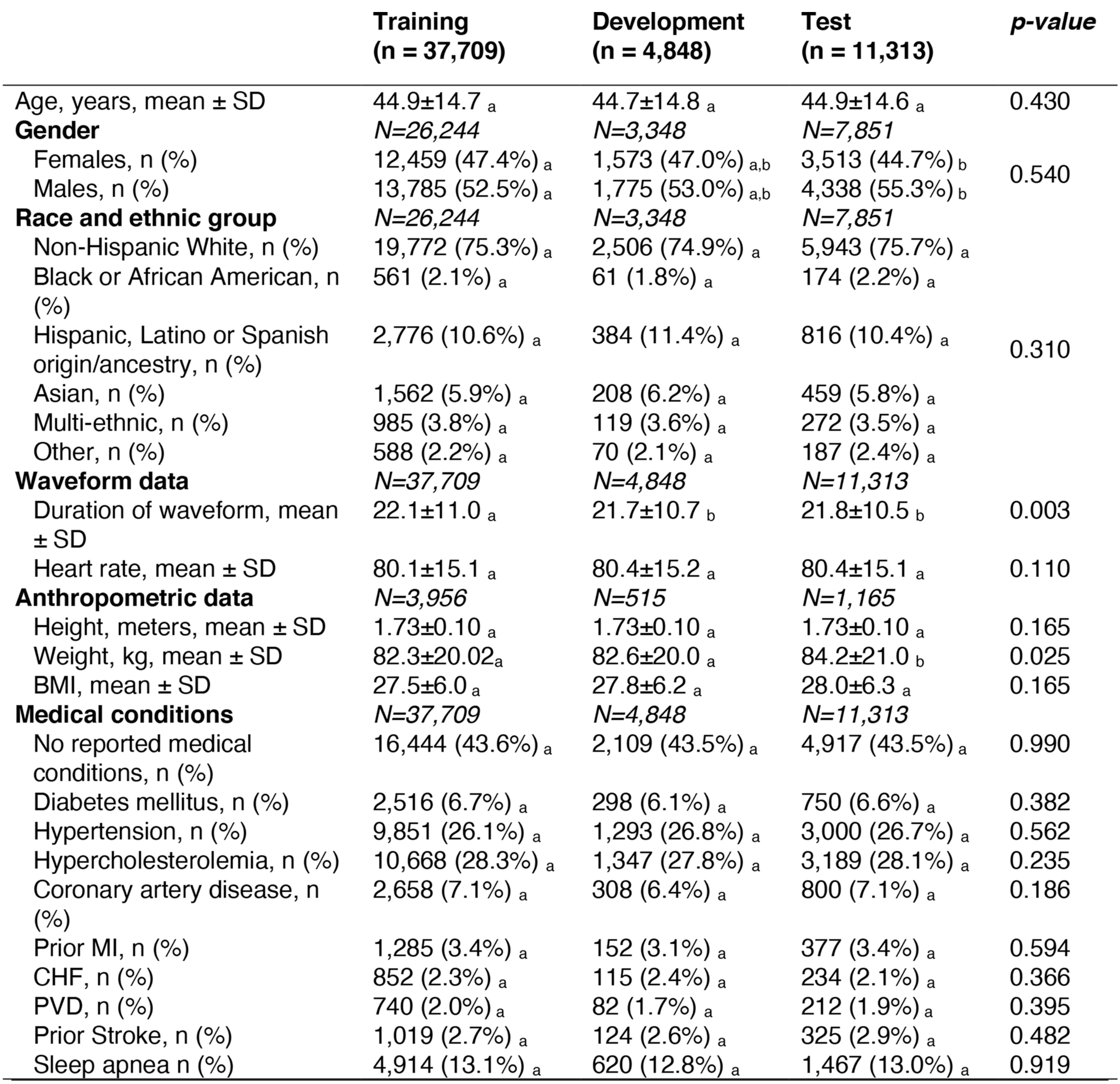
Peer review information Michael Basson was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Reprints and permissions information is available at www.nature.com/reprints.



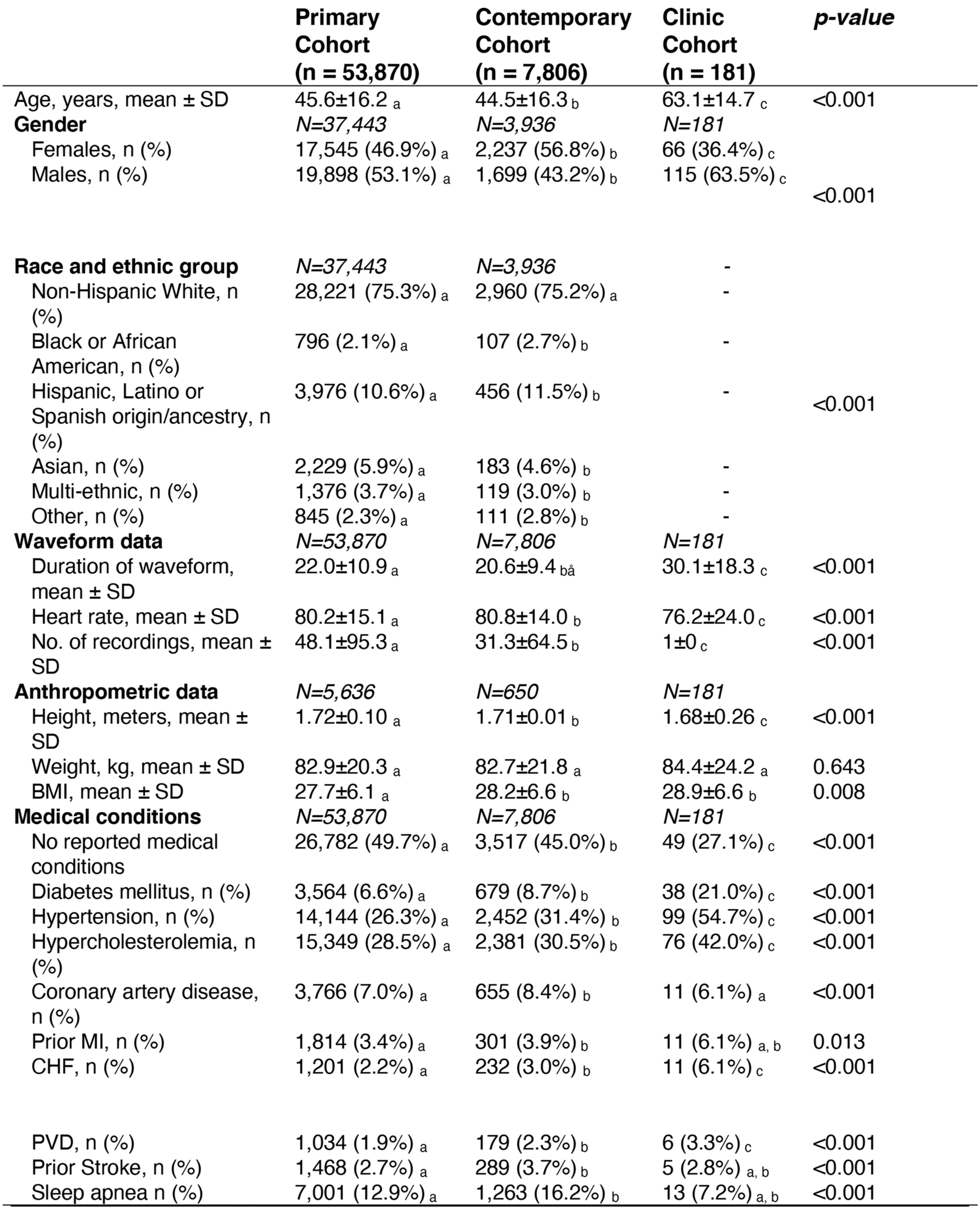
	With Diabetes (n = 3,564)	Without Diabetes (n = 50,306)	*p-value*
Age, years, mean ± SD	54.6±14.7	45.0±15.1	<0.001
Gender	*N=2162*	*N=35,281*	
Females, n (%)	881 (40.7%)	16,664 (47.2%)	<0.001
Males, n (%)	1,281 (59.3%)	18,617 (52.8%)	
Race and ethnic group	*N=2,162*	*N=35,281*	<0.001
Non-Hispanic White, n (%)	1,564 (72.3%)	26,657 (75.6%)	
Black or African American, n (%)	87 (4.0%)	709 (2.0%)	
Hispanic, Latino or Spanish origin/ancestry, n (%)	219 (10.1%)	3,757 (10.6%)	
Asian, n (%)	167 (7.7%)	2,062 (5.8%)	
Multi-ethnic, n (%)	80 (3.7%)	1,296 (3.7%)	
Other, n (%)	45 (2.1%)	800 (2.3%)	
Waveform data	*N=3,564*	*N=50,306*	
Total number of recordings	182,912	2,406,536	-
Number of recordings per user, mean ± SD	51.3±101.3	47.8±94.9	0.046
Duration of waveform, seconds , mean ± SD	21.2±10.2	22.1±10.9	<0.001
Heart rate, bpm, mean ± SD	83.8±14.5	79.9±15.1	<0.001
Anthropometric data	*N=369*	*N=5,267*	
Height, meters, mean ± SD	1.73±0.11	1.73±0.10	0.98
Weight, kg, mean ± SD	96.5±23.6	81.8±19.7	<0.001
BMI, mean ± SD	32.1±7.0	27.3±5.9	<0.001
Medical conditions	*N=3,564*	*N=50,306*	
No reported medical conditions, n (%)	0 (0%)	26,782 (53.2%)	<0.001
Diabetes mellitus, n (%)	3,564 (100%)	0 (0%)	<0.001
Hypertension, n (%)	2,342 (65.7%)	11,802 (23.5%)	<0.001
Hypercholesterolemia, n (%)	2,238 (62.8%)	13,120 (26.1%)	<0.001
Coronary artery disease, n (%)	792 (22.2%)	2,974 (5.9%)	<0.001
Prior MI, n (%)	478 (13.4%)	1,336 (2.7%)	<0.001
CHF, n (%)	384 (10.8%)	817 (1.6%)	<0.001
PVD, n (%)	345 (9.7%)	689 (1.4%)	<0.001
Prior Stroke, n (%)	357 (10.0%)	1,111 (2.2%)	<0.001
Sleep apnea n (%)	1,198 (33.6%)	5,803 (11.5%)	<0.001

Extended Data Fig. 1 | Baseline characteristics of the primary cohort by diabetes status. Primary cohort sample size was 53,870 individual people. Where data was only available for subgroups of the full cohort, subgroup sample size is denoted by N. Differences in means of continuous variables between 2 groups were compared using the two-sample t-test. Differences in proportions of categorical variables between 2 groups were compared using the Chi-Squared test. Tests of significance were 2 sided. Abbreviations: bpm: beats per minute; CAD: Coronary artery disease; CHF: Congestive heart failure; COPD: Chronic obstructive pulmonary disease; HR: Heart rate, MI: Myocardial Infarction; PVD: Peripheral Vascular Disease.



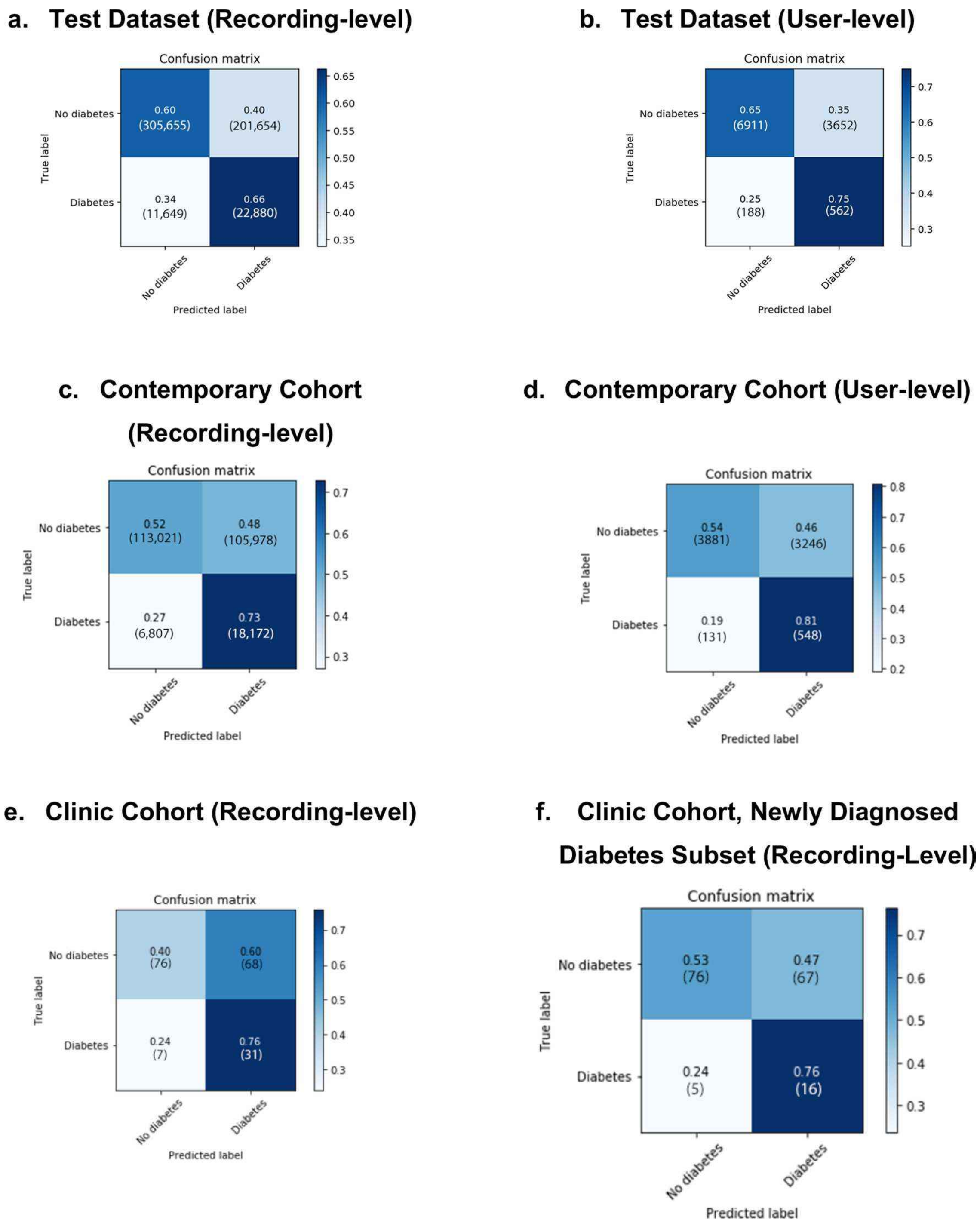
	Training (n = 37,709)	Development (n = 4,848)	Test (n = 11,313)	*p-value*
Age, years, mean ± SD	44.9±14.7 $_a$	44.7±14.8 $_a$	44.9±14.6 $_a$	0.430
Gender	*N=26,244*	*N=3,348*	*N=7,851*	
Females, n (%)	12,459 (47.4%) $_a$	1,573 (47.0%) $_{a,b}$	3,513 (44.7%) $_b$	0.540
Males, n (%)	13,785 (52.5%) $_a$	1,775 (53.0%) $_{a,b}$	4,338 (55.3%) $_b$	
Race and ethnic group	*N=26,244*	*N=3,348*	*N=7,851*	
Non-Hispanic White, n (%)	19,772 (75.3%) $_a$	2,506 (74.9%) $_a$	5,943 (75.7%) $_a$	0.310
Black or African American, n (%)	561 (2.1%) $_a$	61 (1.8%) $_a$	174 (2.2%) $_a$	
Hispanic, Latino or Spanish origin/ancestry, n (%)	2,776 (10.6%) $_a$	384 (11.4%) $_a$	816 (10.4%) $_a$	
Asian, n (%)	1,562 (5.9%) $_a$	208 (6.2%) $_a$	459 (5.8%) $_a$	
Multi-ethnic, n (%)	985 (3.8%) $_a$	119 (3.6%) $_a$	272 (3.5%) $_a$	
Other, n (%)	588 (2.2%) $_a$	70 (2.1%) $_a$	187 (2.4%) $_a$	
Waveform data	*N=37,709*	*N=4,848*	*N=11,313*	
Duration of waveform, mean ± SD	22.1±11.0 $_a$	21.7±10.7 $_b$	21.8±10.5 $_b$	0.003
Heart rate, mean ± SD	80.1±15.1 $_a$	80.4±15.2 $_a$	80.4±15.1 $_a$	0.110
Anthropometric data	*N=3,956*	*N=515*	*N=1,165*	
Height, meters, mean ± SD	1.73±0.10 $_a$	1.73±0.10 $_a$	1.73±0.10 $_a$	0.165
Weight, kg, mean ± SD	82.3±20.02$_a$	82.6±20.0 $_a$	84.2±21.0 $_b$	0.025
BMI, mean ± SD	27.5±6.0 $_a$	27.8±6.2 $_a$	28.0±6.3 $_a$	0.165
Medical conditions	*N=37,709*	*N=4,848*	*N=11,313*	
No reported medical conditions, n (%)	16,444 (43.6%) $_a$	2,109 (43.5%) $_a$	4,917 (43.5%) $_a$	0.990
Diabetes mellitus, n (%)	2,516 (6.7%) $_a$	298 (6.1%) $_a$	750 (6.6%) $_a$	0.382
Hypertension, n (%)	9,851 (26.1%) $_a$	1,293 (26.8%) $_a$	3,000 (26.7%) $_a$	0.562
Hypercholesterolemia, n (%)	10,668 (28.3%) $_a$	1,347 (27.8%) $_a$	3,189 (28.1%) $_a$	0.235
Coronary artery disease, n (%)	2,658 (7.1%) $_a$	308 (6.4%) $_a$	800 (7.1%) $_a$	0.186
Prior MI, n (%)	1,285 (3.4%) $_a$	152 (3.1%) $_a$	377 (3.4%) $_a$	0.594
CHF, n (%)	852 (2.3%) $_a$	115 (2.4%) $_a$	234 (2.1%) $_a$	0.366
PVD, n (%)	740 (2.0%) $_a$	82 (1.7%) $_a$	212 (1.9%) $_a$	0.395
Prior Stroke, n (%)	1,019 (2.7%) $_a$	124 (2.6%) $_a$	325 (2.9%) $_a$	0.482
Sleep apnea n (%)	4,914 (13.1%) $_a$	620 (12.8%) $_a$	1,467 (13.0%) $_a$	0.919

Extended Data Fig. 2 | Baseline characteristics in the primary cohort training, development and test datasets. Primary cohort sample size was 53,870 individual people. Where data was only available for subgroups of the full cohort, subgroup sample size is denoted by N. Differences in means of continuous variables between 2 groups were compared using two-sample t-test. Differences in means of continuous variables between 3+ groups were compared using one-way ANOVA. Differences in proportions of categorical variables between the 2+ groups were compared using Chi-Squared. Tests of significance were 2 sided. $^{a, b, c}$: Each subscript letter denotes a subset of dataset categories whose column proportions do not differ significantly from each other at the 0.05 level. Post-hoc analysis was performed using Fisher's least significant differences to compare means of continuous variables between groups. Abbreviations**:** SD: Standard deviation; CAD: Coronary artery disease; CHF: Congestive heart failure; COPD: Chronic obstructive pulmonary disease; HR: Heart rate, MI: Myocardial Infarction; PVD: Peripheral Vascular Disease.



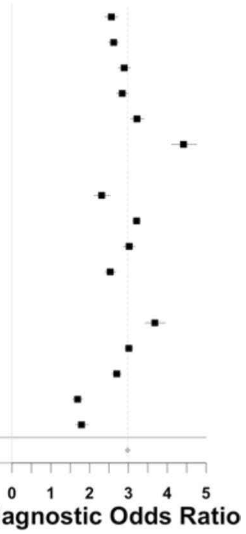
	Primary Cohort (n = 53,870)	Contemporary Cohort (n = 7,806)	Clinic Cohort (n = 181)	*p-value*
Age, years, mean ± SD	45.6±16.2 $_{a}$	44.5±16.3 $_{b}$	63.1±14.7 $_{c}$	<0.001
Gender	*N=37,443*	*N=3,936*	*N=181*	
Females, n (%)	17,545 (46.9%) $_{a}$	2,237 (56.8%) $_{b}$	66 (36.4%) $_{c}$	
Males, n (%)	19,898 (53.1%) $_{a}$	1,699 (43.2%) $_{b}$	115 (63.5%) $_{c}$	<0.001
Race and ethnic group	*N=37,443*	*N=3,936*	-	
Non-Hispanic White, n (%)	28,221 (75.3%) $_{a}$	2,960 (75.2%) $_{a}$	-	
Black or African American, n (%)	796 (2.1%) $_{a}$	107 (2.7%) $_{b}$	-	
Hispanic, Latino or Spanish origin/ancestry, n (%)	3,976 (10.6%) $_{a}$	456 (11.5%) $_{b}$	-	<0.001
Asian, n (%)	2,229 (5.9%) $_{a}$	183 (4.6%) $_{b}$	-	
Multi-ethnic, n (%)	1,376 (3.7%) $_{a}$	119 (3.0%) $_{b}$	-	
Other, n (%)	845 (2.3%) $_{a}$	111 (2.8%) $_{b}$	-	
Waveform data	*N=53,870*	*N=7,806*	*N=181*	
Duration of waveform, mean ± SD	22.0±10.9 $_{a}$	20.6±9.4 $_{\text{bå}}$	30.1±18.3 $_{c}$	<0.001
Heart rate, mean ± SD	80.2±15.1 $_{a}$	80.8±14.0 $_{b}$	76.2±24.0 $_{c}$	<0.001
No. of recordings, mean ± SD	48.1±95.3 $_{a}$	31.3±64.5 $_{b}$	1±0 $_{c}$	<0.001
Anthropometric data	*N=5,636*	*N=650*	*N=181*	
Height, meters, mean ± SD	1.72±0.10 $_{a}$	1.71±0.01 $_{b}$	1.68±0.26 $_{c}$	<0.001
Weight, kg, mean ± SD	82.9±20.3 $_{a}$	82.7±21.8 $_{a}$	84.4±24.2 $_{a}$	0.643
BMI, mean ± SD	27.7±6.1 $_{a}$	28.2±6.6 $_{b}$	28.9±6.6 $_{b}$	0.008
Medical conditions	*N=53,870*	*N=7,806*	*N=181*	
No reported medical conditions	26,782 (49.7%) $_{a}$	3,517 (45.0%) $_{b}$	49 (27.1%) $_{c}$	<0.001
Diabetes mellitus, n (%)	3,564 (6.6%) $_{a}$	679 (8.7%) $_{b}$	38 (21.0%) $_{c}$	<0.001
Hypertension, n (%)	14,144 (26.3%) $_{a}$	2,452 (31.4%) $_{b}$	99 (54.7%) $_{c}$	<0.001
Hypercholesterolemia, n (%)	15,349 (28.5%) $_{a}$	2,381 (30.5%) $_{b}$	76 (42.0%) $_{c}$	<0.001
Coronary artery disease, n (%)	3,766 (7.0%) $_{a}$	655 (8.4%) $_{b}$	11 (6.1%) $_{a}$	<0.001
Prior MI, n (%)	1,814 (3.4%) $_{a}$	301 (3.9%) $_{b}$	11 (6.1%) $_{a, b}$	0.013
CHF, n (%)	1,201 (2.2%) $_{a}$	232 (3.0%) $_{b}$	11 (6.1%) $_{c}$	<0.001
PVD, n (%)	1,034 (1.9%) $_{a}$	179 (2.3%) $_{b}$	6 (3.3%) $_{c}$	<0.001
Prior Stroke, n (%)	1,468 (2.7%) $_{a}$	289 (3.7%) $_{b}$	5 (2.8%) $_{a, b}$	<0.001
Sleep apnea n (%)	7,001 (12.9%) $_{a}$	1,263 (16.2%) $_{b}$	13 (7.2%) $_{a, b}$	<0.001

Extended Data Fig. 3 | Baseline characteristics of the primary, contemporary and clinic cohorts. Where data was only available for subgroups of the full cohorts, subgroup sample size is denoted by N. Differences in means of continuous variables between 2 groups were compared using two-sample t-test. Differences in means of continuous variables between 3+ groups were compared using one-way ANOVA. Differences in proportions of categorical variables between the 2+ groups were compared using Chi-Squared. Tests of significance were 2 sided. $^{a, b, c}$: Each subscript letter denotes a subset of dataset categories whose column proportions do not differ significantly from each other at the 0.05 level. Post-hoc analysis was performed using Fisher's least significant differences to compare means of continuous variables between groups. Abbreviations: SD: Standard deviation; CAD: Coronary artery disease; CHF: Congestive heart failure; COPD: Chronic obstructive pulmonary disease; HR: Heart rate, MI: Myocardial Infarction; PVD: Peripheral Vascular Disease.

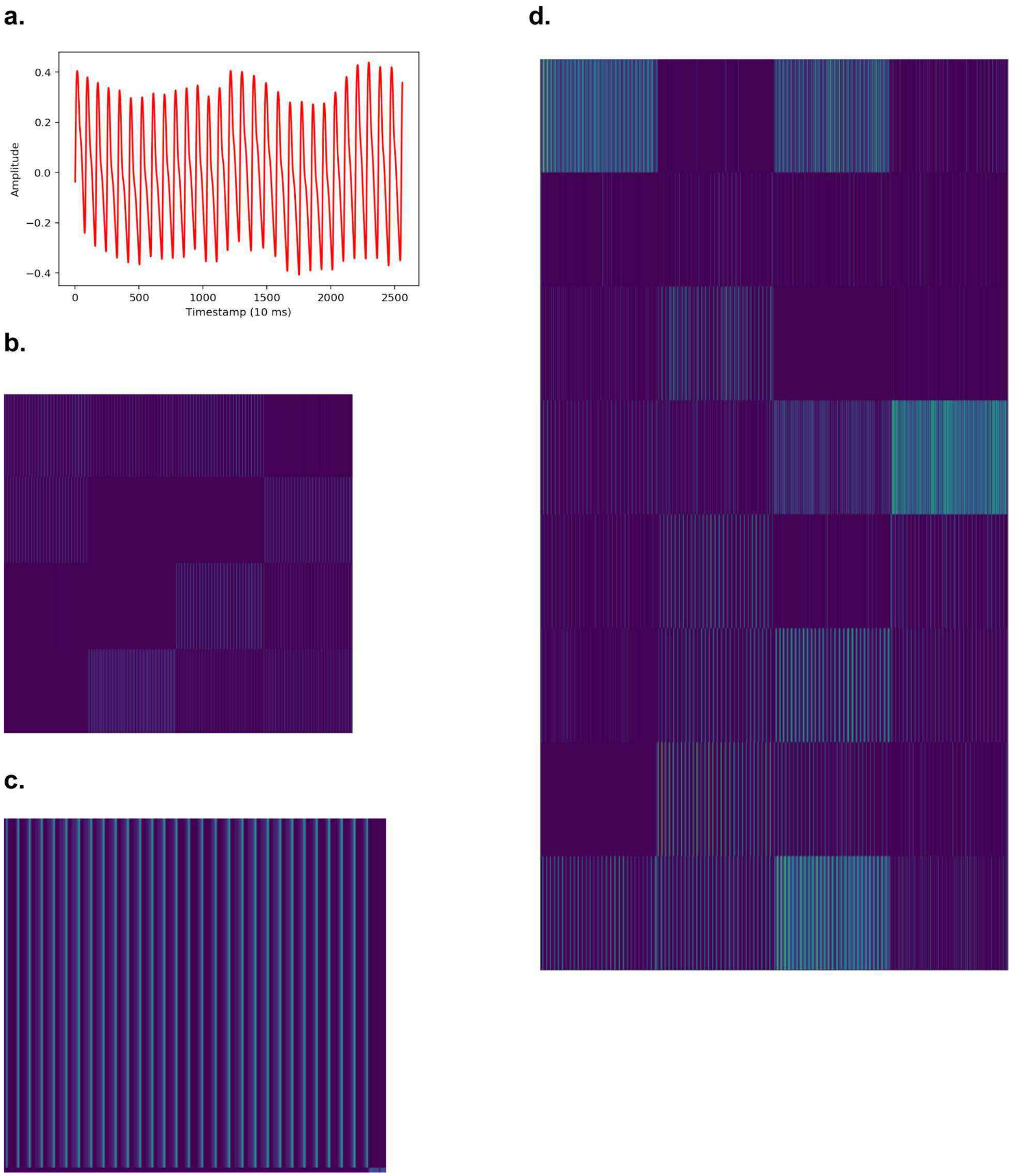


Extended Data Fig. 4 | Confusion matrices for DNN performance in three validation datasets. Confusion matrices for the predictions of the DNN in the Test Dataset (**a**, **b**), Contemporary Cohort (**c**, **d**), and Clinic Cohort (**e**, **f**), at both the recording and user-level. Total number of patients are presented in parentheses. The DNN Score cutoff used was 0.427.

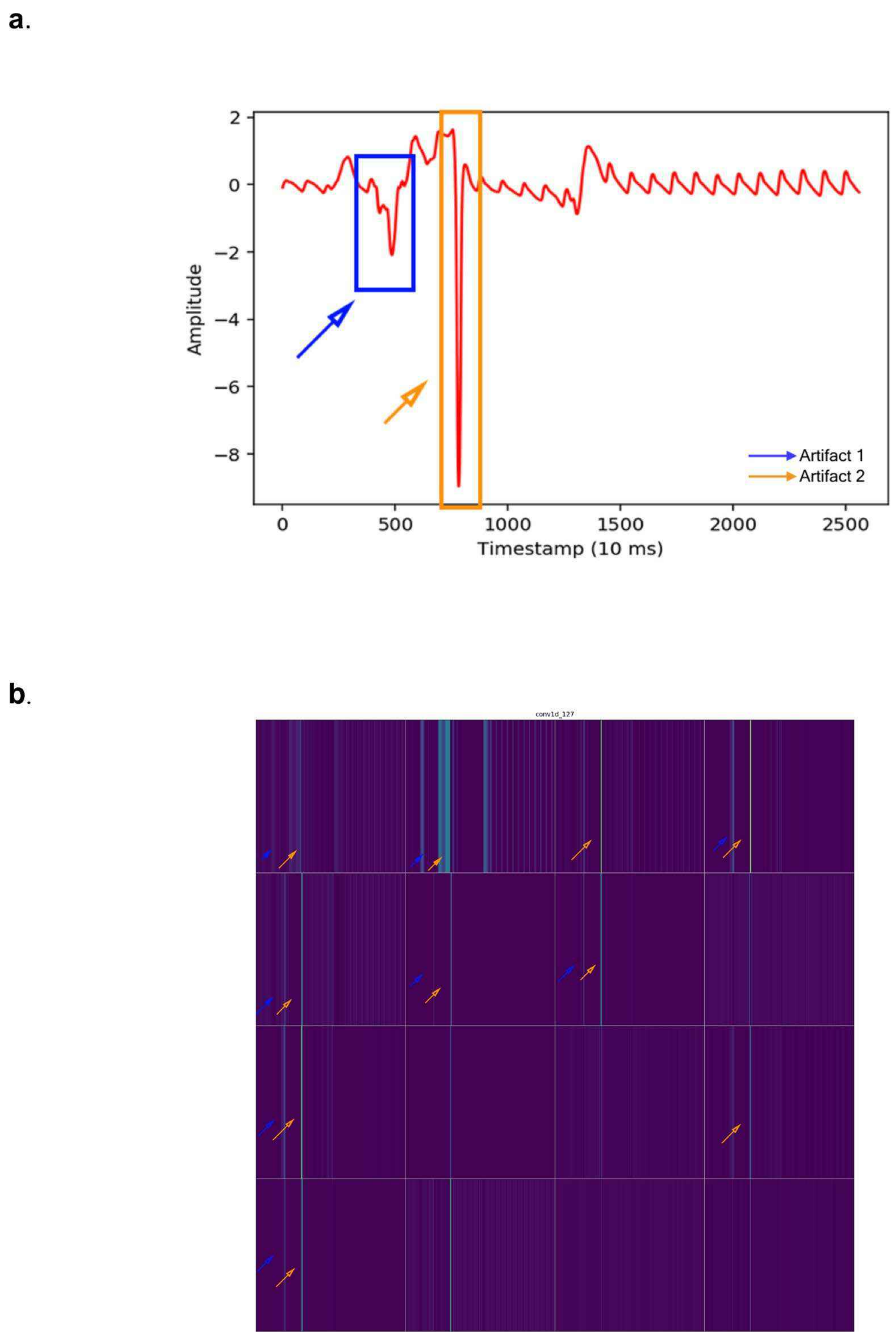
Group	Sensitivity	Specificity	OR (95% CI)	Interaction p-value	AUC
Time of the recording					
0-4 AM	60% (2,463/4,107)	63% (42,167/66,862)	2.56 (2.40-2.73)	0.01	0.66
4-8 AM	64% (4,745/7,390)	59% (54,320/91,565)	2.62 (2.49-2.75)	0.04	0.66
8 AM-12 PM	67% (3,824/5,668)	58% (52,752/90,664)	2.89 (2.72-3.06)	0.78	0.68
12-4 PM	65% (3,992/6,124)	60% (58,524/97,096)	2.84 (2.69-3.00)	0.37	0.67
4-8 PM	70% (4,487/6,446)	58% (49,804/85,206)	3.22 (3.05-3.40)	0.26	0.68
8 PM-12 AM	71% (2,566/3,591)	63% (34,878/54,633)	4.42 (4.10-4.76)	<0.001	0.73
Length of the recording					
5-10 seconds	62% (1,252/2,010)	58% (14,835/25,429)	2.32 (2.11-2.54)	0.01	0.64
10-15 seconds	67% (12,219/18,347)	62% (165,903/268,934)	3.21 (3.11-3.31)	0.04	0.69
15-20 seconds	67% (5,117/7,638)	60% (67,179/112,338)	3.02 (2.87-3.17)	0.03	0.68
>20 seconds	65% (4,067/6,227)	57% (53,441/93,205)	2.53 (2.40-2.67)	<0.001	0.66
Heart rate					
<60 BPM	43% (1,426/3,343)	83% (76,562/92,029)	3.68 (3.43-3.95)	0.05	0.73
60-80 BPM	64% (9,168/14,327)	63% (133,211/211,761)	3.01 (2.91-3.12)	0.53	0.68
80-100 BPM	77% (8,531/11,009)	44% (53,470/121,646)	2.70 (2.58-2.83)	0.53	0.65
100-120 BPM	73% (2,701/3,695)	38% (16,455/42,985)	1.69 (1.53-1.56)	0.01	0.59
≥120 BPM	45% (829/1,848)	68% (21,660/31,485)	1.79 (1.63-1.97)	0.02	0.61
Overall (Recording-level)	**66% (22,880/34,529)**	**60% (305,655/507,309)**	**2.98 (2.91-3.05)**	**-**	**0.68**



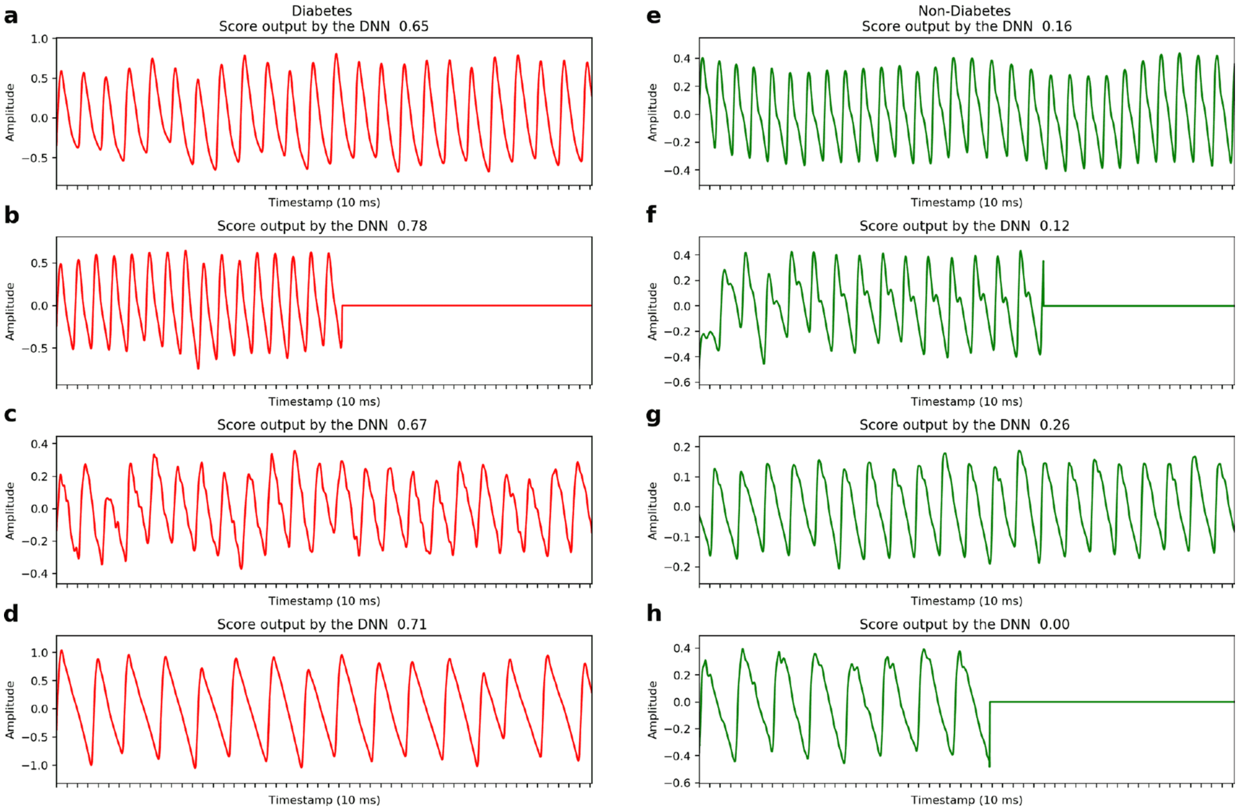
Extended Data Fig. 5 | DNN performance to predict diabetes according to time of day, recording length and heart rate in the test dataset. DNN sensitivity, specificity, diagnostic odds-ratio and AUC to detect prevalent diabetes are presented across strata of age, gender and number of recordings. The Test Dataset sample size is 11,313 individuals. Counts are provided in parentheses for all subgroup metrics. The diagnostic odds-ratio is the ratio of positive likelihood ratio (sensitivity / (1−specificity)) to the negative likelihood ratio ((1−sensitivity)/specificity). The diagnostic odds-ratio is presented at the recording-level with the associated 95% confidence interval. Interaction p-values are two-sided Wald tests for interaction between the DNN Score and the respective covariates for diabetes. Abbreviations**:** DNN: deep neural network; OR: diagnostic odds ratio; AUC: area under the curve; CI: confidence interval; BPM: beats per minute.



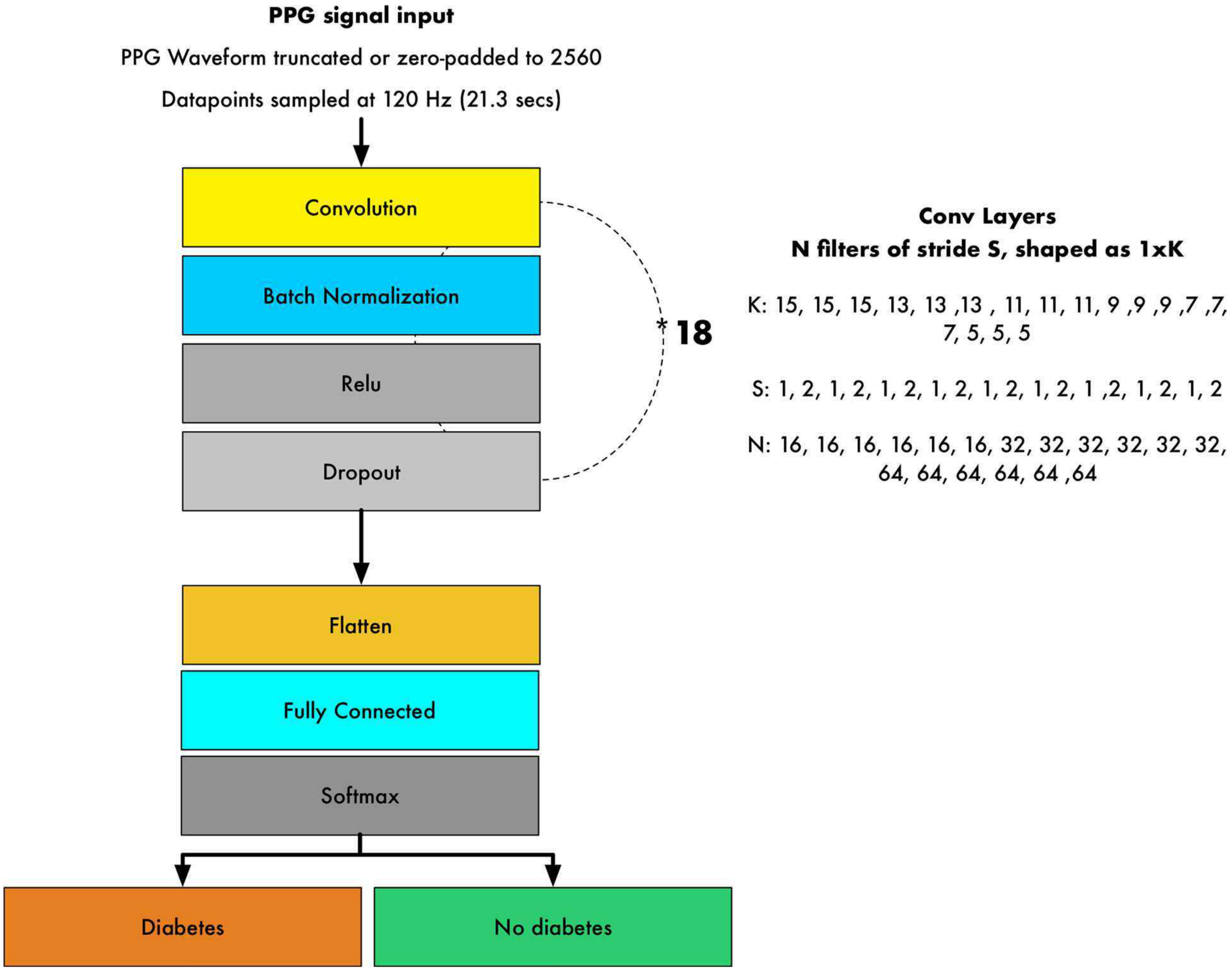
Extended Data Fig. 6 | Activation maps from several hidden convolutional layers of the trained deep neural network (DNN) for one photoplethysmography (PPG) record. a, An example of a PPG recording which serves as the input into the DNN. **b**, The activation map of one example filter (out of 16) from the first convolutional layer of the neural network. This activation map is obtained after the example PPG recording is fed into the trained DNN. Each lighter colored band illustrates "activation" of a model parameter. At this early layer of the neural network, the lighter colored bands correspond directly to each cardiac cycle of the PPG waveform. Thicker lines likely indicate morphological features of the waveform. **c**, Visualization of the activation maps of the 16 filters from the first convolutional layer of the neural network, obtained after the input PPG is fed into the trained DNN. Each of the 16 filters can learn different sets of "features" from the input PPG recording. Filters with more purple bands have more inactive neurons, as compared to those with lighter colors (green being the strongest activation and dark purple being the weakest activation). Six filters appear completely inactivated (all purple), suggesting that the features these filters focus on are not present in this example input PPG. **d**, Visualization of the activation maps of the 7th convolutional layer of the DNN, comprised of 32 filters. Broadly, these activation maps from the 7th layer of the DNN are more complex compared to those from the 1st layer (**b**, **c**), demonstrating how deeper layers of the DNN encode increasingly abstract information representing higher level interactions and complex features.



Extended Data Fig. 7 | Activation maps from hidden convolutional layers of the trained deep neural network (DNN) for an example photoplethysmography (PPG) recording with artifacts. a, An example PPG recording with 2 artifacts (blue and orange rectangles) which serves as the input into the DNN. **b**, Activation maps of the 16 filters from the first convolutional layer of the DNN. Each lighter colored band illustrates "activation" of a model parameter. Orange and blue arrow are placed on filters denoting the location of artifacts, highlighted by orange and blue rectangles (**a**), respectively. Some filters, such as the 4th image in the top row, seem to not have activation at the location of the artifactual beats (hollow orange and blue arrows), suggesting that the DNN is "ignoring" data from these artifact locations. Whereas other filters are have activation, suggested by lighter color bars, in the locations of the artifacts (full orange and blue arrows), such as the 2nd filter from the left in the top row, suggesting that the DNN is using data from these artifact locations. Some filters, such as the 2nd from the left in the bottom row "ignore" the artifactual beats by having uniform activation throughout the signal length (except where there are artifacts) likely representing the cardiac cycle. These findings suggest that the DNN is able to identify artifactual beats and differentiate them from good quality waveforms.



Extended Data Fig. 8 | Example photoplethysmography (PPG) waveforms. a, Examples of raw PPG recordings from individuals with and without diabetes (red/green recordings, respectively), which serve as inputs to the deep neural network. DNN Scores predicted for each recording are shown. PPG recordings are either cropped or zero-padded to the same fixed length (~20.3 seconds) before being input into the DNN. The "flat line" in three examples is a demonstration of zero-padding shorter records to the fixed length. DNN: Deep Neural Network; ms: milliseconds.


PPG signal input
PPG Waveform truncated or zero-padded to 2560
Datapoints sampled at 120 Hz (21.3 secs)
Convolution
Batch Normalization
Relu
Dropout
* 18
Conv Layers
N filters of stride S, shaped as 1xK
K: 15, 15, 15, 13, 13 ,13 , 11, 11, 11, 9 ,9 ,9 ,7 ,7, 7, 5, 5, 5
S: 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1 ,2, 1, 2, 1, 2
N: 16, 16, 16, 16, 16, 16, 32, 32, 32, 32, 32, 32, 64, 64, 64, 64, 64 ,64
Flatten
Fully Connected
Softmax
Diabetes
No diabetes

Extended Data Fig. 9 | Deep neural network architecture. The neural network had 39 layers organized in a block structure, consisting of convolutional layers with an initial filter size of 15 and filter number (N) of 16. The size of the filters decreased, and the number of filters increased as network depth increased, as shown. After each convolutional layer, we applied batch normalization, rectified linear activation and dropout with a probability of 0.2. The final flattened and fully connected softmax layer produced an output distribution across the classes of diabetes/no diabetes. This output distribution is referred to as the DNN Score. PPG: photoplethysmography; DNN: Deep Neural Network; Hz: Hertz.

nature research

Corresponding author(s): Tison, Geoffrey H

Last updated by author(s): 6/1/2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a | Confirmed

□ ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement

□ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

□ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.

□ ☒ A description of all covariates tested

□ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

□ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)

□ ☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted
Give P values as exact values whenever suitable.

☒ □ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

□ ☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

□ ☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The Azumio Instant Heart Rate app is a commercial, freely-available smartphone application by Azumio Inc, which was used to collect PPG recordings. The PPG waveforms were pre-processed by an Azumio algorithm for camera artifact removal, utilizing standard de-trending and low pass filter techniques. Web-based data collection in the Health eHeart study is supported by custom code for the Health eHeart web portal and smartphone application.
Data analysis	The convolutional neural network was built in Python 2.7 using Keras (version 2.0.3) and TensorFlow (version 1.0.1). The logistic regression models and AUC were derived in SPSS v22.0 (IBM). Custom code for data processing and analysis (including the neural network) used in this study is copyright of the Regents of the University of California and can be made available through license.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The data that support the findings of this study can be made available upon reasonable request from the authors, but restrictions apply to the availability of these data which were used under license for the current study and due to their containing information that could compromise participant privacy/consent.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For Primary and Contemporary cohorts, the largest sample size available during the inclusion period was used in order to maximize the sample size for neural network training and validation. Similarly to maximize validation sample size, we aimed for the largest enrollment in the Clinic Cohort given the time and resource constraints associated with prospective in-person enrollment.
Data exclusions	Participants in all cohorts were included if they made at least one PPG recording. Based on our prior data exploration and analysis using similar PPG data (doi:10.1038/s41746-019-0134-9), we prespecified exclusions for low sampling rate (<100 Hz; which is based on smartphone model), invalid date of birth or average PPG heart rate outside of 20-220 beats per minute which we assumed are not physiologic and thus predominantly erroneous/artifactual signals). PPG recordings of <5 seconds were excluded, as were individuals who answered "I don't know" to the self-reported diabetes question.
Replication	We replicated our results in a total of three test datasets, including a hold-out test dataset from the Primary cohort, a temporally distinct cohort enrolled into Health eHeart after initial data-lock ("Contemporary Cohort"), and a prospectively enrolled in-person clinic cohort ("Clinic Cohort").
Randomization	Participants in the Primary cohort were divided into training, development and test datasets randomly, and all datasets are disjoint.
Blinding	Investigators were not blinded to group allocation, however all three test datasets were kept separate and not used during model development. Blinding is not relevant in this study since the exposure of deployment of the trained neural-network on test dataset(s) PPG data occurs without investigator input.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☒	☐	Antibodies
☒	☐	Eukaryotic cell lines
☒	☐	Palaeontology
☒	☐	Animals and other organisms
☐	☒	Human research participants
☒	☐	Clinical data

Methods

n/a	Involved in the study	
☒	☐	ChIP-seq
☒	☐	Flow cytometry
☒	☐	MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	A total of 3,564 participants (6.6%) had self-reported diabetes and 50,306 (93.4%) did not have self-reported diabetes. Compared to those without diabetes, those with diabetes were older (mean±SD): 54.6±14.7 vs 45.0±15.1, p<0.001), more likely male (59.3% vs 52.8%; p<0.001), had a higher BMI (32.1±7.0 vs 27.3±5.9; p<0.001), less likely non-Hispanic whites (Table 1) and had higher HR (83.8.±14.5 vs 79.9±15.1 bpm; p<0.001).
Recruitment	Health eHeart (HeH) Study is a worldwide, internet-based, longitudinal eCohort; English-speaking adults, 18 years or order, with an email address were eligible to join. Participants were actively recruited through a variety of campaigns at UCSF (through clinics and electronically delivered invitations) and by partner organizations (e.g., American Heart Association), and passively recruited through word of mouth and press releases. Existing users of the Azumio Instant Heart Rate app were also invited to join Health eHeart. Since all study participants voluntarily enrolled into Health eHeart and elected to download the Azumio smartphone app, there is likely to be selection bias in our Primary and Contemporary cohorts, as we describe in limitations. The Clinic Cohort enrolled consecutive patients referred to three cardiovascular prevention clinics who consented to participate in the study. We opted to perform our additional validation in the Clinic Cohort precisely because it does not likely

	exhibit the selection bias of the model derivation cohort, providing a test of external generalizability of the neural network performance.
Ethics oversight	The University of California San Francisco Institutional Review Board approved this study.

Note that full information on the approval of the study protocol must also be provided in the manuscript.