Using wearable activity trackers to predict Type-2 Diabetes: A machine learning-based cross-sectional study of the UK Biobank accelerometer cohort

Abstract

Background: Between 2013 and 2015, the UK Biobank collected accelerometer traces using wrist-worn triaxial accelerometers for 103,712 volunteers aged between 40 and 69, for one week each.
This dataset has been used in the past to verify that individuals with chronic diseases exhibit reduced activity levels compared to healthy populations. Yet, the dataset is likely to be noisy, as the devices were allocated to participants without a specific set of inclusion criteria, and the traces reflect uncontrolled free-living conditions.

Objective: To determine the extent to which accelerometer traces can be used to distinguish individuals with Type-2 Diabetes (**T2D)** from normoglycaemic controls, and to quantify their limitations.

Methods: Supervised machine learning classifiers were trained using the different sets of features, to segregate T2D positive individuals from normoglycaemic individuals. Multiple criteria, based on a combination of self-assessment UKBiobank variables and primary care health records linked to the participants in UKBiobank, were used to identify 3,103 individuals in this population who have T2D. The remaining non-diabetic 19,852 participants were further scored on their physical activity impairment severity levels based on other conditions found in their primary care data, and those likely to have been physically impaired at the time were excluded.

Physical activity features were first extracted from the raw accelerometer traces dataset for each participant, using an algorithm that extends the previously developed Biobank Accelerometry Analysis toolkit from Oxford University [1]. These features were complemented by a selected collection of socio-demographic and lifestyle features available from UK Biobank.

Results: Three types of classifiers were tested, with AUC close to[0.86; 95% CI: .85-.87] for all three, and F1 scores in the range [.80,.82] for T2D positives and [.73,.74] for controls. Results obtained using non-physically impaired controls were compared to highly physically impaired controls, to test the hypothesis that non-diabetes conditions reduce classifier performance. Models built using a training set that includes highly impaired controls with other conditions had worse performance: AUC [.75-.77; 95% CI: .74-.78] and F1 in the range [.76-.77] (positives) and [.63,.65] (controls).

Conclusions: Granular measures of free-living physical activity can be used to successfully train machine learning models that are able to discriminate between T2D and normoglycaemic controls, albeit with limitations due to the intrinsic noise in the datasets. In a broader, clinical perspective, these findings motivate further research into the use of physical activity traces as a means to screen individuals at risk of diabetes and for early detection, in conjunction with routinely used risk scores, provided that appropriate quality control is enforced on the data collection protocol in order to improve the signal-to-noise ratio.

Keywords:
Digital phenotype; accelerometer; machine learning; physical activity; self-monitoring; diabetes; classification

Introduction

Objective measures of physical activity can be used to characterise people's free-living movement behaviour to provide the kind of digital phenotype [2] that promises to support a vision of participatory, preventive, personalised healthcare. The largest available dataset of free-living physical activity traces has been collected by the UK Biobank. [3] It includes uncontrolled, raw accelerometry traces collected for 7 days for a random selection of 103,712 out of a total 502,664 UK Biobank participants, (approximately 25%), between February 2013 and December 2015. All the studies cited here, including the one described in this paper, have used a reduced set after performing quality checks.

This dataset has been used in recent studies to quantify differences in physical activity levels across the general UK Biobank population [1], as well as to show that participants with chronic diseases exhibit lower levels of activity than the general UK Biobank cohort [4]. It has also demonstrated associations between cardiometabolic health, multimorbidity, and mortality.[5] [6] However, this dataset has not been used to validate the hypothesis that accelerometer traces measures of physical activity can be used as a predictor for Type 2 Diabetes (T2D) and thus, potentially, as a valid digital phenotype for early detection of T2D. Yet, this is a natural direction to explore, as T2D is linked with low levels of physical activity and increasing age. [7] This disease has become much more prevalent and is rapidly rising globally, especially in parts of the developing world. [8]

Research into the effectiveness of activity monitoring for T2D detection and prevention is motivated by the disproportionately high cost, both economic and social, of treating T2D[9], considering that approximately 90-95 percent of diagnosed diabetes among adults is type-2, and is therefore potentially preventable. In the UK alone, more than 2.7 million people are diagnosed with T2D whilst a further 750,000 people are believed to have the symptoms, but are yet to be diagnosed with the disease. [10]

Studies have been undertaken to use digital phenotypes for early diagnosis, but most are focused on using traditional multi -*omics* approaches[11]. In this study we test the hypothesis that activity profiles, when represented in sufficient detail, differ significantly between individuals with T2D and the general population.

This study begins by defining T2D participants in the UK Biobank using a combination of pre-existing diagnoses collected in the UK Biobank assessment centres, and automated analysis of the participants' Electronic Health Records **(EHR)** follow-up. It then evaluates the extent to which accelerometer traces traces can distinguish individuals with T2D from normoglycaemic controls. The approach employs a combination of traditional machine learning classification models to quantify the predictive power of features extracted from accelerometer traces, and to assess their limitations relative to this task.

Methods

In this paper, we refer to each volunteer's one-week activity recording period as their *wear time*, and to those UK Biobank volunteers as the *accelerometry cohort*.

The dataset used in this study was derived from the collection of activity traces for each of these participants, filtered using the inclusion and exclusion criteria described below. Variables representing physical activity features were extracted from the raw traces. Additionally, a small set of socio-demographic, anthropometric and metabolic variables were added, following recent studies [11], where those same variables were used to characterise the behavioural phenotype of UK Biobank participants relative to cardiovascular disease (CVD) and T2D.

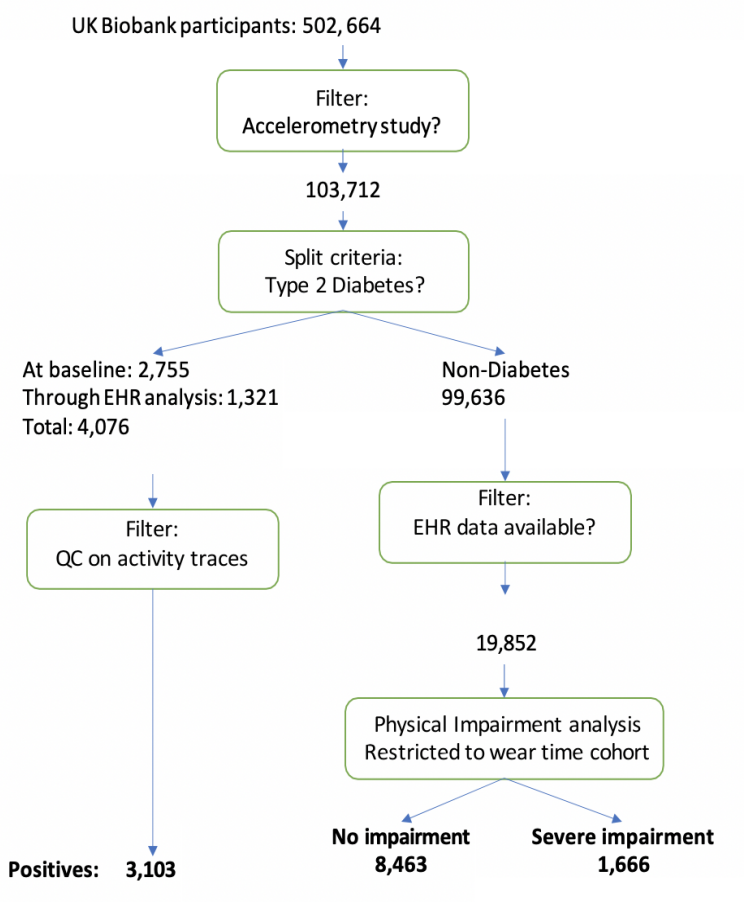

Figure 1 Training set selection criteria for T2D positive and negative individuals

Inclusion and exclusion criteria for T2D positive participants

The criteria described below and the resulting dataset sizes are summarised in Figure 1. T2D participants were identified using a combination of self-reported data collected at the Biobank assessment centre, as well as data from the participants' primary care Electronic Health Records (EHR), including prescriptions. At the time of writing, EHR records are available for about 245,000 out of 502,664 individuals (around 45%) of the UK Biobank population. Inclusion into the T2D group, based on self-reporting, follows the same criteria as in[11], namely individuals with an explicit diagnosis as part of their assessment, based on the UK Biobank Showcase [13].
At the baseline assessment centre, participants who had been entered as having 'diabetes' or 'type 2 diabetes' were selected. Those taking insulin within their first year (variable 2986-0.0), and were less than 35 years of age (variable 2976-0.0) at diagnosis were excluded to reduce the likelihood including individuals with Type-1 Diabetes and monogenic forms of diabetes.[14] This resulted in 2,755 participants from the accelerometry cohort being identified as having T2D.

Primary care EHRs were also used to identify participants who developed T2D after their baseline assessment but before their accelerometer wear time. The incidence of T2D was defined as the occurrence of a Read v2 or Clinical Terms Version 3 (CTV3) code corresponding to T2D after the date of the assessment centre visit. Read v2 code sets developed by Kuan et al were used [15] and equivalent Clinical Terms Version 3 codes were mapped using mapping data provided by the UK Biobank.[4][5]
The low prevalence of T2D in the UK Biobank population is reflected in the very small positive group, compared to an overwhelmingly large non-T2D group control group (99,636 participants). Therefore, it is necessary to

rebalance these classes prior to model learning. Rather than random selection from the control group, a better selection criteria can be adopted.

We observe that the normoglycaemic control group may include individuals with non-diabetes related physical activity impairments. Excluding such individuals is desirable, as it is likely to remove noise from the control group. The controls' selection process described below includes a judgement, grounded in general medical knowledge, of how a broad variety of conditions may have affected a participant's ability to perform normal activities. Although the assessment may not be entirely accurate, to the best of our knowledge, this is the first attempt to select from a control group based on EHR data. The outcome is assumed to be no worse than random selection from the control group. Results show that prediction accuracy improves relative to using a random-controls training set.

The selection process involved a further analysis of EHRs for a period antecedent the wear time, to identify any non-diabetes medical conditions that may have resulted in physical activity impairment. This analysis is limited by the partial availability of EHR (about 20,000 individuals within the cohort). The analysis is described in detail in the **Multimedia Appendix A**. An impairment score is calculated for each individual, by (i) associating a *severity score* with each type of relevant disease reference in the Read v2 catalogue, and (ii) averaging the scores across all occurrences of the disease references in the individual's EHR history, within a period of 6 months before wear time. Records are included for 1 month after wear time, as there may be a delay in recording new conditions. The analysis resulted in two control sub-populations as shown in Figure 1 (bottom right), *Norm-0*, where we expect no impairment (N=8,463) , and *Norm-2*, with expected high impairment (N=1,666). These figures are summarised in Table 1. Both sets were used as part of supervised learning, in separate experiments, as explained next.

It is also acknowledged that there were 151 out of 3,101 T2D positive individuals also with a high impairment severity score for physical activity. This small subset of the T2D positive population was not excluded from the training datasets. T2D is known to be a complex disease which can cause many complications or as a co-morbidity with other conditions, such as Cardiovascular Disease. Therefore, in order to capture all behaviours and activity patterns associated with T2D, it is important to include the severely impaired T2D positive individuals in the overall T2D positive population.

We have also experimentally verified that removing these few individuals from the trianing set does not alter the properties of the resulting model -- see Results.

Training datasets

Using these two control groups, two training sets were formed: **TS1**: T2D vs Norm-0, and **TS2**: T2D vs Norm-2. The first was used to test our main hypothesis that activity levels in the T2D group are significantly different from those in the un-impaired control group. The second was used to quantify the effect of possible non-diabetes activity impairment as a source of noise in the controls. This was achieved by training the same models using TS1 and TS2, then comparing their relative predictive performance.

Impairment Score	**Participants**	
	Total	**With adequate wear time**
Norm-0 (0)	11, 019	8,463
Norm-2 (> 50)	3, 355	1,666

Table 1 Number of participants in each sub-population according to activity impairment severity score

Physical activities features

Raw Axivity accelerometry traces contain triaxial (*x*,*y*,*z*) time series. The open source Accelerometry Analysis toolkit developed at the University of Oxford, available on GitHub in [16], for Axivity *.cwa* formats was used to reconstruct annotated timelines for each raw activity trace [17]. The tool breaks down the time series into 30-second fragments, called *epochs*, and then employs a classifier (Random Forests and Hidden Markov Models) to

annotate a time series where each epoch belongs in one of five activity types: *Sedentary*, *Moderate*, *Walking*, *Sleep*, and *Light tasks*. The tool makes a distinction between walking from sedentary and moderate activities. According to the authors, these activity types correspond to the following MET levels: Walking: 3.2, Sedentary: 1.5, Moderate: 4.9, Light tasks: 2.2, Sleep: 1.0 The feature extraction hierarchy is summarised in Figure 2.

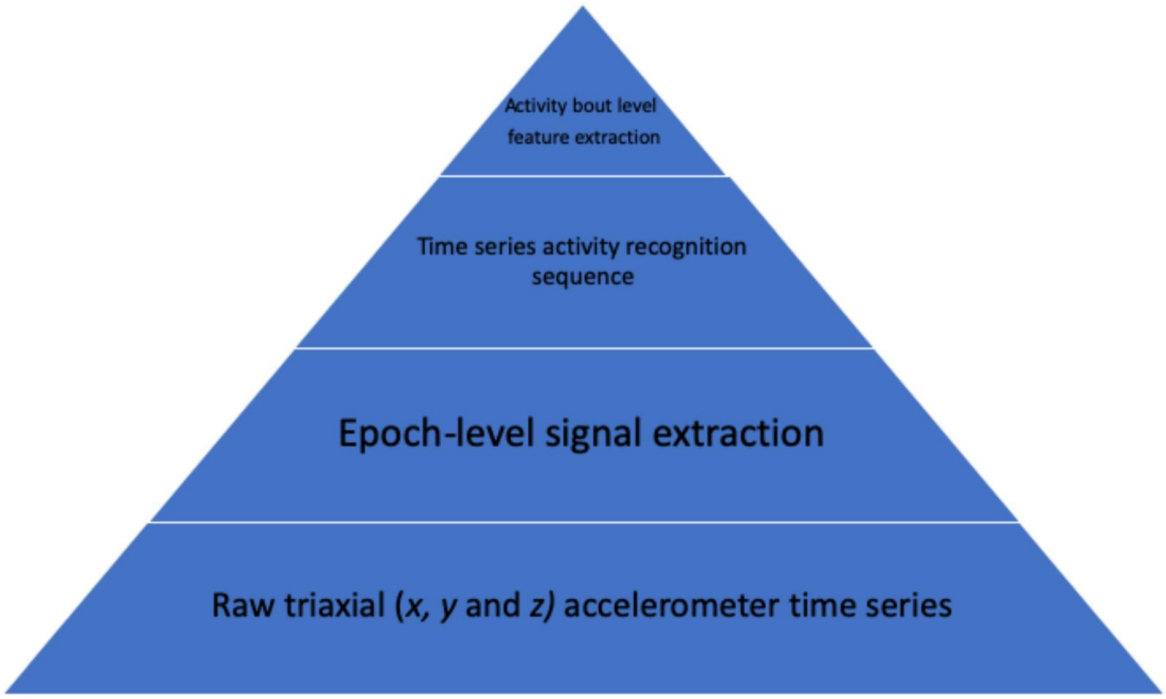

Figure 2 Hierarchy of physical activity representations

The level above the time series activity recognition sequence uses activity bouts. An activity bout is defined as a single epoch or an uninterrupted consecutive series of epochs where a single activity type is performed. The length of a bout refers to how many 30-second epochs each bout is performed for. The features extracted for this study are at the level of activity bouts of each activity type: their frequency, average length and percentage time spent in each, broken down into fractions of a 24-hour day. This choice is inspired by neuroscience research on the effects of cognitive impairment in early stages of Parkinson's on gait, where ambulatory bouts play a key role[18,19]. To extract these features, a personalised analysis of daily activities is performed. Firstly, to accommodate for different sleeping habits, night-sleep time boundaries are identified for each individual. These are defined as the average of the largest nearly-continuous period of sleep activity bouts across a 24 hour period. Then, the remaining period of

the 24-hour day is divided into three phases, denoted as Morning, Afternoon, and Evening. Within each phase, bouts of activities of different types are counted.

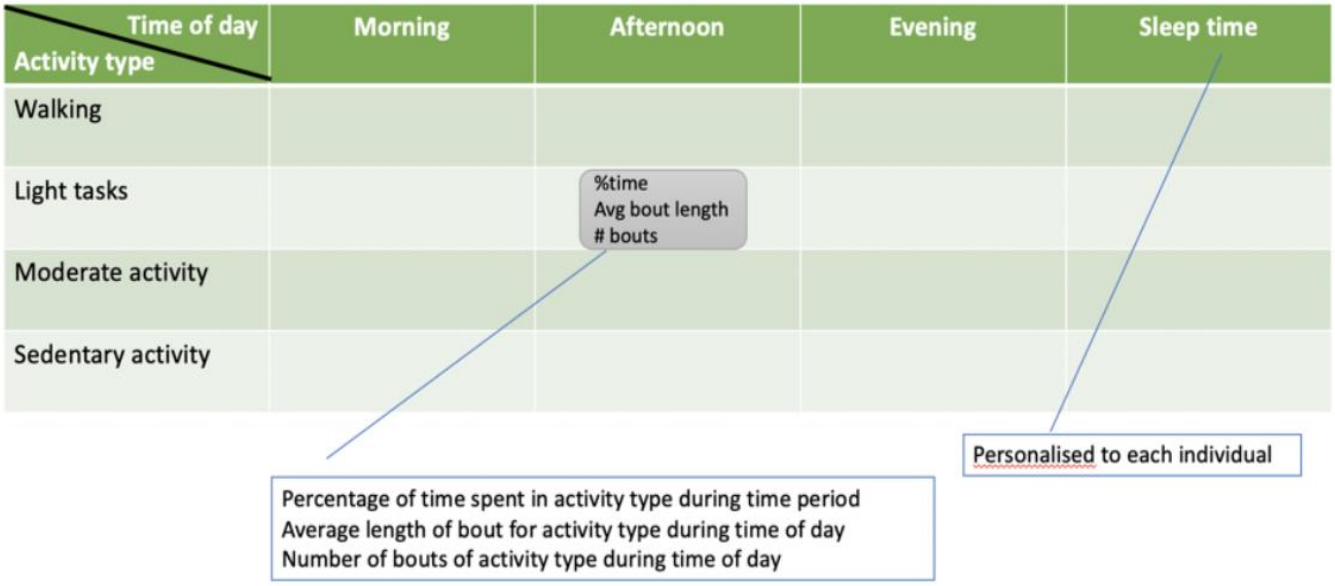

Figure 3 Feature matrix for physical activity bout representation space

This analysis results in a breakdown of 60 bout-level features, organised into a 5x4 matrix, having 4 types of daily activities plus sleep time, which is counted as the longest continuous bout of sleeping activity and 3 times of the remaining day, as shown in Figure 3. Each element in the matrix (type of activity, time of day) has three features: (i) number of bouts for that activity, percentage time spent in the activity, and average length of the bouts. This arrangement results in a total of 60 features per individual. These are then aggregated over the seven days wear time, taking the average for each element in the matrix. This feature space is referred to as the **High-Level Activity-bout Features** in this paper. The code is available on GitHub [20].

Sociodemographic, anthropometric and lifestyle features

To quantify the relative importance of the new high-level activity bout features when used in machine learning, traditional socio-demographic and lifestyle indicators that are commonly associated with incidence of T2D have been added. These are shown in Table 2 and are chosen is based upon prior studies [5,12]. These features are combined with self-reported assessments of physical activity, some of which are not part of the output from the Oxford accelerometer analysis tool, notably vigorous activity. In contrast, the physical activity features in our approach are the high-level activity bout features, obtained from objective accelerometer measurements. Objective physical activity metrics also help to validate subjective measurements[21,22].

Sociodemographic/ Lifestyle/ Anthropometry characteristic	**Description**	**Percentage of data missing (%)**
Sex	Male or female (roughly 50:50 ratio)	0.0
Age at assessment centre	Recruits at baseline are between ages 40 and 69	0.0
Ethnic group	Predominantly White British, with some participants identifying as BAME groups.	0.33

Alcohol drinking status	Participant reports if they are alcohol drinkers in the past, are currently drinking alcoholor have never drank alcohol	0.06
Smoking status	Participants report if they have smoked in the past, are currently smoking or have never smoked	0.24
Body fat percentage	Percentage of fat in total body mass (A better indicator for obesity than BMI)	1.42
Waist circumference	Measurement taken around the abdomen at the level of the umbilicus (belly button)	0.14
Sleep duration	Self-reported average duration of sleep in a day	0.28
Time spent watching television	Self-reported average time spent watching television per day	0.28
Townsend index	Metric for material deprivation within a population	0.10
Duration of walking activity	Self-reported average duration of time spent walking in a day	3.27
Duration of vigorous activity	Self-reported average duration of time spent performing vigorous activities during the day	36.43
Duration of moderate activity	Self-reported average duration of time spent performing moderate activity during the day.	15.43

Table 2 Sociodemographic, lifestyle and anthropometric characteristics selected from the UK Biobank baseline assessment for comparison with high-level activity bout features space

The IPAQ short questionnaire was used for the variables measuring physical activity (including moderate, vigorous, walking), television viewing times and sleep duration in Table 2. As noted in Table 2, some socio-demographic and lifestyle features have significant portions of missing data. This was solved by using a *k*-Nearest Neighbour imputer in scikit-learn [23] which calculates the missing value using the mean of *k*-nearest neighbours found in the training data using Euclidean distances, thus preserving the distribution of the original data.

Binary Classification

This exercise compares a number of classification models, obtained using different learning algorithms and using training set TS1 and TS2, introduced earlier, in separate sets of experiments. Furthermore, different combinations of features were considered for each of the training sets: (1) high-level activity bout features features only, (2) socio-demographic and lifestyle features only, and (3) high-level activity bout features combined with socio-demographic and lifestyle.

These combinations produce a space of 6 datasets on which models are trained. Three learning algorithms were tested on these datasets: Random forests, Logistic Regression and XGBoost. The eXtra Gradient Boosting algorithm, or XGBoost, is a relatively recent and perhaps less known algorithm [24], which has come to prominence thanks to its superior performance, both in terms of training time, and of prediction accuracy, compared for instance with Random Forests. XGBoost makes use of *gradient boosting*, an ensemble method which builds a stronger classifier by adding weaker models on top of each iteratively, until the training data achieves a good level of prediction performance.

Using these combinations of 6 datasets and 3 algorithms, a total of 18 classifier models were trained. Standard 10-fold cross validation was used to avoid overfitting. When learning the classifiers, a random selection of half the Norm-0 T2D negative controls in TS1 only was undertaken to balance the size of the Norm-0 T2D negatives and T2D positives (3,103 individuals). Norm-0 T2D negative individuals still vastly outnumbered the T2D positive population.

Following common practice for binary classifiers, this study reports on F1 scores, precision, recall, and Area Under Receiver Operating Characteristic Curve (AUC) scores. F1 conveys the balance between precision and recall and is a value between 0 and 1, where 1 indicates perfect precision and recall. It is calculated by the harmonic mean of

the precision and recall. The AUC, or AUROC, is a metric, with values between 0 and 1, for how well a classifier is capable of distinguishing between two classes. A value of 1 implies a good measure of discrimination, whereas a measure of 0.5 implies no discrimination capacity.

Based on these performance and evaluation metrics, models were compared to assess (i) the differences in predictive power between the two feature sets using TS1 (ii) the effect of noise in controls, using TS2, and (iii) the best modelling algorithms.

Clustering analysis

Further analysis was undertaken where unsupervised clustering algorithms were used to segregate and identify unlabelled individuals that exhibit similar behaviour with the new high-level activity bout feature space. These clusters are then profiled and interpreted in terms of their anthropometric, lifestyle and sociodemographic characteristics. This analysis is out of the scope for this paper, but is however reported in **Multimedia Appendix C**.

Results

Distribution of physical activity features

To summarise the distribution of the T2D positive and Norm-0 negative populations, the high-level activity bout features are aggregated for a 24-hour day and averaged across both populations. The distribution of average percentage of time spent in each activity type for the individuals in each population are shown in Figure 4 and Figure 5.

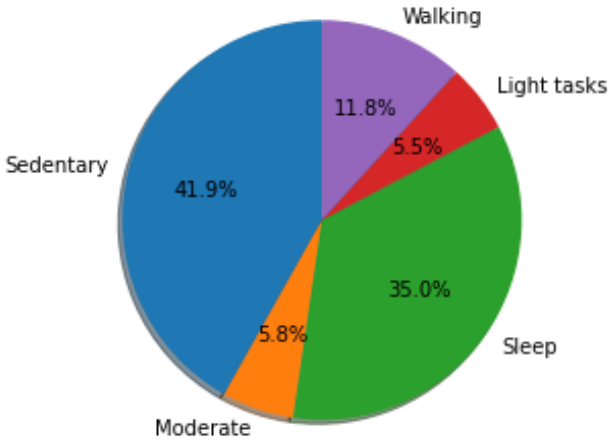

Figure 4 Pie chart for distribution of average daily percentage time spent in each activity type for T2D positive population

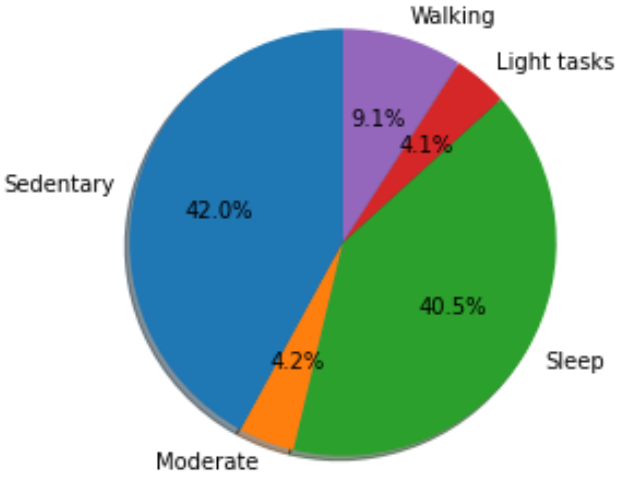

Figure 5 Distribution of daily average percentage time spent in each activity type for T2D negative Norm-0 population

The charts shown in Figure 4 and Figure 5 imply that both populations, on average, do not undertake significantly different quanities of each activity type aggregated to the level of 24- hour day. However, the high-level activity bout features also offer an insight into the regularity and length of activity bouts. The values for these features do offer some discrimination between the T2D positive and Norm-0 T2D negative populations. The histograms below demonstrate an example of this by showing the distribution of daily averages for bout length, number of bouts and the percentage times spent for sleep activity.

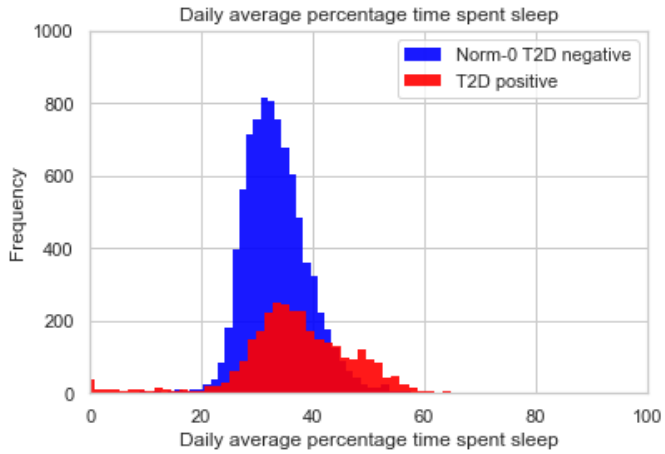

Figure 6 Histogram for daily average percentage times spent asleep

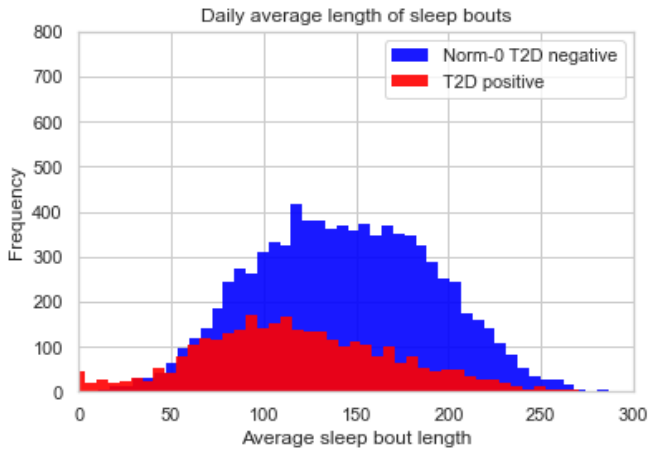

Figure 7 Histogram for daily average length of sleep bouts

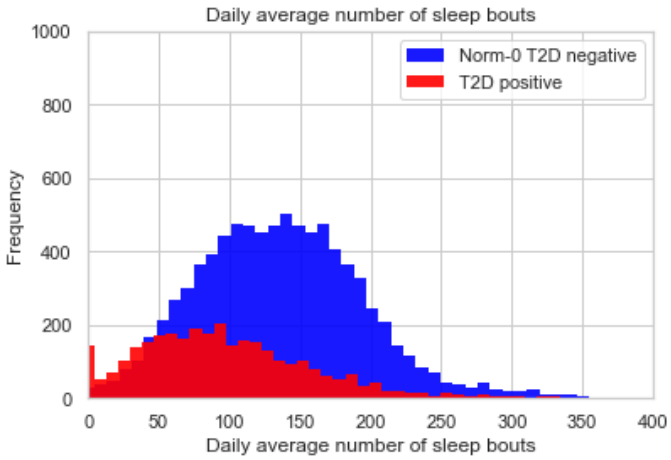

Figure 8 Histogram for daily average number of sleep bouts

The histograms in Figure 6Figure 7Figure 8 show noticeable differences between the two populations in the features that we have developed, when aggregated out to a day. Breaking the daily patterns into four distinct times of day (morning, afternoon, evening and during sleep) would further demonstrate the differences in activity bout patterns for the two populations by virtue of the granularity. The combined effect of all these granular-level activity bout features produce high model accuracy as reported below.

Binary Classification

A summary and performance comparison across the 18 models built for this study is presented in Table 3 and Table 4, where AUC measures are obtained by averaging over 10 models using cross-validation for robustness. The ROC curves, and AUC scores, shown in Figures 9-14 for all models with training and test sets split with 80:20 ratio. More detailed metrics for precision, recall, F1, ROC curves, using 10-fold cross validation, are available in the **Multimedia Appendix B**.

	High-level activity bout features		Socio-demographic and lifestyle		High-level activity bout features+socio-demographic and lifestyle	
Negatives:	**Norm-0**	**Norm-2**	**Norm-0**	**Norm-2**	**Norm-0**	**Norm-2**
Random Forest	.80	.68	.83	.78	**.86**	.77
Logistic Regression	.79	.70	.83	.78	**.86**	.78
XGBoost	.78	.66	.80	.74	**.85**	.75

Table 3 Classification results measured using AUC, contrasting Norm-0 (random) vs Norm-2 (no physical impairment) negatives

		High-level activity bout features		Socio-demographic and lifestyle		High-level activity bout features+socio-demographic and lifestyle	
	Negatives:	Norm-0	Norm-2	Norm-0	Norm-2	Norm-0	Norm-2
Random Forest	T2D negatives	.65 .78	.70 .54	.65 .78	.77 .63	**.73** **.81**	.77 .63
Logistic Regression	T2D negatives	.66 .77	.72 .54	.69 .79	.77 .65	**.74** **.82**	.77 .65
XGBoost	T2D negatives	.66 .77	.68 .52	.67 .76	.74 .62	**.73** **.80**	.76 .63

Table 4 Classification results measured using F1: contrasting Norm-0 (no impairment) vs Norm-2 (severe physical activity impairment) T2D negatives

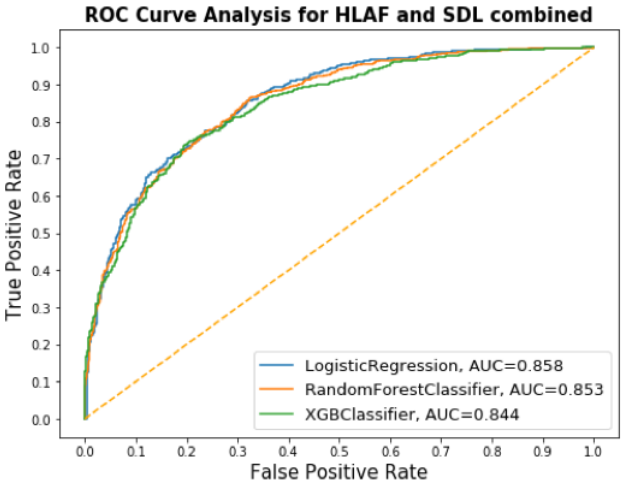

Figure 9 ROC and AUC for T2D vs Norm-0: high-level activity bout features and socio-demographic and lifestyle combined

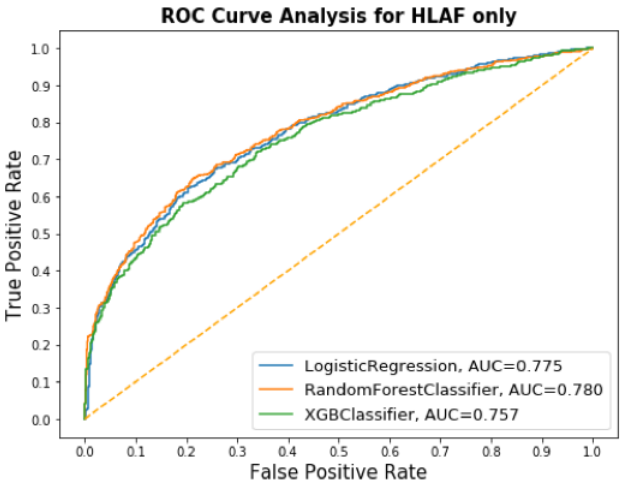

Figure 10 ROC and AUC for T2D vs Norm-0: high-level activity bout features only

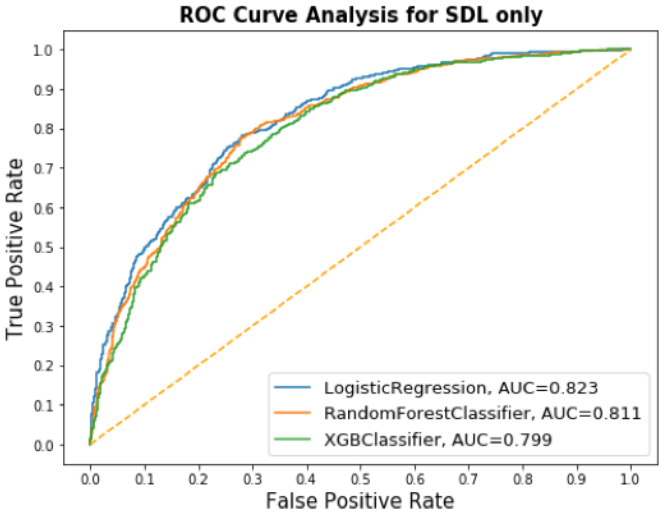

Figure 11 ROC and AUC for T2D vs Norm-0: socio-demographic and lifestyle only

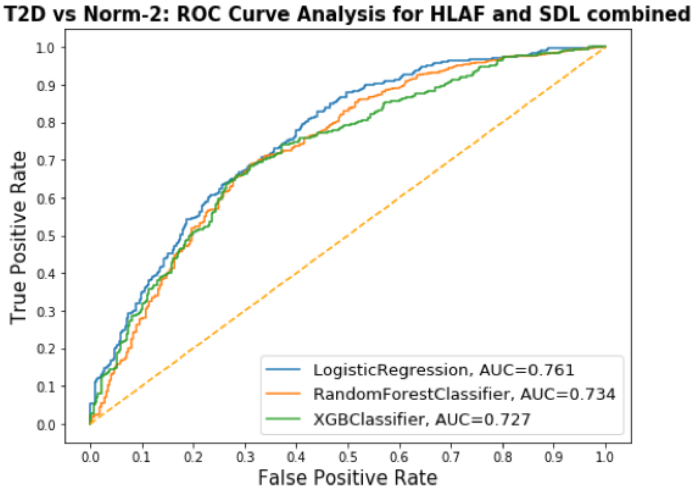

Figure 12 ROC and AUC for T2D vs Norm-2: high-level activity bout features and socio-demographic and lifestyle combined

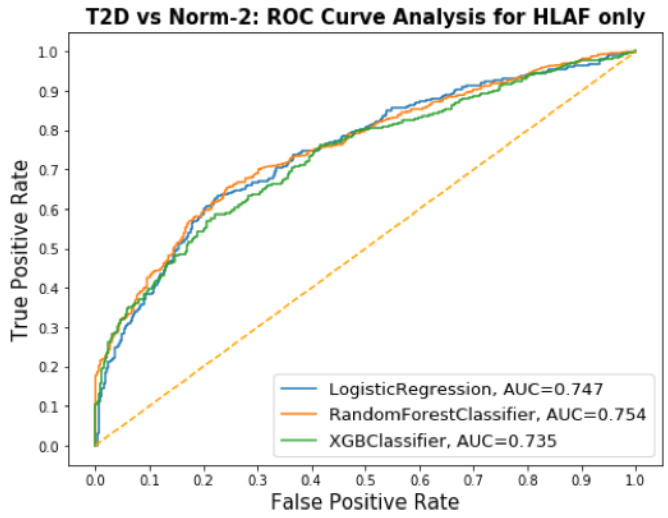

Figure 13 ROC and AUC for T2D vs Norm-2: high-level activity bout features only

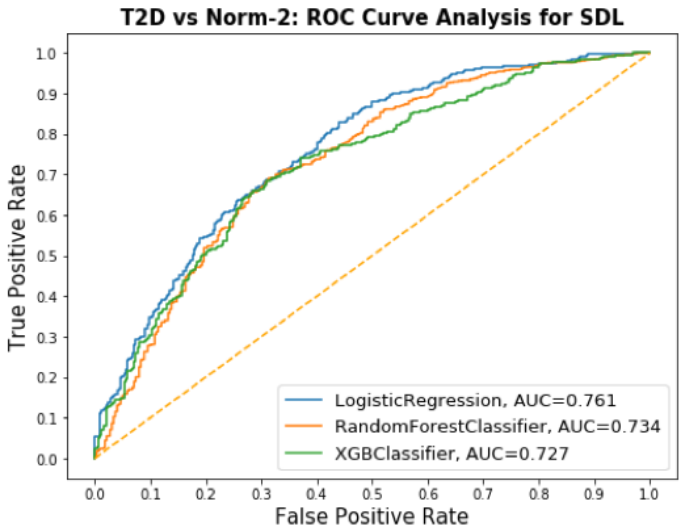

Figure 14 ROC and AUC for T2D vs Norm-2: socio-demographic and lifestyle only

When performance is measured using AUC, stronger results are achieved when using high-level activity bout features and socio-demographic and lifestyle in combination, as expected. Using high-level activity bout features on their own reduces performance (about 7 to 8%). However, high-level activity bout features provide almost the same performance as traditional socio-demographic and lifestyle features on their own.
Models were also generated using alternate training datasets, where the 151 T2D positive individuals with high physical activity impairment severity scores were excluded. As these models exhibit very similar performance as those presented here above, which suggests that physically impaired (Norm-2) T2D positive individuals can be used as part of the T2D positives in the training set.

F1 measures, in **Multimedia Appendix B**, reveal differences in prediction accuracy between T2D against Norm-0 and Norm-2 controls. When using Norm-0 controls, negatives are more accurately predicted than T2D, presumably due to class unbalance (4,178 vs 3,103). It is also clear that excluding physically impaired negatives improves the results.
When Norm-2 are used, however, T2D are more accurately predicted than negatives, perhaps because now Norm-2 are the minority class (1,666 vs 3,103), and also because of potential diversity within the highly impaired control population. This will be investigated in a further study.

In all cases, the combination of high-level activity bout features and socio-demographic and lifestyle variables gives better results than using either set of features on their own, as expected, and are also largely independent of the choice of learning algorithm, as seen by the overlapping ROC curves.

Discussion

Principal Findings

Using data from the UKBiobank, this study supports the hypothesis that individuals with diagnosed T2D exhibit patterns of physical activity that are significantly different from those of normoglycaemic controls, thus providing novel ways to detect T2D, i.e., through appropriate analysis of physical activity patterns. While most prior studies, particularly using UKBiobank, are limited to self-reported physical activity levels, [11][5][25], here we have demonstrated the benefits of extracting a more objective and granular representation of physical activity from raw accelerometry traces data, namely by activity type and time of day / sleep time. Using these features, either on their own, or in combination with a selected set of socio-demographic, anthropometric and lifestyle variables, we

have shown that appropriately trained machine learning models are able to discriminate between the two cohorts with good predictive accuracy.

Practical Significance

These findings suggest that it may be possible to use continuous or periodic self-monitoring of individuals at risk of T2D, specifically those in a pre-diabetes state, for screening and early detection of disease progression. This is particularly important as evidence emerges that reversal of T2D is possible, with a higher success rate when interventions are undertaken within the first 5 years into the disease. [26–28]

Early detection, however, is still an unsolved problem, with recent figures reporting that over 190 million people worldwide live with undiagnosed diabetes[29]
Risk scores that are routinely used for screening, such as the Leicester score, are easy to obtain but not very accurate[30]
This suggests that self-monitoring of physical activity patterns such as those presented in this study, may usefully complement risk scores to help with early detection of T2D, especially in high-risk individuals. Today, this can be achieved at low-cost using readily available technology[31], including internet-enabled data loggers that do not require participants to periodically return the devices. These include smartphones, where however further research is required to establish the quality and significance of physical activity data for this specific purpose.

Limitations

In principle, it may be possible to try and detect early signs of T2D using specific "fingerprint" patterns found in physical activity traces, where an example of pattern may be "a person who takes short bouts of low or moderate activities with frequent sedentary breaks in between". In practice, however, we have found no evidence in the UK Biobank dataset that strong enough correlations exist between specific patterns and T2D. Thus, what the machine learning approach has to offer may be limited to the strong indication demonstrated in this work, namely that granular features extracted from the raw traces, taken together, are indeed good predictors and usefully augment the more traditional socio-demographic set of variables.

While the UK Biobank is the largest known public accelerometry dataset where a T2D cohort can be identified, detecting differences between T2D and controls remains challenging, due to their low prevalence in the population, which is reflected in this study with the relatively small dataset available for training when using supervised machine learning. At the same time, this dataset is subject to noise, for two reasons. Firstly, because no formal quality assurance protocol was enforced during data collection, and secondly, because of the limited knowledge around other, non-2TD-related conditions amongst the controls, which may contribute to reduce physical mobility or a more sedentary routine. We have shown how EHRs can be used to overcome this limitation.

Conclusions

This study motivates further research into the use of granular physical activity measures as a form of *digital phenotype* for type 2 diabetes. It also suggests that more rigorous protocols on wearing physical activity loggers are required to improve the quality of the data and the signal-to-noise ratio, along with stringent inclusion and exclusion criteria or at least comprehensive knowledge of clinical conditions that may affect the signal in the traces, reflecting other studies[32,33]. When such quality criteria are met, it should be possible to repeat the analysis presented here using on datasets from a large-scale deployment of physical activity loggers, to validate the hypothesis that early detection of type 2 diabetes is both scientifically possible, and technically practical.

Acknowledgements
We would like to thank all the participants and data collectors of the UK Biobank for providing this resource that makes this study possible. We would also like to thank Dr. Doherty and his collaborators at Oxford University for making the accelerometer data analysis software libraries available in the public domain.

Authors' Contributions
B.L. and P.M. conceived this study and wrote this manuscript. B.L developed and implemented analysis. P.D and S.B developed training set inclusion/exclusion criteria. S.C and M.C reviewed and edited this manuscript. M.T, M.C and S.C were mostly responsible for making access to the UK Biobank possible.

Conflicts of Interest
None declared

Multimedia Appendices:
Mulitmedia Appendix 1:
Full details of how physical activity impairment severity was scored with EHR data.

Multimedia Appendix 2:
Full details performance metrics and evaluative measures from predictive models (ROC curves with 10-fold cross validation, f-1, precision and recall scores)

Multimedia Appendix 3:
Analysis details and results of unsupervised clustering work.

References:

1. Doherty A, Jackson D, Hammerla N, Plötz T, Olivier P, Granat MH, White T, van Hees VT, Trenell MI, Owen CG, Preece SJ, Gillions R, Sheard S, Peakman T, Brage S, Wareham NJ. Large Scale Population Assessment of Physical Activity Using Wrist Worn Accelerometers: The UK Biobank Study. Buchowski M, editor. PLoS One [Internet] 2017 Feb 1;12(2):e0169649. PMID:28146576
2. Jain SH, Powers BW, Hawkins JB, Brownstein JS. The digital phenotype. Nat Biotechnol [Internet] 2015;33(5):462–463. PMID:25965751
3. UK Biobank homepage [Internet]. [cited 2018 Sep 15]. Available from: https://www.ukbiobank.ac.uk/
4. Barker J, Smith Byrne K, Doherty A, Foster C, Rahimi K, Ramakrishnan R, Woodward M, Dwyer T. Physical activity of UK adults with chronic disease: cross-sectional analysis of accelerometer-measured physical activity in 96 706 UK Biobank participants. Int J Epidemiol England; 2019 Feb;48(4):1167–1174. PMID:30721947
5. Cassidy S, Fuller H, Chau J, Catt M, Bauman A, Trenell MI. Accelerometer-derived physical activity in those with cardio-metabolic disease compared to healthy adults: a UK Biobank study of 52,556 participants. Acta Diabetol [Internet] 2018 May; PMID:29808390
6. Strain T, Wijndaele K, Dempsey PC, Sharp SJ, Pearce M, Jeon J, Lindsay T, Wareham N, Brage S. Wearable-device-measured physical activity and future health risk. Nat Med [Internet] 2020 Sep;26(9):1385–1391. PMID:32807930
7. Tarp J, Child A, White T, Westgate K, Bugge A, Grøntved A, Wedderkopp N, Andersen LB, Cardon G, Davey R, Janz KF, Kriemler S, Northstone K, Page AS, Puder JJ, Reilly JJ, Sardinha LB, van Sluijs EMF, Ekelund U, Wijndaele K, Brage S, International Children's Accelerometry Database (ICAD) Collaborators. Physical activity intensity, bout-duration, and cardiometabolic risk markers in children and adolescents. Int J Obes (Lond) [Internet] 2018;42(9):1639–1650. PMID:30006582
8. Chen L, Magliano DJ, Zimmet PZ. The worldwide epidemiology of type 2 diabetes mellitus--present and future perspectives. Nat Rev Endocrinol [Internet] 2011 Nov 8;8(4):228–36. PMID:22064493
9. Meng Y-Y, Pickett MC, Babey SH, Davis AC, Goldstein H. Diabetes tied to a third of California hospital stays, driving health care costs higher. Policy Brief UCLA Cent Health Policy Res United States; 2014 May;(PB2014-3):1–7. PMID:24912203
10. Type 2 Diabetes- 90% of People with Diabetes have Type 2 Diabetes [Internet]. [cited 2020 Aug 20]. Available from: https://www.diabetes.co.uk/type2-diabetes.html
11. Schüssler-Fiorenza Rose SM, Contrepois K, Moneghetti KJ, Zhou W, Mishra T, Mataraso S, Dagan-Rosenfeld O, Ganz AB, Dunn J, Hornburg D, Rego S, Perelman D, Ahadi S, Sailani MR, Zhou Y, Leopold SR, Chen J, Ashland M, Christle JW, Avina M, Limcaoco P, Ruiz C, Tan M, Butte AJ, Weinstock GM, Slavich GM, Sodergren E, McLaughlin TL, Haddad F, Snyder MP. A longitudinal big data approach for precision health. Nat Med [Internet] 2019 May 8;25(5):792–804. PMID:31068711
12. Cassidy S, Chau JY, Catt M, Bauman A, Trenell MI. Low physical activity, high television viewing and poor sleep duration cluster in overweight and obese adults; a cross-sectional study of 398,984 participants from the UK Biobank. Int J Behav Nutr Phys Act [Internet] British Medical Journal Publishing Group; 2017 Apr 28;14(1):57. PMID:28454540
13. UKB: Resource 592 [Internet]. [cited 2020 Jul 9]. Available from: http://biobank.ndph.ox.ac.uk/showcase/refer.cgi?id=592
14. Cassidy S, Chau JY, Catt M, Bauman A, Trenell MI. Cross-sectional study of diet, physical activity, television viewing and sleep duration in 233,110 adults from the UK Biobank; the behavioural phenotype of cardiovascular disease and type 2 diabetes. BMJ Open [Internet] British Medical Journal Publishing Group; 2016 Mar 15 [cited 2019 May 29];6(3):e010038. PMID:27008686
15. Kuan V, Denaxas S, Gonzalez-Izquierdo A, Direk K, Bhatti O, Husain S, Sutaria S, Hingorani M, Nitsch D, Parisinos CA, Lumbers RT, Mathur R, Sofat R, Casas JP, Wong ICK, Hemingway H, Hingorani AD. A chronological map of 308 physical and mental health conditions from 4 million individuals in the English National Health Service. Lancet Digit Heal [Internet] The Author(s). Published by Elsevier Ltd. This is an Open Access article under the CC BY 4.0 license; 2019;1(2):e63–e77. PMID:31650125
16. GitHub- activityMonitoring/biobankAccelerometerAnalysis [Internet]. [cited 2019 Jun 19]. Available from: https://github.com/activityMonitoring/biobankAccelerometerAnalysis
17. Willetts M, Hollowell S, Aslett L, Holmes C, Doherty A. Statistical machine learning of sleep and physical

activity phenotypes from sensor data in 96,220 UK Biobank participants. Sci Rep [Internet] 2018;8(1):7961. PMID:29784928
18. Del Din S, Godfrey A, Galna B, Lord S, Rochester L. Free-living gait characteristics in ageing and Parkinson's disease: impact of environment and ambulatory bout length. J Neuroeng Rehabil [Internet] BioMed Central; 2016 May 12;13(1):46. PMID:27175731
19. Weiss A, Herman T, Giladi N, Hausdorff JM. New evidence for gait abnormalities among Parkinson's disease patients who suffer from freezing of gait: insights using a body-fixed sensor worn for 3 days. J Neural Transm Austria; 2015 Mar;122(3):403–410. PMID:25069586
20. Lam B, Missier P. GitHub- Activity bout feature extraction [Internet]. 2019. Available from: https://github.com/bplam88/P4-NU_Public
21. Cleland C, Ferguson S, Ellis G, Hunter RF. Validity of the International Physical Activity Questionnaire (IPAQ) for assessing moderate-to-vigorous physical activity and sedentary behaviour of older adults in the United Kingdom. BMC Med Res Methodol [Internet] 2018;18(1):176. PMID:30577770
22. O'Donnell J, Smith-Byrne K, Velardo C, Conrad N, Salimi-Khorshidi G, Doherty A, Dwyer T, Tarassenko L, Rahimi K. Self-reported and objectively measured physical activity in people with and without chronic heart failure: UK Biobank analysis. Open Hear 2020;7(1). PMID:32153787
23. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: Machine Learning in {P}ython. J Mach Learn Res 2011;12:2825–2830.
24. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. Proc 22nd acm sigkdd Int Conf Knowl Discov data Min 2016. p. 785–794.
25. Koivula RW, Atabaki-Pasdar N, Giordano GN, White T, Adamski J, Bell JD, Beulens J, Brage S, Brunak S, De Masi F, Dermitzakis ET, Forgie IM, Frost G, Hansen T, Hansen TH, Hattersley A, Kokkola T, Kurbasic A, Laakso M, Mari A, McDonald TJ, Pedersen O, Rutters F, Schwenk JM, Teare HJA, Thomas EL, Vinuela A, Mahajan A, McCarthy MI, Ruetten H, Walker M, Pearson E, Pavo I, Franks PW. The role of physical activity in metabolic homeostasis before and after the onset of type 2 diabetes: an IMI DIRECT study. Diabetologia Diabetologia; 2020;63(4):744–756. PMID:32002573
26. McCombie L, Leslie W, Taylor R, Kennon B, Sattar N, Lean MEJ. Beating type 2 diabetes into remission. BMJ [Internet] BMJ Publishing Group Ltd; 2017;358. PMID:28903916
27. Taylor R, Al-Mrabeh A, Sattar N. Understanding the mechanisms of reversal of type 2 diabetes. lancet Diabetes Endocrinol [Internet] England; 2019 Sep;7(9):726–736. PMID:31097391
28. Xin Y, Davies A, Briggs A, McCombie L, Messow CM, Grieve E, Leslie WS, Taylor R, Lean MEJ. Type 2 diabetes remission: 2 year within-trial and lifetime-horizon cost-effectiveness of the Diabetes Remission Clinical Trial (DiRECT)/Counterweight-Plus weight management programme. Diabetologia 2020 Oct;63(10):2112–2122. PMID:32776237
29. Chatterjee S, Khunti K, Davies MJ. Type 2 diabetes. Lancet (London, England) England; 2017 Jun;389(10085):2239–2251. PMID:28190580
30. Barber SR, Dhalwani NN, Davies MJ, Khunti K, Gray LJ. External national validation of the Leicester Self-Assessment score for Type 2 diabetes using data from the English Longitudinal Study of Ageing. Diabet Med England; 2017 Nov;34(11):1575–1583. PMID:28744894
31. Staite E, Bayley A, Al-Ozairi E, Stewart K, Hopkins D, Rundle J, Basudev N, Mohamedali Z, Ismail K. A wearable technology delivering a web-based diabetes prevention program to people at high risk of type 2 diabetes: Randomized controlled trial. JMIR mHealth uHealth 2020;8(7):1–14. PMID:32459651
32. Whelan ME, Orme MW, Kingsnorth AP, Sherar LB, Denton FL, Esliger DW. Examining the use of glucose and physical activity self-monitoring technologies in individuals at moderate to high risk of developing type 2 diabetes: Randomized trial. JMIR mHealth uHealth 2019;7(10):1–16. PMID:31661077
33. Kondama Reddy R, Pooni R, Zaharieva DP, Senf B, El Youssef J, Dassau E, Doyle FJ, Clements MA, Rickels MR, Patton SR, Castle JR, Riddell MC, Jacobs PG. Accuracy of wrist-worn activity monitors during common daily physical activities and types of structured exercise: Evaluation study. JMIR mHealth uHealth 2018;6(12). PMID:30530451

Abbreviations:
T2D: Type-2 Diabetes
EHR: Electronic Health Record
AUC: Area Under Curve