Longitudinal cardio-respiratory fitness prediction through free-living wearable sensors

Dimitris Spathis¹,*, Ignacio Perez-Pozuelo²+, Tomas I. Gonzales², Soren Brage², Nicholas Wareham², and Cecilia Mascolo¹

¹Department of Computer Science and Technology, University of Cambridge, UK
²MRC Epidemiology Unit, School of Clinical Medicine, University of Cambridge, UK
*corresponding author: ds806@cam.ac.uk
+these authors contributed equally to this work

SUMMARY

Background Cardiorespiratory fitness is an established predictor of metabolic disease and mortality. Fitness is directly measured as maximal oxygen consumption (VO₂max), or indirectly assessed using heart rate response to a standard exercise test. However, such testing is costly and burdensome, limiting its utility and scalability. Fitness can also be approximated using resting heart rate and self-reported exercise habits but with lower accuracy. Modern wearables capture dynamic heart rate data which, in combination with machine learning models that find latent patterns in high-dimensional sensor signals, could improve fitness prediction.

Methods In this work, we analyze movement and heart rate signals from wearable sensors in free-living conditions from 11,059 participants who also underwent a standard exercise test, along with a longitudinal repeat cohort of 2,675 participants. We design algorithms and models that convert raw sensor data into cardio-respiratory fitness estimates, and validate these estimates’ ability to capture fitness profiles in a longitudinal cohort over time while subjects engaged in real-world (non-exercise) behavior. Additionally, we validate our methods with a third external cohort of 181 participants who underwent maximal VO₂max testing, which is considered the gold standard measurement because it requires reaching one’s maximum heart rate and exhaustion level.

Findings Our results show that the developed models yield high correlation (r = 0.82, 95CI 0.80-0.83), when compared to the ground truth in a holdout sample. These models outperform conventional non-exercise fitness models and traditional bio-markers using measurements of normal daily living without the need of a specific exercise test. Additionally, we show the adaptability and applicability of this approach for detecting fitness change over time in the longitudinal subsample who repeated measurements after 7 years.

Interpretation These results demonstrate the value of wearables for the longitudinal assessment of fitness that today can be measured only with laboratory tests.

Funding Partly funded by the University of Cambridge, Jesus College Cambridge, EPSRC, and GSK i-CASE scholarships.

Introduction

Cardiorespiratory fitness (CRF) is one of the strongest known predictors of cardiovascular disease (CVD) risk and is inversely associated with many other health outcomes. CRF is also a potentially stronger predictor of CVD outcomes when compared to other risk factors like hypertension, type 2 diabetes, high cholesterol, and smoking. Despite its prognostic value, routine CRF assessment remains uncommon in clinical settings because maximal oxygen consumption (VO₂max), the gold-standard measure of CRF, is challenging to directly measure. A computerised gas analysis system is needed to monitor ventilation and expired gas fractions during exhaustive exercise on a treadmill or cycle ergometer. Additional equipment may be needed to monitor other biosignals, such as heart rate (HR). These equipment require trained research personnel to operate, and an attending physician is a requisite for exercise testing in some scenarios. Several criteria must also be achieved to verify that exhaustion has been reached, including leveling-off of VO₂, achieving a percentage of age-predicted maximal HR, and surpassing a peak respiratory exchange ratio threshold. The costs of VO₂ measurement and risks of exhaustive exercise not only limit direct CRF assessment in clinical settings, but also restrict research of CRF at the population level. Thus, our understanding of differences in CRF within populations, across geographic regions, and over time is lacking.

Non-exercise prediction models of VO₂max are an alternative to exercise testing in clinical settings. These models are usually regression based and incorporate variables like sex, age, body mass index (BMI), resting heart rate (RHR), and self-reported physical activity. We have recently shown that RHR alone can be used to estimate VO₂max, however the validity of estimates from this approach are considerably lower than those achieved with exercise testing. Wearable devices such as activity trackers and smartwatches can monitor not only RHR and physical activity but other biosignals in free-living conditions, potentially enabling more precise estimation of VO₂max without exercise testing. Recent attempts to use wearable devices to estimate VO₂max are difficult to
Research in Context

Evidence before this study
Large-scale assessment of physical activity and heart rate data could enable training robust models that anticipate and predict fine-grained changes of fitness over time. Current fitness estimation models, however, have limitations. We searched the PubMed database for publications up to April 27 2022, for studies about VO2max estimation through wearables, using the search criteria "Vo2max AND wearable" and "Vo2max AND machine learning". We should note that the intersection of the three terms "Vo2max AND machine learning AND wearable" yielded only one result. We systematically screened all 34 results (21, 12, 1 from each respective search) by reviewing abstracts, and identified five original research studies which were relevant to the study aims. These studies were either small-scale assessments of VO2max or required data collected from non-free-living environments. We discuss these studies in further detail in the respective sections of this paper.

Added value of this study
In this study, deep learning models were developed for remote and scalable estimation of cardio-respiratory fitness through wearable devices. Our study expands on previous studies by being the first large-scale study integrating a large cohort, a longitudinal validation cohort, and a third gold standard cohort. The deep learning model takes statistical features extracted from the raw sensor timeseries covering movement and heart rate attributes, and returns fine-grained predictions of VO2max, either in the present or the future. We independently validated the models on several holdout test sets (different participants from the set used when training the AI model). Our results show that using passive sensor data from everyday life can improve fitness prediction significantly, opening the door to population-scale assessment of fitness, moving beyond traditional predictors of individuals’ health such as weight or BMI.

Implications of all the available evidence
The performance of our deep learning-based sensor analysis model suggests that it can potentially be used for fitness estimation with similar accuracy to a laboratory-measured test. In the future, wearables could include additional sensors which could further improve fitness prediction, such as blood pressure, skin conductance, and oximetry. Future prospective studies will help to identify the impact of personalized fitness predictions by suggesting interventions on individuals through lifestyle, nutrition, and exercise recommendations.

Evidence before this study
Large-scale assessment of physical activity and heart rate data could enable training robust models that anticipate and predict fine-grained changes of fitness over time. Current fitness estimation models, however, have limitations. We searched the PubMed database for publications up to April 27 2022, for studies about VO2max estimation through wearables, using the search criteria "Vo2max AND wearable" and "Vo2max AND machine learning". We should note that the intersection of the three terms "Vo2max AND machine learning AND wearable" yielded only one result. We systematically screened all 34 results (21, 12, 1 from each respective search) by reviewing abstracts, and identified five original research studies which were relevant to the study aims. These studies were either small-scale assessments of VO2max or required data collected from non-free-living environments. We discuss these studies in further detail in the respective sections of this paper.

Added value of this study
In this study, deep learning models were developed for remote and scalable estimation of cardio-respiratory fitness through wearable devices. Our study expands on previous studies by being the first large-scale study integrating a large cohort, a longitudinal validation cohort, and a third gold standard cohort. The deep learning model takes statistical features extracted from the raw sensor timeseries covering movement and heart rate attributes, and returns fine-grained predictions of VO2max, either in the present or the future. We independently validated the models on several holdout test sets (different participants from the set used when training the AI model). Our results show that using passive sensor data from everyday life can improve fitness prediction significantly, opening the door to population-scale assessment of fitness, moving beyond traditional predictors of individuals’ health such as weight or BMI.

Implications of all the available evidence
The performance of our deep learning-based sensor analysis model suggests that it can potentially be used for fitness estimation with similar accuracy to a laboratory-measured test. In the future, wearables could include additional sensors which could further improve fitness prediction, such as blood pressure, skin conductance, and oximetry. Future prospective studies will help to identify the impact of personalized fitness predictions by suggesting interventions on individuals through lifestyle, nutrition, and exercise recommendations.

Results
Baseline measurements were collected from 12,435 healthy adults from the Fenland study in the United Kingdom, where all required data for the present analysis were available in 11,059 participants (Fenland I, baseline timepoint referred to as "current" in our evaluation). A subset of 2,675 participants were assessed again after a median (interquantile range) of 7 (5-8) years (Fenland II, referred to as "future" in our evaluation). Descriptive characteristics of the two analysis samples are presented in Table 1. We present the characteristics of the longitudinal cohort in both temporal snapshots in in Table 1 ("present" and "future"). Mean and standard deviations for each characteristic are presented in this table. An overview of the study design and the three experimental tasks is provided in Figure 1.
STUDY DESIGN

Fenland I
(N = 11,059)

Fenland II
(N = 2,675)

VO\textsubscript{2max} Test

BASELINE VISIT (DAY 1)

FREE-LIVING PHASE (DAY 1-7)

Anthropometrics

Questionnaires

Movement and heart data from wearables

~ 6 days

MODELING FITNESS

FEATURES

USING

DEEP NEURAL NETWORK

1. Fine-grained cardiorespiratory fitness prediction in the present
2. Inferring fitness direction and magnitude 7 years in the future
3. Validating adaptability and change with new sensor data

Figure 1. Study and experimental design. Fenland is a cohort study, including 11,059 participants with treadmill and wearable sensor data at baseline (Fenland I, 2005-2015) and a longitudinal subsample of 2,675 participants who were retested approximately 7 years later (Fenland II). During both phases participants underwent a variety of tests during the baseline clinic visit comprising anthropometric measurements, questionnaires, and a submaximal VO\textsubscript{2max} test. Following this baseline clinic visit, participants were fitted with a combined heart rate and movement sensing device which they wore during free-living conditions for approximately 6 days. In this work we develop deep learning models utilising wearable data and common biomarkers as input to predict VO\textsubscript{2max} and achieve strong performance, both in the present and when replicated at the second timepoint.

Table 1. Characteristics for the study analytical sample: The Fenland I and II studies. Data is in mean (std). Values with asterisk(*) indicate that this variable comes from Fenland II sensor data which is a smaller cohort (N=2071) due to data filtering (see Figure 6–Panel 3). The values in FII (future) cohort correspond to the second assessment (7 years later). In Task 1, the training set is FI (present) and the testing set is FII (present) so as to make sure that they come from similar distributions.
Fine-grained fitness prediction from wearable sensors

We first developed and externally validated several non-exercise VO_{2max} estimation models as a regression task using features commonly measured by wearable devices (anthropometry, resting heart rate (RHR), physical activity (PA); see Table 2). Here our goal was to explore how conventional non-exercise approaches to VO_{2max} estimation could be enhanced by features from free-living PA data. We split participant data into independent training and test sets. The training set (n=8384, participants with baseline data only) was used for model development. The test set (n=2675, participants with baseline and followup data) was used to externally validate each model. Models using anthropometry or RHR alone had poor external validity, but validity improved when combined in the same model. The best performance (R^2 of 0.67) was attained using a deep neural network model combining wearable sensors, RHR, and anthropometric data (Figure 2). For reference, we compare these results to traditional non-model equations, which rely on Body Mass, RHR, and Age. Using a popular equation (as proposed in 14,15) shows poor validity (R^2 of -3.2 and Correlation of 0.389), a performance lower than using anthropometrics only in our setup (see Methods for details). This motivates the use of machine learning which captures better covariate interactions.

Deep neural networks can learn feature representations that are suitable for clustering tasks, such as population stratification by implicit health status, but are difficult to reveal using linear dimension-reduction techniques. We used t-distributed stochastic neighbor embedding (tSNE), a non-linear dimension-reduction technique, to visualise learned feature representations from our model and their relationship to participant VO_{2max} and HR levels (Figure 7). Clustering and coloring by VO_{2max} and HR levels was shown to be inversely related and more apparent in the learned latent space compared to the original observation space. For example, participants with higher VO_{2max} were clustered similarly to those with lower HR levels, and vice versa.

Predicting magnitude and direction of fitness change in the future

The second group of tasks evaluated our model on the subset of participants who returned for Fenland II \(\approx 7\) years later (referred to as future in our evaluations). For these experiments we carried out three evaluations. Following the process described earlier, we re-trained a model to predict future VO_{2max} using only information from the present as input (Table 3). This model yielded a slightly lower accuracy than Fenland I, achieving a R^2 of 0.49 and a correlation of 0.72. This lower performance is expected since the model has no indication of the behavior of the individuals 7 years later. We also trained a model to directly predict the difference (or delta) of current-future VO_{2max}, which reached a correlation of 0.23.

Further, motivated by the moderate predictability of the fine-grained delta of VO_{2max}, we formulated this problem as a classification task. A visual representation of this task can be found in Figure 3a. By inspecting the distribution of the difference (delta) of current-future VO_{2max} on the training set, we split it to 2 halves (50% quantiles) and set these as prediction outcomes. The purpose of this task is to assess the direction of individual change of fitness. We report an area under the curve (AUC) of 0.61 in predicting the direction of change (\(N = 2675\)). We also focused on the tails of the change distribution which indicates participants who had substantial and dramatic change in fitness over the period of time between Fenland I and Fenland II (\(\approx 7\) years). In this case, the distribution was split into 80%/20% (substantial) and 90%/10% (dramatic) quantiles. The results from these experiments show that the models can distinguish between substantial fitness change with an AUC of 0.72 (\(N = 1068\)) and between dramatic fitness change with an AUC of 0.74 (\(N = 535\)). All AUC curves can be found in Figure 3b.

Enabling adaptive cardiorespiratory fitness inferences

For the final task, we assessed whether the trained models can pick up change using new sensor data from Fenland II. Considering that getting new wearable data is relatively easy since these devices are becoming increasingly pervasive. The intuition behind this task is to evaluate the generalizability of the models over time. We first matched the populations that provided sensor data for both cohorts (\(N = 2,042\)) and applied the trained model from Task 1 in order to produce VO_{2max} inferences. We then compared the predictions with the respective ground truth (current and future VO_{2max}). The true and predictive distributions are shown in Figures 4c and 4d. Through this procedure, we found that the model achieves an \(r = 0.84\) for VO_{2max} future prediction and an \(r = 0.82\) for VO_{2max} current prediction (validating our Task 1 results). In other words, if we have access to wearable sensor data and other information from the future time, we can use the already trained model from Fenland I to accurately infer fitness with minimal loss of accuracy over time, even though this is new sensor data from a completely separate (future) week.

Last, we calculated the delta of the predictions and compared it to the actual delta of fitness over the years. This task showed that the models tend to focus mostly on positive change and under-predict when participants’ fitness deteriorates over the years (Figures 4a, 4b). The overall correlation between the delta of the predictions with the ground truth is significant (\(r=0.57,\ p<0.005\)).

Discussion

Cardiorespiratory fitness declines with age independently of changes to body composition, and low cardiorespiratory fitness is associated with poor health outcomes1,17–19. As such, having the capacity to predict whether CRF would decline in excess of natural aging could be valuable to clinicians when tailoring therapeutic interventions. Here we have developed...
Figure 2. Fine-grained fitness prediction. Comparing the predicted and true VO2max coming from the best performing comprehensive model (Sensors + RHR + Anthro.) trained with Fenland I. (a) Distribution of predicted and true VO2max. The plot combines a kernel density estimate and histogram. (b) Correlation of predicted and true VO2max ($r = 0.82$, $p < 0.005$, see Table 2). The gray line denotes a linear regression fit. Transparency has been applied to the datapoints to combat crowding.

Table 2. Evaluation of predicting fine-grained VO2max with the Fenland I cohort. Comparison between traditional anthropometrics, common biomarkers (RHR), and passively collected data over a week (wearable sensors). Best performance in bold. The units of VO2max are measured in $\text{ml}O_2/\text{min}/\text{kg}$. Results reported from the testing set.

| Data modality                                      | Evaluation Metrics [95% CI] | N (train+val / test set) |
|---------------------------------------------------|-----------------------------|--------------------------|
|                                                   | $R^2$ Corr RMSE              |                          |
| **Anthropometrics**                               | 0.362 [0.332-0.391] 0.604 [0.579-0.627] 4.043 [3.924-4.172] |                          |
| Age/Sex/Weight/BMI/Height                         | 0.374 [0.344-0.403] 0.615 [0.589-0.639] 4.007 [3.891-4.117] |                          |
| **Resting Heart Rate**                            | 0.616 [0.588-0.641] 0.785 [0.767-0.802] 3.138 [3.031-3.237] | 11059                    |
| RHR (Sensor-derived)                              |                            |                          |
| **Anthropometrics + RHR**                          | 0.671 [0.649-0.692] 0.822 [0.808-0.835] 2.903 [2.801-3.003] | 8384/2675                |
| Age/Sex/Weight/BMI/Height/RHR                     |                            |                          |
| **Wearable Sensors + RHR + Anthro.**              |                            |                          |
| Acceleration/HR/HRV/MVPA                          |                            |                          |
| Age/Sex/Weight/BMI/Height/RHR                     |                            |                          |
a deep learning framework for predicting CRF and changes in CRF over time. Our framework estimates VO₂max by combining learned features from heart rate and accelerometer free-living data extracted from wearable sensors with anthropometric measures. To evaluate our framework’s performance, VO₂max estimates were compared with VO₂max values derived from a submaximal exercise test. Free-living and exercise test data were collected at a baseline investigation in 11,059 participants (Fenland I). A subset of those participants (n=2,675) completed another exercise test at a follow-up investigation approximately seven years later (Fenland II). This study design allowed us to address three questions: 1) Do baseline estimates of VO₂max from the deep learning framework agree with VO₂max values measured from exercise testing at baseline?, 2) Can the framework learn features from heart rate and accelerometer free-living data collected at baseline that predict VO₂max measured at follow-up?, and 3) Can the framework be used to predict the magnitude of change in VO₂max from baseline to follow-up?

In the VO₂max estimation tasks, our model demonstrated strong agreement with VO₂max measured from the submaximal exercise test at baseline (Pearson’s correlation coefficient (PCC): 0.82) as well as for the longitudinal, follow-up visit (PCC: 0.72). We were also able to distinguish between substantial and dramatic changes in CRF (AUCs 0.72 and 0.74, respectively). Finally, we further evaluated the initial model on new input data by feeding Fenland II free-living data along with updated heart rate and anthropometrics to the model, showing that it is able to adapt and monitor change over time. We evaluated the inference capabilities of the model in the difference (delta) between the current (Fenland I) and future (Fenland II) VO₂max for those participants that came back approximately 7 years later. For this last task, the model produced outcomes that translated to a 0.57 correlation between the delta of predicted and delta of true VO₂max.

The application of our work to other cohort and longitudinal studies is of particular importance as serial measurement of cardiorespiratory fitness has significant prognostic value in clinical practice. Small increases in fitness are associated with reduced cardiovascular disease mortality risk and better clinical outcomes in patients with heart failure and type 2 diabetes. Nevertheless, routine measurement of fitness in clinical practice is rare due to the costs and risks of exercise testing. Non-exercise based regression models can be used to estimate changes in fitness in lieu of serial exercise testing. It is unclear, however, the extent to which changes in fitness detected with such models reflect true changes in exercise capacity. Here, we relied on the relationship between CRF and heart rate responses to different levels of physical activity at submaximal, real-life conditions captured through wearable sensors. Using deep learning techniques, we have developed a non-exercise based fitness estimation approach that can be used not only to accurately infer current VO₂max, but also to do so in the settings of a future cohort, where the model did not require any retraining, just influx of new data. Further, we show that the model can also be used to infer the changes in CRF that occurred during the ≈ 7 year time span between Fenland I and II.

Our proposed deep learning approach outperforms traditional non-exercise models, which are the state-of-the-art in the field and rely on simple variables inputted to a linear model. Importantly, our model is able to take week-level in-
In this paper, we developed deep learning models utilising wearable data and other bio-markers to predict the gold standard of fitness (VO$_{2max}$) and achieved strong performance compared to other traditional approaches. Cardio-respiratory fitness is a well-established predictor of metabolic disease and mortality and our premise is that modern wearables capture non-standardised dynamic data which could improve fitness prediction. Our findings on a population of 11,059 participants showed that the combination of all modalities reached an $r^2 = 0.82$, when compared to the ground truth in a holdout sample. Additionally, we show the adaptability and applicability of this approach for detecting fitness change over time in a longitudinal subsample ($n = 2,675$) who repeated measurements after 7 years. Last, the latent representations that arise from this model pave the way for fitness-aware monitoring and interventions at scale. It is often said that "If you cannot measure it, you cannot improve it". Cardio-fitness is such an important health marker, but until now we did not have the means to measure it at scale. Our findings could have significant implications for population health policies, finally moving beyond weaker health proxies such as the BMI.

**Methods**

**Study description**

The Fenland study is a population-based cohort study designed to investigate the independent and interacting effects of environmental, lifestyle and genetic influences on the development of obesity, type 2 diabetes and related metabolic disorders. Exclusion criteria included prevalent diabetes, pregnancy or lactation, inability to walk unaided, psychosis or terminal illness (life expectancy of $\leq 1$ year at the time of recruitment).

The Fenland study has two distinct phases. Phase I, during which baseline data was collected from 12,435 participants, took place between 2005 and 2015. Phase II was launched in 2014 and involved repeating the measurements collected during Phase I, alongside the collection of new measures. All participants who had consented to being re-contacted after their Phase I assessment were invited to participate in Phase II. At least 4 years must have elapsed between visits. As a result of this stipulation, recruitment to Phase II is ongoing. A flowchart of the analytical sample by each one of the study tasks is provided in Figure 6.

After a baseline clinic visit, participants were asked to wear a combined heart rate and movement chest sensor Actiheart, CamNtech, Cambridgeshire, UK) for 6 complete days. For this study, data from 11,059 participants was included after excluding participants with insufficient or corrupt data or missing covariates as shown in Figure 6. A subset of 2,675 of the study participants returned for the second phase of the study, after a median (interquartile range) of 7 (5–8) years, and underwent a similar set of tests and protocols, including wearing the combined heart rate and movement sensing for 6 days. All participants provided written informed consent and the study was approved by the University of Cambridge, NRES
Committee - East of England Cambridge Central committee. All experiments and data collected were done in accordance with the declaration of Helsinki.

**Study procedure**

Participants wore the Actiheart wearable ECG which measured heart rate and movement recording at 60-second intervals. The Actiheart device was attached to the chest at the base of the sternum by two standard ECG electrodes. Participants were told to wear the monitor continuously for 6 complete days and were advised that these were waterproof and could be worn during showering, sleeping or exercising. During a lab visit, all participants performed a treadmill test that was used to establish their individual response to a submaximal test, informing their VO\textsubscript{2max} (maximum rate of oxygen consumption measured during incremental exercise). RHR was measured with the participant in a supine position using the Actiheart device. HR was recorded for 15 minutes and RHR was calculated as the mean heart rate measured during the last 3 minutes. Our RHR is a combination of the Sleeping HR measured by the ECG over the free-living phase and the RHR as described above.

**Cardiorespiratory fitness assessment**

VO\textsubscript{2max} was predicted in study participants using a previously validated submaximal treadmill test. Participants exercised while treadmill grade and speed were progressively increased across several stages of level walking, inclined walking, and level running. The test was terminated if one of the following criteria were met: 1) the participant wanted to stop, 2) the participant reached 90% of age-predicted maximal heart rate (208-0.7*age), 3) the participants exercised at or above 80% of age-predicted maximal heart rate for 2 minutes. Details about the fitness characteristics of the cohort and the validation of the submaximal test are provided elsewhere.

To further validate the models trained on submaximal VO\textsubscript{2max}, we employ the external cohort UK Biobank Validation Study (BBVS). We recruited 105 female (mean age: 54.3y±7.3) and 86 male (mean age: 55.0y±6.5) validation study participants and VO\textsubscript{2max} was directly measured during an independent maximal exercise test, which was completed to exhaustion. Some maximal exercise test data were excluded because certain direct measurements were anomalous due to testing conditions (N=10). BBVS participants completed the same free-living protocol as in Fenland and we collected similar sensor and anthropometrics data which were processed with the same way as in Fenland (see next section).

**Free-living wearable sensor data processing**

Participants were excluded from this analysis if they had less than 72 hours of concurrent wear data (three full days of recording) or insufficient individual calibration data (treadmill test-based data). All heart rate data collected during free-living conditions underwent pre-processing for noise filtering. Non-wear detection procedures were applied and any of those non-wear periods were excluded from the analyses. This algorithm detected extended periods of non-physiological heart rate concomitantly with extended (>90 minutes) periods that also registered no movement through the device’s accelerometer. We converted movement these intensities into standard metabolic equivalent units (METs), through the conversion 1 MET = 71 J/min/kg (3.5 ml O\textsubscript{2}·min\textsuperscript{-1}·kg\textsuperscript{-1}). These conversions where then used to determine intensity levels with ≤1.5METs classified as sedentary behaviour, activities between 3 and 6 METs were classified as moderate to vigorous physical activity (MVPA) and those >6METs were classified as vigorous physical activity (VPA).

Since the season can have a big impact on physical activity considering on how it affects workouts, sleeping patterns, and commuting patterns, we encoded the sensor timestamps using cyclical temporal features \(T_f\). Here we encoded the month of the year as \((x,y)\) coordinates on a circle:

\[
T_{f1} = \sin\left(\frac{2 \pi f t}{\text{max}(t)}\right) \quad (1) \quad T_{f2} = \cos\left(\frac{2 \pi f t}{\text{max}(t)}\right) \quad (2)
\]

where \(t\) is the relevant temporal feature (month). The intuition behind this encoding is that the model will "see" that e.g. December (12th) and January (1st) are 1 month apart (not 11). Considering that the month might change over the course of the week, we use the month of the first time-step only. Additionally, we extracted summary statistics from the following sensor time-series: raw acceleration, HR, HRV, Aceleration-derived Euclidean Norm Minus One, and Acceleration-derived Metabolic Equivalents of Task. Then, for every time-series we extracted the following variables which cover a diverse set of attributes of their distributions: mean, minimum, maximum, standard deviation, percentiles (25%, 50%, 75%), and the slope of a linear regression fit. The rest of variables (anthropometrics and RHR) are used as a single measurement.

In total, we derived a comprehensive set of 68 features using the Python libraries Pandas and Numpy. A detailed view of the variables is provided in Table 4.

**Deep learning models**

We developed deep neural network models that are able to capture non-linear relationships between the input data and the respective outcomes. Considering the high-sampling rate of the sensors (1 sample/minute) after aligning HR and Acceleration modalities, it is impossible to learn patterns with such long dependencies (a week of sensor data includes more than 10,000 timestamps). Even the most well-tuned recurrent neural networks cannot cope with such sequences and given the size of the training set (7,545 samples), the best option was to extract statistical features from the sensors and represent every participant-week as a row in a feature vector (see Fig. 1). This feature vector was fed to fully connected neural network layers which were trained with backpropagation.

**Data preparation.** For Task 1 (see Figure 6), we matched the sensor data with the participants who had eligible lab tests. Then we split into disjoint train and test sets, making sure that
Figure 4. Assessing the robustness of the model to pick up change using new sensor data from Fenland II repeats. By matching the populations who provided sensor data for both cohorts (N=2,042) we passed them through our trained model from Task 1 to predict VO2max. (a-b) We then calculated the difference (Δ) of the predictions juxtaposed with the true difference of fitness over the years. Distribution of Δ of predicted and true VO2max. Correlation of Δ of predicted and true VO2max (r = 0.57, p < 0.005). The gray line denotes a linear regression fit. Transparency has been applied to the datapoints to combat crowding. (c-d) Comparison of predicted and true VO2max using FI and FII covariates (sensors, RHR, anthro.), respectively. The distribution plots combine a kernel density estimate and histogram with bin size determined automatically with a reference rule.
participants from Fenland I go to the train set, while those from Fenland II go to the test set (see Figure 5). This would allow to re-use the trained model from Task 1, with different sensor data from Fenland II participants. Intuitively, we train a model on the big population, and we evaluate it with two snapshots of another longitudinal population over time (Task 1 & 3). After splitting, we normalize the training data by applying standard scaling (removing the mean and scaling to unit variance) and then denoise it by applying Principal Components Analysis (PCA), retaining the components that explain 99.99% of the variance. In practice, the original 68 features are reduced to 48. We save the fitted RCA projection and scaler and we apply them individually to the test-set, to avoid information leakage across the sets. The same projection and scaler are applied to all downstream models (Task 2 and 3) to leverage the knowledge of the big cohort (Fenland I).

Model architecture and training. The main neural network (used in Task 1) receives a 2D vector of [users, features] and predicts a real value. For this work, we assume $N$ users and $F$ features of an input vector $\mathbf{x} = (x_1, ..., x_N) \in \mathbb{R}^{N \times F}$ and a target $\mathbf{y} = (y_1, ..., y_N) \in \mathbb{R}^N$. The network consists of two densely-connected feed forward layers with 128 units each. A dense layer works as follows: $output = activation(input \cdot kernel + bias)$, where activation is the element-wise activation function (the exponential linear unit in our case), kernel is a learned weights matrix with a Glorot uniform initialization, and bias is a learned bias vector. Each layer is followed by a batch normalization operation, which maintains the mean output close to 0 and the output standard deviation close to 1. Also, dropout of 0.3 probability is applied to every layer, which randomly sets input units to 0 and helps prevent overfitting. Last, the final layer is a single-unit dense layer and the network is trained with the Adam optimizer to minimize the Mean Squared Error (MSE) loss, which is appropriate for continuous values, we computed the root mean squared error (RMSE) as the waist circumference, which however was not recorded in our cohorts.

Evaluation
To evaluate the performance of the deep learning models which predict continuous values, we computed the root mean squared error (RMSE) as the waist circumference, which however was not recorded in our cohorts.

Statistical analyses
We performed a number of sensitivity analyses to investigate potential sources of bias in our results. Full results of these sensitivity analyses are shown in the main text and corresponding Tables. In particular, we use bootstrapping with replacement (1000 samples) to calculate 95% Confidence Intervals when we report the performance of the models in the hold-out sets. Wherever we report p-values, we use the recently proposed strict threshold of $p < 0.005$.

Data availability
All Fenland data used in our analyses is available from the MRC Epidemiology Unit at the University of Cambridge upon reasonable request (http://www.mrc-epid.cam.ac.uk/research/studies/fenland/)
Code availability

The source code of this work will be available on Github upon acceptance.

Acknowledgements

DS is supported by the Embiricos Trust Scholarship of Jesus College Cambridge and the Engineering and Physical Sciences Research Council through grant DTP (EP/N509620/1). IP-P is supported by GlaxoSmithKline and Engineering and Physical Sciences Research Council through an iCase fellowship (17100053).

Author contributions statement

DS, IP, SB, NW, CM designed the study. SB, NW collected the data on behalf of the Fenland Study. DS, IP sourced, selected, and pre-processed the data. DS developed the models and produced the figures. DS, IP performed the statistical analysis. IP, DS wrote the first draft of the manuscript. All authors (DS, IP, TG, SB, NW, CM) critically reviewed, contributed to the preparation of the manuscript, and approved the final version. All authors vouch for the data, analyses, and interpretations.

Competing interests

The authors declare that they have no competing financial interests.

References

1. Mandsager, K. et al. Association of cardiorespiratory fitness with long-term mortality among adults undergoing exercise treadmill testing. *JAMA network open* 1, e183605–e183605 (2018).
2. Swain, D. P., Brawner, C. A., of Sports Medicine, A. C. et al. ACSM’s resource manual for guidelines for exercise testing and prescription (Wolters Kluwer Health/Lippincott Williams & Wilkins, 2014).
3. Cao, Z.-B. et al. Predicting v’o2max with an objectively measured physical activity in Japanese women. *Medicine & Sci. Sports & Exerc.* 42, 179–186 (2010).
4. Gonzales, T. I. et al. Resting heart rate as a biomarker for tracking change in cardiorespiratory fitness of uk adults: the fenland study. *medRxiv* (2020).
5. Gonzales, T. I. et al. Estimating maximal oxygen consumption from heart rate response to submaximal ramped treadmill test. *medRxiv* (2020).
6. Nes, B. M. et al. Estimating v’o2peak from a nonexercise prediction model: the hunt study, norway. *Medicine & Sci. Sports & Exerc.* 43, 2024–2030 (2011).
7. Shcherbina, A. et al. Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort. *J. personalized medicine* 7, 3 (2017).
8. Passler, S., Bohrer, J., Blöchinger, L. & Senner, V. Validity of wrist-worn activity trackers for estimating vo2max and energy expenditure. *Int. journal environmental research public health* 16, 3037 (2019).
9. Cooper, K. D. & Shafer, A. B. Validity and reliability of the polar a300’s fitness test feature to predict vo2max. *Int. journal exercise science* 12, 393 (2019).
10. Altini, M., Casale, P., Penders, J. & Amft, O. Cardiorespiratory fitness estimation in free-living using wearable sensors. *Artif. intelligence medicine* 68, 37–46 (2016).
11. Richardson, K., Levett, D., Jack, S. & Grocott, M. Fit for surgery? perspectives on preoperative exercise testing and training. *BJA: Br. J. Anaesth.* 119, i34–i43 (2017).
12. Leon-Ferre, R., Ruddy, K. J., Staff, N. P. & Loprinzi, C. L. Fit for chemo: Nerves may thank you. *JNCI: J. Natl. Cancer Inst.* 109 (2017).
13. Lindsay, T. et al. Descriptive epidemiology of physical activity energy expenditure in uk adults (the fenland study). *Int. J. Behav. Nutr. Phys. Activity* 16, 1–13 (2019).
14. Uth, N., Sørensen, H., Overgaard, K. & Pedersen, P. K. Estimation of vo2max from the ratio between hrmax and hrrest—the heart rate ratio method. *Eur. journal applied physiology* 91, 111–115 (2004).
15. Tanaka, H., Monahan, K. D. & Seals, D. R. Age-predicted maximal heart rate revisited. *J. american college cardiology* 37, 153–156 (2001).
16. Gaspar, H. A. & Breen, G. Probabilistic ancestry maps: a method to assess and visualize population substructures in genetics. *BMC bioinformatics* 20, 1–11 (2019).
17. Schmid, D. & Leitzmann, M. Cardiorespiratory fitness as predictor of cancer mortality: a systematic review and meta-analysis. *Annals oncology* 26, 272–278 (2015).
18. Schuch, F. B. et al. Are lower levels of cardiorespiratory fitness associated with incident depression? a systematic review of prospective cohort studies. *Prev. Medicine* 93, 159–165 (2016).
19. Laukkonen, J. A., Kurl, S., Salonen, R., Rauramaa, R. & Salonen, J. T. The predictive value of cardiorespiratory fitness for cardiovascular events in men with various risk profiles: a prospective population-based cohort study. *Eur. heart journal* 25, 1428–1437 (2004).
20. Jakicic, J. M. et al. Four-year change in cardiorespiratory fitness and influence on glycemic control in adults with type 2 diabetes in a randomized trial: the look ahead trial. *Diabetes care* 36, 1297–1303 (2013).
21. Qui, S., Cai, X., Sun, Z., Wu, T. & Schumann, U. Is estimated cardiorespiratory fitness an effective predictor for cardiovascular and all-cause mortality? a meta-analysis. *Atherosclerosis* 330, 22–28 (2021).
22. Apple. Using apple watch to estimate cardio fitness with vo2 max (2021). [Online; posted 2021, note =
23. Lipton, Z., Wang, Y.-X. & Smola, A. Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*, 3122–3130 (PMLR, 2018).

24. Brage, S., Brage, N., Franks, P. W., Ekelund, U. & Wareham, N. J. Reliability and validity of the combined heart rate and movement sensor actiheart. *Eur. journal clinical nutrition* **59**, 561–570 (2005).

25. Brage, S. *et al.* Hierarchy of individual calibration levels for heart rate and accelerometry to measure physical activity. *J. Appl. Physiol.* **103**, 682–692 (2007).

26. Gonzales, T. I. *et al.* Descriptive epidemiology of cardiorespiratory fitness in uk adults: The fenland study. *medRxiv* (2022).

27. Yosinski, J., Clune, J., Nguyen, A., Fuchs, T. & Lipson, H. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579* (2015).

28. Van der Maaten, L. & Hinton, G. Visualizing data using t-sne. *J. machine learning research* **9** (2008).

29. Wattenberg, M., Viégas, F. & Johnson, I. How to use t-sne effectively. *Distill* **1**, e2 (2016).

30. Faurholt-Jepsen, M., Brage, S., Kessing, L. V. & Munkholm, K. State-related differences in heart rate variability in bipolar disorder. *J. psychiatric research* **84**, 169–173 (2017).

31. White, T., Westgate, K., Wareham, N. J. & Brage, S. Estimation of physical activity energy expenditure during free-living from wrist accelerometry in uk adults. *PLoS One* **11**, e0167472 (2016).

**Supplementary appendix**
Figure 5. Distribution of VO2max in the training and test sets in Fenland I cohort. Both sets display similar ranges of values, making sure that inferences based on the test set are robust. This plot refers to Task’s 1 train and test sets.

Table 3. Evaluation of predicting fine-grained VO2max in the present and the future with the Fenland II repeats cohort using covariates of Fenland I. Neural network results. (*the Delta outcome is in a different unit and hence a direct comparison with raw VO2max results might not apply)

| Outcomes               | Evaluation Metrics [95% CI] | N (train+val / test set) |
|------------------------|----------------------------|-------------------------|
|                        | R² | Corr | RMSE |                  |
| **Wearable Sensors + RHR + Anthro.** |    |      |      |                  |
| Current VO2max         | 0.652 [0.606-0.695]   | 0.815 [0.783-0.846] | 2.959 [2.742-3.201] | 2675 (2140/535) |
| Future VO2max          | 0.499 [0.431-0.55]    | 0.721 [0.67-0.759]     | 3.673 [3.421-3.916] |
| Delta (Current - Future)* | 0.081 [0.02-0.078] | 0.233 [0.159-0.307] | 3.175 [2.923-3.41] |
Table 4. Description of the features/variables used in our analysis as inputs to the models. The features with asterisks (*) are time-series and therefore we have extracted the following statistical variables: \textit{mean, minimum, maximum, standard deviation, percentiles (25\%, 50\%, 75\%), and the slope of a linear regression fit}. The final set of features is 68.

| Features/Variables | Description |
|--------------------|-------------|
| **Sensors**        |             |
| Acceleration*      | Acceleration measured in mg |
| Heart rate (HR)*   | Mean HR resampled in 15sec intervals, measured in BPM |
| Heart Rate Variability (HRV)* | HRV calculated by differencing the second-shortest and the second-longest inter-beat interval (as seen in\textsuperscript{30}), measured in ms |
| Acceleration-derived Euclidean Norm Minus One (ENMO)* | ENMO-like variable (Acceleration/0.0060321) + 0.057) (as seen in\textsuperscript{31}) |
| Acceleration-derived Metabolic Equivalents of Task (METs)* | If Accelerometer <1, take daily count and average |
| Sedentary*         | If Accelerometer >= 1, take daily count and average |
| Moderate to Vigorous* | If Accelerometer >= 4.15, take daily count and average |
| Vigorous*          |             |
| **Anthropometrics** |             |
| Age                | Age, measured in years |
| Sex                | Sex is binary (female/male) |
| Weight             | Weight, measured in kilograms |
| Height             | Height, measured in meters.centimeters |
| Body Mass Index (BMI) | BMI is calculated by Weight/(Height\(^2\)), measured in kg/m\(^2\) |
| **Resting Heart Rate** |             |
| Wearable-derived RHR | RHR is calculated by averaging the 4th, 5th, and 6th minute of the baseline visit and adding to that the Sleeping Heart Rate that has been inferred by the wearable device\textsuperscript{4} |
| **Seasonality**    |             |
| Month of year      | The month number is used along with a coordinate encoding that allows the models to make sense of their cyclical sequence. |
The first task trains a model to predict fitness using the large cohort (Fenland I). The second task is using the smaller cohort of repeats in Fenland I (called Fenland II) and trains further models to predict fitness now and in the future (and their delta). The third task evaluates the original model trained in Task 1 by feeding new sensor data (Fenland II sensors and anthropometrics) to assess the adaptability of the model to pick up change. (*Training set is 90% of the 80% remaining dataset after splitting to testing set. Validation set is 10% of the training set)
Figure 7. t-distributed stochastic neighbor embedding (tSNE) projection of the original feature vector (Fenland I testing set, Sensors + RHR + Anthro.) compared to the model’s latent space after training. (a-b) The original data presents some clusters but the outcome is not clearly linearly separable. The model activations on the penultimate layer of the neural network capture the continuum of low-high VO$_{2\text{max}}$ both locally and globally. (c-d) A similar assessment to VO$_{2\text{max}}$ is presented by coloring with the mean HR of the week of each participant. In the learned space, participants with low HR (high fitness) are placed in the same clusters as in VO$_{2\text{max}}$, unlike the original space. In all plots a 50% transparency has been applied to combat crowding and the colorbar is centered on the median value to illustrate extreme cases. The VO$_{2\text{max}}$ label is used only for color-coding purposes (the projection is label-agnostic). Every participant is a dot.

Figure 8. External validation of Fenland I model with maximal (peak exercise) test data using the BBVS cohort. (a) Distribution of the predicted vs the true VO2max (RMSE=8.998) using all participants (N=181). (b) Distribution of the predicted vs the true VO2max (RMSE=5.19) by matching BBVS to have similar VO2max (mean±std) to the training set of Fenland I, using a subset of participants (N=82). Please see the main text for interpretation of these results.