Longitudinal cardio-respiratory fitness prediction through wearables in free-living environments

Dimitris Spathis, Ignacio Perez-Pozuelo, Tomas I. Gonzales, Yu Wu, Soren Brage, Nicholas Wareham and Cecilia Mascolo

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 stronger predictor of CVD outcomes when compared to other risk factors like hypertension, diabetes, high cholesterol, and smoking. Despite its prognostic value, routine CRF assessment remains uncommon in clinical settings because it requires specialized equipment such as treadmills and oxygen masks, limiting its utility. Modern wearables capture dynamic real-world data which could improve fitness prediction. In this work, we design algorithms and models that convert raw wearable sensor data into cardiorespiratory fitness estimates. We validate these estimates' ability to capture fitness profiles in free-living conditions using the Fenland Study (N=11,059), along with its longitudinal cohort (N = 2675), and a third external cohort using the UK Biobank Validation Study (N = 181) who underwent maximal VO2max testing, the gold standard measurement of fitness. Our results show that the combination of wearables and other biomarkers as inputs to neural networks yields a strong correlation to ground truth in a holdout sample (r = 0.82, 95CI 0.80–0.83), outperforming other approaches and models and detects fitness change over time (e.g., after 7 years). We also show how the model's latent space can be used for fitness-aware patient subtyping paving the way to scalable interventions and personalized trial recruitment. These results demonstrate the value of wearables for fitness estimation that today can be measured only with laboratory tests.

npj Digital Medicine (2022)5:176; https://doi.org/10.1038/s41746-022-00719-1

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 it requires specialized equipment such as treadmills and oxygen masks, limiting its utility. Modern wearables capture dynamic real-world data which could improve fitness prediction. In this work, we design algorithms and models that convert raw wearable sensor data into cardiorespiratory fitness estimates. We validate these estimates' ability to capture fitness profiles in free-living conditions using the Fenland Study (N=11,059), along with its longitudinal cohort (N = 2675), and a third external cohort using the UK Biobank Validation Study (N = 181) who underwent maximal VO2max testing, the gold standard measurement of fitness. Our results show that the combination of wearables and other biomarkers as inputs to neural networks yields a strong correlation to ground truth in a holdout sample (r = 0.82, 95CI 0.80–0.83), outperforming other approaches and models and detects fitness change over time (e.g., after 7 years). We also show how the model's latent space can be used for fitness-aware patient subtyping paving the way to scalable interventions and personalized trial recruitment. These results demonstrate the value of wearables for fitness estimation that today can be measured only with laboratory tests.

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 externally evaluate, however, because their estimation methods tend to be non-transparent and lack scientific validation. Although certain wearable devices show promise, they tend to rely on detailed physical activity intensity measurements, GPS-based speed monitoring, and require users to reach near-maximal HR values, which limits their use to fitter individuals. Some studies attempt to estimate VO2max from data collected during free-living conditions, but these are typically from small-scale cohorts and use contextual data from treadmill activity, which restricts their application in population settings.

Here, we use data from the largest study of its kind, by over two orders of magnitude, and use purely free-living data to predict VO2max, with no requirement for context-awareness. This work substantially advances previous non-exercise models for predicting CRF by introducing an adaptive representation learning approach to physiological signals derived from wearable sensors in a large-scale population with free-living condition data. We employ a deep neural network model that utilizes feedforward non-linear layers to learn personalized fitness representations. We demonstrate that these models yield better performance than traditional and state-of-the-art non-exercise models. We illustrate how these models can be used to predict the magnitude and direction of change in CRF. Moreover, we show that they can adapt in time given behavioural changes by showcasing strong performance in a subset of the same
population who were retested seven years later. This has implications for the estimation of population fitness levels and lifestyle trends, including the identification of sub-populations or areas in particular need of intervention. Such models can improve both population health and personalized medicine applications. For instance, the ability of the patient to undergo certain treatments like surgery or chemotherapy can be assessed through wearables before the procedure takes place and therefore reduce postoperative complications.

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 2675 participants was assessed again after a median (interquantile range) of 7 (5–8) years (Fenland II, referred to as “future” in our evaluations). Descriptive characteristics of the two analysis samples are presented in Fig. 1. We present the characteristics of the longitudinal cohort in both temporal snapshots in Fig. 1 (“current” 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 Fig. 2.

Fine-grained fitness prediction from wearable sensors
We first developed and externally validated several non-exercise VO2max estimation models as a regression task using features...
commonly measured by wearable devices (anthropometry, resting heart rate (RHR), physical activity (PA); see Table 1). Here our goal was to explore how conventional non-exercise approaches to VO2max estimation could be enhanced by features from free-living PA data. We split participant data into independent training and test sets. The training set \((n=8384, \text{participants with baseline data only})\) was used for model development. The test set \((n=2675, \text{participants with baseline and follow-up data})\) was used to externally validate each model. Starting with linear regression models, we used anthropometry or RHR alone which yielded poor external validity \((R^2\text{ of } 0.35–0.37)\), but validity improved when combined in the same model \((R^2\text{ of } 0.61)\). The best performance \((R^2\text{ of } 0.67)\) was attained using a deep neural network model combining wearable sensors, RHR, and anthropometric data (Fig. 3). 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 refs. 17,18) we obtained poor validity \((R^2\text{ of } -3.2\text{ and Correlation of } 0.389)\), 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.

To understand the limits of the models, we conducted a number of post-hoc sensitivity analyses by investigating subgroup performance in terms of sex (male/female), age, weight, BMI, and height (Table 3), as well as investigating model errors on Bland-Altman agreement plots (Fig. Suppl. 2). We found that the comprehensive model is robust to most subgroups, showing minimal differences in most cases, with exceptions in weight and age. In particular, we found no difference between male and female participants even though the performance was lower for...
each group compared to the mixed set ($R^2$ of 0.59). Further, the models perform better on participants of lower age ($R^2$ of 0.68), higher weight ($R^2$ of 0.69). No effects were observed on height or BMI differences (overlapping CIs). The best performing subgroup is "higher weight" and the worst performing "sex-male". Last, Bland-Altman plots showed that the Dense model has better upper difference compared to the linear model, where the lower and mean difference were similar (Fig. Suppl. 2).

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 nonlinear dimension-reduction technique, to visualise learned feature representations from our model and their relationship to participant VO$_{2\text{max}}$. Clustering and coloring by VO$_{2\text{max}}$ was shown to be inversely related and more apparent in the learned latent space compared to the original feature space. Further, we show how this latent space can be used for patient subtyping through embedding neighbours. Starting from an initial patient (query), we retrieved the five nearest neighbours in the latent and original space. In a case study with three randomly selected participants, we found that the total euclidean distance of the query to all neighbours is higher in the original than the latent space, pointing to better semantic clustering (Fig. 4, bottom panel).

**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$_{2\text{max}}$ using only information from the present as input (Table 2). This model yielded a slightly lower accuracy than Fenland I, achieving an $R^2$ of 0.49 and a correlation of 0.72. This lower performance is expected since the model has no information on the behavior of the individuals 7 years later. We also trained a model to directly predict the difference (or delta) of current-future VO$_{2\text{max}}$, which reached a correlation of 0.23. Motivated by the moderate predictability of the fine-grained delta of VO$_{2\text{max}}$, we formulated this problem as a classification task. A visual representation of this task can be found in Fig. 5a. By inspecting the distribution of the difference (delta) of current-future VO$_{2\text{max}}$ on the training set, we split it into two halves (50% quantiles) of equally balanced data 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 (or delta) of current-future VO$_{2\text{max}}$, which reached a correlation of 0.23. 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 Fig. 5b.

**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 obtaining new wearable data is relatively easy.
Fig. 4  **Fitness subtyping through latent neighbours.** tSNE projection of the original feature vector (Fenland I testing set, Sensors + RHR + Anthro.) and the latent space of the Dense model after training. Top: 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 VO₂max both locally and globally. Bottom: Starting from a query participant (+) we retrieve the five nearest neighbours in the original and latent space and list their details on the tables on the right. The total distance of each query to each neighbour is listed in each subplot. Transparency has been applied to combat crowding and the colorbar is centered on the median value to illustrate extreme cases. The VO₂max label is used only for color-coding purposes (the projection is label-agnostic). Every participant is a dot.
when participants' body composition, and low cardiorespiratory fitness has significant prognostic value in

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\textsuperscript{23}. 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 VO\textsubscript{2}max, 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 \textasciitilde7 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 information from each participant and combine it with various anthropometrics and bio-markers such as the RHR, providing a truly personalized approach for CRF inference generation. The approach we present here outperforms traditional non-exercise models, which are considered state-of-the-art methods for longitudinal monitoring and highlights the potential of wearable sensing technologies for digital health monitoring. An additional application of our work is the potential routine estimation of VO\textsubscript{2}max in clinical settings, given the strong association between estimated CRF levels and CVD health outcomes\textsuperscript{24}.

This study has several limitations worthy of recognition. First, the validity of the deep learning framework was assessed by comparing estimated VO\textsubscript{2}max values with VO\textsubscript{2}max values derived from a submaximal exercise test. Ideally, one would use VO\textsubscript{2}max values directly measured during a maximal exercise test to establish the ground truth for cardiorespiratory fitness comparisons. Maximal exercise tests, however, are problematic when used in large population-based studies because they may be unsafe for some participants and, consequently, induce selection bias. The submaximal exercise test used in the Fenland Study was well tolerated by study participants and demonstrated acceptable validity against direct VO\textsubscript{2}max measurements\textsuperscript{25}. Submaximal tests are also utilized to validate popular wearable devices such as the Apple Watch\textsuperscript{25}. We are therefore confident that VO\textsubscript{2}max values estimated from the deep learning framework reflect true cardiorespiratory fitness levels.

To investigate this limitation, we validate our models with 181 participants from the UK Biobank Validation Study (BBVS) who were recruited from the Fenland study. These participants completed an independent maximal exercise test, where VO\textsubscript{2}max was directly measured. Taking into account that the BBVS cohort is less fit (VO\textsubscript{2}max = 32.9 ± 7) compared to the cohort the model was trained on (Fenland I, VO\textsubscript{2}max = 39.5 ± 5), we observe that the model over-predicts with a mean prediction of VO\textsubscript{2}max = 39.9, RMSE = 8.998. This is expected because the range of VO\textsubscript{2}max seen during training did not include participants with VO\textsubscript{2}max below 25. Even when looking at women in isolation—who perform lower than men in these tests—, they had a VO\textsubscript{2}max = 37.4 ± 4.7 in Fenland I (see Fig. 1), which is still significantly higher than the average participant of BBVS. This is a common distribution/label shift issue where the model encounters an outcome which is out of its training data range and is still an open problem in statistical modelling\textsuperscript{26}. To partially mitigate this issue, we match BBVS’s fitness to have similar statistics to the validation set of original model (Fenland I: VO\textsubscript{2}max = 39 ± 5, matched BBVS: VO\textsubscript{2}max = 39 ± 4) and observe an RMSE = 5.19 (see Fig. Suppl. 1). This result is still not on par with our main results in Fenland but shows the impact of including very low-fitness participants in this validation study. To improve future CRF models, we believe that population-scale studies should focus on including low-fitness participants along with the general population.
Putting our results in context, we are confident that the level of accuracy of our methods is acceptable for use in population scale or even commercial wearables. For example, in Table 3, we observe that the MAE of all subgroups is between 2.1 and 2.3. Future directions to improve these results include transfer learning and domain adaptation, particularly tailored to the problem of distribution shift, as discussed in the previous paragraph. Compared to smaller studies such as the Apple Watch study which was conducted in more constrained environments that required users to log their workouts, we see that the reported MAE is 1.4. We believe this difference is reasonable due to the completely free-living data we incorporate, bringing our study’s evaluation setup closer to the real world.

In this paper, we developed deep learning models utilising wearable data and other bio-markers to predict the gold standard of fitness (VO₂max) 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 $r = 0.82$, when compared to the ground truth in a holdout sample. Additionally, we show the adaptability and applicability of this approach for detecting fitness changes over time in a longitudinal subsample ($n = 2675$) who repeated measurements after 7 years. Last, the latent representations that arise from this model pave the way for 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, 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 ≤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 be 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 Fig. 1.

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 were included after excluding participants with insufficient or corrupt data or missing covariates as shown in Fig. 1. A subset of 2675 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.

| Table 3: Sensitivity analysis of the comprehensive Dense model with regards to anthropometrics. |
| --- |
| **Features** | **Evaluation Metrics [95% CI]** | **R²** | **MAE** | **RMSE** | **MAE** |
| **Sex** | | | | | |
| Male (N = 1303) | | | | | |
| Female (N = 1337) | | | | | |
| **Age** | | | | | |
| ≤ 48 years (M = 1160) | | | | | |
| > 48 years (M = 1270) | | | | | |
| **Weight** | | | | | |
| ≤ 75 kg (M = 1345) | | | | | |
| > 75 kg (M = 1335) | | | | | |
| **BMI** | | | | | |
| ≤ 25.5 Kg/m² (M = 1338) | | | | | |
| > 25.5 Kg/m² (M = 1337) | | | | | |
| **Height** | | | | | |
| ≤ 1.70 m (M = 1343) | | | | | |
| > 1.70 m (M = 1332) | | | | | |
| Breakdown of sex, age, weight, BMI, and height based on either the median value of the testing set or the available categories. Bold font denotes best values (low for Error metrics, high for $R^2$ and Cor). Note: the groups are not balanced because the median values were assigned to the first (5) group.
Study procedure
Participants wore the Actiheart wearable ECG which measured heart rate and movement recording at 60-s intervals\textsuperscript{27}. 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{2 max} (maximum rate of oxygen consumption measured during incremental exercise)\textsuperscript{28}. RHR was measured with the participant in a supine position using the Actiheart device. HR was recorded for 15 min and RHR was calculated as the mean heart rate measured during the last 3 min. 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{2 max} was predicted in study participants using a previously validated submaximal treadmill test\textsuperscript{9}. 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)\textsuperscript{18}, 3) the participants exercised at or above 80% of age-predicted maximal heart rate for 2 min. Details about the fitness characteristics of the cohort and the validation of the submaximal test are provided elsewhere\textsuperscript{29}. To further validate the models trained on submaximal VO\textsubscript{2 max}, we employ the external cohort UK Biobank Validation Study (BBVS)\textsuperscript{29}. 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{2 max} 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 h 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 min) 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 = 71.5 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 ≤1.5 METs classified as sedentary behaviour, activities between 3 and 6 METs were classified as moderate to vigorous physical activity (MVPA) and those >6 METs 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\textsuperscript{7,10}. Here we encoded the month of the year as (x<sub>y</sub>, y<sub>y</sub>) coordinates on a circle:

\[ T_{x,y} = \sin \left( \frac{2 \pi \times n \times t}{\max(t)} \right) \]  
\[ T_{y,y} = \cos \left( \frac{2 \pi \times n \times t}{\max(t)} \right) \]  

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, Acceleration-derived Euclidean Norm Minus One, and Acceleration-derived Metabolic Equivalents of Task\textsuperscript{30,31}. 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 the 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 Suppl. 2.

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/min) 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 time-steps). Even the well-tuned recurrent neural networks cannot cope with such sequences and given the size of the training set (7545 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. 2). This feature vector was fed to fully connected neural network layers which were trained with backpropagation. All deep learning models are implemented in Python using Tensorflow/Keras.

Data preparation. For Task 1 (see Fig. 1), 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 participants from Fenland I go to the train set, while those from Fenland II go to the test set (see Fig. 7). 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 and 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 PCA 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 \( X = (x_1,...,x_N) \in \mathbb{R}^{N \times F} \) and a target \( \text{VO}_2\text{max}_N = (y_1,...,y_N) \in \mathbb{R}^N \). The network consists of two densely-connected feed-forward layers with 128 units each. A dense layer works as
In all other cases, a smaller network with one Dense layer of 128 units and a sigmoid activation is used to study the delta distribution as a binary problem. When we cannot be comparable with the models predicting the present future (and the $\delta$ prediction task corresponds to $\text{VO}_2\text{max}$ in the future (and the $\text{VO}_2\text{max}$ present-future). We note that the delta prediction task cannot be comparable with the models predicting the present and future outcomes. Essentially, the delta model predicts the difference between these two timepoints, which results in a range of values roughly from $-10$ to $+10$. This distribution is not normally distributed (Shapiro-Wilk test $p = 0.002$) and hence both linear and neural models cannot approximate the error distributions in regression tasks. For the subgroup sensitivity analysis, we additionally employed the Mean Squared Error (MSE), Mean Absolute Error (MAE) and its standard deviation (STD of MAE), and the Mean Absolute Percentage Error (MAPE). In most regression metrics, $y$ and $\hat{y}$ are the measured and predicted $\text{VO}_2\text{max}$ and is the mean. For the binary models, we used the Area under the Receiver Operator Characteristic (AUROC or AUC) which evaluates the probability of a randomly selected positive sample to be ranked higher than a randomly selected negative sample.

**Visualizing the latent space**

The activations of the trained model allow us to understand the inner workings of the network and explore its latent space. We first pass the test set of Task 1 through the trained model and retrieve the activations of the penultimate layer\cite{19}. This is a 2D vector of $[2675, 128]$ size, considering that the layer size is 128 and the participants of the test set are 2675. Intuitively, every participant corresponds to a 128-dimensional point. In order to visualize this embedding, we perform t-SNE\cite{20}, an algorithm for dimensionality reduction. For its optimization, we use a perplexity of 50, as it was suggested recently\cite{21}. We calculated the k-nearest neighbours on both the original and latent spaces (using $k = 5$) in a case study presented in Fig. 4. We also calculated the total euclidean distance of each query participant across all their neighbours, as a means of quantifying the proximity in the high-dimensional space.

**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 (500 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 suggested threshold of $p < 0.005$ for human studies\cite{22}.

**Reporting summary**

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

**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/).
