EAI Endorsed Transactions
on Bioengineering and Bioinformatics

Research Article **EAI.EU**

Well-being Forecasting using a Parametric Transfer-Learning method based on the Fisher Divergence and Hamiltonian Monte Carlo

Eirini Christinaki[1,*], Tasos Papastylianou[1,*], Sara Carletto[2], Sergio Gonzalez-Martinez[3], Luca Ostacoli[4], Manuel Ottaviano[3], Riccardo Poli[1], Luca Citi[1,§]

[1]School of Computer Science and Electronic Engineering, University of Essex, Colchester, UK
[2]Department of Neuroscience "Rita Levi Montalcini", Università degli Studi di Torino, Turin, Italy.
[3]Life Supporting Technologies, Universidad Politécnica de Madrid, Madrid, Spain.
[4]Department of Clinical and Biological Sciences, Università degli Studi di Torino, Turin, Italy.

Abstract

INTRODUCTION: Traditional personalised modelling typically requires sufficient personal data for training. This is a challenge in healthcare contexts, e.g. when using smartphones to predict well-being.

OBJECTIVE: A method to produce incremental patient-specific models and forecasts even in the early stages of data collection when the data are sporadic and limited.

METHODS: We propose a parametric transfer-learning method based on the Fisher divergence, where information from other patients is injected as a prior term into a Hamiltonian Monte Carlo framework. We test our method on the NEVERMIND dataset of self-reported well-being scores.

RESULTS: Out of 54 scenarios representing varying training/forecasting lengths and competing methods, our method achieved overall best performance in 50 (92.6%) and demonstrated a significant median difference in 45 (83.3%).

CONCLUSION: The method performs favourably overall, particularly when long-term forecasts are required given short-term data.

Received on 09 June 2020; accepted on 01 October 2020; published on 16 October 2020

Keywords: Transfer Learning, MCMC, Bayesian Inference, Well-being Prediction, Personalised Modelling, NEVERMIND

doi:10.4108/eai.16-10-2020.166661

1. Introduction

Clinical depression is a psychiatric disorder affecting mood in a pervasive manner, leading to a reduced quality of life and daily functioning for the patient [1]. It is characterised by sadness, loss of interest or pleasure, feelings of guilt or low self-esteem, disturbed sleep or appetite, feelings of tiredness, poor concentration and even medically unexplained symptoms. It often co-exists with symptoms characteristic of anxiety disorders [2], and there is also evidence of a strong bi-directional association with physical illness [3]. In other words, having a physical illness is a strong risk factor for developing clinical depression, while depression is also a risk factor for developing or exacerbating existing physical illness [4, 5], and is linked to early death [6].

Smartphones and wearable sensors are increasingly used for prediction and management in healthcare contexts, including depression. However, few systems are based on patient-specific models, despite the fact that these are known to perform better compared to

[*]Eirini Christinaki and Tasos Papastylianou contributed equally to this work and should be regarded as joint first authors.
[§]Corresponding author. Email: lciti@essex.ac.uk

general models. In addition, many of the proposed systems and their purported benefits are often not properly backed up by evidence obtained from appropriate scientific research or clinical studies [7]. Research on constructing models to predict future mood-states in patients with depression has shown that, apart from the expected variables that describe patients' historical mood, the pertinent variables that determine model performance tend to be diverse and patient-specific [8].

Personalisation requires the learning and tuning of a model to an individual user. This is a non-trivial task for two reasons. Firstly, in an ideal situation, predictions should be provided from day one; in other words, the moment a patient starts using the system, the model will be expected to start making meaningful predictions for that individual, despite the fact that no, or very limited patient-specific data will be available initially. While it is possible to train/update a model incrementally on the data available at a given time, it is difficult to give reliable predictions when little data are available. Secondly, such datasets are also likely to be 'sparse' (in the broader sense), both because of the nature of the data acquired, and because users may have the option to refuse, or postpone their interaction with the system (e.g. when prompted with a questionnaire, or asked to wear a sensor), thereby exacerbating the problem.

Challenges faced when training a model on such sporadic, limited data with traditional machine learning algorithms include overfitting, difficulties in handling outliers, and inappropriate assumptions of equivalence between training and test data distributions — a concept known as dataset shift [9]. As a result, models trained on such 'sparse' datasets run the risk of being meaningless or unreliable in practice. In order to account for the uncertainty surrounding data sparsity, one could allow models to be of sufficient generality, but then one would run the risk of creating models that are of no practical value. Similarly, for any given model, sparsity complicates assessing the effective fit to the data, since there are not enough sample points to help meaningfully differentiate between more specific and more generalised models.

The above conundrum was one that was also faced in the NEVERMIND project [10]. NEVERMIND is an EU-funded project, which includes a randomised-controlled trial, and is tasked with helping individuals at risk of developing depressive symptoms following a primary clinical event such as cancer, myocardial infarction, amputation, and kidney failure [11]. It does so by providing effective, smartphone/wearable-based self-management tools, which in turn complement and guide the patients' clinical support plan. In this setting, both subjective as well as multimodal biomedical data are collected — including a collection of physiological signals, body movement, and speech — through the use of a lightweight sensorised T-shirt and patients' smartphones. For these tools to be effective, predictions need to be produced on a daily basis. However, particularly in the early stages of a patient's enrolment, the obtained data exhibit all the problematic behaviours mentioned earlier, making such personalised daily predictions a very challenging task.

Our approach to overcoming these challenges has been the following. In the early stages of a patient's enrolment, when data are limited and sporadic, rather than risk overfitting with an inappropriately personalised approach, our model relies on a more generalised prediction, borne from prior knowledge; however, as more personal data become available, the model slowly starts transitioning to more personalised predictions, in an organic, data-driven manner. In this way, patients can still benefit from useful, general information early on. The prior knowledge required for the initial generalised predictions was obtained from the other participants in the study, and used to inform the prediction of a specific patient through an appropriate transfer learning mechanism.

Transfer Learning (TL) methods are a recent class of techniques, which enable one to work around the strict requirement that the test and training data should necessarily conform to the same probability distribution [12]. These methods can use data from unrelated or partially related tasks [13], and allow the domains, tasks, and distributions used in training and testing to be different up to a certain point, without having to build a completely new model from scratch [14]. They rely on the basic assumption that the source and target domains, while not necessarily of the same underlying distribution, may still be related in other ways, i.e. via an explicit or implicit relationship between the feature space of the two domains. The goal of TL is to improve learning in the target domain by leveraging previously acquired knowledge gained in the source domain. These methodologies have also been employed to improve performance in the presence of scarce data [15], and techniques that take into account population heterogeneity have been proposed in domains involving sequential data modelling [16]; for a more general overview of the history, taxonomy, and state of the art in transfer learning methods for classification, regression, and clustering problems, see [14, 17–21].

In this paper, we propose a new parametric TL algorithm, addressing the challenge of creating user-specific models and making predictions of self-reported well-being scores, when training is performed incrementally on limited, sporadic data that become slowly available over time. The proposed method makes use of the Fisher divergence to make predictions about a specific individual, leveraging general information available from other patients in the form of priors.

The main contribution of this paper is the following: we propose a personalised Bayesian inference method making use of TL in the context of Hamiltonian Monte Carlo sampling, which allows a population prior to be directly represented in the sampling process through the use of the Fisher divergence. We demonstrate the effectiveness of the above technique in the context of personalised prediction of self-reported well-being scores, using data from the NEVERMIND project [10, 11]. Our method allows for a seamless transition from generalised to highly personalised models, as data become gradually available.

This paper is an extended version of the work presented in [22]. We extend our previous work in three important ways: first, by providing a more comprehensive literature review, particularly with respect to the theoretical background underpinning our approach; second, we have included information on diagnostics, detailing how we assessed the quality and convergence of the Markov chains; and third, we present a more rigorous mathematical derivation of the models involved, which was beyond the scope of [22].

The rest of the paper is organised as follows: Sec. 2 provides a brief overview of relevant work in the field of well-being forecasting, and the motivation for our method. Sec. 3 provides a brief description of the predictive model used in the NEVERMIND project and describes the proposed TL pipeline. Sec. 4 describes the NEVERMIND dataset in some detail, and the experimental protocol used in this work. Sec. 5 shows a rigorous evaluation of the model and an analysis of the factors affecting predictive performance. Finally, Sec. 6, lays out our conclusions and provides some indications of promising future research directions.

2. Background and motivation

There is evidence to suggest that ‘well-being’ — i.e. the subjective presence, absence, fluctuation, intensity, and nature of mood-states as perceived and reported by the individual — is at least as sensitive a predictor for the risk of adverse outcomes as formal clinical assessment of depression based on established diagnostic criteria [23–28]. There has therefore been a significant research interest over the past decade in monitoring and predicting mood states through non-invasive means. While some of these approaches involve laboratory-based techniques such as electroencephalography (e.g. [29]), or elaborate, bespoke equipment (e.g. the ‘Smart Mirrors’ project [30]), smartphones and wearable devices due to their ubiquity and convenience have now become the predominant research focus for the non-invasive collection of information and signals for this purpose (e.g. [31–37]).

However, the majority of studies in this field focus on mood detection and classification, and only few focus on the more challenging problem of long-term forecasting. In the latter case, studies commonly employ neural network methods for this purpose. E.g. Spathis *et al.* [36], used smartphones to collect a sequence of self-reported mood states over three weeks, by asking users to select a point from a two-dimensional grid corresponding to ‘valence’ and ‘arousal’ dimensions. They then trained a multi-task encoder-decoder recurrent neural network to produce a sequence of valence/arousal forecasts (expressed as points on the same grid) for up to 7 days. Their model performed well, though the authors noted performance was less reliable in participants with high mood variability. Similarly, Yu *et al.* [37] used data from the SNAPSHOT study [38], which involved detailed data from 251 college students, including data from surveys, mobile phones, wearables and weather information. These were used to define mood, health, and stress scores, on which they compared a series of multi-task learning approaches including Regularized Linear Models and several varieties of Neural Networks, in next-day, and up to 7-day forecast scenarios. Their findings showed good performance for next-day scenarios; however, the authors noted that even after selecting the best-performing algorithm, there was a significant reduction in accuracy when it came to 7-day forecasts.

One limitation of neural-network based approaches like the above, is that any transfer learning component learnt is typically applied as an initial point estimate of the network’s parameters. This is typically then either used as an initialization point for subsequent fine-tuning, or the topmost layers are ‘frozen’, meaning that they are excluded from subsequent training [39]. While this approach can achieve a significant initial speed-up in terms of learning, it is less robust, in that it does not allow for any uncertainty present in the transfer domain to be propagated to the prediction. Expressing the transfer learning component as a prior probability in the context of Bayesian inference methods [40] allows us to make use of this information. However, this is not necessarily a straightforward thing to do: when dealing with complicated distributions defined in high-dimensional spaces, obtaining posterior parameter estimates expressed in closed form is typically not feasible, as the integrals involved in the inference process tend to be computationally intractable.

Markov Chain Monte Carlo (MCMC) techniques are an elegant way to work around this problem. MCMC refers to a very general and powerful framework that allows sampling from a large class of distributions and which scales well with the dimensionality of the sample space [41]. It can be used to empirically obtain information about complicated distributions, and is particularly useful in estimating posterior distributions in Bayesian inference, even when complicated distributions in high-dimensional spaces are involved.

Therefore, the main motivation for this work, is the use of a suitable prior probability transfer-learning component within a Bayesian inference framework, for the long-term prediction of mood states from sporadic/limited data. This prior probability is obtained empirically through the use of MCMC, where the target distribution is obtained through an optimisation process involving the minimization of the Fisher Divergence over all participants in the transfer domain. We explain the steps involved in more detail below.

3. Method

3.1. Linear Dynamical System Model

In NEVERMIND, we propose to model participants and predict their self-reported well-being scores using a *Linear Dynamical System* (LDS) model [42] and a Bayesian TL approach relying on MCMC sampling [43]. The method assumes that the well-being of the user can be represented by a state vector, and that its dynamics can be captured by an LDS of the following form:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{x}(t-1) + \mathbf{B}\mathbf{u}(t) + \varepsilon_x(t), \tag{1a}$$

$$\mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) + \boldsymbol{\mu}_y + \varepsilon_y(t), \tag{1b}$$

where $\mathbf{x}(t) \in \mathbb{R}^{n_\mathbf{x}}$ is the latent state for the model, reflecting the user's underlying state of well-being; $\mathbf{y}(t) \in \mathbb{R}^{n_\mathbf{y}}$ is a vector of observations corresponding to the measurements collected from the user (including biomedical signal features and self-reported well-being scores); $\mathbf{u}(t) \in \mathbb{R}^{n_\mathbf{u}}$ is the input vector (representing external interventions, or influences from the external environment, e.g. weather or day of the week); and $\boldsymbol{\mu}_y$ is the baseline value of the observation vector. Finally, ε_x and ε_y represent noise (i.e. uncertainty) over the state and observation vectors, and are assumed to be distributed as $\varepsilon_x(t) \sim \mathcal{N}(0, \mathbf{S}_x)$ and $\varepsilon_y(t) \sim \mathcal{N}(0, \mathbf{S}_y)$ respectively. The parameters of this model will be collectively referred to as $\boldsymbol{\theta}$.

The above LDS model's latent state at any time t can be extended to describe an auto-regression of arbitrary order, simply by extending the state-vector to include its most recent values, e.g. by writing $\mathbf{x}(t) = [\xi(t), \xi(t-1), \xi(t-2)]^\intercal$, where $\xi(t)$ is the original, 'base' latent state, and $\mathbf{x}(t)$ is the extended one.

3.2. Hamiltonian Monte Carlo

Our approach requires that we sample from the posterior distribution of the parameters, i.e. our beliefs about the parameters after having seen the data for a given participant. This can be achieved using an MCMC sampler, which constructs a Markov chain of samples (i.e. parameter sets), having as their equilibrium distribution the target posterior distribution.

In our previous work [43], we used the affine invariant ensemble sampler for MCMC (emcee) proposed in [44]. emcee was chosen as an easy to use, well tested, pure Python module, where the underlying algorithm also has an affine invariance property that allows it to perform equally well under all linear transformations, and therefore be insensitive to covariances among parameters. It is also an ensemble method which relies on multiple walkers (the members of the ensemble) sampling in parallel. The main idea behind emcee's proposal strategy is that, for any given walker in the ensemble, their next position is proposed by randomly selecting another walker's location and performing a 'stretch-move' step towards it, i.e. performing a 'jump' in the proposal space in the direction of the randomly selected walker's location, such that the length of the jump is determined by the relative likelihood of the two proposal values. However, there are cases where the affine-invariant ensemble sampler may not perform well or shows unusual and undesirable properties. In particular, when the target density is a multi-modal landscape, the walkers can become stuck in different modes [45] or in lower dimensional subspaces. Furthermore, in high dimensions, the chains can show insufficient convergence and slow mixing, or appear to have converged when they have not [46].

For these reasons, in this work, we decided to utilise a *Hamiltonian Monte Carlo* (HMC) sampler [47], written in the Stan language [48], and more specifically its adaptive extension, the *No-U-Turn Sampler* (NUTS) [49].

The HMC approach exploits Hamiltonian dynamics in order to propose future states in the Markov chain. Effectively, the system simulates the movement of particles over a surface, such that the overall energy of the system is conserved, and can be expressed as the sum of two energy components — a 'kinetic energy', and a 'potential energy' component. The kinetic energy component is generated via a pre-determined probability distribution, and thus plays the part of the proposal component in MCMC, whereas the potential energy component maps directly to the underlying probability distribution we are trying to sample from. Standard HMC algorithms generally depend on, and are sensitive to an appropriate choice of hyperparameters, namely the step-size and number of steps to use during exploration of the domain; the NUTS variant modifies the proposal component of the base algorithm slightly, in that it evolves the initial system both forwards and backwards in time to form a balanced binary tree. The system then stops automatically when the algorithm detects that the sampler has started retracing earlier steps (i.e. making a "U-turn"), thus eliminating the need to define a pre-determined number of steps; at the same time, the step-size parameter is adapted on the fly, completely eliminating the need to hand-tune HMC.

The HMC algorithm is advantageous, in that it organically makes use of gradient information, enabling it to move faster toward regions of high probability and explore the parameter space more efficiently compared to standard random walks. Consequently, with this sampler we obtain faster convergence in high-dimensional target distributions, while the resulting Markov chain is less correlated. In addition, like in `emcee`, multiple chains can be allowed to run in parallel. Finally, the use of HMC allows for straightforward scaling up of models to even higher dimensionality and complexity, which may be required in future work.

According to Bayes' Theorem, given a vector of observations $\mathbf{y}$, and a vector of parameters $\boldsymbol{\theta}$, the posterior probability $p(\boldsymbol{\theta}\,|\,\mathbf{y})$ is related to the likelihood term $p(\mathbf{y}\,|\,\boldsymbol{\theta})$ and the prior term $p(\boldsymbol{\theta})$ via

$$p(\boldsymbol{\theta}\,|\,\mathbf{y}) \propto p(\mathbf{y}\,|\,\boldsymbol{\theta})\,p(\boldsymbol{\theta}). \tag{2}$$

Therefore, given a way to compute the product $p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$, the HMC sampler allows one to generate K random vectors $\boldsymbol{\theta}_k$, distributed according to $p(\boldsymbol{\theta}|\mathbf{y})$. One can then use this fact to estimate a posterior expectation $\mathbb{E}_{\boldsymbol{\theta}|\mathbf{y}}\Big[h(\boldsymbol{\theta})\Big]=\int h(\boldsymbol{\theta})\,p(\boldsymbol{\theta}|\mathbf{y})\,\mathrm{d}\boldsymbol{\theta}$ with respect to an arbitrary function $h(\boldsymbol{\theta})$, as the sample average $\frac{1}{K}\sum_{k=1}^{K} h(\boldsymbol{\theta}_k)$, evaluated at the posterior samples $\boldsymbol{\theta}_k$ [40, Sec. 10.1].

In our case, the likelihood term $p(\mathbf{y}|\boldsymbol{\theta})$ in (2) is the marginal likelihood of the LDS model in Sec. 3.1, marginalised over the latent state, i.e.,

$$p(\mathbf{y}\,|\,\boldsymbol{\theta}) = \int p(\mathbf{y}\,|\,\mathbf{x},\boldsymbol{\theta})\;p(\mathbf{x}\,|\,\boldsymbol{\theta})\;\mathrm{d}\mathbf{x}. \tag{3}$$

This likelihood term can be readily obtained from the LDS model using a Kalman filter applied to the participant's data (see [42]).

Additionally, we specify a prior probability distribution $p(\boldsymbol{\theta})$ to inform and constrain our model. In this work, we use a simplified version of the LDS model described by (1), which ignores the influences from the external environment; in other words, the inputs $\mathbf{u}(t)$ are absent and, therefore, the matrix $\mathbf{B}$ is unused. Furthermore, the observation vector $\mathbf{y}(t)$ is limited to reflect only self-reported well-being scales. Given the above, we consider a unit-root, third-order autoregressive model, which can be represented by the LDS model in (1) with:

$$\mathbf{A} = \begin{bmatrix} 1-a_1-a_2 & a_1 & a_2 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \qquad \mathbf{C} = \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix},$$

$$\mathbf{S}_x = \begin{bmatrix} s_x & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad \mathbf{S}_y = \begin{bmatrix} s_y & 0 & 0 \\ 0 & s_y & 0 \\ 0 & 0 & s_y \end{bmatrix},$$

$\boldsymbol{\mu}_y$=$[\mu_y, \mu_y, \mu_y]^\intercal$ and $\mathbf{x}(0) = [\xi_1, \xi_1, \xi_1]^\intercal$. The estimates of the unknown model matrices are parametrised through $\boldsymbol{\theta}$, where $\boldsymbol{\theta} = [a_1, a_2, c_{11} \dots c_{33}, s_x, s_y, \mu_y, \xi_1]$.

3.3. Bayesian Transfer Learning

When making a prediction, we want to take into consideration the information coming from both the individual patient, as well as more general information available from other patients acting as prior knowledge in a general sense. Formally, we want to obtain the posterior predictive distribution $p(\widetilde{\mathbf{y}}\,|\,\mathbf{y},\mathbf{Y}_N)$ for a given patient (without loss of generality we consider the one with index N+1), where $\widetilde{\mathbf{y}}$ is the desired prediction, $\mathbf{y}$ represents the patient's existing observations, and $\mathbf{Y}_N = \{\mathbf{y}_1, \dots, \mathbf{y}_N\}$ corresponds to the information coming from all other N participants' observations. In theory, this expression can be obtained by marginalising $\boldsymbol{\theta}$ out as follows:

$$p(\widetilde{\mathbf{y}}\,|\,\mathbf{y},\mathbf{Y}_N) = \int_{\boldsymbol{\theta}} p(\widetilde{\mathbf{y}}\,|\,\mathbf{y},\boldsymbol{\theta})\;p(\boldsymbol{\theta}\,|\,\mathbf{y},\mathbf{Y}_N)\;\mathrm{d}\boldsymbol{\theta}. \tag{4}$$

Unfortunately, in practice this integral is generally intractable. Both our previous TL approach based on *Bayesian Model Averaging* (BMA) [43] and the one proposed in this paper, are essentially machine-learning techniques for estimating the intractable integral in (4); however, they do so in a substantially different way. The two approaches are contrasted and explained in more detail in the sections below.

3.4. Transfer Learning based on BMA

To explain our motivation for developing a new approach, we first briefly review our previous approach based on BMA (see [43] for more details). Assuming conditional independence with respect to $\boldsymbol{\theta}$ across data coming from different participants, we expanded (4) as:

$$\begin{aligned} p(\widetilde{\mathbf{y}}\,|\,\mathbf{y},\mathbf{Y}_N) &= \int p(\widetilde{\mathbf{y}}\,|\,\mathbf{y},\boldsymbol{\theta})\,\frac{p(\mathbf{y}\,|\,\boldsymbol{\theta})\,p(\boldsymbol{\theta}\,|\,\mathbf{Y}_N)}{\int p(\mathbf{y}\,|\,\boldsymbol{\theta}')\,p(\boldsymbol{\theta}'|\mathbf{Y}_N)\,\mathrm{d}\boldsymbol{\theta}'}\,\mathrm{d}\boldsymbol{\theta} \\ &\approx \sum_{k=1}^{K} p(\widetilde{\mathbf{y}}\,|\,\mathbf{y},\boldsymbol{\theta}_k)\,\frac{p(\mathbf{y}\,|\,\boldsymbol{\theta}_k)}{\sum_{j=1}^{K} p(\mathbf{y}\,|\,\boldsymbol{\theta}_j)}, \end{aligned} \tag{5}$$

with each of the K samples in $\{\boldsymbol{\theta}_k\}_{k=1}^{K}$ distributed according to $p(\boldsymbol{\theta}\,|\,\mathbf{Y}_N)$, which represents our beliefs about the parameters for patient $N+1$ after observing the data from the other N participants, but prior to observing any data for the new patient. Effectively, $p(\boldsymbol{\theta}\,|\,\mathbf{Y}_N)$ embodies the knowledge transfer from the other N participants to the new one. In our BMA-based approach, $p(\boldsymbol{\theta}\,|\,\mathbf{Y}_N)$ was approximated by running the MCMC sampler on each of the N participants and then pooling together the resulting samples. Under this scheme, assuming each run creates M samples (i.e. M vectors of model parameters), we obtain the K vectors used in (5) via uniform sampling from the mixed sample pool of $N{\cdot}M$ vectors of model parameters. This corresponds to sampling from the mixture distribution obtained by combining the posterior distributions of the N previous participants. The probabilities $p(\widetilde{\mathbf{y}}\,|\,\mathbf{y},\boldsymbol{\theta}_k)$ and $p(\mathbf{y}\,|\,\boldsymbol{\theta}_k)$ are then obtained by using the Kalman filter as described earlier. The fractional term in the summation shown in (5) represents the probability that, out of the K models considered, the given model $\boldsymbol{\theta}_k$ generated the observed data $\mathbf{y}$. Therefore, using (5) to estimate (4) corresponds to performing BMA [50] over the K candidate models.

As reported in [43], this method showed an improvement over previous work, which relied on Maximum Likelihood parameter estimation using a standard Expectation Maximisation approach [42]. However, such a BMA approach does not exploit the full potential of the MCMC sampler with respect to estimating the integral in (4). In [43], the BMA approach makes use of MCMC, simply in order to explore and generate samples from $p(\boldsymbol{\theta}\,|\,\mathbf{Y}_N)$; these samples are then weighted using $p(\mathbf{y}\,|\,\boldsymbol{\theta}_k)$ as shown in (5), to obtain the probability $p(\boldsymbol{\theta}\,|\,\mathbf{y},\mathbf{Y}_N)$ required to approximate (4). To make fuller use of the potential of the MCMC sampler for the estimation of the integral in (4), we formulated a new algorithm, which allows MCMC to explore and sample from $p(\boldsymbol{\theta}\,|\,\mathbf{y},\mathbf{Y}_N)$ directly. This improved algorithm is the main contribution of this paper and is described in the next section.

3.5. Transfer Learning based on the Fisher divergence

The approach delineated in this paper is a parametric TL method based on the *Fisher divergence*, which can be used to fit a sample of data points to given probabilistic models defined up to a normalisation constant [51–53]. In this approach, the HMC sampler uses the data from each participant directly, to create chains of parameter vectors reflecting the posterior probability distribution of their personalised models; however, in doing so, the MCMC process itself makes use of a TL component internally, in the sense that the generated chains are obtained based on a modified prior, where this prior accounts for the knowledge available from all other participants, in a manner reminiscent of empirical Bayes approaches. Stated formally, we estimate (4) as

$$p(\widetilde{\mathbf{y}}\,|\,\mathbf{y},\mathbf{Y}_N) \approx \frac{1}{K}\sum_{k=1}^{K} p(\widetilde{\mathbf{y}}\,|\,\mathbf{y},\boldsymbol{\theta}_k), \tag{6}$$

where the samples $\boldsymbol{\theta}_k \sim p(\boldsymbol{\theta}\,|\,\mathbf{y},\mathbf{Y}_N)$ are obtained via MCMC using the likelihood $p(\mathbf{y}|\boldsymbol{\theta})$ from (3) and a prior $p(\boldsymbol{\theta}\,|\,\mathbf{Y}_N)$ obtained as a mixture of the posterior distributions of the N previous participants:

$$p(\boldsymbol{\theta}\,|\,\mathbf{Y}_N) = \frac{1}{N}\sum_{n} p(\boldsymbol{\theta}\,|\,\mathbf{y}_n), \tag{7}$$

that we approximate in parametric form as:

$$p(\boldsymbol{\theta}\,|\,\mathbf{Y}_N) \approx q_\beta(\boldsymbol{\theta})\,p(\boldsymbol{\theta}), \tag{8}$$

where $q_\beta(\boldsymbol{\theta})$ is a function governed by a vector of hyperparameters $\boldsymbol{\beta}$. Note that while $p(\boldsymbol{\theta}\,|\,\mathbf{Y}_N)$ in (7) is, from a theoretical point of view, the same as in the BMA approach, the fact that we now consider an approximation of it in parametric form allows us to explore it fully using the HMC sampler, rather than being constrained to using only the $N{\cdot}M$ samples previously obtained from the other participants.

In this work, the specific $q_\beta(\boldsymbol{\theta})$ used is an exponentiated quadratic w.r.t. a non-linear mapping of $\boldsymbol{\theta}$:

$$q_\beta(\boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2}\,g(\boldsymbol{\theta})^{\mathsf{T}}\mathbf{Q}_\beta\,g(\boldsymbol{\theta}) - \mathbf{v}_\beta^{\mathsf{T}} g(\boldsymbol{\theta})\right), \tag{9}$$

where $\boldsymbol{\beta} = \mathrm{vec}([\mathbf{Q}_\beta,\mathbf{v}_\beta])$ and g is a vector function such that its i-th element is $\log(\theta_i)$ if θ_i is a parameter representing a variance (e.g. s_x) and θ_i otherwise. The quadratic parameter $\mathbf{Q}_\beta$ is chosen in the set of positive semi-definite matrices so that $q_\beta(\boldsymbol{\theta})$ is bounded and $q_\beta(\boldsymbol{\theta})\,p(\boldsymbol{\theta})$ is a proper prior. The hyperparameters $\boldsymbol{\beta}$ leading to the best approximation (8) can then be found by minimising the Fisher divergence from $q_\beta(\boldsymbol{\theta})\,p(\boldsymbol{\theta})$ to $p(\boldsymbol{\theta}\,|\,\mathbf{Y}_N)$.

The Fisher divergence. The *Fisher divergence* from a distribution $q(\mathbf{x})$ to a distribution $p(\mathbf{x})$, denoted $D_F(p\|q)$, is defined as:

$$\begin{aligned} D_F(p\|q) &= \int_{\mathbf{x}} p(\mathbf{x}) \left\| \frac{\nabla_{\mathbf{x}} p(\mathbf{x})}{p(\mathbf{x})} - \frac{\nabla_{\mathbf{x}} q(\mathbf{x})}{q(\mathbf{x})} \right\|^2 \mathrm{d}\mathbf{x} \\ &= \int_{\mathbf{x}} p(\mathbf{x}) \left\| \nabla_{\mathbf{x}} \log p(\mathbf{x}) - \nabla_{\mathbf{x}} \log q(\mathbf{x}) \right\|^2 \mathrm{d}\mathbf{x}. \end{aligned} \tag{10}$$

Much like its better known counterpart — the *Kullback-Leibler divergence* (also known as *relative entropy*) — the Fisher divergence can be understood as an asymmetric measure of distance between a target distribution p, and an approximating distribution q serving as a model for p. In the same manner that the Kullback-Leibler divergence is tightly linked to the concept of entropy (in that it corresponds to the entropy difference between p and q) the definition of the Fisher divergence is similarly tightly linked to the concept of the *Fisher information*, defined[1] by $J(p) = \int p(\mathbf{x}) \left\| \nabla_{\mathbf{x}} \log p(\mathbf{x}) \right\|^2 \mathbf{dx}$.

A practical disadvantage of the Kullback-Leibler divergence is that, for the result to be meaningful, it requires that both the target and the approximating function be expressed as appropriately *normalised* probability density functions. However, when one only has unnormalised quantities to work with, the computation of an appropriate normalisation constant, whose proper evaluation requires integration over the entire domain of the function, tends to be intractable in the context of high-dimensional problems. By contrast, the Fisher divergence obviates the need for computing such a normalisation constant, since the fractional nature of the calculation with respect to both the target *and* the approximating function, means that a normalization constant would cancel out from either of those two terms anyway, and therefore lack of appropriate normalization does not affect the final result. This makes the Fisher divergence an advantageous measure of distance to use when dealing with high-dimensional, unnormalised probability density functions; this is indeed the case in our TL approach, since both our $q_\beta(\theta)$ model, and any distribution represented by the output of an MCMC sampler, will necessarily represent unnormalised quantities.

Minimising the Fisher divergence of the mixture distribution. As shown in appendix A, for a mixture distribution, the Fisher divergence of the mixture is simply the weighted sum of the individual divergences from $q_\beta(\boldsymbol{\theta})\,p(\boldsymbol{\theta})$ to the mixture components. From a collection of samples distributed according to $p(\boldsymbol{\theta}\,|\,\mathbf{y}_n)$, obtained by running MCMC separately for each one of the N "prior" patients, we can derive the Fisher divergence for each mixture component as follows:

$$\begin{aligned} F_n(\beta) &= D_F\Big(p(\boldsymbol{\theta}|\mathbf{y}_n) \,\Big\|\, q_\beta(\boldsymbol{\theta})\,p(\boldsymbol{\theta})\Big) \\ &= \int_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{y}_n) \left\| \nabla_{\boldsymbol{\theta}} \left[\log \frac{p(\mathbf{y}_n|\boldsymbol{\theta})\,p(\boldsymbol{\theta})}{p(\mathbf{y}_n)} \right] \right. \\ &\qquad\qquad \left. - \nabla_{\boldsymbol{\theta}} \left[\log\Big(q_\beta(\boldsymbol{\theta})\,p(\boldsymbol{\theta})\Big) \right] \right\|^2 \mathrm{d}\boldsymbol{\theta} \qquad (11) \\ &\approx \frac{1}{K} \sum_{k=1}^{K} \left\| \nabla_{\boldsymbol{\theta}}[\log p(\mathbf{y}_n|\boldsymbol{\theta}_k)] - \nabla_{\boldsymbol{\theta}}[\log q_\beta(\boldsymbol{\theta}_k)] \right\|^2 \end{aligned}$$

with $\boldsymbol{\theta}_k \sim p(\boldsymbol{\theta}|\mathbf{y}_n)$. Therefore, it becomes possible to obtain an optimal value β^* by solving the following constrained optimisation problem:

$$\beta^* = \underset{\beta\,:\,\mathbf{Q}_\beta \in S_+}{\arg\min}\; \frac{1}{N} \sum_{n=1}^{N} F_n(\beta); \qquad (12)$$

where S_+ is the set of symmetric positive semi-definite matrices. The problem in (12) is an instance of a cone quadratic program, which we solve efficiently using the cvxopt library [54].

The prior $q_{\beta^*}(\boldsymbol{\theta})\,p(\boldsymbol{\theta})$ so obtained is then used alongside the likelihood provided by the LDS model in the context of MCMC, to produce the K samples required for (6), thus giving rise to our final prediction as the average of K individual prediction components.

The mean and variance of the overall prediction can be obtained at each future timepoint by pooling the means and variances of the individual prediction components (i.e. μ_k and σ_k^2 respectively) as follows:

$$\mu(t) = \frac{1}{K} \sum_{k=1}^{K} \mu_k(t), \qquad (13a)$$

$$\sigma^2(t) = \frac{1}{K} \sum_{k=1}^{K} \left\{ \sigma_k^2(t) + \Big[\mu_k(t) - \mu(t)\Big]^2 \right\}. \qquad (13b)$$

4. Experimental Study

4.1. Dataset

For the purposes of this work, we used data collected over 112 participants in the context of the NEVERMIND randomised controlled trial [11]. This dataset consists of participants aged 18 or older, who have received a diagnosis of a severe somatic disease, including myocardial infarction, breast cancer, prostate cancer, kidney failure and lower limb amputation. The data were collected in Pisa, Turin and Lisbon, with appropriate informed consent obtained from the patients in writing, and experiments approved by local ethical committees.

The NEVERMIND dataset consists of subjective data in the form of questionnaires, as well as other multimodal data, collected over time from individual subjects via a smartphone and a specialised lightweight sensorised T-shirt. The full dataset includes a collection of physiological signals, accelerometer data, and voice recordings; however, for the purposes of this work, as mentioned earlier, we only consider the three self-reported well-being scales that the user is prompted to provide on a daily basis. The resulting daily scores from each scale are fed into the LDS model as the observation vector $\mathbf{y}(t)$. Each scale's numerical input is obtained from the participant via a sliding scale, which takes values from 1.0 to 6.0 (at 0.2 increments),

[1] As also noted in [51, 53], while the Fisher information can be defined with respect to *any* parameter, this particular formulation is specifically defined with respect to a hypothetical location parameter.

where lower values represent better outcomes.[2] The three scales correspond to the following questions:

- *"How are you feeling today?"* — the *Feel* score: a measure of the participant's subjective assessment of their morning / waking mood;
- *"How was your sleep?"* — the *Sleep* score: a measure of the participant's subjective assessment of sleep quality for the night before; and
- *"How was your day?"* — the *Day* score: a measure of the participant's subjective assessment of the quality of (potentially stressful) events over the course of the day.

Each question is prompted daily, and the participants may refuse to provide an answer, contributing to the sporadic nature of the dataset; human review may be triggered if no significant interaction has occurred for a certain time interval, according to the clinical protocol used in the randomised controlled trial. Participants for whom there were no available data (e.g. patients who had already been enrolled in NEVERMIND, but had not yet started using the system), or whose total data recording-length was less than two weeks, were excluded from the analysis carried out here.

4.2. Experimental protocol

This section describes the specific values and implementation of the model described above, as used in the experiments, as well as the approach used to validate the method.

We have chosen to use weakly informative priors on model parameters, expressing vague or general information; this has the effect that model selection is primarily driven by the likelihood function, such that in the presence of adequate data, the specific choice of prior has a minimal effect on the final inference, relative to the data.

Specifically, with regard to the parameters described in Sec. 3.2, we placed a diffuse Gaussian prior over the elements $a_1, a_2 \sim \mathcal{N}(0, 0.5^2)$ of the transition matrix $\mathbf{A}$, and over the coefficients $c_i \sim \mathcal{N}(0, 1)$ of the observation matrix $\mathbf{C}$. A diffuse Gaussian prior was also placed over the initial state vector $\mathbf{x}(0)$ where $\xi_1 \sim \mathcal{N}(1, 2)$; this distribution was centred away from zero to break the symmetry of the problem and reduce the occurrence of multiple equivalent modes. We further placed an inverse gamma prior over the non-zero diagonal element of the state-noise matrix $\mathbf{S}_x$, as $s_x \sim \Gamma^{-1}(\alpha, \beta)$ with shape parameter α=2 and scale parameter $\beta \approx 0.06$. Small values of α lead to wide distributions and in particular α=2 corresponds to a prior with infinite variance, thus allowing the inference mechanism to explore values of s_x as large as needed. The mode of the inverse gamma distribution is given by $s_x^* = \beta(\alpha+1)^{-1}$. We also set the baseline value of the observation vector to the fixed value μ_y=3, and the parameter s_y to the fixed value of 0.04 (i.e. 0.2^2); the latter was chosen empirically, by estimating the variance of the error made by the subjects when they provide answers to the questionnaires, given the fact that they use a slider in order to do so, and that this results in the scales being quantised at a resolution of 0.2.

The HMC sampler was set to compute Markov chains using 8 walkers working in parallel, such that each sample corresponds to a vector $\boldsymbol{\theta}$ consisting of the scalar parameters described in Sec. 3.2 (i.e. a_1, s_x, etc). Each walker was set to create 275 samples, where the first 150 obtained samples were discarded as 'burn-in', leaving 125 representative samples per chain. The individual chains generated from each walker were then combined into a single larger chain, having a total of 1000 samples.

We further monitored the convergence of the chains, by computing the *potential scale reduction factor on split chains*, typically referred to more concisely as the 'split-$\hat{R}$' measure [40, Sec. 11.4]. The split-$\hat{R}$ provides a measure of convergence and mixing quality of the chains in an MCMC simulation, which can be used to gain insight into the rate and degree of convergence, as well as in terms of detecting non-stationarity, allowing for better evaluation of the underlying algorithms. We also obtained the log-posterior density (denoted by the 'lp__' variable in Stan) and summary-statistics for each model parameter, including means, standard deviations (SD) and various quantiles computed from the draws.

The summary also reports the Monte Carlo Standard Errors ($\mathrm{SE_{mean}}$), and the effective sample sizes (n_{eff}). The Monte Carlo Standard Error is the uncertainty about a statistic in the sample due to sampling error; the smaller the standard error, the closer the mean estimate of the posterior draws of the parameter is expected to be to the true value. The effective sample size, n_{eff}, measures the amount by which autocorrelation in samples increases uncertainty (standard errors) relative to an independent sample; if the samples are independent, the effective sample-size equals the actual sample-size. It is particularly important in terms of gauging the reliability of the split-$\hat{R}$ measure, as a small n_{eff} can lead to unreliable values for $\hat{R}$. Table S1 (see Supplementary Material) presents an indicative example of a summary for the parameters of interest, as estimated from a collection of samples corresponding to one of the participants in the study. The results show that all values for the split-$\hat{R}$ are approximately 1.0

[2] Initially, during the early stages of the trial, the upper limit of the scale was set at 6.8; however, this was later capped at 6.0, following interface and user design considerations.

(above 0.9 and below 1.1) and n_{eff} is well above the minimum recommended value of 100 effective samples per chain [55], indicating that the chains had mixed well and the model had successfully converged.

The predictive performance of the current TL approach — namely the *Fisher-divergence minimization* approach (referred to as the M_{FD} model henceforth) — was evaluated in relation to a number of competing models. In the first instance, we compared this model against the BMA one used in [43] (model M_{BMA}). To ensure a fairer comparison between M_{FD} and M_{BMA}, the chains for M_{BMA} were created using HMC rather than emcee as previously done in [43]. Additionally, we also compared M_{FD} against a Maximum A Posteriori (MAP) model (M_{MAP}) and 4 'baseline' models: *a)* A *patient-average prediction* model (M_{A}), *b)* A *population-average* model (M_{P}), *c)* A *last-datapoint* model (M_{L}), and *d)* An *ordinary least squares regression* model (M_{R}). The M_{MAP} model was obtained by running Stan in 'optimization' mode instead of 'sampling' mode, which uses the Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) optimization algorithm under the hood [48]; this directly provides a single 'best' estimate for the parameter vector $\boldsymbol{\theta}$, corresponding to the MAP estimate of the posterior distribution, as computed by Stan.

Given that new data arrive incrementally, it is necessary to rebuild each patient's chains on a daily basis in order to keep the models up-to-date, and ensure that all observations available for that participant are being used. It is important to note that, for any given time-interval, given the fact that the patients are allowed to refuse to answer some or all questions on any particular day, the number of observations present within that time-interval may well differ between patients.

We define the following notation. Let:

- L_{tr} be the length of the 'training period', i.e. the number of weeks used for training (regardless of the number of actual observations that happen to be contained within), chosen from the set $\{1, 2, ..., 10\}$,
- L_{fc} be the length of the 'forecasting period', i.e. the number of future datapoints (one per day) to be forecasted, chosen from the set $\{1, 3, 7\}$,
- μ_t and σ_t^2 be the mean and variance of the forecasted prediction at timepoint t,
- r_t be the corresponding *actual* value of a well-being score observed at time t, (which may be missing if no answer was provided), used as a target value for validation.

In our experiments we evaluated the forecasting scenarios that result from all possible combinations of training and forecasting period-length pairs $\{L_{\text{tr}}, L_{\text{fc}}\}$; however, for brevity, we only show here a representative subset from these experiments, as explained further in Sec 5. Please note that in each case, the number of participants for which it is possible to obtain predictions, depends on the choice of L_{tr} and L_{fc}.

Since our goal is to obtain predictions with an associated measure of uncertainty (i.e. how much we can trust the prediction itself), the quality of our predictive algorithms must be assessed using a measure of accuracy that takes both the mean prediction accuracy and the estimated uncertainty into account.

One such measure is the *Log Likelihood* (*LL*), which, for a predictive model M_i, is given by

$$LL = \sum_t \log p(r_t \,|\, \mu_t, \sigma_t^2) = \sum_t \log \left[\frac{1}{\sqrt{2\pi\sigma_t^2}} e^{-\frac{(r_t-\mu_t)^2}{2\sigma_t^2}} \right] \tag{14}$$

where t corresponds to timepoints within the forecasting period for which an actual observation is available. The higher the *LL* measure, the better the probabilistic predictions are. Note that, while the output of M_{FD} and M_{BMA} is not strictly speaking a Gaussian, for the purposes of obtaining an *LL* measure, we represent them as Gaussians of their respective mean and variance.

While we believe *LL* is the correct measure to account for both mean prediction accuracy and accuracy in the uncertainty around the prediction, we also calculate the actual forecasting error for the predictions of each individual, using the more traditional *Root Mean Squared Error* (*RMSE*) calculated as $RMSE = \left[\frac{1}{n}\sum_t (r_t - \mu_t)^2\right]^{\frac{1}{2}}$, where t here again corresponds to timepoints within the forecasting period, n is the number of targets present when missing values are excluded, and where $r_t - \mu_t \overset{\text{def}}{=} 0$ when r_t is missing. Note that the *RMSE* ignores how accurate our estimates of the prediction variance are.

Furthermore, for any pair of competing models, we can calculate a 'winning percentage', as a measure of predictive superiority for one model over another. This is computed as:

$$\begin{aligned} \mathsf{wins} &= \mathsf{games} - \mathsf{ties} - \mathsf{losses}, \\ \mathsf{winning\%} &= [\mathsf{wins} + (\mathsf{ties}/2)] \,/\, \mathsf{games}, \end{aligned} \tag{15}$$

where games represents the total number of participants for whom it was possible to obtain predictions given a specific $\{L_{\text{tr}}, L_{\text{fc}}\}$ pair, and wins corresponds to the subset of those participants, for whom the model in question performed *better* than its counterpart, with respect to a particular performance measure (i.e. either *LL* or *RMSE*). Furthermore, we used the exact Wilcoxon Signed-Rank test-statistic [56] to make pairwise comparisons for these methods, effectively investigating the extent to which the winning percentages represent a genuine and statistically significant improvement, per pair of competing models.

5. Results and Analysis

5.1. Model output

The output of the model at each timepoint is a 3-dimensional probability distribution expressing a probabilistic prediction for the values of the three questions involved, i.e. the *Feel*, *Sleep*, and *Day* scores. For a different value of the length-of-training hyperparameter L_{tr}, a different model output is obtained over both the training and forecasting period. For visualisation purposes, we graph the individual questions independently as three separate graphs, each reporting score as a function of time t.

Fig. 1 shows a typical example of the means and variances of the model's probabilistic outputs per timepoint as learned by our model, along with the sporadic self-reported well-being scores for one of our participants, with L_{tr}=1 and L_{fc}=7. In the figure, the mean prediction is represented as a dashed line, and the uncertainty around the prediction is indicated by a shaded $\pm\sigma$ area around the mean. As expected, most (but not all) reported scores fall within the shaded region, even in the forecast phase, where, however, as expected σ grows progressively bigger due to the absence of inputs to the LDS. The *LL* and *RMSE* measures (see Sec. 4.2) corresponding to the model outputs over the timepoints in Fig. 1 were -4.41 and 0.5447, respectively.

5.2. Comparison against competing models

We compared the performance of the M_{FD} method against the competing models outlined in Sec. 4.2 in two ways: *a)* by analysing the distribution of performance differences directly; *b)* by analysing 'winning percentages' as per (15). Both analyses were performed using the *LL* and *RMSE* measures, separately.

A representative example of the first type of analysis can be seen in Fig. 2 for the *LL* differences and Fig. 3 for the *RMSE* differences between models. The results were obtained for L_{tr}=3 and L_{fc}=7, which was the most conservative choice for comparing M_{FD} and M_{BMA} (more on this later). It is clear that, for both performance measures, M_{FD} performs better across the board. Also, among all other competitors, M_{BMA} is the one with the least spread in terms of pairwise differences over all patients. Similar results were obtained with other values of L_{tr} and L_{fc}.

Figs. 4 and 5 look at the same predictions using the second type of analysis, again when training with 3 weeks of past data, and predicting 7 days ahead. These show that the M_{FD} model scores significantly more wins (at the 5% level, using a one-tailed hypothesis) compared to all of its competitors, when both the accuracy *and* the uncertainty of the predictions is taken into account (i.e. when using the *LL* as the performance measure). When only the *RMSE* is used, M_{FD} scores significantly better than four out of the six competitors, but shows no significant difference compared to its M_{BMA} and M_{A} competitors.

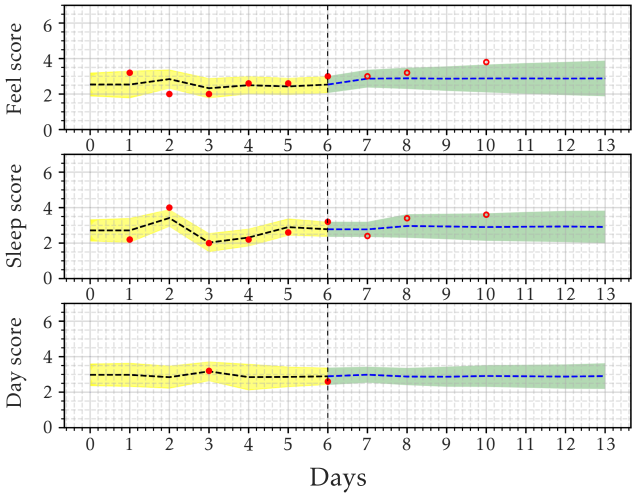

Figure 1. An example of self-reported well-being score modelling and prediction. The dashed vertical lines mark the last time point available to the model for training and visually separate past observations from future predictions. The solid red circles mark the reported scores that were used by the model, while the empty ones are reported to visually assess the prediction accuracy. The dashed black and blue lines and the associated yellow and green shadows represent the mean and standard deviation, respectively, of the distribution of the model outputs.

5.3. Effect of training / testing period length on performance

One would generally expect that increasing training-period length would improve performance over all models, and that predictions further away from the last training timepoint would diminish in accuracy. An analysis was therefore conducted to confirm this, for training periods $L_{\text{tr}}\in\{1, 3, 7\}$ and forecasting periods $L_{\text{fc}}\in\{1, 3, 7\}$. Table 1 shows the results obtained with respect to the *LL* evaluation measure[3] The top half of the table shows the median *LL* values, obtained over all patients for whom data was available for the corresponding $\{L_{\text{tr}}, L_{\text{fc}}\}$ pair; values closer to zero represent better performance. The bottom half shows pairwise median differences with respect to M_{FD} versus

[3] The corresponding table for the *RMSE* is not shown as it was very similar.

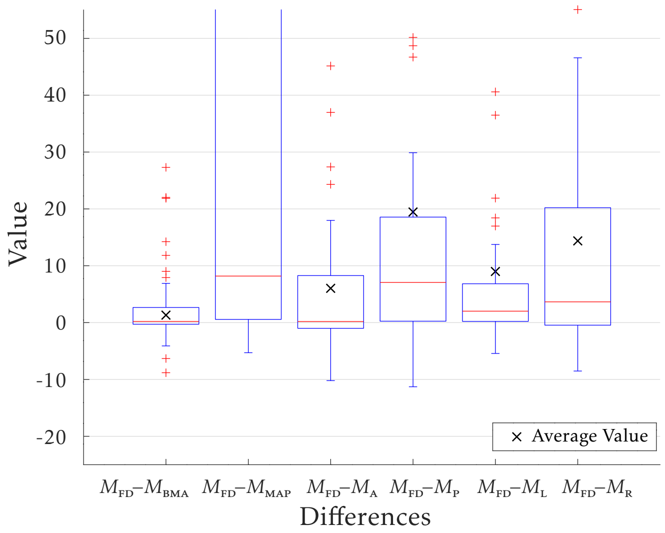

Figure 2. Box and whisker graph plots showing the median, interquartile range, and extreme cases of the *LL* differences when training with 3 weeks of past data and predicting 7 days ahead. Values above 0 represent cases where M_{FD} is better than its competitors.

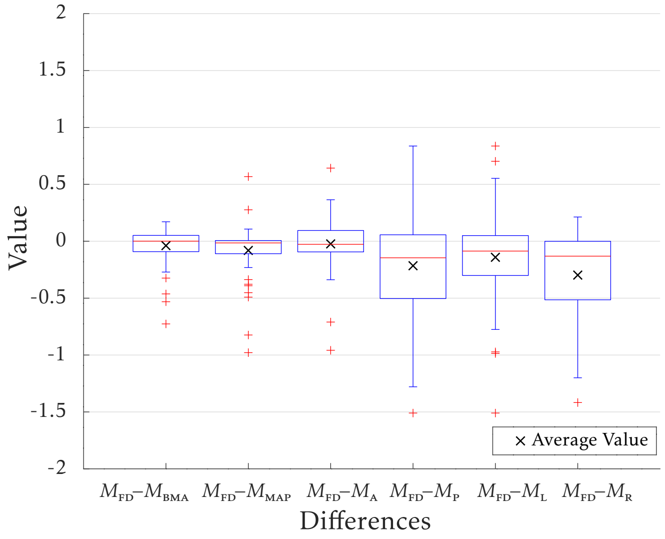

Figure 3. Like Fig. 2 but for *RMSE*. Values below 0 represent cases where M_{FD} is better than its competitors.

competing models. Note that because of the nature of these intervals, each $\{L_{tr}, L_{fc}, \}$ pair will consist of a different number of 'valid' participants (i.e. the participants who have data within this period), and therefore it is important to note that the above medians are calculated over different sets of patients. The last row shows the number of such valid participants per $\{L_{tr}, L_{fc}\}$ pair. Also note that, pairwise median differences are generally not equivalent to pairwise differences between medians.

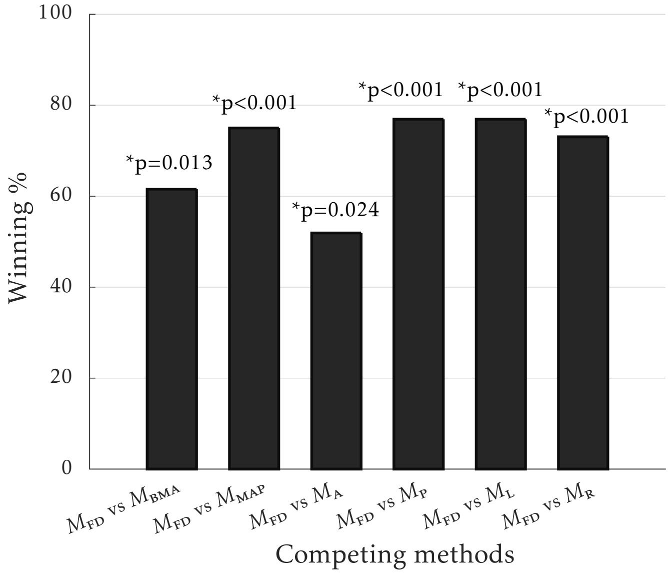

Figure 4. Comparison of the winning results based on *LL* for the M_{FD} against all competing models when training with 3 weeks of past data and predicting 7 days ahead. The * indicates statistical significance using a one-tailed Wilcoxon Signed-Rank test.

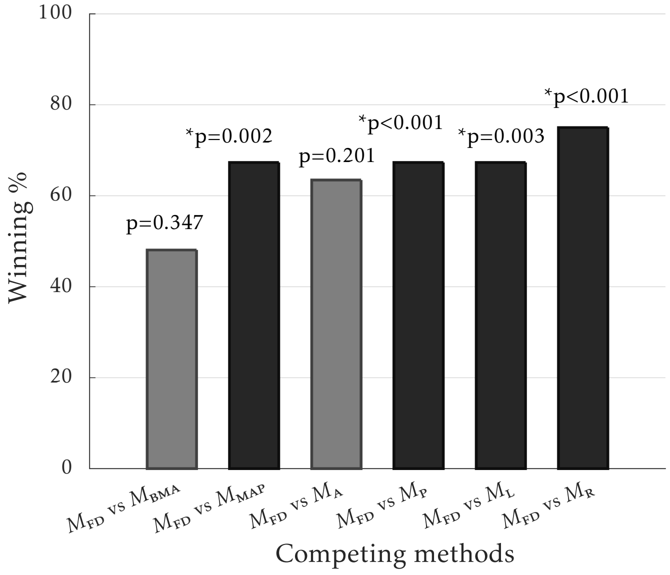

Figure 5. Like Fig. 4 but for *RMSE*.

As shown in the table, some of the methods under comparison initially struggle to make acceptable 7-day predictions, when only one week or three weeks of data are available to them (e.g. M_{MAP}, M_{P} and M_{R}); it is not until 7 weeks of training that most models can predict one or three days ahead with reasonable accuracy. By contrast, looking at M_{FD}, we see that, not only is it able to make predictions from week one, but it can even make 7-day predictions with remarkable stability for any number of training weeks. Similarly, as expected and consistent with the findings by Yu

Table 1. Medians of the *LL* (top) and of *LL*-differences (bottom) for various durations of training and forecast periods.

Model Type	$L_{tr} = 1$			$L_{tr} = 3$			$L_{tr} = 7$		
	L_{fc}=1	L_{fc}=3	L_{fc}=7	L_{fc}=1	L_{fc}=3	L_{fc}=7	L_{fc}=1	L_{fc}=3	L_{fc}=7
M_{FD}	-0.15	-2.12	-5.07	-0.76	-2.40	-5.54	-0.72	-1.09	-3.20
M_{BMA}	-1.23	-2.63	-4.87	-2.07	-3.59	-6.66	-1.23	-2.15	-3.61
M_{MAP}	0.49	-8.25	-26.60	-1.44	-5.80	-17.57	-1.41	-3.10	-6.32
M_A	-0.98	-3.67	-8.40	-1.80	-2.60	-11.03	-1.11	-1.40	-3.45
M_P	-5.97	-8.02	-14.80	-7.61	-7.93	-13.85	-2.14	-4.45	-12.80
M_L	-2.07	-4.14	-10.90	-1.95	-4.11	-10.03	-1.56	-3.37	-6.59
M_R	-1.33	-7.87	-33.61	-1.74	-4.13	-14.73	-1.14	-2.08	-3.76
Models Compared									
M_{FD} vs M_{BMA}	0.88*	0.60*	0.30*	0.71*	0.25*	0.20*	0.58*	0.57*	0.69*
M_{FD} vs M_{MAP}	-0.75	4.28*	22.96*	0.95*	1.78*	8.18*	0.17	0.16*	0.40*
M_{FD} vs M_A	0.54*	0.64*	2.27*	0.58*	0.05	0.19*	0.50*	0.23	-0.12
M_{FD} vs M_P	5.35*	4.24*	8.75*	5.77*	4.95*	7.07*	1.07*	2.87*	4.49*
M_{FD} vs M_L	1.66*	1.83*	4.70*	0.89*	0.50*	2.01*	0.73*	1.56*	1.32*
M_{FD} vs M_R	1.73*	6.95*	26.70*	0.29	0.24*	3.65*	0.29	0.09	0.09
participants	24	37	46	24	40	52	27	40	48

Annotations (bottom): Asterisks (*) denote entries where a statistically significant difference (p<0.05) was observed on one-tailed Wilcoxon Signed-Rank testing. Entries in blue indicate winning percentages above 50% (i.e. M_{FD} scored better than its corresponding competitor more than 50% of the time); entries in red indicate winning percentages above 90%; entries in black indicate winning percentages equal to or below 50%.

et al. [37], there is some reduction in accuracy going from next-day forecasts to 7-day forecasts, regardless of training length. However, it can be seen that the reduction in accuracy in the M_{FD} model tends to be relatively small compared to the competitors. Finally, as shown in the bottom half of Table 1, we can see that M_{FD} is superior to most methods in most conditions, with the median *LL* difference being significantly greater than zero in 45 out of 54 comparisons (83.3%), and scoring more 'wins' than its competitors in 50 out of 54 comparisons (92.6%). In fact, focusing on the M_{FD} vs M_{BMA} row, we see that the new method is also superior to its predecessor, under all scenarios considered; the choice L_{tr}=3 and L_{fc}=7 corresponding to the graphs shown earlier in Secs. 5.1 and 5.2 showing a relatively narrow interquartile range between the two, was simply selected on the basis that it represented the worst-case scenario for M_{FD} in relation to M_{BMA}.

6. Conclusions

The use of smartphones and wearable sensors for quantifying and predicting well-being states through personalised predictions, is an actively growing field, with potential applications in the prevention and self-management of depression and other disorders. Personalisation refers to the ability to learn a model, which is specifically tuned with respect to the individual it is intended to be applied to. However, a major obstacle to success in this field so far has been that the more traditional machine learning approaches to this — which typically require the availability of large datasets of uniformly sampled data — are generally not applicable to this domain. There are two main reasons for this; the first is that the kind of data provided by patients through smartphones and wearable sensors tend to be sporadic or intermittently available. The second is that, realistically, for such a personalised system to be useful, users need predictions virtually from day one, whereas in a typical situation the data available to the system for personalisation will initially be very limited, and acquired incrementally over time. As has been demonstrated in the sections above, it could take *weeks* before a decent amount of data has accumulated to guarantee reliable prediction.

In this paper, we have proposed a parametric transfer learning approach based on the Fisher divergence in the context of Hamiltonian Monte Carlo sampling and Bayesian inference to address these challenges. Our approach makes it possible to create patient-specific models and make useful predictions of self-reported well-being scores, even when the data available for initial training are sporadic and limited, such that training is performed incrementally as more data become slowly available over time. This approach allows us to make informed predictions even in the early stages of data collection, by leveraging external information coming from other patients, in the form of a prior used within a Markov-Chain Monte Carlo process.

We demonstrated this approach on data obtained by the NEVERMIND clinical trial, and measured its performance against previous work (e.g. the BMA method introduced in [43]), and a number of baseline approaches. Our results show that this approach yields a significant improvement over its competitors, and is particularly useful in difficult training/forecasting scenarios, e.g. when one requires a distant, patient-specific forecast, with only a limited initial amount of patient-specific data available for training.

One limitation of this study is that the Linear Dynamical System model used is limited to a single latent state, and the observations reflected questionnaire responses only. The addition of physiological signals in the observation vector could strengthen the model's predictive abilities further, albeit at the cost of increased complexity. Similarly, a single latent state reflecting well-being in a general sense may not be powerful enough to capture the underlying complexity of depressive states. It is possible that an appropriately extended latent state might allow a richer representation of the underlying biological states, as well as allow linking latent states to observations directly. All this will be addressed in future work.

Additionally, in this study we found that the background patient population acting as the source domain for the transfer learning component was informative, as shown by the performance of TL. However, a limitation is that we made no attempt to quantify, or investigate ways in which this background knowledge could be made more informative. Future work will focus on investigating whether applying preprocessing strategies that promote individuals in the population that are known to be similar in some way to the person being modeled, enhance transfer learning.

Finally, we recognise that our method has wider applicability to other domains, such as finance, recommender systems, training initiatives, etc, and generally any scenario where limited or sporadic data arrive in a sequential manner, and a seamless transition from generalised to personalised models is required. Therefore in the first instance, future work will also focus on verifying the performance and generality of this approach, both on the complete NEVERMIND dataset (which will become available at the end of the clinical trial), as well as other known external datasets (such as the MIMIC-III critical care database [57]).

In the meantime, the tools we have produced can already be used by clinicians and carers to monitor patients. However, the broader aim is to also use them in the context of a self-management tool. It is therefore important to recognise that however accurate such tools may be in the absence of user feedback, they also need to be evaluated when users are provided with the system's predictions. Since it is likely that patients will be affected by the system's forecasts, this has not only practical implications (which we believe the learning element of the system will likely be able to deal with automatically) but also ethical ones. It is issues of this type that the randomised controlled trial leg of the NEVERMIND project will be called to address.

Appendix A.

In this appendix we derive a relationship between the Fisher divergence $D_F(p\|q)$ from a given distribution q to a mixture distribution p and the Fisher divergences $D_F(p_i\|q)$ from q to p's mixture components. Given a mixture distribution $p(\mathbf{x}) = \sum_i w_i p_i(\mathbf{x})$ with $\sum_i w_i = 1$ and $w_i \geq 0$, the Fisher divergence from q to p can be computed as:

$$\begin{aligned} D_F(p\|q) &= \int_{\mathbf{x}} p(\mathbf{x})\, \|\nabla_{\mathbf{x}} \log p(\mathbf{x}) - \nabla_{\mathbf{x}} \log q(\mathbf{x})\|^2 \,\mathrm{d}\mathbf{x} \\ &= \int_{\mathbf{x}} p(\mathbf{x}) \left\| \frac{\nabla_{\mathbf{x}} p(\mathbf{x})}{p(\mathbf{x})} - \nabla_{\mathbf{x}} \log q(\mathbf{x}) \right\|^2 \mathrm{d}\mathbf{x} \\ &= \int_{\mathbf{x}} \left\{ \frac{\|\nabla_{\mathbf{x}} p(\mathbf{x})\|^2}{p(\mathbf{x})} - 2\, \nabla_{\mathbf{x}} p(\mathbf{x})^{\mathrm{T}} \nabla_{\mathbf{x}} \log q(\mathbf{x}) \right. \\ &\qquad \left. + \, p(\mathbf{x}) \|\nabla_{\mathbf{x}} \log q(\mathbf{x})\|^2 \right\} \mathrm{d}\mathbf{x}. \qquad \text{(A.1)} \end{aligned}$$

We rearrange equation (A.1) by adding and subtracting a convenient term, then breaking up the mixture distribution and regrouping, obtaining:

$$\begin{aligned} D_F(p\|q) &= \int_{\mathbf{x}} \left\{ \frac{\|\nabla_{\mathbf{x}} p(\mathbf{x})\|^2}{p(\mathbf{x})} - \sum_i w_i \frac{\|\nabla_{\mathbf{x}} p_i(\mathbf{x})\|^2}{p_i(\mathbf{x})} \right. \\ &\qquad + \sum_i w_i \left[\frac{\|\nabla_{\mathbf{x}} p_i(\mathbf{x})\|^2}{p_i(\mathbf{x})} - 2\, \nabla_{\mathbf{x}} p_i(\mathbf{x})^{\mathrm{T}} \nabla_{\mathbf{x}} \log q(\mathbf{x}) \right. \\ &\qquad \left. \left. + \, p_i(\mathbf{x}) \|\nabla_{\mathbf{x}} \log q(\mathbf{x})\|^2 \right] \right\} \mathrm{d}\mathbf{x} \\ &= J(p) - \sum_i w_i\, J(p_i) + \sum_i w_i\, D_F(p_i\|q), \qquad \text{(A.2)} \end{aligned}$$

where $J(p) = \int_{\mathbf{x}} p(\mathbf{x}) \|\nabla_{\mathbf{x}} \log p(\mathbf{x})\|^2 \mathrm{d}\mathbf{x}$ is the Fisher information [53] of p while $J(p_i)$ is that of p_i. This is especially useful when looking for the best approximation q_β to a mixture distribution p. Since $J(p)$ and $J(p_i)$ do not depend on β, the best approximation to the mixture distribution can be computed by minimising a weighted sum of the Fisher divergences between the approximant and the mixture components:

$$\hat{\beta} = \arg\min_{\beta} D_F(p\|q_\beta) = \arg\min_{\beta} \sum_i w_i\, D_F(p_i\|q_\beta). \qquad \text{(A.3)}$$

Acknowledgements. This work was supported by the European Commission under the NEVERMIND project (H2020-PHC-689691). The authors would also like to thank the NEVERMIND consortium partners [10] for providing the data and the relevant infrastructure, and the participants who took part in this study.

References

[1] American Psychiatric Association (2013) *Diagnostic and statistical manual of mental disorders (DSM-5®)* (American Psychiatric Pub).

[2] Jong, P.J.D., Sportel, B.E. and Hullu, E.D. (2012) Co-occurrence of social anxiety and depression symptoms in adolescence: differential links with implicit and explicit self-esteem? *Psychological Medicine* **42**(3): 475–484.

[3] Clarke, D.M. and Currie, K.C. (2009) Depression, anxiety and their relationship with chronic diseases: a review of the epidemiology, risk and treatment evidence. *Medical Journal of Australia* **190**(7): S54.

[4] Kang, H.J., Kim, S.Y., Bae, K.Y., Kim, S.W., Shin, I.S., Yoon, J.S. and Kim, J.M. (2015) Comorbidity of depression with physical disorders: research and clinical implications. *Chonnam medical journal* **51**(1): 8–18. doi:10.4068/cmj.2015.51.1.8.

[5] Katon, W., Lin, E.H.B. and Kroenke, K. (2007) The association of depression and anxiety with medical symptom burden in patients with chronic medical illness. *General Hospital Psychiatry* **29**(2): 147–155. doi:10.1016/j.genhosppsych.2006.11.005.

[6] Wulsin, L., Vaillant, G., Medicine, V.W.P. and undefined 1999 (1999) A systematic review of the mortality of depression. *Psychosomatic medicine* **61**(1).

[7] Shen, N., Levitan, M.J., Johnson, A., Bender, J.L., Hamilton-Page, M., Jadad, A.A.R. and Wiljer, D. (2015) Finding a depression app: a review and content analysis of the depression app marketplace. *JMIR mHealth and uHealth* **3**(1): e16.

[8] Pastor, J. and Van Breda, W. (2015) Analyzing and Predicting Mood of Depression Patients.

[9] Moreno-Torres, J.G., Raeder, T., Alaiz-RodríGuez, R., Chawla, N.V. and Herrera, F. (2012) A unifying view on dataset shift in classification. *Pattern recognition* **45**(1): 521–530.

[10] (2018), NEurobehavioural predictiVE and peRsonalised Modelling of depressIve symptoms duriNg primary somatic Diseases with ICT-enabled self-management procedures, Online at http://www.nevermindproject.eu.

[11] Carli, V., Wasserman, D., Hadlaczky, G., Petros, N.G., Carletto, S., Citi, L., Dinis, S. *et al.* (2020) A protocol for a multicentre, parallel-group, pragmatic randomised controlled trial to evaluate the nevermind system in preventing and treating depression in patients with severe somatic conditions. *BMC psychiatry* **20**(1): 1–10.

[12] Sousa, R., Silva, L.M., Alexandre, L.A. and Santos, J. (2014) Transfer Learning: Current Status, Trends and Challenges. In *20th Portuguese Conference on Pattern Recognition, RecPad*: 57–58. doi:10.1109/72.788640.

[13] Roy, D. and Kaelbling, L. (2007) Efficient Bayesian Task-Level Transfer Learning. *IJCAI* **7**: 2599–2604.

[14] Lu, J., Behbood, V., Hao, P., Zuo, H., Xue, S. and Zhang, G. (2015) Transfer learning using computational intelligence: A survey. *Knowledge-Based Systems* **80**: 14–23. doi:10.1016/j.knosys.2015.01.010.

[15] Maxhuni, A., Hernandez-Leal, P., Sucar, L.E., Osmani, V., Morales, E.F. and Mayora, O. (2016) Stress modelling and prediction in presence of scarce data. *Journal of Biomedical Informatics* **63**: 344–356. doi:10.1016/J.JBI.2016.08.023.

[16] Jaini, P., Chen, Z., Carbajal, P., Law, E., Middleton, L., Regan, K., Schaekermann, M. *et al.* (2017) Online Bayesian Transfer Learning for Sequential Data Modeling. In *ICLR 2017*.

[17] Taylor, M.E. and Stone, P. (2009) Transfer Learning for Reinforcement Learning Domains: A Survey. *Journal of Machine Learning Research* **10**: 1633–1685.

[18] Pan, S.J. and Yang, Q. (2010) A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering* **22**(10): 1345–1359. doi:10.1109/TKDE.2009.191.

[19] Cook, D., Feuz, K.D. and Krishnan, N.C. (2013) Transfer learning for activity recognition: A survey. *Knowledge and Information Systems* **36**(3): 537–556. doi:10.1007/s10115-013-0665-3.

[20] Wang, P., Lu, J., Zhang, B. and Tang, Z. (2015) A review on transfer learning for brain-computer interface classification. In *2015 5th International Conference on Information Science and Technology, ICIST 2015* (IEEE): 315–322. doi:10.1109/ICIST.2015.7288989.

[21] Weiss, K., Khoshgoftaar, T.M. and Wang, D. (2016) A survey of transfer learning. *Journal of Big Data* **3**(1): 9. doi:10.1186/s40537-016-0043-6.

[22] Christinaki, E., Papastylianou, T., Poli, R. and Citi, L. (2019) Parametric transfer learning based on the fisher divergence for well-being prediction. In *Proceedings - 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering, BIBE 2019* (Institute of Electrical and Electronics Engineers Inc.): 288–295. doi:10.1109/BIBE.2019.00059.

[23] Finkelstein, F.O. and Finkelstein, S.H. (2000) Depression in chronic dialysis patients: assessment and treatment. *Nephrology Dialysis Transplantation* **15**(12): 1911–1913.

[24] Bush, D.E., Ziegelstein, R.C., Tayback, M., Richter, D., Stevens, S., Zahalsky, H. and Fauerbach, J.A. (2001) Even minimal symptoms of depression increase mortality risk after acute myocardial infarction. *The American journal of cardiology* **88**(4): 337–341.

[25] Whooley, M.A., De Jonge, P., Vittinghoff, E., Otte, C., Moos, R., Carney, R.M., Ali, S. *et al.* (2008) Depressive symptoms, health behaviors, and risk of cardiovascular events in patients with coronary heart disease. *Jama* **300**(20): 2379–2388.

[26] Giese-Davis, J., Collie, K., Rancourt, K.M., Neri, E., Kraemer, H.C. and Spiegel, D. (2011) Decrease in depression symptoms is associated with longer survival in patients with metastatic breast cancer: a secondary analysis. *Journal of clinical oncology* **29**(4): 413.

[27] Zuidersma, M., Conradi, H.J., van Melle, J.P., Ormel, J. and de Jonge, P. (2013) Self-reported depressive symptoms, diagnosed clinical depression and cardiac morbidity and mortality after myocardial infarction. *International Journal of Cardiology* **167**(6): 2775 – 2780. doi:https://doi.org/10.1016/j.ijcard.2012.07.002, URL http://www.sciencedirect.com/science/article/pii/S0167527312009394.

[28] Roest, A.M., Heideveld, A., Martens, E.J., de Jonge, P. and Denollet, J. (2014) Symptom dimensions of anxiety following myocardial infarction: Associations with depressive symptoms and prognosis. *Health Psychology* **33**(12): 1468.

[29] Sani, O.G., Yang, Y., Lee, M.B., Dawes, H.E., Chang, E.F. and Shanechi, M.M. (2018) Mood variations decoded from multi-site intracranial human brain activity. *Nature Biotechnology* **36**(10): 954–961. doi:10.1038/nbt.4200.

[30] Henriquez, P., Matuszewski, B.J., Andreu, Y., Bastiani, L., Colantonio, S., Coppini, G., D'Acunto, M. *et al.* (2017) Mirror Mirror on the Wall... An Unobtrusive Intelligent Multisensory Mirror for Well-Being Status Self-Assessment and Visualization. *IEEE Transactions on Multimedia* **19**(7): 1467–1481. doi:10.1109/TMM.2017.2666545.

[31] LiKamWa, R., Liu, Y., Lane, N.D. and Zhong, L. (2013) MoodScope: building a mood sensor from smartphone usage patterns. In *Proceeding of the 11th annual international conference on Mobile systems, applications, and services - MobiSys '13* (New York, New York, USA: ACM Press): 389–402. doi:10.1145/2462456.2464449, URL http://dl.acm.org/citation.cfm?doid=2462456.2464449.

[32] Saeb, S., Zhang, M., Schueller, S., Christopher, ., Karr, J., Stephen, M.., Schueller, M. *et al.* (2015) Mobile Phone Sensor Correlates of Depressive Symptom Severity in Daily-Life Behavior: An Exploratory Study. *Article in Journal of Medical Internet Research* **17**(7): e175. doi:10.2196/jmir.4273.

[33] Taylor, S., Jaques, N., Nosakhare, E., Sano, A., Klerman, E.B. and Picard, R.W. (2017) Importance of Sleep Data in Predicting Next-Day Stress, Happiness, and Health in College Students. *Journal of Sleep and Sleep Disorders Research* **40**(suppl_1): A294–A295.

[34] Taylor, S.A., Jaques, N., Nosakhare, E., Sano, A. and Picard, R. (2017) Personalized Multitask Learning for Predicting Tomorrow's Mood, Stress, and Health. *IEEE Transactions on Affective Computing* doi:10.1109/TAFFC.2017.2784832.

[35] Suhara, Y., Xu, Y. and Pentland, A.S. (2017) DeepMood: Forecasting Depressed Mood Based on Self-Reported Histories via Recurrent Neural Networks. In *Proceedings of the 26th International Conference on World Wide Web - WWW '17* (New York, New York, USA: ACM Press): 715–724. doi:10.1145/3038912.3052676.

[36] Spathis, D., Servia-Rodriguez, S., Farrahi, K., Mascolo, C. and Rentfrow, J. (2019) Sequence Multi-task Learning to Forecast Mental Wellbeing from Sparse Self-reported Data. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Association for Computing Machinery): 2886–2894. doi:10.1145/3292500.3330730.

[37] Yu, H., Klerman, E.B., Picard, R.W. and Sano, A. (2019) Personalized Wellbeing Prediction using Behavioral, Physiological and Weather Data. In *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)* (IEEE): 1–4. doi:10.1109/BHI.2019.8834456.

[38] Sano, A., Taylor, S., McHill, A.W., Phillips, A.J., Barger, L.K., Klerman, E. and Picard, R. (2018) Identifying Objective Physiological Markers and Modifiable Behaviors for Self-Reported Stress and Mental Health Status Using Wearable Sensors and Mobile Phones: Observational Study. *Journal of medical Internet research* **20**(6): e210. doi:10.2196/jmir.9410.

[39] Yosinski, J., Clune, J., Bengio, Y. and Lipson, H. (2014) How transferable are features in deep neural networks? In *Advances in neural information processing systems*: 3320–3328.

[40] Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A. and Rubin, D.B. (2013) *Bayesian data analysis* (Chapman and Hall/CRC).

[41] Bishop, C.M. (2006) *Pattern recognition and machine learning* (Springer), chap. 11. Sampling Methods, 537–538.

[42] Li, X., Poli, R., Valenza, G., Scilingo, E.P. and Citi, L. (2017) Self-reported well-being score modelling and prediction: Proof-of-concept of an approach based on linear dynamic systems. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (IEEE): 2205–2208.

[43] Christinaki, E., Poli, R. and Citi, L. (2018) Bayesian Transfer Learning for the Prediction of Self-reported Well-being Scores. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (IEEE): 41–44. doi:10.1109/EMBC.2018.8512255.

[44] Goodman, J. and Weare, J. (2010) Ensemble Samplers With Affine Invariance. *Communications in Applied Mathematics and Computational Science* **5**(1): 65–80. doi:10.2140/camcos.2010.5.65.

[45] Foreman-Mackey, D., Hogg, D.W., Lang, D. and Goodman, J. (2013) emcee : The MCMC Hammer. *Publications of the Astronomical Society of the Pacific* **125**(925): 306–312. doi:10.1086/670067.

[46] Huijser, D., Goodman, J. and Brewer, B.J. (2015) Properties of the Affine Invariant Ensemble Sampler in high dimensions. *arXiv preprint arXiv:1509.02230* 1509.02230.

[47] Neal, R.M. (2011) MCMC using Hamiltonian dynamics. In Steve Brooks, Andrew Gelman, Galin Jones, X.L.M. [ed.] *Handbook of Markov Chain Monte Carlo* (CRC Press), chap. 5, 113–162.

[48] Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. *et al.* (2017) Stan: A probabilistic programming language. *Journal of statistical software* **76**(1).

[49] Hoffman, M. and Gelman, A. (2014) The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research* **15**(1): 1593–1623.

[50] Hoeting, J.A., Madigan, D., Raftery, A.E. and Volinsky, C.T. (1999) Bayesian Model Averaging: A Tutorial. *Statistical Science* **14**(4): 382–401. doi:10.2307/2676803.

[51] Hyvärinen, A. (2005) Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research* **6**(Apr): 695–709.

[52] Hyvärinen, A. (2007) Some extensions of score matching. *Computational statistics & data analysis* **51**(5): 2499–2512.

[53] Lyu, S. (2009) Interpretation and generalization of score matching. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence* (AUAI Press): 359–366.

[54] Andersen, M., Dahl, J., Liu, Z. and Vandenberghe, L. (2011) Interior-point methods for large-scale cone programming. In Sra, S., Nowozin, S. and Wright, S. [eds.] *Optimization for Machine Learning* (MIT Press), chap. 3.

[55] Vehtari, A., Gelman, A., Simpson, D., Carpenter, B. and Bürkner, P.C. (2019) Rank-normalization, folding, and localization: An improved $\widehat{r}$ for assessing convergence of mcmc. *arXiv preprint arXiv:1903.08008* .

[56] Hollander, M., Wolfe, D.A. and Chicken, E. (2013) *Nonparametric statistical methods*, **751** (John Wiley & Sons).

[57] Johnson, A.E., Pollard, T.J., Shen, L., Li-wei, H.L., Feng, M., Ghassemi, M., Moody, B. *et al.* (2016) Mimic-iii, a freely accessible critical care database. *Scientific data* **3**: 160035.

Supplementary Material

Table S1. Summary of results using stan for the parameters of interest estimated by the samples for a single participant

Parameter	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
a_{01}	0.0863	0.0029	0.0871	–0.0770	0.0278	0.0856	0.1426	0.2599	882.9309	1.0010
a_{02}	0.0032	0.0061	0.1497	–0.3080	–0.0944	0.0076	0.1004	0.2788	604.7551	0.9982
a_{03}	0.1771	0.0144	0.3475	–0.5423	–0.0506	0.1996	0.4243	0.8121	585.1530	1.0006
a_{04}	0.0006	0.0027	0.0855	–0.1759	–0.0538	0.0027	0.0603	0.1637	1019.4277	1.0004
a_{05}	0.1319	0.0054	0.1451	–0.1602	0.0338	0.1338	0.2388	0.3981	717.8639	1.0029
a_{11}	0.0536	0.0041	0.1454	–0.2280	–0.0340	0.0480	0.1435	0.3494	1257.8595	1.0007
a_{12}	–0.0532	0.0079	0.2459	–0.5414	–0.2140	–0.0448	0.1095	0.4344	959.2048	1.0040
a_{13}	0.5468	0.0124	0.3718	–0.1838	0.2800	0.5573	0.7887	1.2924	892.5546	1.0028
a_{14}	0.5410	0.0056	0.1582	0.2074	0.4403	0.5527	0.6531	0.8261	802.2508	1.0018
a_{15}	–0.0049	0.0079	0.2330	–0.4417	–0.1745	–0.0086	0.1596	0.4523	861.1409	1.0042
a_{21}	0.0718	0.0046	0.1517	–0.2087	–0.0301	0.0621	0.1736	0.3867	1081.4613	1.0019
a_{22}	–0.2019	0.0107	0.2570	–0.6839	–0.3797	–0.2204	–0.0287	0.3285	573.6596	1.0013
a_{23}	0.3631	0.0119	0.3585	–0.3241	0.1276	0.3621	0.6134	1.0868	908.1137	0.9985
a_{24}	–0.1932	0.0049	0.1449	–0.4598	–0.2937	–0.1999	–0.0907	0.0972	887.2103	0.9997
a_{25}	–0.1025	0.0065	0.1899	–0.4588	–0.2386	–0.1038	0.0286	0.2722	847.3869	1.0032
$S_{x_{00}}$	0.2353	0.0032	0.0917	0.0819	0.1731	0.2233	0.2862	0.4436	811.2320	1.0041
$S_{x_{11}}$	1.3516	0.0138	0.4259	0.7136	1.0479	1.2784	1.5717	2.3702	949.6742	0.9984
$S_{x_{22}}$	0.5590	0.0118	0.3038	0.0943	0.3412	0.5125	0.7272	1.3054	665.3888	1.0022
ξ_1	1.0116	0.0381	1.0741	–1.5606	0.4478	1.1017	1.7275	2.8530	796.3929	1.0065
ξ_2	1.2758	0.0478	1.5653	–2.1297	0.2680	1.4053	2.3020	4.2206	1073.7744	0.9990
ξ_3	0.7028	0.0703	1.8407	–3.2101	–0.5245	0.7903	1.9692	4.0550	685.7082	1.0085
$lp\_\_$	-262.1961	0.1446	3.0299	-268.7097	-264.1561	-261.8284	-259.9745	-257.0868	438.9218	1.0004

Note: Rows correspond to model parameters, and columns to the various summary metrics. **mean** denotes the posterior mean, **se_mean** denotes the Monte Carlo standard error, and **sd** denotes the posterior standard deviation. The numbers 2.5%, 25%, 50%, 75%, and 97.5% denote quantiles. **n_eff** denotes the effective sample size, and **Rhat** denotes the split-$\hat{R}$ statistic.