Meaningless comparisons lead to false optimism in medical machine learning

Orianna DeMasi[1]*, **Konrad Kording[2,3]**, **Benjamin Recht[1]**

1 Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, California, United States of America, **2** Department of Bioengineering, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America, **3** Department of Neuroscience, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America

* odemasi@eecs.berkeley.edu

Abstract

A new trend in medicine is the use of algorithms to analyze big datasets, e.g. using everything your phone measures about you for diagnostics or monitoring. However, these algorithms are commonly compared against weak baselines, which may contribute to excessive optimism. To assess how well an algorithm works, scientists typically ask how well its output correlates with medically assigned scores. Here we perform a meta-analysis to quantify how the literature evaluates their algorithms for monitoring mental wellbeing. We find that the bulk of the literature ($\sim$77%) uses meaningless comparisons that ignore patient baseline state. For example, having an algorithm that uses phone data to diagnose mood disorders would be useful. However, it is possible to explain over 80% of the variance of some mood measures in the population by simply guessing that each patient has their own average mood—the patient-specific baseline. Thus, an algorithm that just predicts that our mood is like it usually is can explain the majority of variance, but is, obviously, entirely useless. Comparing to the wrong (population) baseline has a massive effect on the perceived quality of algorithms and produces baseless optimism in the field. To solve this problem we propose "user lift" that reduces these systematic errors in the evaluation of personalized medical monitoring.

Introduction

Health care should be tailored to individuals to maximize their wellbeing and health [1]. There is considerable hope that data collected from emerging data sources, such as smartphones and smartwatches, can be used to extract medical information and thus improve the tailoring of monitoring, diagnostics, and treatments for personalizing health care [2]. In particular, mental health care could particularly benefit from automated monitoring, as many mental health conditions need long-term monitoring and clinical monitoring is expensive, but automatically tracking a user with ubiquitous sensors is cheap [3–5].

Machine learning algorithms are commonly being used in an attempt to extract medical information from easy to collect data sources [6–9]. These algorithms are attractive as, by

automating information extraction, they promise to provide rich analyses cheaply and objectively based on collected data. Machine learning works by taking data that are easy to collect, building a model, and then using the model to make predictions for data that are harder to collect [10]. As an example, social media posts may be used to predict individuals' depressive symptoms or future suicidal ideation [11, 12]. However, without sufficient evaluation, the outputs of algorithms may be meaningless and mislead clinicians.

Whenever algorithms are used to make predictions, they must carefully be evaluated to ensure that their predictions meaningfully represent medically relevant information. Evaluation must be specified for each problem [13]. For example, if an algorithm is being used to predict one of two things, such as whether a patient is depressed, then it could be evaluated by the percent of predictions that are correct [9]. Alternatively, it could be evaluated by the percent of times that it correctly identified depression, i.e. sensitivity or true positive rate, and ascribe less importance to false positives [14]. In contrast, if an algorithm is trying to predict a value, such as someone's level of depressive symptoms, one could consider the degree to which predictions differ, i.e., the mean squared error. There are myriad additional methods for evaluating algorithms because, without sufficiently evaluating algorithms, it is easy to generate misplaced optimism about the utility of algorithms [15–18].

Regardless of how the correctness of an algorithm is quantified, algorithms must be compared to a baseline approach that simply makes guesses to prove that the algorithm makes better predictions than guessing. For example, if an algorithm is trying to predict a rare event, such as a mental breakdown or suicide, an approach that simply guesses that the event never happens will usually be correct and thus will have high accuracy [19–21]. However, such an approach is entirely useless for medicine. If algorithms are not compared with reasonable guesses, the accuracy of the algorithm's predictions can appear to be good, when in reality the algorithm is doing no better than guessing and is thus medically useless.

Here we review, for modeling of longitudinal individual state, what baselines algorithms are commonly compared against and how much of the apparent success of algorithms can be ascribed to poor comparisons. We focus on the example of mental wellbeing and demonstrate in two popular datasets that individuals exhibit little variance over time. Typical wellbeing prediction algorithms seem to work well, but we find that this is simply because they are basically always guessing individuals' personal average states. This example highlights how falsely optimistic results can easily be obtained by comparing machine learning with population as opposed to personal baselines. We perform a systematic literature review and find that most studies ($\sim$77%) compare with the population baseline. By not comparing with personal baselines, studies are prone to making falsely optimistic conclusions that can unintentionally mislead researchers' perspectives and delay progress on important medical applications. We argue for a new measure, "user lift," that measures the benefit of an algorithm relative to the single-person model.

Methods

Algorithm evaluation

There are many ways to evaluate how good an algorithm's predictions are [10, 13, 16, 18]. The general approach is a two step process of measuring an algorithm's error, or how inaccurate its predictions are, and then comparing the algorithm's error with the error of simply guessing answers. These guesses form a baseline approach and could be specific to each patient, or they could use other trivial factors, e.g. the time of the day. Regardless, because it is totally useless for medicine to simply guess answers based on subject and other trivial factors, algorithms must have lower error than such baselines to be of any use.

Algorithm error

To evaluate how well an algorithm predicts a binary outcome, e.g., whether an individual is having a happy vs. sad or stressed vs. relaxed day, we consider the classical measure of prediction error. Prediction error is the percent of observations that were incorrectly predicted (percent incorrect). To evaluate how well an algorithm predicts an individual's level of happiness or stress, we consider the root mean squared error (RMSE), which considers how different the predicted levels are from the true reported levels [10]. With both prediction error and RMSE, lower values indicate that an algorithm is doing better at predicting an individual's state. Higher values indicate more significant prediction error.

Baselines

We consider two baseline methods that simply guess how an individual is doing: personal baselines and population baselines. Both baselines are simple approaches that always guess individuals are at the same state. The personal baseline always guesses that each individual is at a constant state, but that state can differ between individuals. The population baseline predicts that all individuals are always at the same state.

When an algorithm is attempting to predict whether an individual is having a stressed (or happy) day or not, we consider personal baseline error, which is the prediction error of always guessing that each individual is always at their most frequently reported state (mode). We also consider the population baseline error, which is the error of estimating that all individuals are always at the most frequently reported state of the population (mode).

When an algorithm is attempting to predict an individual's level of happiness or stress, we consider the personal baseline RMSE, which is how far predicted levels were from always guessing each individual to be at their average level of stress or happiness. RMSE indicates a model with higher error and thus worse predictions. We also consider the population baseline RMSE, which is the RMSE of estimating that all individuals are always at the average state of the population.

User lift

We propose the measure of user lift as a way to evaluate whether an algorithm is making better predictions than simply guessing an individual's state. The user lift is the improvement of an algorithm's predictions over the personal baseline, or the amount that error is decreased by adding better features and a model. User lift is the difference between personal baseline error and model error in RMSE or in prediction error (personal baseline error—model error). The user lift can be thought of as the increase in accuracy of an algorithm over the null accuracy of guessing an individual to be at their average state. The average user lift is the mean user lift across the individuals in the dataset.

User lift framework

As a stricter measure of whether algorithms have any utility, we suggest the user lift framework instead of comparison with a single weak baseline, such as the population baseline. With this framework, researchers calculate user lift for each study participant. The user lift quantifies whether an algorithm is better than the simple personal baseline on each user. We propose then reporting descriptive statistics on the distribution of user lift and utilizing statistical tests to determine whether the average user lift is greater than zero. Nonparametric permutation tests are appropriate and powerful tests for considering whether a single sample, such as of user lifts on study participants, has a mean greater than zero. A permutation test is appropriate

here so that no assumptions on distributions are needed. While other nonparametric tests, such as the paired Wilcoxon signed-rank test, may be appropriate for comparing two samples, permutation tests have been reported to be more reliable than paired non-parametric tests [22, 23].

Machine learning example: Predicting subjective state from location and mobility

We present an example of how falsely optimistic conclusions can be reached about algorithms' performance. For this example, we follow previous works that have used and suggested that smartphone GPS location data can predict individuals' mental wellbeing [7, 9, 24]. We follow these studies' methodologies only for constructing features, or describing individuals' daily behavior from their GPS location logs. The dataset, testing methodology, and prediction tasks that we explore differ from these prior works.

Datasets

We consider two well established datasets that are freely available. Both datasets collected individuals' smartphone data, specifically GPS location, and their stress and happiness levels. The StudentLife dataset [25] followed a cohort of students at an American university during the course of a semester. Data that was collected included daily measures of stress on a five point Likert scale. Of the initial 48 students with data accessible, we consider data for the 15 students who had sufficient data available: stress level and at least 35 GPS location observations for at least 30 days of the study period.

The second dataset we consider, the MIT Friends and Family dataset [26], resulted from a project that collected various types of data on a cohort of university affiliates and their families at another American university. The data collected included daily wellbeing measures. Here we consider the nine point Likert scale of happiness and seven point Likert scale of stress that were collected. Of the 116 participants included in the available dataset we consider data for the 31 individuals who had measurements of stress or happiness, respectively, for at least 30 study days and at least 35 GPS location measurements on those days.

Data processing: Location and mobility features

To derive meaningful features of location and mobility, we follow three previous studies [7, 9, 24]. All features from these studies that were reproducible (due to the data available) were included. Before constructing features, we used two preprocessing methods.

The first preprocessing method fit a Gaussian Mixture Model (GMM) to all of the location samples for each participant collected to identify locations frequented by participants [24]. The number of clusters was chosen to be the number, up to twenty maximum, that minimized the Bayesian Information Criterion [27]. It was assumed that participants would frequent at most twenty locations during the course of the study. The home location of a participant was determined to be the location where the participant spent the majority of their time during the evening hours (11pm—6am) and the work location was similarly determined to be where the participant spent the majority of their time during working hours (11am—4pm). In contrast to prior work, we did not interpolate the location observations to a regular time sampling, as we did not find this beneficial to prediction accuracy [24]. We consider this first set of clusters to be the full clustering.

The second preprocessing method used K-means clustering on stationary points only [9, 28]. The StudentLife dataset included a prediction of whether the participant was moving or stationary at each observation, but the Friend and Family dataset did not. To determine

whether participants in the Friends and Family dataset were stationary at each observation, we approximated movement speed with the time derivative at each observation and used a threshold. We attempted to set the threshold to be about 1km/h [9]. We consider this second set of clusters as the stationary clustering and the night cluster to be the cluster where each individual spent the most time between midnight and 6am.

To protect participants' anonymity, the GPS location data in the Friends and Family dataset was subjected to an affine transform before being released. Because this transform purposefully changes the space, but collinearity should be preserved, we approximated features in one dimension on the Friends and Family dataset.

Utilizing the two set of location clusters that resulted from the GMMs and Kmeans, the features of mobility and location that we derived for each participant each day of the study are as follows:

1. The fraction of a day that a participant spent not stationary.

2. The average displacement of a participant between two observations during the day, i.e., average speed.

3. The standard deviation of displacements between points.

4. The location variance (on log scale), i.e., the sum of the variance of location coordinates in each dimension.

5. The "circadian movement" of a participant [9], which we adapted to our daily monitoring setting as the Euclidean distance of the vector of fraction of time a participant spent in each of their stationary location clusters with the participant's mean location distribution. The mean location distribution of a participant was calculated as the average fraction of a day that a participant would spend in each stationary location cluster during the study.

6. The location entropy, which was calculated as the entropy of the vector where each entry represented the fraction of the day that a participant spent in each stationary location cluster.

7. The radius of minimum circle enclosing the participant's location samples.

8. The fraction of time a participant spent at their GMM home cluster.

9. The fraction of time a participant spent at their GMM work cluster.

10. The fraction of time a participant spent at their stationary night cluster.

11. The log likelihood of a day from the GMM to estimate how routine the day was.

12. The AIC and BIC of the GMM evaluated with the day's coordinates, to also determine how typical the day was.

13. The number of GMM clusters visited in a day.

14. The number of stationary clusters visited in a day.

Experimental framework

We present two prediction tasks:

1. Predicting whether a participant was happy or stressed or not on a given day.

2. Predicting the average level of happiness or stress that a participant reported on a given day.

To construct levels of stress or happiness on a given day, we average all the Likert scale responses that a participant reported on that day. Whether the participant was happy (or stressed) or not is defined by a threshold on the daily average on a value to distinguish when students reported any stress versus no stress. For the StudentLife user inputs, we use "A little stressed" as the threshold. For the Friends and Family dataset, we use the middle value of the Likert scale as the threshold, as the Friends and Family scales were defined from negative to positive values, where the middle value was supposed to indicate a neutral state.

For both problems, we attempt to predict the stress or happiness from the location and mobility data with a variety of standard machine learning methods. For regression we consider: linear regression with an Elastic Net penalty, and Lasso regression [29, 30]. For the binary classification task, we consider: logistic regression with L2 penalty, support vector machines with radial basis function kernels, and random forests [10, 31]. Hyperparameters were chosen with 10-fold cross-validation on the training data. The methods that return the lowest error are presented.

We consider both population models, which could also be referred to as global, general, or all-user models and utilize all the individuals' data to make predictions, and personal models, which use only a single individual's data to make predictions for that individual. Prediction error is measured with leave-one-out cross-validation, which is commonly used for estimating an algorithm's prediction error [10]. To perform leave-one-out cross-validation on population models, we combine data from all of the participants into a single set. Then one observation is withheld, a model is trained on all of the other observations, and then that model is used to make a prediction for the held out observation. The process is repeated until every observation has been withheld exactly once. The model error reported from this process is the average error across the predictions for each data point. Population models assume that some of each participant's data is seen during training, in addition to data from other participants [32]. For personal models, we similarly hold out one observation, but we only train on the remaining observations of that individual's data and then repeat only for the number of observations that we have on that individual. Personal models only attempt to extrapolate predictions for an individual from their own data. Alternative cross validation schemes, such as N-fold, offered no benefit to the results, so are omitted for brevity.

Literature review

In addition to an example on two real datasets of how false machine learning results can be arrived at by comparing to weak baseline models, we perform a systematic literature review to investigate how algorithms are commonly evaluated and whether baselines are sufficiently reported. Our literature review took three steps:

1. Find relevant literature.

2. Establish whether a baseline (personal or population) was compared with.

3. Identify the error of the baselines and the best reported machine learning algorithms.

Finding relevant literature

While baselines are needed to evaluate all machine learning algorithms on personal data, we make our literature review tractable by focusing on studies similar to the machine learning example we present. We utilize GoogleScholar to find publications that attempted to automatically infer an individual's subjective states, similar to the example we presented. The studies we include meet the following criteria:

- Relate to subjective personal data, as denoted by having one of the following words in the title: depression, depressive, stress, mood, mental, happiness, or wellness.

- Attempt a machine learning prediction task and report prediction accuracy by having the following word "accuracy" somewhere in the text of the publication.

- Attempt prediction on participants' longitudinal data, where personal baselines are defined, by containing the words "participant" or "user".

- Collect data from sensors by requiring one of the following words to be included somewhere in the text: smartphones, sensor, sensors, or sensing.

- Were published since 2010.

Because of a particularly strong focus on stress in previous work, we break the query into two queries: one that requires the word "stress" to be in the title and another search that requires any of the other wellbeing words to be in the title. We perform this joint search with the following GoogleScholar queries:

(participant OR user) accuracy (sensor OR sensors OR smartphones OR sensing)intitle: stress (from 2010)

(participant OR user) accuracy (sensor OR sensors OR smartphones OR sensing) (intitle: mental OR intitle:depression OR intitle:depressive OR intitle:mood OR intitle:happiness) (from 2010)

To be considered relevant, studies need to attempt to predict user input data of users' subjective state from other collected data, i.e. sensors. Examples of studies that are returned by our query, but are excluded from our analysis are:

- Correlational analyses that reveal certain data or behaviors are correlated with subjective state.

- Studies of one-time user surveys (in contrast to repeated prompts) or where the goal is to separate subjects, i.e., each subject was a data point.

- Literature summaries or reviews.

- Randomized control studies of intervention efficacy.

- Other evaluations of treatments on subjective state.

- Collection and presentation of a dataset collected without a prediction task.

- Measurements of behaviors without attempting prediction.

- Descriptions of tools and systems implemented with user reviews of the systems.

- Non-peer reviewed publications, such as reports and book chapters.

- Prediction of non-subjective states, e.g., prediction of labels coded by researchers who intuit what state the user was in from observational data, or labels of stimulus exposure when studies attempted to induce a given emotional state such as stress.

We only consider studies where labels are for multiple observations of a participant's subjective input state.

Establishing comparison with a baseline

Some studies do not report any baseline model for comparison, so we begin by noting which studies reported a baseline model. For studies that provided sufficient detail, we did the following:

- When baseline models are reported we recorded the baseline performance metrics directly from the text and the type of baseline used, e.g., a population baseline or a user baseline.

- When baseline models are not provided, but confusion matrices are provided we manually calculate the baseline performance.

- When individuals baselines are reported, we take the average user baseline performance.

- When only mean squared error are reported, we note whether the mean squared error is also provided for a constant baseline.

Comparing the accuracy of different models

There are a wide variety of performance metrics authors report when evaluating their models. We extract model prediction error for multi-class classification problems according the to following criteria:

- When results are broken down for personal models by individual, the average is used.

- When accuracy results are given for multiple objectives, e.g., different dimensions of mood, the best results for each objective is recorded.

- When multiple feature sets and models were tried, only the best performing model is considered. Models that utilize user input as features were excluded when possible.

- The number of folds in the cross validation scheme used is not incorporated into our analysis. We considered 10-fold, leave-one-out, and leave-user-out cross validation schemes to all be "population" models. Both 10-fold and leave-one-out cross validations on personal data only are considered to be "personal" models.

- The uniform baseline was calculated by noting the number of classes that the study reported using in their measurement scale.

Results

We want to consider to what extent the choice of baselines matter in medical machine learning and how baselines are used in practice. To quantify the importance of baselines, we use two publicly available datasets and compare the performance of machine learning algorithms to two different baselines: a population baseline and a personal baseline. More specifically, we use the StudentLife [25] and Friends and Family [26] datasets and analyze machine learning predictions of stress and happiness, which we compare to both personal and population baselines. To understand how the field generally uses baselines, we perform a systematic literature review. These two complementary analyses will allow us to meaningfully inform the debate about machine learning in medicine.

Initially, we find, that individual subjects have little variance over time, relative to the variance across the population, i.e. low personal baseline error relative to higher population baseline error (Fig 1). Thus, comparing learned models with population baselines can obscure whether a model is better (lower error) on individuals than constant personal baseline models.
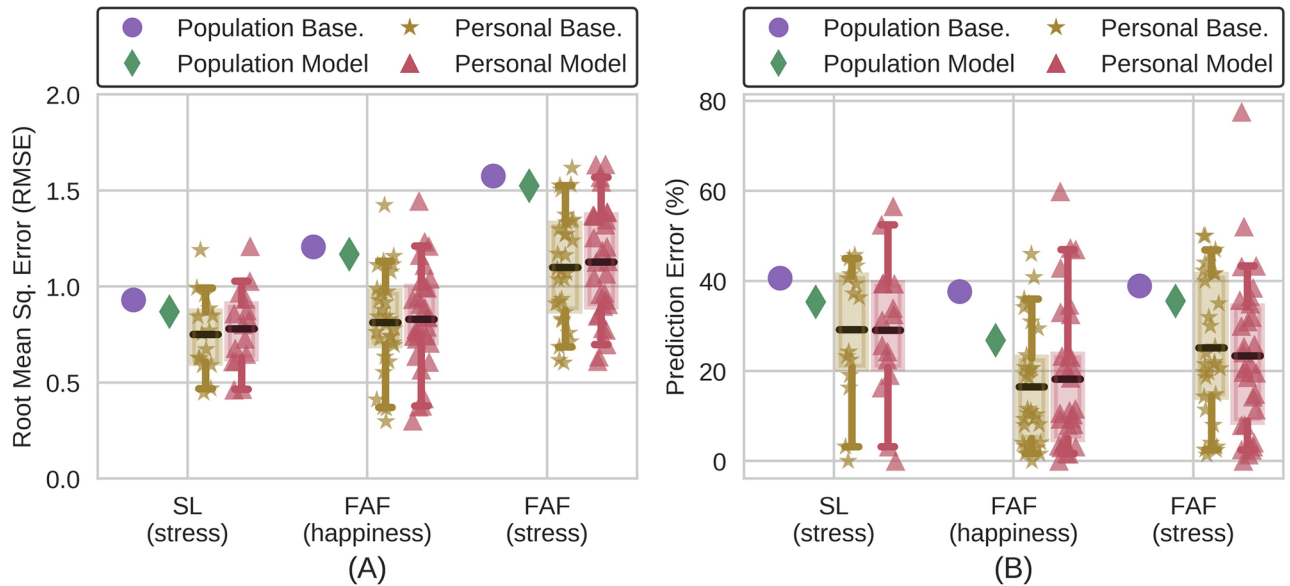
Fig 1. Results of machine learning models on StudentLife (SL) and Friends and Family (FAF) datasets. Bars represent the 5th and 95th percentiles, black lines indicate means, and boxes indicate the 1st and 3rd quartiles of error incurred on individuals. Personal models yield lower error than population models and population baselines, which often leads researchers to the conclusion that personal models are successful. Comparing personal models with personal baselines reveals that their error is no lower, so algorithms are doing no better than predicting individuals to be their most frequently reported state. The models presented are those with lowest error.

https://doi.org/10.1371/journal.pone.0184604.g001

We find the same pattern when we ask about RMSE and binary predictions. This gives us an intuition that guessing each subject's mean value should produce relatively low errors.

Motivated by the intuition that there is little within-subject and more across-subject variance, we now ask how machine learning algorithms compare to the two baselines. In line with prior literature [7, 24], the algorithms predict whether an individual was having a particularly stressful or happy day from their GPS location and mobility data (Fig 1). Our binary results are comparable (max difference = 6%) to past studies predicting binary stress or emotion from similar datasets [24, 33–38], as are the errors of the personal models [7, 37, 39–44]. Similarly, the difference between RMSE of personal models and personal baselines is comparable to the differences reported in prior publications [45, 46]. Our algorithms are much better than the population baseline and population models. They are not, however, lower than the personal baselines. This shows how good performance relative to the population baseline can be entirely meaningless.

To prevent comparing with the wrong baseline and to control against obscuring the range of how well algorithms do on individuals with aggregate statistics, we propose using statistical tests with the metric of user lift to prove that an algorithm is doing significantly better than the personal baseline. User lift is the difference of the personal model with the personal baseline, as described above. Positive user lift indicates that a model is better than the personal baseline, that the algorithm's predictions are more accurate than always assuming an individual is at their average state. Indeed, user lift shows that our naive model is useless while our moderately careful model at least adds something (Table 1). Using the wrong baseline, may make bad machine learning with a performance that is by any meaningful definition useless seem impressive underscoring the importance of meaningful baselines.

To understand how algorithms are typically evaluated, we perform a systematic literature review of related studies that attempted to predict emotion and stress from sensed data, such

Table 1. Statistical significance for user lift of personal models in Fig 1. The user lifts are the differences of personal baselines with personal models, in terms of prediction error or RMSE. The p-values are for permutation tests considering whether the user lifts were larger than zero. In every case the user lifts are not significantly greater than zero—the models are not doing better than constant personal baselines.

Dataset	Problem	Model	Avg. Personal Baseline Error	Avg. Personal Model Error	Avg. User Lift (Error)	p-value
SL—Stress	binary	Log.Reg.	29.19%	29.09%	**0.10**	.481
FaF—Happiness	binary	SVM(rbf)	16.51%	18.67%	**-2.17**	.967
FaF—Stress	binary	SVM(rbf)	25.17%	23.35%	**1.82**	.240
SL—Stress	regression	Elastic Net	0.75	0.78	**-0.03**	.988
FaF—Happiness	regression	Elastic Net	0.81	0.83	**-0.02**	.999
FaF—Stress	regression	Elastic Net	1.10	1.13	**-0.03**	1.000

https://doi.org/10.1371/journal.pone.0184604.t001

as from smartphones or smartwatches (Fig 2). Just like in our example datasets, participants report surprisingly little variation (Fig 3A). As a result, guessing that an individual was at the same state incurred low personal baseline error and machine learning algorithms typically had only slightly lower error than the personal baseline (Fig 3B).

Studies that do report personal averages sometimes have negative user lift (Fig 4A). When personal baselines are reported, they are usually reported in aggregate, which can be misleading by obscuring negative user lift on some individuals. Aggregation also precludes statistical tests on user lift and the only study that did report a statistical rank test on improvement across individuals found that there algorithms were no better than a naive model (using an historical averages of individuals' states) [45]. However, the bulk of studies only report population baselines making it impossible to know if they have any user lift (Fig 4B). As such, it seems that the bulk of papers have questionable results, at best.

Discussion

We have shown, with examples of stress and happiness on two popular datasets how easily machine learning algorithms can appear promising when compared with meaningless baselines. Individuals report surprisingly little variation in state, so always guessing that an individual is at their most frequently reported state is correct most of the time. As a result, when an algorithm is compared with a population baseline that always predicts all users are always at
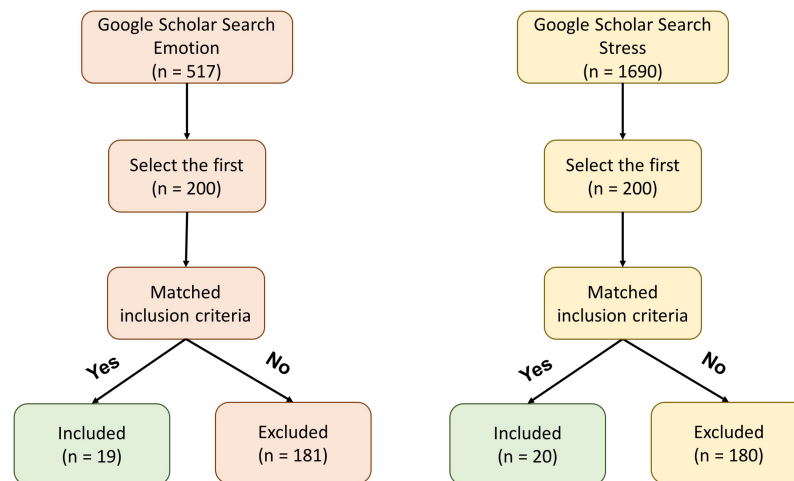
Fig 2. Diagram of literature review process.

https://doi.org/10.1371/journal.pone.0184604.g002
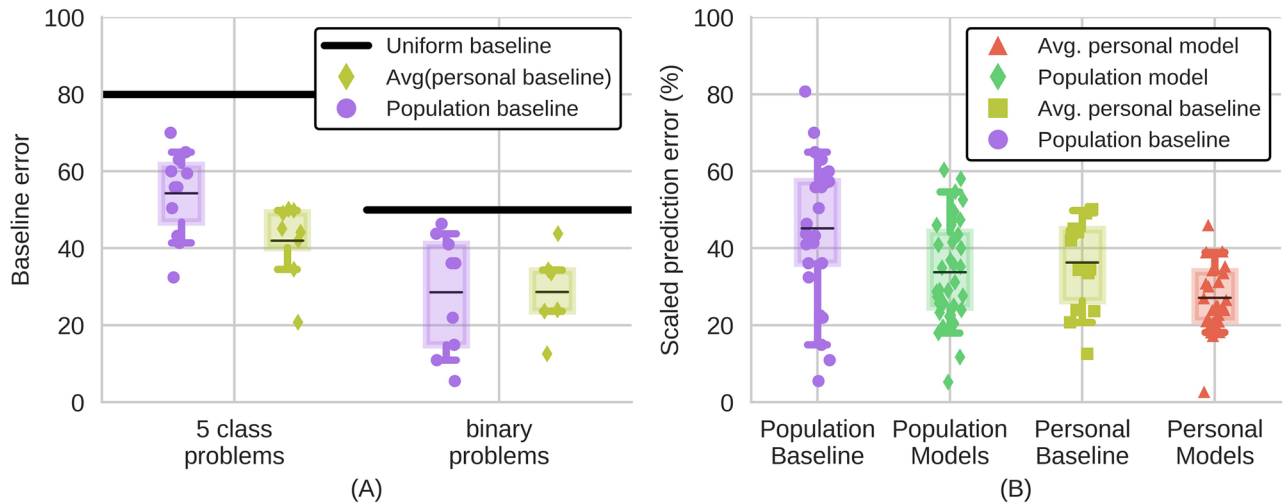
Fig 3. Participant variability and model performance reported in related studies. Reported results reveal little participant variability and that models do not dramatically improve upon personal baselines. (A) Population and personal baselines reported by studies that had participants report their state on two point and five point scales. The black bars indicate the what the baseline would have been if participants were to report every state equally often, e.g., happy and sad each half the time. (B) Population and personal baselines and model error reported in literature reviewed. Performance (prediction error) is scaled by the minimum class imbalance to compare studies that asked participants to report their states on scales with different numbers of points. In both figures boxes denote 1st and 3rd quantiles, bars indicate 5th and 95th percentiles, and lines the average of the markers.

the same state, the algorithm's predictions can seem accurate even if they are no better than predicting each individual to be at their most frequently reported state. Despite the possibility for falsely optimistic results, we found in a systematic literature review that population baselines are commonly compared with in roughly 77% of publications reviewed. We also find that when personal baselines are reported that the algorithms often add little or nothing over these baselines (and in fact they sometimes do worse).
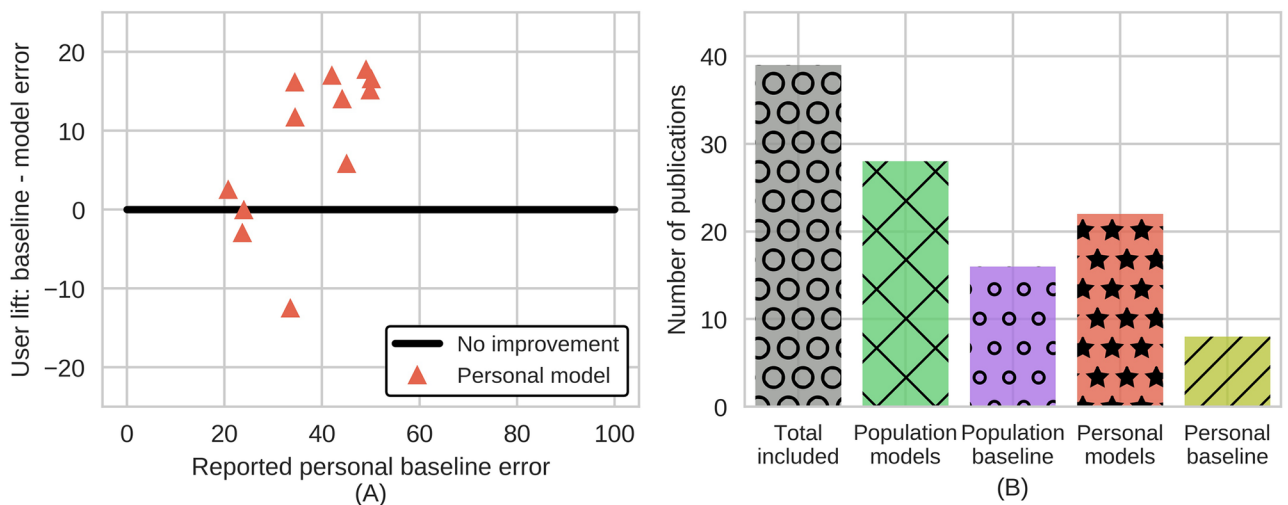
Fig 4. Calculated user lift and prevalence of baselines reported in studies reviewed. (A) User lift calculated for studies where error of baselines and algorithms were both reported. Algorithms sometimes have no improvements over baseline guessing, and these figures are biased to studies that reported sufficient information. (B) Population baselines are reported in roughly half the publications reviewed while personal baselines are infrequently reported (approximately 23% of publications).

A limitation of the datasets that we explored, and most of the literature we reviewed, was that the study cohorts were not clinical populations, the sample size was small, and the study duration was limited. However, the study characteristics of the datasets presented are characteristic of many studies. While target populations for the monitoring we have discussed are typically individuals with mood disorders, study cohorts are frequently small in size and from the general population. It is possible that individuals with mood disorders would report more variability in state than the general public. More variability would reduce the likelihood of falsely optimistic results, but our proposed evaluation method would still be appropriate for showing that algorithms are an improvement over always predicting that individuals are at their average state. Finally, the user lift evaluation framework that we suggest would complement a larger dataset, despite being demonstrated on fewer subjects here.

While we reviewed a representative portion of relevant literature, we had to focus the scope and present a reproducible search that aligned with public datasets. We constructed general search queries to include pertinent studies, but inconsistencies in terminology between communities made it impossible to included all relevant studies and some known related works were not covered.

In addition to coverage, there were a variety of features that we could not control in the literature review. Studies recruited from disparate populations and had different study protocols. In addition to collecting different data and conducting different analyses, studies reported results in an variety of ways. We did our best to standardize across studies and present results favorably and comparably.

The proposed user lift evaluation framework is more generally applicable to predicting longitudinal patient state than we have shown here. We have focused our review on a narrow, important application of mental wellbeing, as this is a nascent and exciting application for machine learning algorithms. However, user lift would apply to any application predicting longitudinal data, such as monitoring blood sugar level, body weight, or daily sleep duration. The importance of statistical tests on user lift becomes greater for applications where individuals are expected to exhibit less variation and descriptive statistics must also be reported to quantify the size of any statistically significant user lift.

While we have calculated personal baselines here over the entire dataset, in principle this is not necessary. Because personal baselines are calculated with respect to individuals' most common state, personal baselines are easy to quickly approximate, with minimal sampling. Personal baselines could potentially vary over an extended period of study, but such scales are outside the scope of most studies and require further investigation.

An ability to predict meaningful personal signals for medical monitoring, such as mental wellbeing, could greatly improve personalized medicine by enabling novel approaches to just in time and personalized interventions. However, we have highlighted some pitfalls of evaluating algorithms for this application that can easily result in falsely optimistic results and unintentionally provide baseless optimism. To reduce alsely optimistic results, we have suggested an alternative evaluation framework using statistical tests on our proposed metric of user lift, which takes an individual-centric approach. As was shown, there is a range of model predictive capability across individuals, so we suggest statistically testing for significant improvement on the population. This framework of evaluation can help researchers to focus efforts and thus help advance progress on this application.

Author Contributions

Conceptualization: Orianna DeMasi, Konrad Kording, Benjamin Recht.

Data curation: Orianna DeMasi.

Formal analysis: Orianna DeMasi.

Funding acquisition: Konrad Kording, Benjamin Recht.

Investigation: Orianna DeMasi.

Methodology: Orianna DeMasi, Konrad Kording, Benjamin Recht.

Resources: Benjamin Recht.

Software: Orianna DeMasi.

Supervision: Konrad Kording, Benjamin Recht.

Validation: Orianna DeMasi.

Visualization: Orianna DeMasi.

Writing – original draft: Orianna DeMasi.

Writing – review & editing: Orianna DeMasi, Konrad Kording, Benjamin Recht.

References

1. Collins FS, Varmus H. A new initiative on precision medicine. New England Journal of Medicine. 2015; 372(9):793–795. https://doi.org/10.1056/NEJMp1500523 PMID: 25635347

2. Darcy AM, Louie AK, Roberts LW. Machine learning and the profession of medicine. Jama. 2016; 315(6):551–552. https://doi.org/10.1001/jama.2015.18421 PMID: 26864406

3. Calvo RA, Dinakar K, Picard R, Maes P. Computing in mental health. In: Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems. ACM; 2016. p. 3438–3445.

4. Marzano L, Bardill A, Fields B, Herd K, Veale D, Grey N, et al. The application of mHealth to mental health: opportunities and challenges. The Lancet Psychiatry. 2015; 2(10):942–948. https://doi.org/10.1016/S2215-0366(15)00268-0 PMID: 26462228

5. Olff M. Mobile mental health: A challenging research agenda. European journal of psychotraumatology. 2015; 6. https://doi.org/10.3402/ejpt.v6.27882

6. Ben-Zeev D, Scherer EA, Wang R, Xie H, Campbell AT. Next-generation psychiatric assessment: Using smartphone sensors to monitor behavior and mental health. Psychiatric rehabilitation journal. 2015; 38(3):218. https://doi.org/10.1037/prj0000130 PMID: 25844912

7. Canzian L, Musolesi M. Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing. ACM; 2015. p. 1293–1304.

8. Pantelopoulos A, Bourbakis NG. A survey on wearable sensor-based systems for health monitoring and prognosis. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews). 2010; 40(1):1–12. https://doi.org/10.1109/TSMCC.2009.2032660

9. Saeb S, Zhang M, Karr CJ, Schueller SM, Corden ME, Kording KP, et al. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. Journal of medical Internet research. 2015; 17(7):e175. https://doi.org/10.2196/jmir.4273 PMID: 26180009

10. Friedman J, Hastie T, Tibshirani R. The elements of statistical learning. vol. 1. Springer series in statistics Springer, Berlin; 2001.

11. De Choudhury M, Counts S, Horvitz EJ, Hoff A. Characterizing and predicting postpartum depression from shared facebook data. In: Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing. ACM; 2014. p. 626–638.

12. De Choudhury M, Kiciman E, Dredze M, Coppersmith G, Kumar M. Discovering shifts to suicidal ideation from mental health content in social media. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. ACM; 2016. p. 2098–2110.

13. Witten IH, Frank E, Hall MA, Pal CJ. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann; 2016.

14. Boyko EJ. Ruling Out or Ruling In Disease with the Most sensitiue or Specific Diagnostic Test Short Cut or Wrong Turn? Medical Decision Making. 1994; 14(2):175–179. https://doi.org/10.1177/0272989X9401400210 PMID: 8028470

15. Fawcett T. An introduction to ROC analysis. Pattern recognition letters. 2006; 27(8):861–874. https://doi.org/10.1016/j.patrec.2005.10.010

16. He H, Garcia EA. Learning from imbalanced data. IEEE Transactions on knowledge and data engineering. 2009; 21(9):1263–1284. https://doi.org/10.1109/TKDE.2008.239

17. Powers DM. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. 2011;.

18. Weiss GM. Mining with rarity: a unifying framework. ACM SIGKDD Explorations Newsletter. 2004; 6(1):7–19. https://doi.org/10.1145/1007730.1007734

19. Kessler R, Stein M, Petukhova M, Bliese P, Bossarte R, Bromet E, et al. Predicting suicides after outpatient mental health visits in the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS). Molecular Psychiatry. 2016;.

20. Murphy GE. The prediction of suicide: why is it so difficult? American Journal of Psychotherapy. 1984;. PMID: 6385733

21. Tran T, Phung D, Luo W, Harvey R, Berk M, Venkatesh S. An integrated framework for suicide risk prediction. In: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM; 2013. p. 1410–1418.

22. Kempthorne O, Doerfler T. The behaviour of some significance tests under experimental randomization. Biometrika. 1969; 56(2):231–248. https://doi.org/10.1093/biomet/56.2.231

23. Smucker MD, Allan J, Carterette B. A comparison of statistical significance tests for information retrieval evaluation. In: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management. ACM; 2007. p. 623–632.

24. Jaques N, Taylor S, Azaria A, Ghandeharioun A, Sano A, Picard R. Predicting students' happiness from physiology, phone, mobility, and behavioral data. In: Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on. IEEE; 2015. p. 222–228.

25. Wang R, Chen F, Chen Z, Li T, Harari G, Tignor S, et al. Studentlife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing. ACM; 2014. p. 3–14.

26. Aharony N, Pan W, Ip C, Khayal I, Pentland A. Social fMRI: Investigating and shaping social mechanisms in the real world. Pervasive and Mobile Computing. 2011; 7(6):643–659. https://doi.org/10.1016/j.pmcj.2011.09.004

27. Schwarz G, et al. Estimating the dimension of a model. The annals of statistics. 1978; 6(2):461–464. https://doi.org/10.1214/aos/1176344136

28. Arthur D, Vassilvitskii S. k-means++: The advantages of careful seeding. In: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics; 2007. p. 1027–1035.

29. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B (Methodological). 1996; p. 267–288.

30. Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2005; 67(2):301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

31. Breiman L. Random forests. Machine learning. 2001; 45(1):5–32. https://doi.org/10.1023/A:1010933404324

32. Saeb S, Lonini L, Jayaraman A, Mohr DC, Kording KP. Voodoo Machine Learning for Clinical Predictions. bioRxiv. 2016;

33. Bogomolov A, Lepri B, Ferron M, Pianesi F, Pentland AS. Daily stress recognition from mobile phone data, weather conditions and individual traits. In: Proceedings of the 22nd ACM international conference on Multimedia. ACM; 2014. p. 477–486.

34. Bogomolov A, Lepri B, Ferron M, Pianesi F, Pentland AS. Pervasive stress recognition for sustainable living. In: Pervasive Computing and Communications Workshops (PERCOM Workshops), 2014 IEEE International Conference on. IEEE; 2014. p. 345–350.

35. Carroll EA, Czerwinski M, Roseway A, Kapoor A, Johns P, Rowan K, et al. Food and mood: Just-in-time support for emotional eating. In: Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on. IEEE; 2013. p. 252–257.

36. Jaques N, Taylor S, Sano A, Picard R. Multi-task, Multi-Kernel Learning for Estimating Individual Wellbeing. In: Proc. NIPS Workshop on Multimodal Machine Learning, Montreal, Quebec; 2015.

37. Lu H, Frauendorfer D, Rabbi M, Mast MS, Chittaranjan GT, Campbell AT, et al. Stresssense: Detecting stress in unconstrained acoustic environments using smartphones. In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing. ACM; 2012. p. 351–360.

38. Moturu ST, Khayal I, Aharony N, Pan W, Pentland A. Sleep, mood and sociability in a healthy population. In: 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE; 2011. p. 5267–5270.

39. Deng Y, Wu Z, Chu CH, Yang T. Evaluating feature selection for stress identification. In: Information Reuse and Integration (IRI), 2012 IEEE 13th International Conference on. IEEE; 2012. p. 584–591.

40. Hernandez J, Morris RR, Picard RW. Call center stress recognition with person-specific models. In: International Conference on Affective Computing and Intelligent Interaction. Springer; 2011. p. 125–134.

41. Sandulescu V, Andrews S, Ellis D, Bellotto N, Mozos OM. Stress detection using wearable physiological sensors. In: International Work-Conference on the Interplay Between Natural and Artificial Computation. Springer; 2015. p. 526–532.

42. Valenza G, Gentili C, Lanatà A, Scilingo EP. Mood recognition in bipolar patients through the PSYCHE platform: Preliminary evaluations and perspectives. Artificial intelligence in medicine. 2013; 57(1): 49–58. https://doi.org/10.1016/j.artmed.2012.12.001 PMID: 23332576

43. Valenza G, Nardelli M, Lanata A, Gentili C, Bertschy G, Paradiso R, et al. Wearable monitoring for mood recognition in bipolar disorder based on history-dependent long-term heart rate variability analysis. IEEE Journal of Biomedical and Health Informatics. 2014; 18(5):1625–1635. https://doi.org/10.1109/JBHI.2013.2290382 PMID: 24240031

44. Wu M, Cao H, Nguyen HL, Surmacz K, Hargrove C. Modeling perceived stress via HRV and accelerometer sensor streams. In: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE; 2015. p. 1625–1628.

45. Asselbergs J, Ruwaard J, Ejdys M, Schrader N, Sijbrandij M, Riper H. Mobile Phone-Based Unobtrusive Ecological Momentary Assessment of Day-to-Day Mood: An Explorative Study. Journal of Medical Internet research. 2016; 18(3). https://doi.org/10.2196/jmir.5505 PMID: 27025287

46. LiKamWa R, Liu Y, Lane ND, Zhong L. Moodscope: Building a mood sensor from smartphone usage patterns. In: Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services. ACM; 2013. p. 389–402.