Behavioral Phenotyping of Digital Health Tracker Data

Thomas Quisel* tquisel@evidation.com

Luca Foschini* lfoschini@evidation.com

Alessio Signorini* asignorini@evidation.com

Abstract

With the surge in popularity of wearable technologies a large percentage of the US population is now tracking activities such as sleep, diet and physical exercise. In this study we empirically evaluate the ability to predict metrics (e.g., weekly alcohol consumption) directly related to health outcomes from densely sampled, multi-variate time series of behavioral data. Our predictive models are based on temporal convolutional neural networks and take as input the raw historical time series of daily step counts, sleep duration, and weight scale usage sourced from an online population of thousands of digital trackers. The prediction accuracy achieved outperforms several strong baselines that use hand-engineered features and indicates that tracker data contains valuable information on individuals' lifestyles even for behavioral aspects seemingly unrelated to the measured quantities. We believe that this insight can be applied to the design of new digital interventions and enable future large-scale preventive care strategies.

Keywords Digital health, activity tracking, behavioral phenotyping, mHealth, temporal convolutional neural networks.

1 Introduction

It is estimated that 69% of the U.S. population keeps track of their weight, diet, or exercise routine, and 20% of trackers claim to leverage technology such as digital health devices and apps to perform self-monitoring [9, 28]. With tech giants like Apple and Google entering the arena of wearable technologies, the market for activity trackers and wearable devices is projected to increase to more than $50 billion by 2018 [33].

Not only has the number of digital health trackers surged in recent years, the breadth of measures these devices can quantify has also dramatically expanded. The last Consumer Electronic Conference held every year in Las Vegas [1] featured consumer-grade sensors able to continuously capture hemoglobin, arterial oxygen saturation (SpO2), pulse rate (PR), perfusion index (PI), and Plethysmograph Variability Index (PVI). With these new additions, the digital tracker ecosystem starts resembling the capabilities of the sensor arrays found in ICU rooms [19], and constitutes a significant step forward from pedometers and calorie counters that have become prevalent in smartphones and watches.

While it is disputed whether digital health tracking alone can lead to healthier behavior in the adopter [24], it is clear that the wealth of information provided by the trackers, however inaccurate [20], can be predictive of lifestyle. In our recent study [29] we provided evidence of this fact by showing that changes in an individual's adherence to weight tracking and food logging are predictive of weight change over time.

A large body of empirical evidence demonstrates that lifestyle plays an important role in long term health outcomes [8, 25, 32]. An illustrative example for the case of cardiovascular diseases is the Harvard Healthy Heart Score survey [2], which calculates a Cardiovascular Lifestyle Risk Score based on lifestyle habits such as smoking, physical activity, and diet. Some of the questions on the survey, such as, "During the past year, what was your average time per week spent on walking (slower than 3 miles per hour)," can be immediately answered by the step count reported by a pedometer. Other questions, such as the lifestyle ones pertaining to alcohol consumption habits, cannot be directly inferred from tracker summary statistics.

That said, the temporally dense information recorded by digital trackers contain more complex patterns. For example, a decrease in sleep duration on Friday nights and corresponding lower step counts in the following day may correlate with a weekly habit of partying–a pattern that might go undetected when only looking at summary statistics of the sleep duration or the step counts taken in isolation–and may be a good predictor of increased weekly alcohol consumption.

1.1 Contribution In this work we extend the analysis of Pourzanjani et al. [29] in the pursuit of closing the gap between behavioral phenotyping and health outcomes. From an outcome perspective, we focused on metrics known to be important predictors of future health:

- Their (measured) Body-Mass Index;
- The (self reported) frequency of weekly alcohol

[*] Evidation Health - Menlo Park, CA

consumption;

- The (measured) propensity to increase their level of physical exercise as a result of a digital intervention.

The Body-Mass Index (BMI) is strongly correlated with other aspects of an individual's health and abnormalities are estimated to cost 21% ($190.2 billion) of annual medical spending in the United States [6]. Similarly, immoderate alcohol consumption and lack of physical exercise are associated with unfavorable health outcomes [8, 25, 30].

From a methods perspective, we present a model based on a temporal Convolutional Neural Network (CNN) that allows for prediction of the outcome variables from the raw time series recorded by the digital health trackers: daily step count, sleep duration, and weight scale utilization (i.e., whether or not the individual has weighed themselves on a given day).

We show that the CNN approach matches or outperforms several strong baselines that leverage hand-engineered features, in line with the same groundbreaking advances that representation learning and unsupervised feature discovery via deep learning have brought to image processing [16], speech recognition [10], and natural language processing [22].

Finally, we show that the performance of the CNN model is robust to the imputation strategy used for the time series, in line with the hypothesis of Razavian et al. [31] who argue that missing values do not constitute a major concern in temporally dense time series, such as the ones under study.

2 Related work

The task of deriving observable physiological traits from clinical data is generally termed phenotyping [26]. Although phenotyping has become an established practice in medical machine learning, to the best of our knowledge this is the first attempt at extracting phenotypes from behavioral data to predict health-related outcomes. In [29], Pourzanjani et al. showed that frequency of weight tracking and gaps in tracking behavior are predictive of an individual's weight change. The methodologies used in their work only considered temporal summary statistics such as the frequency and gaps between reported measurements, computed separately on a single time series, and predicted a single outcome. On the contrary, the method presented in this paper uses as input the raw multivariate time-series of digital health measurements and considers several diverse health-related outcome variables. From a methods perspective, the present work shares commonalities with the machine learning research focused on phenotyping of medical data, but while in general medical settings observations such as vital signs, lab test results, and subjective assessments are sampled irregularly [21], behavioral data recorded by digital health trackers is *dense* and recorded at least with daily frequency.

In the medical machine learning community, several recent works have addressed the topic of phenotyping of clinical data. In their recent work [31], Razavian et al. use a multi-resolution CNN to perform early detection of multiple diseases from irregularly measured sparse lab values. We benefit from the same ease of interpretability of the learned model brought about by the temporal convolutional approach, however, as Razavian et al. argue in their paper, their method focuses more on devising a highly refined imputation strategy to cope with missing data, a problem far less common on digital health data.

Another very recent work by Lipton et al. [19], uses Long Short-Term Memory (LSTM) networks, a variant of Recurrent Neural Networks (RNNs), to identify patterns and classify 128 diagnoses from multivariate time series of 13 frequently but irregularly sampled clinical measurements. As pointed out in [31], it is not clear whether the long-term dependencies that RNNs very effectively model are necessary in contexts similar to the one under study.

Neural networks in general have a long history of applications in the medical domain [3, 5]. More recently, deep learning has been applied to assess Parkinsons Disease [12] and feed-forward networks have been applied to medical time series for gout, leukemia, and critical illness classification [7, 17]. Finally, non-neural-network based techniques have been leveraged to perform classification of multi-variate time-series in the medical domain. See [23] for a review.

3 Data

The source of our data is AchieveMint[1], a consumer rewards platform for healthy activities powered by Evidation Health[2]. The AchieveMint platform automatically collects data (e.g., step counts) from its users' digital trackers and aggregates it into their accounts rewarding health related activities (e.g., a run) with points. We considered binary classification tasks on three datasets. Each dataset is composed of pairs of multivariate time series and binary labels, each pair associated with a different individual. The multivariate time series for a given individual contained a history (different lengths were used in different datasets) of daily step counts, sleep durations, and interactions with a connected scale (a binary indicator whose value

[1] http://www.achievemint.com

[2] http://www.evidation.com

is 1 if the user weighed themselves through a connected scale, and 0 otherwise). All the time series measurements were passively recorded by the relevant tracker (i.e., pedometer, sleep trackers, scale); none of them was self-reported. A detailed description of each dataset and prediction task is provided below:

Uptake The dataset consists of 1,996 users who took part in an IRB-approved study designed to increase the level of physical activity through small monetary incentives. Over the two-week intervention period, the groups were offered the same average incentives for physical activity. We considered a subset of the users in the experimental arms (the control group did not undergo the intervention) that have a history of measurements of at least 147 days. We assigned a positive label to users whose median daily step count during the intervention period had shown an uptake of more than 2,000 steps/day[3] compared to the median pre-intervention. This resulted in 22% positive labels. A visual representation of the daily step count histories for a few hundred users is shown in Figure 1.

BMI The dataset consists of 1,978 AchieveMint users who have shared their BMI measurements (weight reported by a connected scale, height self-reported). We assigned a positive label to users with BMI higher than a chosen clinically relevant threshold [34], which resulted in 44% positive labels.

Alcohol The dataset consists of 815 users that agreed to participate in a one-click survey answering the lifestyle question "On average, do you have more than one drink per week?" inspired by the Healthy Heart Survey [2]. We assigned a positive label to the users who answered the question positively, which resulted in 33% positive labels.

4 Methods

We learned a binary classifier to generate estimates $\hat{y_u}$ of the true labels y_u for each user u from the multivariate time series of observations for the time period T, $X_u = x_1, \ldots, x_T$. Each observation x_t is a vector of size K, representing one of the K behavioral biomarkers (in our case, $K = 3$: step count, sleep duration in hours, and binary indicator of weight measurement) recorded for a given day.

Our temporal convolution model is shown in Figure 2. The input to the model can be raw (un-imputed) observations, imputed observations, or the concatenation of imputed data and the binary observation pattern.

Following Zheng et al. [35] each time series is fed separately to a two-stage univariate feature extraction, where each stage is composed of a convolution filter followed by a Max-pooling layer and sigmoid layer (unlike [35], which uses avg-pooling).

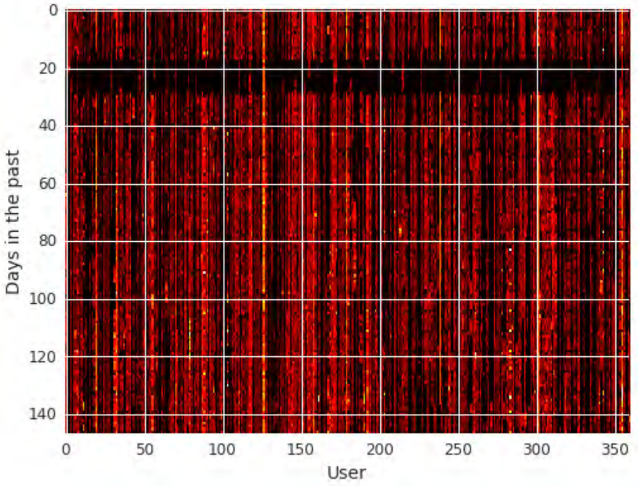

Figure 1: Heatmap for a few hundred users of the Uptake dataset. Brighter pixels correspond to higher step counts. The darker band on top represent marks a period of low activity during the winter holidays

The output of the feature extraction layers is flattened and fed to a standard fully connected MLP with hidden layer for classification [18].

The specific architectural choices for the shared part of the prediction network is as follows: we set the number of filters to be 8 for the first convolution layer and 4 for the second, with a kernel length of 7 for the first layer and 5 for the second layer and step size of 1 for all convolutional layers. Each Max-pooling layer has a length of 2 with step size of 2 (i.e. no overlap). Each convolution layer is followed by a Sigmoid nonlinearity ($sigmoid(t) = \frac{1}{1+e^{-t}}$).

We added 1 fully connected hidden layer with 300 nodes after the concatenation of outputs of all convolution layers. After the last Sigmoid layer corresponding to the output of the shared part of the network we added a fully connected layer (of the size of 2 nodes corresponding to binary outcome) and a Log Softmax Layer in this order. We use ADAM [15] instead of SGD for parameter updates.

The loss function for each label is the negative log likelihood of the true label: $L = -\sum_{u \in U} \sum_{c \in 0,1} y_{c,u} \log \hat{y_{c,u}}$. Each gradient is backpropagated throughout the entire prediction network.

We set the momentum to 0.9, use a fixed lr of 0.005, and set a 0.003 weight decay. The Alcohol task required a 0.006 weight decay to avoid overfitting,

[3] Considered an increased in activity that when sustained in the long term can provide health benefits [30]

since it is a smaller dataset.

We implemented our model in the Caffe [14] environment.

5 Results

The CNN model is fed the imputed times series (using linear imputation). We found that mean-centering each day of a time series before imputation, so that the mean across users is zero for each given day, significantly improved the results.

To test the robustness of our model to missing values, we considered a variant of the model, CNN-U, in which each input time series is augmented with its utilization signal: a time series of binary indicators encoding whether the data for a given day was missing and had been imputed. Input time series that are already utilization signals, such as the weight measurement one, are not augmented.

In Table 1 we reported the mean area under the ROC curve (AUC) over 4 cross-validated folds for the three datasets. Given the small size of our datasets, a 4-fold cross-validation mean AUC provides more robust and stable results. We compared the two convolutional neural network approaches with several baseline models (logistic regression, random forest (RF) and SVM classifiers) trained on hand-engineered features. Following [7, 19, 21] the features we computed for each variable are the mean, standard deviation, median, quartiles, minimum, maximum, and a count of non-missing values. Hyperparameters for the baseline models trained on the hand-engineered features were tuned using random search [4]. The SVM hyperparameter search space was derived from [13].

	CNN	CNN-U	logistic	RF	SVM
Uptake	**0.699**	0.698	0.629	0.622	0.611
BMI	0.640	0.639	0.653	**0.654**	0.648
Alcohol	0.549	**0.552**	0.526	0.551	0.526

Table 1: 4-fold cross-validated AUC for the three datasets. CNN is the temporal convolutional model that takes as input the linearly imputed time series. CNN-U takes the step count and sleep utilization time series as additional inputs. Logistic, random forest (RF) and SVM models are trained on hand-engineered features.

We observed that the CNN models significantly outperform the baseline ones on the Uptake dataset and slightly on the Alcohol dataset. We also note that the AUC values reported demonstrate that daily recordings of step counts, sleep duration, and scale usage, however inaccurate, are predictive of an individual's overall behavior, even for health-related properties not directly related to the observed variables.

Unlike other neural network based models, CNNs provide direct interpretability of the learned models. The weekly trends learned by the CNN in the first layer convolutional filters for the step count biomarker are reported in Figure 3.

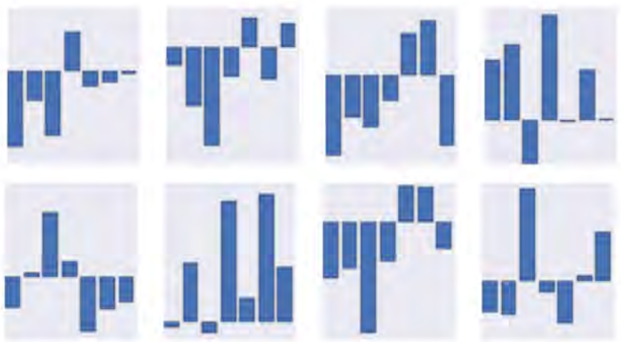

Figure 3: Convolution weights learned by the CNN on the step count time series of the Uptake dataset. Each graph shows the 7 learned weights for each of the nodes in the first convolutional layer for step counts.

Since our dataset is small when compared to datasets found in common deep learning tasks, regularization heavily affects the results. Figure 4 shows the learning curves for both CNN models and demonstrates that the regularization parameters used successfully avoid overfitting.

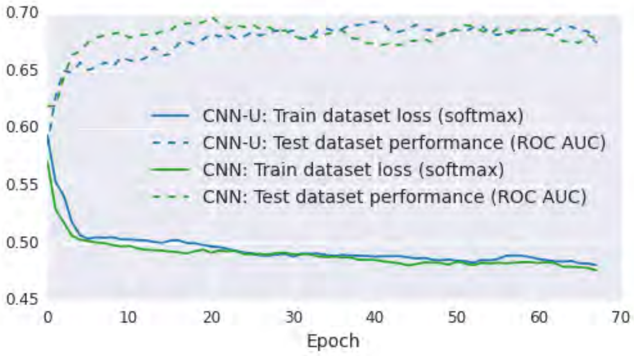

Figure 4: Training set softmax loss and testing set AUC vs. training epocs for the Uptake dataset. The curves demonstrates that the regularization employed successfully prevents overfitting. In addition, the negligible difference between CNN and CNN-U highlights the robustness of the model to imputation.

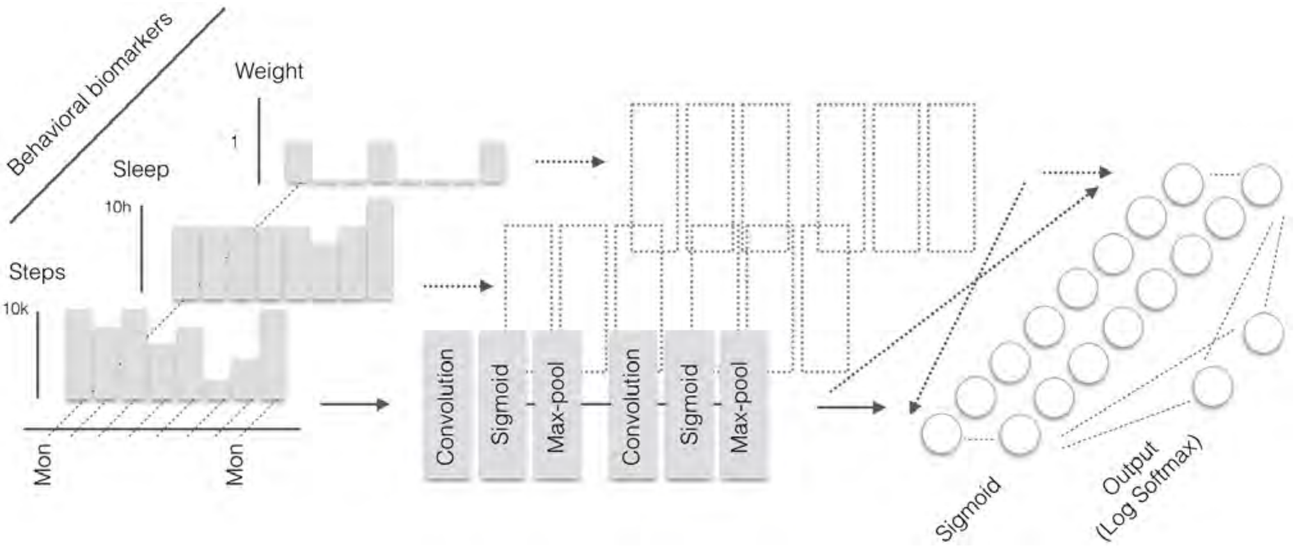

Figure 2: The temporal CNN architecture. Each behavioral biomarker time series is fed through the convolutional layer separately. The output layer, the only one label-specific, is a Log SoftMax classifier.

6 Conclusion and Future work

The results presented in this paper indicate that a temporal convolutional neural network learned on raw time series of data streamed directly from digital health trackers can accurately predict important health-related variables.

Our models' automatic behavioral phenotyping has many potential applications: (1) it can be used to passively infer lifestyle choices of individuals with the goal of complementing, or even replacing, surveys (e.g., [2]) that must actively acquire such data to determine its relationship with disease risk factors; (2) it can be used to augment models based on more traditional medical data sources [31] to further improve medical decision making; and (3) the learned phenotypes can be used to optimize behavior-changing interventions [11, 27] with the goal of proactively addressing high-risk behaviors. The high accuracy achieved on predicting the propensity of individuals to increase their physical activity as a result of a digital intervention can improve targeting decisions for interventions. More-sophisticated, higher-cost interventions (e.g., in-person coaching) can be targeted to individuals identified as less inclined to improve, while simpler and more cost-effective strategies (e.g., an email reminder) can be sufficient for those who display a higher propensity to change.

Possible extensions to our approach include improving the performance of the network by further tuning its architecture and testing it on a larger set of input variables, including fixed-time ones (e.g., demographics). The model could also benefit from more sophisticated imputation strategies (e.g., [31]) and modules that encode longer terms dependencies (e.g., by multiresolution approach as in [31] or using RNN-like techniques, such as in [19]).

References

[1] Ces 2016: Running list of health and wellness devices. `http://mobihealthnews.com/content/ces-2016-running-list-health-and-wellness-devices`. Accessed: 2016-01-13.

[2] Harvard healthy heart score. `https://healthyheartscore.sph.harvard.edu/`. Accessed: 2016-01-17.

[3] W. G. Baxt. Application of artificial neural networks to clinical medicine. *The lancet*, 346(8983):1135–1138, 1995.

[4] J. Bergstra and Y. Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13.

[5] R. Caruana, S. Baluja, T. Mitchell, et al. Using the future to" sort out" the present: Rankprop and multitask learning for medical risk evaluation. *Advances in neural information processing systems*, pages 959–965, 1996.

[6] J. Cawley and C. Meyerhoefer. The medical care costs of obesity: an instrumental variables approach. *Journal of health economics*, 31(1):219–230, 2012.

[7] Z. Che, D. Kale, W. Li, M. T. Bahadori, and Y. Liu. Deep computational phenotyping. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 507–516. ACM, 2015.

[8] S. E. Chiuve, M. L. McCullough, F. M. Sacks, and E. B. Rimm. Healthy lifestyle factors in the primary prevention of coronary heart disease among men benefits among users and nonusers of lipid-lowering and antihypertensive medications. *Circulation*, 114(2):160–167, 2006.

[9] S. Fox and M. Duggan. *Tracking for health.* Pew Research Center's Internet & American Life Project, 2013.

[10] A. Graves and J. Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5):602–610, 2005.

[11] D. Halpern. *Inside the Nudge Unit.* Random House, 2015.

[12] N. Y. Hammerla, J. M. Fisher, P. Andras, L. Rochester, R. Walker, and T. Plötz. Pd disease state assessment in naturalistic environments using deep learning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[13] C.-W. Hsu, C.-C. Chang, C.-J. Lin, et al. A practical guide to support vector classification, 2003.

[14] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, pages 675–678. ACM, 2014.

[15] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[17] T. A. Lasko, J. C. Denny, and M. A. Levy. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. *PloS one*, 8(6):e66341, 2013.

[18] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–48. Springer, 2012.

[19] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzell. Learning to diagnose with LSTM recurrent neural networks. *CoRR*, abs/1511.03677, 2015.

[20] C. MA, B. HA, V. KG, and P. MS. Accuracy of smartphone applications and wearable devices for tracking physical activity data. *JAMA*, 313(6):625–626, 2015.

[21] B. M. Marlin, D. C. Kale, R. G. Khemani, and R. C. Wetzel. Unsupervised pattern discovery in electronic health care data using probabilistic clustering models. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pages 389–398. ACM, 2012.

[22] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[23] R. Moskovitch and Y. Shahar. Classification-driven temporal discretization of multivariate time series. *Data Min. Knowl. Discov.*, 29(4):871–913, July 2015.

[24] P. MS, A. DA, and V. KG. Wearable devices as facilitators, not drivers, of health behavior change. *JAMA*, 313(5):459–460, 2015.

[25] K. J. Mukamal, S. E. Chiuve, and E. B. Rimm. Alcohol consumption and risk for coronary heart disease in men with healthy lifestyles. *Archives of Internal Medicine*, 166(19):2145–2150, 2006.

[26] A. Oellrich, N. Collier, T. Groza, D. Rebholz-Schuhmann, N. Shah, O. Bodenreider, M. R. Boland, I. Georgiev, H. Liu, K. Livingston, et al. The digital revolution in phenotyping. *Briefings in bioinformatics*, page bbv083, 2015.

[27] P. Olson. A massive social experiment on you is under way, and you will love it. `http://www.forbes.com/sites/parmyolson/2015/01/21/jawbone-guinea-pig-economy/`. Accessed: 2016-01-13.

[28] W. Plank. The future of wearables market. `http://www.wsj.com/articles/the-future-of-the-wearables-market-1452736738`. Accessed: 2016-01-20.

[29] A. Pourzanjani, T. Quisel, and L. Foschini. Adherent use of activity trackers is associated with weight loss. *PLOS ONE*, Submitted.

[30] K. E. Powell, A. E. Paluch, and S. N. Blair. Physical activity for health: What kind? how much? how intense? on top of what? *Public Health*, 32(1):349, 2011.

[31] N. Razavian and D. Sontag. Temporal convolutional neural networks for diagnosis from lab tests. *CoRR*, abs/1511.07938, 2015.

[32] A. A. Thorp, N. Owen, M. Neuhaus, and D. W. Dunstan. Sedentary behaviors and subsequent health outcomes in adults: a systematic review of longitudinal studies, 1996–2011. *American journal of preventive medicine*, 41(2):207–215, 2011.

[33] T. Wang. The future of biosensing wearables. rock health. `http://rockhealth.com/2014/06/future-biosensing-wearables`. Accessed: 2016-01-13.

[34] World Health Organization. BMI Classification, Global Database on Body Mass Index, 2006.

[35] Y. Zheng, Q. Liu, E. Chen, Y. Ge, and J. L. Zhao. Time series classification using multi-channels deep convolutional neural networks. In *Web-Age Information Management*, pages 298–310. Springer, 2014.