Observation Time vs. Performance in Digital Phenotyping

Tom Quisel
Evidation Health
510 State St. Suite 200
Santa Barbara, CA 93101
tquisel@evidation.com

Wei-Nchih Lee
Evidation Health
11 N. Ellsworth Avenue
San Mateo, CA 94401
wlee@evidation.com

Luca Foschini
Evidation Health
510 State St. Suite 200
Santa Barbara, CA 93101
luca@evidation.com

ABSTRACT

Mobile health (mHealth) technologies enable frequent sampling of physiological and psychological signals over time. In our recent work we used a convolutional neural network (CNN) model to predict self-reported phenotypes of chronic conditions from step and sleep data recorded from passive trackers in free living conditions [9]. We investigated the impact of the time-granularity of the collected data and showed that training the models on higher-resolution (minute-level) data improved classification performance on conditions related to mental health and nervous system disorders, as compared to using only day-level totals. In the present work we shift the focus from the time resolution of the observation window to its duration. We study how the performance of the best-performing model on the highest-resolution data changes as the length of the data collection window is varied from 3 to 147 days for each user. We found that for mental health and nervous system disorders, a model trained on 30 days of mHealth data attains the same performance as using the full 147-day window of data, in terms of AUC increase over a baseline model that uses only demographics, height, and weight. Additionally, for the same cluster of conditions, only 7 days of data are sufficient to realize 62% of the maximum increase in AUC over baseline attainable using the full window. The results suggest that for some conditions health-related digital phenotyping in free-living conditions can potentially be performed in a relatively short amount of time, imposing minimal disruptions on user habits.

CCS Concepts

•**Computing methodologies** → **Neural networks;** *Uncertainty quantification;*

1. INTRODUCTION

Consumer mobile health (mHealth) applications run the gamut of wearable devices capturing physiological data to personalized diaries enabling the user to track calorie intake, physical activity or mental health states [8]. The in-depth metrics these devices can provide, such as minute-level step, heart rate, or sleep data, open an opportunity to explore the effects of day-to-day behaviors on genetic and environmental determinants of disease [7].

DigitalBiomarker '17, June 23, 2017, Niagara Falls, NY, USA.

 ISBN 978-1-4503-4963-5/17/06…$15.00

DOI: http://dx.doi.org/10.1145/3089341.3089347

Specifically, the ability to capture measurements continuously over time improves the quality and breadth of inferences that can be made on health and disease states. First, for measured quantities whose frequency of variation over time is significantly lower than that at which they are sampled, repeated measurements provide a way to reduce noise and obtain more accurate estimations of the underlying quantity [13]. Second, continuous monitoring provides the ability to detect changes and trends, enabling first-order analyses on disease progression and health maintenance trajectories [11]. Finally, prolonged observation times increase the probability of capturing rare anomalous signals that could nevertheless be highly informative of an individual's underlying clinical state, such as an uncommon or sporadic heart arrhythmia [5].

The combined benefits derived from reducing noise, capturing trends, and increasing the likelihood of detecting rare events are expected to grow monotonically with the length of the observation window. It is thus expected that a more extended observation period would improve the quality of inferences that can be made from data, but little is known about the impact that different timescales of data collection have on the information gained. Literature on digital biomarker discovery and behavioral phenotyping rarely provide sensitivity analyses on the data collection window. Rather, studies generally report the best-performing results, which are invariantly obtained when the totality of available data is used.

However, longer observation periods come with greater costs. In settings of free-living conditions where mHealth technologies have the ability to penetrate, longer observation periods may be unfeasible due to circumstances such as usability concerns, loss of engagement, and limited device lifetime. In even more restrictive use cases, such as clinical studies or remote clinical monitoring where patients may not be familiar with the mobile technologies involved, minimizing the burden of wearing or using the device becomes paramount.

Contribution.

In previous work [9] we reported findings on the phenotyping of self-reported chronic conditions from passively collected steps and sleep data from consumer-grade devices. We showed that a higher time granularity increases predictive power for two predefined sets of chronic conditions on a population of 7,261 users. In the current study we build on that work by exploring how the performance of the best-performing model on the most granular data identified varies as we increase the length of the observation window on which the model is trained.

2. DATA

The dataset used in this study is the same as the one presented in [9]. The dataset is composed of data shared from users of a commercial reward-based wellness platform. Users were recruited to participate in an IRB-approved online health survey, which

captured information on self-reported demographics, basic health metrics, and a selected medical diagnoses. Participants shared the history of their step/weight/sleep data from passive activity trackers between 5/8/2016 and 10/1/2016. All values for the step/weight/sleep data were passively recorded by the relevant tracker (i.e., pedometer, sleep trackers, scale) over a 147 day data collection window; none were self-reported. To be included in the analysis, participants must have completed the survey, reported at least 10 days of step data during the data collection window and reported at least one day of minute-level data for step and sleep. As detailed in [9], gaps in per-day values are imputed using per-user linear interpolation after censoring suspected partially-reported values. We found that partially-reported values are likely on the day immediately preceding and the day immediately following a tracking gap.

Overall, 9,486 people responded to the health survey, and 7,261 met the requirements for inclusion in the study analysis. The users in the data set are 75% female with an average age of 38 years. In the survey, participants reported a binary label for whether they had been diagnosed with each condition. We used a subset of these labels as the target labels for the classification tasks. We considered two clusters of conditions: a mental health/nervous system (MH/NS) cluster composed of 6 conditions: anxiety, depression, other mental illness, chronic pain, insomnia, and sleep apnea and a metabolic/circulatory (M/C) cluster composed of 3 conditions: hypertension, type 2 diabetes, and dyslipidemia.

The main goal of the research in [9] was to quantify the effect of higher time-resolution data in phenotyping performance. For that reason, we considered models trained on inputs derived from data collected at increasing time granularities. At the lowest time granularity, we had the `demographics` category, including time-invariant features such as age, gender, ethnicity, education level, and parental status. Slowly-changing features where collectively grouped into the `basic health` category, which included weight, max weight in the past, height, and BMI. More frequently-changing data was captured by `day-level activity` features, such as per-day step counts, sleep durations, weight measurements, and binary utilization indicators for step, sleep and weight devices. Finally, the categories `minute-level step` and `minute-level sleep` contained pre-specified daily summary statistics computed from data collected at minute-level granularity. The full set of features considered contained 31 per-day summary statistics computed from minute-level step data and 18 per-day summary statistics computed from minute-level sleep data. A sample of the daily summaries collected is reported in table 1.

Table 1: Example of per-day summary statistics computed from minute-level data, from [9]

Kind	Description
Step	Longest streak with mean of 30+ steps/minute
Step	Time of day for the first step taken
Step	Max step count over all 6 minute windows in a day
Sleep	Number of restless sleep periods during the night
Sleep	Time user goes to bed

3. METHODS

We cast the phenotyping of chronic medical conditions as a set of *sequence classification* tasks. Given multivariate time series $X=[\boldsymbol{x}_1,...,\boldsymbol{x}_T]$ of tracked behavioral data for T days for a user, we estimate the conditional probability $p(y \mid X)$ of the target y (e.g., a binary label indicating an anxiety diagnosis). In the full data set, $\boldsymbol{x}_t \in \mathbb{R}^{74}$ contains all per-day features computed for all input time-resolutions: `demographics`, `basic health`, `day-level activity`, `minute-level step`, and `minute-level sleep`, as described in Section 2.

3.1 Model Description

In prior work [9] we identified the multi-task convolutional neural network (MT-CNN) as the best performing of the sequence classification models considered. The architecture of the temporal convolution neural net defined in [9] for a subset of input channels is shown in Figure 1. Each individual sequence of the 74 features $X_u^{(k)}=[x_1^{(k)},...,x_T^{(k)}]_u$ (e.g., demographics, daily step counts, time of day of first step taken) is fed separately to a two-stage univariate feature extractor, where each stage consists of a 1D temporal convolution followed by a non-linear activation function and a pooling operation. In contrast to previous work on time series CNN [12], we use a hyperbolic tangent non-linearity (tanh) and max-pooling (vs. a sigmoid and average pooling) and dropout of probability 0.5 before each pooling layer. The output of the feature extraction layers is flattened and fed to a standard fully connected multilayer perceptron (MLP) with one hidden layer [4]. The hidden layer uses a rectified linear (ReLU) activation function and dropout of probability 0.5 before the final sigmoid output. The CNN is trained using gradient descent with back-propagation to minimize the negative log likelihood of the true label y: $\text{loss}(y,\hat{y})=-(y\log\hat{y}+(1-y)\log(1-\hat{y}))$ where $\hat{y}=p(y \mid X)$. The final CNN architecture includes two convolutional layers of 8 and 4 filters with kernels of width 7 and 5, respectively. Both filters use step size of length 2, and are followed by max pooling with width 2 and step size 2. The fully-connected hidden layer has 300 nodes. The CNN was implemented and trained using the open source Keras package [2].

3.1.1 Multi-task Training

We used multi-task learning to solve the predictive tasks correspondent to all chronic medical conditions considered simultaneously.

Multi-task training can improve performance on individual tasks, especially in the absence of large labeled data sets and when the tasks are related [1, 6]. The specific case of multi-task neural nets provides a flexible way to perform multi-task learning where sharing can happen in earlier feature learning layers. This constitutes a major advantage as compared to other multi-task techniques, such as multi-task linear models. We believe that the multi-task frameworks such as the one described hold promise in learning behavioral features transferable across data sets.

To train a single neural net to solve C different predictive tasks simultaneously, we add a separate output (with its own output weights $\boldsymbol{w}_c$) for each task c. The training loss for a single example with label vector $\boldsymbol{y}=[y_1,...,y_C]$ is the average over the individual task losses $\text{loss}(\boldsymbol{y},\hat{\boldsymbol{y}})=(1/C)\sum\text{loss}(y_c,\hat{y}_c)$.

3.1.2 Effect of Input Time Resolution

The main result of our prior work is that the best classification performance obtained by the MT-CNN across all 9 conditions was achieved by feeding the model with all features up to the highest time resolution: `demographic`,`basic health`, `day-level activity`, `minute-level step`, `minute-level sleep`. The results are summarized in Figure 2a. Averaged across all tasks from both disease clusters, the respective AUCs are .660, .686, .703, .707, and .719. The addition of the activity-tracker layers provides the largest incremental AUC improvement for the MH/NS condition cluster. In contrast, metabolic/circulatory conditions were better predicted by `demographic` and `basic health` data only.

3.2 Selection of Data Collection Windows

The main goal of this research is to study the impact of the data collection window length on phenotyping performance. To this

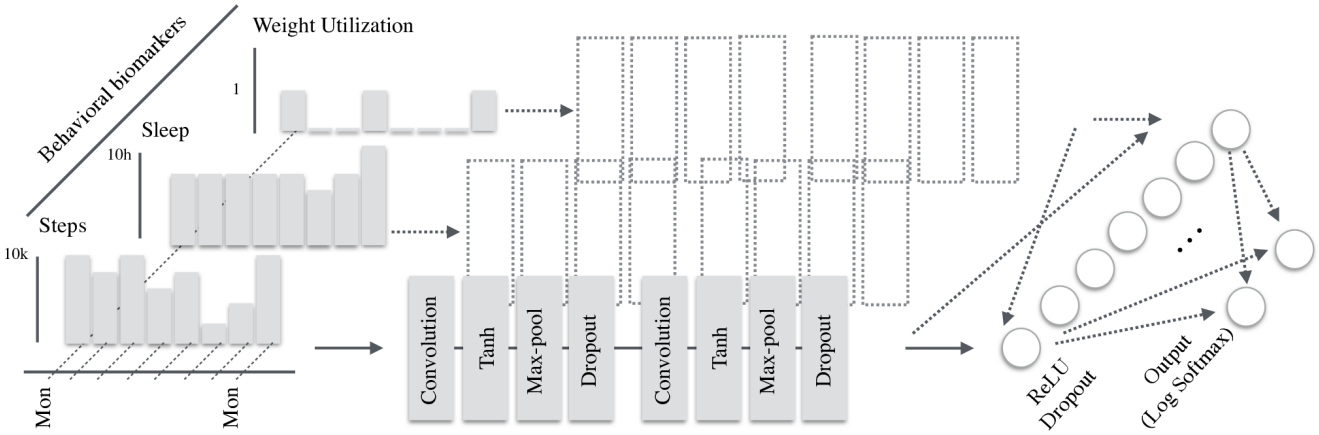


Figure 1: The temporal CNN architecture used in [9]. A subset of 3 (daily step counts, sleep duration, weight utilization) of the 74 possible input channels is shown.

end, we consider the MT-CNN trained on data up to the highest time resolution from [9] and report performances on classification tasks when features are computed on collection windows that are truncated at increasing lengths: 3, 7, 30, 60, and 147 days from the end of the data collection window. In other words, we report on the performance of the model as if it had available for training only the most recent 3,7,...,147 days of data.

4. RESULTS

We measure classifier performance using area under the ROC curve (AUC), averaged across four cross-validation folds. Each fold consists of a training (50%), validation (25%), and test (25%) dataset. The classifiers are trained on the training set, hyperparameters are tuned on the validation set, and all AUCs were computed on the held-out test set. Significance tests are performed using paired t-tests across all tasks and folds.

Results are reported in Figure 2b. The AUC reported for the 0-day observation length is the baseline obtained by training the model on only static and slowly-changing features: `demographics` and `basic health`. Only the mental health/nervous system condition cluster is reported for varying mHealth observation windows length, as the model for metabolic/circulatory cluster does not see any improvement in performance over baseline even at maximum window length (147 days). (See Figure 2a and [9] for details.) For the mental health/nervous system condition cluster, the 3-day window length achieves .656 AUC, which is a significant ($p<.01$) improvement over baseline. The next significant ($p<.01$) gain of AUC happens at 7 days of collection data, which yield 62% of the overall 4.7-point AUC increment over baseline achieved by using the full 147-day window. Further increasing the window from 7 to 30 days yields another significant ($p<.05$) additional improvement over baseline, achieving a AUC that is within error bars of the maximum one attainable with 147 days of data. No additional significant improvements are seen when increasing the mHealth observation window length beyond 30 days.

The fact that the MH/NS cluster sees steady improvements in AUC as the length of observation increases, taken together with the fact that higher time-resolution provides an increased added benefit in the MH/NS cluster (see Figure 2a), may indicate that discriminative patterns of mental health/nervous system-related conditions are more subtle than those related to M/C conditions, thus requiring both higher time resolution and longer observation periods to be detected.

5. CONCLUSIONS

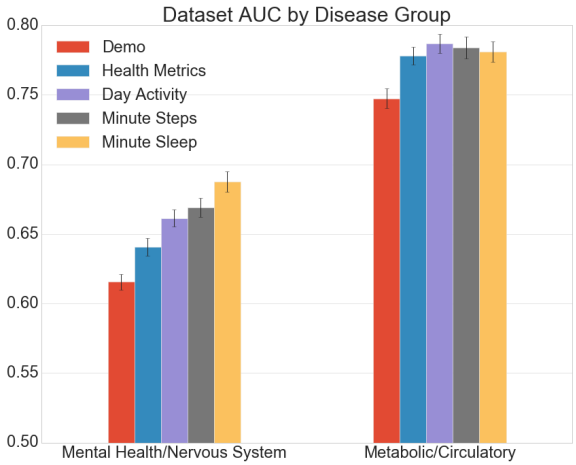
In this work we studied how the performance of phenotyping self-reported chronic-conditions from passively collected mHealth data varies as a function of the length of the data collection window. Our results show that for the mental health/nervous system disease cluster, a model trained on just three days of minute-level mHealth data significantly improves prediction performance over a baseline model trained only on demographics and height/weight. Furthermore, we show that using more than 30 days of data does not yield any additional marginal improvement over baseline.

It is important to stress that the results we report are specific to the kind of phenotyping considered, that is predicting self-reported chronic conditions. However, we believe that these findings provide indications that high-frequency activity data can reveal information related to health status relatively quickly, and ultimately help inform the design of studies, analyses, and interventions relying on the use of wearable devices. Given that many of the new generation trackers feature a battery life between 3 and 7 days, the results indicate that a significant amount of information could be captured even within a single charge cycle. This could help mitigate the risk of drop off and loss to follow up, considering that removing the wearable to charge the battery is a notorious point of friction in free-living conditions.

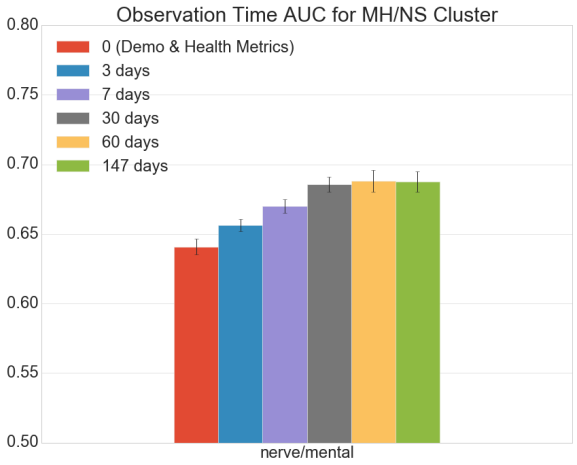
Consistent with our prior work, we found that the CNN trained on raw temporal time-series attained good performance across the different data collection windows. The effect is enhanced when tasks are predicted jointly via multi-task learning [10]. We surmise that the MT setting allows the CNN to learn feature representations over more common conditions and then leverage those representations when classifying rarer conditions. Although the generalizability of these models outside the current dataset needs further investigation, we believe they hold promise to enable more personalized solutions in medicine and health care [3].

We also found that the added value of different activity collection windows appears to depend on the medical condition of interest. This has important implications in the design and costs of population studies using consumer mobile health devices. Further studies are needed to delineate data collection needs for different chronic conditions.

Finally, it is to be noted that the current research presents average results over the population under study, but one can expect significant variability across individuals in the time it takes to build informative phenotypes. Patterns of individual behavior that are more habitual may be quicker to learn. On the other hand, more habitual individuals may take longer to reveal discriminative



(a) AUC by input data time resolution for the MT-CNN



(b) AUC by length of the observation period for the MT-CNN

Figure 2: Classification performance by condition cluster. Error bars are standard error of the mean.

patterns that lie outside their routine. Further research should focus on estimating individual-level "learning" times, with the goal of developing personalized, adaptive algorithms aimed at minimizing the time and burden of data collection on each individual, while maintaining guarantees on the quality of the information gained.

6. ACKNOWLEDGEMENTS

The authors would like to thank David Stück for the insightful comments and suggestions.

7. REFERENCES

[1] R. Caruana, S. Baluja, T. Mitchell, et al. Using the future to "sort out" the present: Rankprop and multitask learning for medical risk evaluation. In *Advances in Neural Information Processing Systems (NIPS) 8*, pages 959–965, 1996.

[2] F. Chollet. Keras. https://github.com/fchollet/keras, 2015.

[3] L. Hood and S. H. Friend. Predictive, personalized, preventive, participatory (p4) cancer medicine. *Nature Reviews Clinical Oncology*, 8(3):184–187, 2011.

[4] 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.

[5] X. Li, J. Dunn, D. Salins, G. Zhou, W. Zhou, S. M. S.-F. Rose, D. Perelman, E. Colbert, R. Runge, S. Rego, et al. Digital health: Tracking physiomes and activity using wearable biosensors reveals useful health-related information. *PLoS Biology*, 15(1):e2001402, 2017.

[6] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzel. Learning to diagnose with LSTM recurrent neural networks. In *Proceedings of the 2016 International Conference on Learning Representations (ICLR)*, 2016.

[7] Precision Medicine Initiative Working Group et al. Report to the advisory committee to the director: The precision medicine initiative cohort program—building a research foundation for 21st century medicine. *Washington, DC: National Institutes of Health*, 2015.

[8] Quantified Self Labs. The quantified self. www.quantifiedself.com, 2016. Accessed: 2016-05-20.

[9] T. Quisel, D. C. Kale, and L. Foschini. Intra-day activity better predicts chronic conditions. In *30th Conference on Neural Information Processing Systems (NIPS 2016)*, 2016.

[10] B. Ramsundar, S. Kearnes, P. Riley, D. Webster, D. Konerding, and V. Pande. Massively multitask networks for drug discovery. *arXiv preprint arXiv:1502.02072*, 2015.

[11] P. Schulam and S. Saria. A framework for individualizing predictions of disease trajectories by exploiting multi-resolution structure. In *Advances in Neural Information Processing Systems*, pages 748–756, 2015.

[12] 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.

[13] A. Zięba and P. Ramza. Standard deviation of the mean of autocorrelated observations estimated with the use of the autocorrelation function estimated from the data. *Metrology and Measurement Systems*, 18(4):529–542, 2011.