Development of machine learning: derived mhealth composite biomarkers for trial@home clinical trials

Zhuparris, A.

APPENDICES

Summary

Introduction

The traditional methods of monitoring Central Nervous System (CNS) diseases often rely on sporadic in-person clinical assessments conducted under clinical settings, which may offer an incomplete or distorted representation of a patient's condition.[1,2] This episodic and in-person approach can miss fluctuations in a patient's condition and doesn't capture a complete picture of their daily living. However, advances in mobile health (mHealth) technologies, including smartphones, wearables, and tablets, offer a potential solution for addressing these limitations by enabling continuous, real-time data collection on a patient's daily living.[3] These mHealth technologies can monitor a variety of health metrics, like heart rate, sleep patterns, and daily physical activity throughout the day and night, regardless of the patient's location. Using mHealth technologies to remotely collect data unobtrusively can provide a clinician a more complete overview of a patient's clinical status. The integration of mHealth and ML into clinical trials should be viewed as a complement to, rather than a replacement for, traditional clinical methodology. The clinical expertise of humans, which includes clinical experience and human rapport remains irreplaceable. As both mHealth technologies, ML, and clinical practices continue to evolve, this integrated approach allows for a more dynamic and data-driven approach, which may ensure that the design of clinical trials remain at the forefront of both technological and medical advancements.

The sheer volume and complexity of data generated through mHealth devices can present new challenges. It's not merely the size but the heterogeneity of the data that makes manual analysis not just labor-intensive but also difficult to model.[4,5] This is where Machine Learning (ML) comes into play. **Chapter 2** underscores the potential for ML algorithms to develop validated mHealth-based biomarkers that can be deployed in clinical trials.[6] ML algorithms can efficiently sift through vast and multifaceted datasets to identify patterns or correlations that may aid the clinical interpretation of the data. By combining ML algorithms with mHealth data to create remotely monitored biomarkers, we can potentially create novel mHealth biomarkers that can be used for diagnosis classification, symptom severity estimation, and quantification of treatment effects. These biomarkers can potentially generate novel insights that may be missed by the clinical gold standard assessments, making it possible to gain a deeper understanding of disease states.[4] However, this relatively young field still requires further research and standardization to encourage adoption of these technologies into clinical trials.

In the following sections, I will summarize the findings and discussions presented in my previous thesis chapters that explore the varied applications and challenges of mHealth biomarkers in clinical trials. I will address how these biomarkers can be developed and applied for diagnosis classification, and as a result offer novel insights into disease-related behavioural profiles that may be elusive in conventional clinical settings. Additionally, the role of mHealth biomarkers in estimating symptom severity will be discussed, and I will examine the importance of developing mHealth biomarkers that are reliable across different conditions and populations. I will also speak to how these biomarkers can be designed for treatment detection, setting the stage for longitudinal monitoring of treatment efficacy. Finally, I will delve into the limitations of mHealth biomarkers, identifying areas that warrant further research and standardization.

Disease Classification

In the context of clinical trials, disease severity classification biomarkers not only offer a quantifiable measure to assess the baseline severity of a disease among trial participants, but it can also act as a reference to track disease progression over time. When evaluating the effectiveness of investigational drugs, these biomarkers become invaluable. If the drug aims to influence the trajectory of a disease, a change in the biomarker's course over time can be indicative of the drug's effect. As a result, leveraging disease severity classification biomarkers can enhance the precision

and reliability of clinical trial outcomes, ensuring that potential treatments are assessed both for their immediate impact and their influence on the longer-term progression of the disease.

Chapter 3 investigated the feasibility of classifying Facioscapulohumeral dystrophy (FSHD) patients and healthy controls using the CHDR's Trial@Home platform. Key features, such as sleep activity and location patterns, were identified that distinguished between FSHD patients and controls.[9] This suggests that significant variances observed in sleep and location patterns might serve as potential novel clinical biomarkers as they currently are not captured by the gold standard assessments of FSHD.[10] These biomarkers, in turn, can be essential in guiding the process of drug development, potentially offering a targeted approach for drug interventions in treating or managing the associated conditions. [11]

Achieving optimal classification accuracy requires a delicate balance between the quantity of features and the duration of monitoring. Introducing a broader range of features from various sensors, such as those from smartwatches and smartphone GPS systems, can improve the precision of the predictions. However, increasing the amount of information into a model also adds complexity to the clinical understanding of these mHealth biomarkers and increases the patient's burden of increased data collection.[12,13]

SYMPTOM SEVERITY ESTIMATION

mHealth biomarkers, when utilized for symptom severity estimation, offer an innovative approach to assessing the effects of drug interventions in clinical trials. As researchers assess new drugs in Phase 2 trials, understanding the relationship between a drug, its dosage, and its resultant effects over time is pivotal.[14] mHealth biomarkers can provide a clear picture of this relationship, aiding in establishing a safe and effective dosage range. mHealth biomarkers also have the potential to serve as immediate indicators of a drug's efficacy. They can quantify symptom fluctuations over time, offering a more comprehensive view compared to labor-intensive methods like clinical interviews. This frequent monitoring can be especially valuable in discerning even the most subtle changes in symptom severity, which is fundamental for early identification of the efficacy of a treatment. By continuously monitoring changes in the biomarkers, researchers can gain valuable feedback on whether the drug is having its intended effect, which is especially crucial during Phase 2 trials where therapeutic effects are under scrutiny. For these biomarkers to be regarded as clinically valid, it is imperative that they correlate with recognized clinical endpoints. Whether those endpoints concern disease progression, symptom relief, or other clinically relevant measures, a strong association assures that the biomarker is a trustworthy measure of the drug's impact.

Chapter 4 investigated the performance of multi-task models to simultaneously estimate the scores of two clinical assessments, the FSHD clinical score and the Timed Up and Go (TUG) test.[15] Traditional single-task models, while they may be effective for predicting a single outcome, may fall short when applied to the multi-dimensional symptom profiles that often encountered in clinical settings. Therefore, the principal advantage of multi-task models over their single-task counterparts is their ability to leverage shared representations and insights across multiple clinical assessments.[16–18] Moreover, the ability of multi-task models to generalize from one clinical assessment to another can be critical in evaluating disease severity across a spectrum of assessments. For example, if the model identifies a deterioration in the FSHD clinical score, it might also predict a parallel decline in the TUG score. Finally, multi-task models can offer a more holistic view of patient health, encompassing various facets of disease severity in a single, unified framework. By enabling the parallel assessment of multiple assessments, these models can provide a fuller, more nuanced picture of disease status, thus guiding more targeted and effective interventions.

In **Chapter 5,** the significance of self-reported outcomes, specifically the Depression Anxiety Stress Scale (DASS) and the Positive and Negative Affect Schedule (PANAS), emerged as decisive features for the depression models. Their inclusion served as a robust indicator for subjective

psychological states, highlighting the irreplaceable value of patient input in capturing the nuances of mental health conditions. Interestingly, even though passively collected features like walking speed and location were not as predictive as DASS and PANAS, they still made valuable contributions to the overall effectiveness of the models. This finding also underscores the importance of integrating real-world, passively collected data, as it appears to reveal patterns and insights that might be overlooked in more controlled clinical settings. Additionally, the models' capacity to accurately represent the full spectrum of depression severity was augmented by the inclusion of healthy controls. This inclusion not only enhanced the robustness of the models but also extended the representation of the potential remission states of depression in the models. This multidimensional approach, combining both active and passive data collection, thus provides a more comprehensive and nuanced understanding of mental health conditions.

Estimating symptom severity using mHealth biomarkers presents specific challenges, particularly when considering the inherent variability in both the devices and the patients themselves. One significant concern is the inter-device variability.[2] Difference in mHealth devices may produce slightly varied measurements, leading to inconsistencies in the collected data. This variation can introduce noise into analyses, potentially skewing results or diminishing the precision of symptom severity estimations. Additionally, symptom severity and expression itself can vary within and between patients, adding another layer of complexity to modelling efforts. External factors that cannot be controlled or accounted for can also confound readings. For instance, while an mHealth device might detect an increased heart rate as a potential symptom of a health condition, however this elevation could be attributed to external influences such as anxiety, physical exercise, or other non-medical causes. Thus, distinguishing genuine symptom fluctuations from these external factors remains a challenge in leveraging mHealth biomarkers for accurate symptom severity estimation.

Treatment effects

For detecting treatment effects, mHealth biomarkers need to demonstrate their ability to detect changes in disease activity following a drug intervention. In essence, this approach to designing and validating mHealth biomarker can make them valuable tools not just for understanding a disease but also for tailoring and evaluating treatment strategies. Here, the focus isn't solely on the biomarker as a predictive or diagnostic tool but also on its sensitivity and efficacy in detecting treatment effects relative to the gold standard. By demonstrating sensitivity to treatment-induced changes, these biomarkers can serve as more dynamic endpoints in trials, which can facilitate more immediate and accurate assessments of a treatment's impact.

Chapter 8 discusses the development of mHealth biomarkers for monitoring the effects of antiparkinsonian drugs and estimating Parkinson's disease symptom severity.[19] The alternative index finger tapping (IFT) biomarker was found to be more predictive and sensitive to treatment effects in motor function than the traditional MDS-UPDRS III score, both in terms of accuracy and clinical significance. Treatment effects were detected at 45 minutes for the thumb–index finger tapping (TIFT) biomarker and at 60 minutes for the IFT composite biomarkers. This coincides well with the mean onset of action for the drug L-dopa/carbidopa, which is around 50 minutes. The findings suggest that IFT and TIFT are sensitive tools for assessing motor function in the context of symptomatic treatments for conditions like Parkinson's disease, potentially identifying small and early changes missed by traditional measures. The large effect sizes also found in this study could reduce the sample size requirements and enhance the statistical power for future studies involving tapping tasks. This pilot study can advance the understanding of how to accurately detect and measure treatment effects on fine motor function, particularly in conditions like Parkinson's disease. It not only validates the efficacy of new biomarkers but also provides methodological guidance for validating novel biomarkers in future research focus on investigating drug effects.

Repeatability of predictions over time and settings

In the context of clinical research, the term 'repeatability' refers to the ability of a test, measurement, or algorithm to yield consistent results when it is performed multiple times under the same conditions.[20,21] In both clinical and home settings, consistent monitoring is vital for tracking the progression or alleviation of symptoms. For instance, if a cough detection algorithm is used to monitor the effectiveness of a new asthma medication in children, inconsistent results would compromise the integrity of the research and could lead to incorrect conclusions. For algorithms designed to monitor biological signals or events—such as coughs or cries—repeatability across different data collection settings and across patients is a key attribute that underscores the algorithm's reliability.[20] In the fields of computer science and ML, repeatability can be interchanged with 'robustness' and 'external validity.' Essentially, these terms—repeatability, robustness, and external validity—point towards an algorithm's consistent performance across varying conditions and datasets. **Chapter 6** and **Chapter 7** focused on the development of a smartphone-based algorithm for automated cough and cry detection among infants and children.[22,23] Both algorithms show strong repeatability, which is crucial for consistent monitoring over time. The cry algorithm appears robust against different types of physical barriers and can be used at various distances, making it flexible for real-world applications. While both algorithms show some level of inter-device variability, it is within an acceptable range that does not severely compromise their utility. Both algorithms are affected by background noise, albeit to varying extents. This points to an area for potential improvement. These findings suggest both algorithms are robust enough for potential use in monitoring cries and coughs in a clinical setting or for home-based care, although adjustments may be needed depending on the device or environmental conditions used.

Limitations

Many conditions, like mental health disorders or chronic diseases, are multifaceted and may not be fully captured by a single gold standard assessment or a single device. In these cases, both the gold standard and the mHealth devices may not capture the complexity of the disease, leading to discrepancies when comparing the true and predicted clinical scores. These discrepancies can be the result of three causes. First, limitations of mHealth devices to capture all clinically relevant behaviors. For instance, the mHealth devices failed to capture and therefore failed to predict the upper arm functionality of FSHD's patients, as seen in **Chapter 3** and **4**.[9,15] Second, shortcomings of the gold standards in capturing all clinically relevant behaviors. As seen in **Chapter 5**, we found that walking and travel behaviors are predictive of MDD, however, these characteristics are not addressed by the SIGH-D IDSC. Further, the gold standard's limitations, such as inter-rater variability or a failure to capture the full complexity of a disease, may introduce biases affecting the biomarker's reliability. In some cases, the gold standard involves human assessment, which can vary depending on the rater's expertise or even day-to-day conditions. For instance, in **Chapter 8**, the finger tapping tasks that tracks multiple tapping-related characteristics could offer insights into motor functionality that might be more comprehensive than traditional Parkinson's Disease studies that solely rely on clinical observation.[19] Third, there may be disparities between the objective behavioral biomarkers and subjective endpoints. For example, a depressed patient may report feeling more restless when in bed, but the objective sleep data captured by the smartwatch shows that the patient slept for 8 hours. As a result, the objective measure of sleep may not correlate well with the subjective experience of sleep as seen in **Chapter 5**. Therefore, it's crucial to consider both objective measurements and subjective experiences when evaluating the effectiveness of mHealth devices for monitoring and managing conditions like depression. The objective measurements may not always be a representative endpoint for subjective experiences.

The discrepancies between mHealth sensors and the gold standard can affect how reliable clinicians and researchers perceive these sensors to be. For a new technology to be integrated into clinical trials, it must either closely match the gold standard or clearly exhibit its superiority. It's worth noting that a lower correlation between mHealth biomarkers and the gold standard might not indicate poor clinical validity of the novel biomarker; instead, the mHealth system could be capturing aspects overlooked by traditional methods. Therefore, understanding the limitations and biases inherent in both mHealth biomarker and gold standards is critical for making accurate clinical decisions. If clinicians are aware of these factors, they can make more nuanced interpretations of the data.

Conclusion

In conclusion, mHealth biomarkers and ML can be expected to cause a paradigm shift in the monitoring and management of CNS diseases. These advanced technologies, facilitated by smartphones, wearables, and tablets, can provide a more immediate, continuous, and accurate assessment of disease. Therefore, these mHealth biomarkers could transform traditional episodic evaluations into nuanced, longitudinal data-driven analyses. The research findings demonstrate the robust predictive capabilities, accuracy, reliability, and clinical relevance of these developed biomarkers. However, it's important to acknowledge the need for further research, development, and standardization, to fully realize the benefits of these innovations. Ultimately, these advancements not only offer a more comprehensive understanding of disease severity and progression but also provide better tools to determine the potential efficacy of pharmacological interventions.

REFERENCES

1 Dobkin BH, Dorsch A. The Promise of mHealth. *Neurorehabil Neural Repair*. 2011;25(9):788-798. doi:10.1177/1545968311425908

2 Kakkar A, Sarma P, Medhi B. mHealth technologies in clinical trials: Opportunities and challenges. *Indian J Pharmacol*. 2018;50(3):105. doi:10.4103/ijp.IJP_391_18

3 WHO. *MHealth New Horizons for Health through Mobile Technologies*. Vol 3.; 2011. doi:10.4258/hir.2012.18.3.231

4 Liang Y, Zheng X, Zeng DD. A survey on big data-driven digital phenotyping of mental health. *Information Fusion*. 2019;52(July 2018):290-307. doi:10.1016/j.inffus.2019.04.001

5 L'Heureux A, Grolinger K, Elyamany HF, Capretz MAM. Machine Learning with Big Data: Challenges and Approaches. IEEE *Access*. 2017;5:7776-7797. doi:10.1109/ACCESS.2017.2696365

6 ZhuParris A, de Goede AA, Yocarini IE, Kraaij W, Groeneveld GJ, Doll RJ. Machine Learning Techniques for Developing Remotely Monitored Central Nervous System Biomarkers Using Wearable Sensors: A Narrative Literature Review. *Sensors*. 2023;23(11):5243. doi:10.3390/s23115243

7 Kruizinga MD, Stuurman FE, Exadaktylos V, et al. Development of Novel, Value-Based, Digital Endpoints for Clinical Trials: A Structured Approach Toward Fit-for-Purpose Validation. *Pharmacol Rev*. 2020;72(4):899-909. doi:10.1124/pr.120.000028

8 Potter WZ. Optimizing early Go/No Go decisions in CNS drug development. *Expert Rev Clin Pharmacol*. 2015;8(2):155-157. doi:10.1586/17512433.2015.991715

9 Maleki G, Zhuparris A, Koopmans I, et al. Objective Monitoring of Facioscapulohumeral Dystrophy During Clinical Trials Using a Smartphone App and Wearables: Observational Study. *JMIR Form Res*. 2022;6:1-13. doi:10.2196/31775

10 Hamel J, Johnson N, Tawil R, et al. Patient-Reported Symptoms in Facioscapulohumeral Muscular Dystrophy (PRISM-FSHD). *Neurology*. 2019;93(12):E1180-E1192. doi:10.1212/WNL.0000000000008123

11 Williams JBW. A Structured Interview Guide for the Hamilton Depression Rating Scale. *Arch Gen Psychiatry*. 1988;45(8):742-747. doi:10.1001/archpsyc.1988.01800320058007

12 Rowland SP, Fitzgerald JE, Holme T, Powell J, McGregor A. What is the clinical value of mHealth for patients? *NPJ Digit Med*. 2020;3(1):4. doi:10.1038/s41746-019-0206-x

13 Wang F, Preininger A. AI in Health: State of the Art, Challenges, and Future Directions. *Yearb Med Inform*. 2019;28(1):16-26. doi:10.1055/s-0039-1677908

14 Lipsmeier F, Taylor KI, Kilchenmann T, et al. Evaluation of smartphone-based testing to generate exploratory outcome measures in a phase 1 Parkinson's disease clinical trial. *Movement Disorders*. 2018;33(8):1287-1297. doi:10.1002/mds.27376

15 ZhuParris A, Maleki G, Koopmans I, et al. Estimation of the clinical severity of Facioscapulohumeral Muscular Dystrophy (FSHD) using smartphone and remote monitoring sensor data. In: FSHD *International Research Congress*. FSHD international research congress; 2021.

16 Li Y, Tian X, Liu T, Tao D. Multi-task model and feature joint learning. *IJCAI International Joint Conference on Artificial Intelligence*. 2015;2015-Jan ua(Ijcai):3643-3649.

17 Yoon H, Gaw N. A novel multi-task linear mixed model for smartphone-based telemonitoring. *Expert Syst Appl*. 2021;164(September 2019):113809. doi:10.1016/j.eswa.2020.113809

18 Lu J, Shang C, Yue C, et al. Joint Modeling of Heterogeneous Sensing Data for Depression Assessment via Multi-task Learning. *Proc ACM Interact Mob Wearable Ubiquitous Technol*. 2018;2(1):1-21. doi:10.1145/3191753

19 ZhuParris A, Thijssen E, Elzinga W, et al. Detection of treatment and quantification of Parkinson's Disease motor severity using finger-tapping tasks and machine learning. In: *9th Dutch Bio-Medical Engineering Conference*. 9th Dutch Bio-Medical Engineering Conference; 2023.

20 Kruizinga MD, Heide N van der, Moll A, et al. Towards remote monitoring in pediatric care and clinical trials—Tolerability, repeatability and reference values of candidate digital endpoints derived from physical activity, heart rate and sleep in healthy children. Harezlak J, ed. *PLoS One*. 2021;16(1):e0244877. doi:10.1371/journal.pone.0244877

21 Makai-Bölöni S, Thijssen E, van Brummelen EMJJ,

Groeneveld GJ, Doll RJ. Touchscreen-based finger tapping: Repeatability and configuration effects on tapping performance. Virmani T, ed. *PLoS One*. 2021;16(12):e0260783. doi:10.1371/journal.pone.0260783

22 ZhuParris A, Kruizinga MD, Gent M van, et al. Development and Technical Validation of a Smartphone-Based Cry Detection Algorithm. *Front Pediatr*. 2021;9:262. doi:10.3389/fped.2021.651356

23 Kruizinga MD, Zhuparris A, Dessing E, et al. Development and technical validation of a smartphone-based pediatric cough detection algorithm. *Pediatr Pulmonol*. 2022;57(3):761-767. doi:10.1002/ppul.25801