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

Zhuparris, A.

PART V

DISCUSSION

CHAPTER 9

General discussion

Introduction

This discussion chapter will unpack the motivation behind the development and adoption of MHEALTH biomarkers for clinical diagnosis, symptom severity estimation, and treatment effect detection. As with any novel biomarker, there are multiple implications and limitations spanning the ethical, privacy, and practical domains. These considerations, especially for clinicians and their potential broader applicability to other CNS disorders, will be discussed. Moreover, I will discuss the potential of MHEALTH composite biomarkers for future clinical trials. The conclusion will provide a clear grasp of the present state, obstacles, and potential future of MHEALTH biomarkers in clinical environments.

MHEALTH biomarkers: from research to clinical application

Central Nervous System (CNS) diseases have profound impacts on various facets of daily functioning. Traditionally, the evaluation of disease severity is largely reliant on temporally confined assessments conducted indirectly by clinicians who only intermittently engage with patients, potentially supplemented by auxiliary information sourced from patient's close acquaintances, such as spouses. Consequently, the current approaches inherently yield a relatively episodic and potentially distorted view of disease progression. Traditionally, the evaluation of disease severity is largely reliant on temporally confined assessments conducted indirectly by clinicians who only intermittently engage with patients, potentially supplemented by auxiliary information sourced from patient's close acquaintances, such as spouses. Consequently, the current approaches inherently yield a relatively episodic and potentially distorted view of disease progression. In contrast, objective evaluation of Activities of Daily Living (ADL) facilitated by smartphone, wearables, and tablets offers a more immediate, continuous, and accurate portrayal of a patient's condition. By capturing real-time data on a patient's everyday functioning, these devices can provide a nuanced, longitudinal view of disease severity, which, in turn, allows for the potential to track the symptomatic impact of therapeutic interventions. Thus, the utilization of these mobile technologies for the objective quantification of ADLs not only offers a more direct, reliable, and comprehensive measure of disease severity but also illuminates the dynamics of disease progression and the potential efficacy of pharmacological interventions.

As illustrated by the literature review in **Chapter 2**, these mobile health (MHEALTH) biomarkers offer a multi-faceted and data-driven approach towards monitoring disease status, disease progression, and treatment responses, which enables a better understanding and management of these neurological and psychiatric disorders. These MHEALTH biomarkers involve the integration of multiple MHEALTH features ranging from data from smartphone, tablets, wearables, and clinical measures. Machine Learning (ML) can be valuable when there is an ambiguity or a lack of consensus regarding which features are relevant (or to what extent they are relevant) in predicting an outcome. Such novelty and ambiguity are inherent when dealing with MHEALTH data, due to the diversity of sensors used for data collection, as well as the complex interactions between disease profiles, lifestyles, environmental factors, social interactions, and other uncontrolled external factors. While the current scientific literature and clinicians' understanding of disease profiles can aid the identification of relevant features, the interplay between these features for a given individual or population can be difficult for experts to discern. Given this difficulty, clinicians may be less enthusiastic about including these new measures into clinical trials. This thesis proposes that for MHEALTH devices and ML to truly benefit healthcare, they must provide substantial benefits to patients and clinicians beyond a digitized gold standard measurement. This thesis argues that these MHEALTH biomarkers can provide a nearly continuous, remote, unobtrusive profile of disease in a way that traditional gold standard measurements, digital or not, cannot.

Classifying a diagnosis

Evaluating the classification performance of a MHEALTH composite biomarker in distinguishing patients from healthy controls is a crucial factor in assessing its suitability for the intended purpose. The magnitude of difference between the two groups can provide insights into the level of change in disease activity and aid in estimating sample sizes for future clinical trials.[1] However, the premise that a specific treatment will render a patient with a CNS more like a healthy individual is not always viable, especially in the context of CNS disorders, thus comparison to healthy controls is not always necessary or meaningful. Instead, a crucial factor lies in identifying differences between someone with mild symptoms and someone at a more advanced stage of the disease. Nevertheless, for the initial development and validation process, we have created classifiers capable of distinguishing between control subjects and patients. If successful classification is achieved, the MHEALTH features used to develop the composite biomarkers can provide valuable information for understanding disease activity. This information can further inform the development of targeted interventions and monitoring strategies for patients with these conditions.

For a biomarker to have clinical utility, it must demonstrate clinical validity. Clinical validity refers to the ability of a biomarker to accurately identify, predict, or estimate the presence or severity of a disease or condition. MHEALTH biomarkers currently aim to approximate clinicians' decisions based on the available training data. While a clinical diagnosis has long been the gold standard, the diagnostic potential of MHEALTH biomarkers may offer novel insights into disease and treatment activities. The selection of an appropriate reference gold standard measurement significantly influences the clinical validation process of MHEALTH biomarkers, as the biomarker's performance is inherently tied to the quality and validity of the chosen gold standard. The reliance on a gold standard measure with limited validity or substantial interrater variability can introduce potential biases and undermine the accuracy and reliability of the biomarker. For FSHD, a genetic test is required for a diagnosis,[23] while a MDD patient would be diagnosed if they persistently demonstrate five or more depressive symptoms (such as depressed mood, anhedonia, lack of energy, poor concentration, or sleep disturbances).[5] The subjective and descriptive nature of the MDD clinical scales reduces its sensitivity to subtle psychomotor symptoms. **Chapters 3** successfully developed classification models that could distinguish between Facioscapulohumeral dystrophy (FSHD) patients and healthy controls. This study leveraged remotely collected multi-faceted data, including information on social interactions, location, and sleep activity, to classify a clinical diagnosis that was assessed on genetic, functional, or behavioural factors. This innovative approach expands our knowledge beyond the limited measurements obtained within the confines of a clinical setting. By harnessing the power of MHEALTH technologies and data analytics, we can now capture real-life experiences and behaviours that were previously unexplored. However, it is crucial to assess the clinical validity of these biomarkers to ensure their effectiveness and accuracy in real-world applications.

Given that MHEALTH devices mainly collect real-world data, these biomarkers may be influenced by real-world factors, such as location, weather, life-style factors, and concomitant drug use.[1] Individual variations in behaviour can potentially affect the reliability of the biomarkers. If a composite biomarker can accommodate the inherent variability observed in real-world settings, while consistently producing reliable results, it can be considered a viable and validated measurement. Thus, longitudinal studies and test-retest reliability analyses can help determine the stability and consistency of these biomarkers. As addressed in **Chapter 2**, research on the consistency and repeatability of a composite biomarker, as well as its ability to account for long-term variability, is currently limited. To ensure that the biomarkers developed in this thesis were reliable and consistent, **Sections 2 to 4** explored the composite biomarkers' ability to consistently achieve consistent and repeatable results across subjects and time windows. Specifically, **Chapters 3 to 5**

demonstrated that using the first week of data for the development of a ML-biomarker allowed for consistent and stable prediction of symptom severity for the remainder of the trial period. This finding highlights the importance of collecting enough data for the development of a reliable composite biomarker and at least one week of data appears to be necessary for the accurate estimation of clinical severity and the monitoring of disease activity outside the clinic. **Chapters 6 and 7** demonstrated consistent intra- and inter-device reliability of the cough and cry biomarkers across different audio recording settings. **Chapter 8** illustrated that training the composite biomarkers on a single timepoint enabled repeatable and reliable estimations of treatment effects and MDS-UPDRS III scores across other time points. In conclusion, the studies included in this thesis, conducted under different settings and with different clinical populations, suggest that composite MHEALTH biomarkers show promise regarding measurement validity.

Estimating symptom severity

Symptom severity estimation based on composite biomarkers provides an objective and standardized measurement for tracking disease progression and treatment response. The development and validation of composite biomarkers for the estimation of symptom severity in clinical trials play a crucial role in determining if the composite biomarker can serve as a meaningful endpoint in clinical trials. The robust relationship between the composite biomarker's predicted symptom severity score and the gold standard score indicates the relative effectiveness of the biomarker in capturing and quantifying symptom severity, thereby supporting its utility in clinical trials. While a perfect correlation may never be achieved due to the nature of the data collected, further research should determine if the observed discrepancy is acceptable and if the cause of the discrepancy is due to the limitations of the composite biomarker or of the gold standard. **Chapters 4, 5, and 8** were aimed at developing composite biomarkers that could estimate the symptom severity of patients with FSHD, MDD, and Parkinson's Disease (PD). While the composite biomarkers demonstrated in each of these chapters showed a certain degree of promise and applicability, their alignment with the gold standards was not perfect. This highlights potential gaps for investigation and areas for refinement in measurement and predictive accuracy. Based on the studies addressed in thesis, there may be three causes for the discrepancy.

First, the MHEALTH sensors cannot monitor all behaviours that are assessed by the gold standard. For example, **in Chapter 4**, the MHEALTH sensors may have failed to capture arm, abdominal, and scapular weaknesses (which are assessed by the FSHD Clinical Score).[6] The identified limitation underscores the importance of discerning the specific aspects of disease activity that can and cannot be effectively monitored using MHEALTH sensors. However, despite this limitation, the study demonstrated the potential of MHEALTH-derived biomarkers in measuring the extent of disease severity beyond the confines of the clinical setting. This capability offers valuable insights into the manifestation of disease activity and its impact on a patient's daily quality of life.

Secondly, objectively monitored behaviour and subjective perception of behaviour are not always correlated. As shown in **Chapter 5**, the daily, detailed, and objective measures of sleep were not well-correlated with the subjective and weekly reported sleep quality. Several factors can influence the subjective reporting of sleep, including mood at the time of awakening,[7] insomnia, impaired memory, and negative bias.[8] Previous studies have also confirmed that objective sleep assessments do not correlate with subjective reports of sleep.[9,10] This indicates that while objective measures may provide more accurate and reliable data about disease activity, subjective reports may still provide valuable insights into an individual's perception and experience of their own behaviours.

Thirdly, it is conceivable that the composite biomarker offers superior capabilities in measuring disease activity than the gold standard or at least captures distinct dimensions of disease activity that are not quantified by the gold standard. The tapping composite biomarkers presented in **Chapter 8** offer a more objective, nuanced, and comprehensive

depiction of a PD patient's fine finger movement than the MDS-UPDRS III. It is important to acknowledge that composite biomarkers may exhibit advantages over the gold standard in terms of sensitivity and specificity. Through the utilization of MHEALTH data and ML, these composite biomarkers have the potential to identify subtle disease markers that may be overlooked or missed by conventional clinical observations. By leveraging these advanced approaches, researchers can gain deeper insights into the complexities of disease activity and potentially enhance the precision and effectiveness of monitoring disease activity and treatment effects.

Further studies are needed to bridge the gap between MHEALTH sensors and traditional clinical assessments. Understanding the relationship between objective data, the gold standards, and patient feedback is pivotal. Additionally, refining composite biomarkers will drive more precise clinical monitoring. These steps are crucial for seamlessly integrating MHEALTH tools in clinical trials.

Detecting treatment effects

To evaluate if the composite biomarker is fit-for-purpose for assessing treatment effects, the biomarker needs to be evaluated for its ability to respond to changes in disease activity in response to a treatment. **Chapter 8** explored the ability of a tablet-based composite finger tapping biomarker to detect anti-parkinsonian (dopaminergic) treatment effects among PD patients. This study investigated if a composite biomarker demonstrates comparable or superior performance to the gold standard in the detection of treatment effects. The approach taken in this chapter introduces a unique perspective compared to previous chapters, as the gold standard measurement was not the predicted outcome itself. Instead, the focus was on comparing the sensitivity and efficacy of the biomarker in relation to the gold standard in the detection of treatment effects. This novel approach presents a fresh methodology for evaluating the validity of a biomarker in clinical trials as it offers a broader perspective on biomarker evaluation, going beyond the traditional notion of a biomarker as solely a predictive or diagnostic tool. This focus shifts towards providing an additional layer of evidence of the biomarkers' unique ability to capture clinically relevant changes and potentially highlighting the limitations of the gold standard.

Limitations of mhealth composite biomarkers

The nature of the MHEALTH devices used raises questions regarding the accuracy and reliability of the data, as factors such as device quality, sensor reliability, data collection protocols, and user adherence can lead to inconsistent or complete data. In turn, this can affect the reliability and validity of the composite biomarkers, and their subsequent predictions. To overcome these issues, this thesis proposes two main methodologies.

First, given that MHEALTH data is collected under free-living environments and requires patients' consent and engagement, seamless integration of MHEALTH data collection tools into existing clinical workflows is crucial. The tools should be user-friendly, compatible with the patient's lifestyle and mobile phone, and should be able to provide consistent, and formative results to the clinicians. Hence, it's crucial to report the quantity of missing data for each study and if possible, as shown in **Chapters 3**, report the study participants' experience with the remote monitoring platform to understand the causes of the missing or poor-quality data.

Second, a large and representative dataset is necessary to build a robust and generalizable biomarker. With a larger sample size, the model can capture a wider range of patterns, relationships, and variations in the data, leading to improved accuracy and generalizability of predictions. The larger sample size reduces the variability in the performance estimates, providing more reliable assessments of the model's strengths and weaknesses. Further, it provides a broader range of instances for the model to learn from, facilitating the identification of more intricate and subtle relationships between features. A representative dataset would reflect a true distribution of the target population, including various demographic factors, characteristics, and potential confounding variables. By

incorporating diverse samples, the model becomes more robust to variations and biases present in the data, ensuring its predictions are reliable across different subgroups or settings.

Reflecting on the chapters in this thesis, to estimate the minimum dataset size for MHEALTH-based clinical trials, consider the desired effect size, statistical power, variability in the specific outcome, type of outcome (e.g., classification vs. severity), potential data collection issues, and the complexity introduced by external factors and free-living conditions. Adjustments should be made based on real-world constraints and the quality of MHEALTH data. For example, in a follow-up study, the objective would be to detect a 10% improvement in FSHD symptoms under free-living conditions. We recognize that sleep activity can affect the FSHD assessments, and hence a larger sample size would be needed to account for the sleep variability. If the study spans a long period, environmental or behavioral factors such as seasons, physiotherapy sessions, or living conditions may affect the physical activity measurements. Therefore, researchers may choose to stratify their sample based on seasons, therapy, or living conditions to account for these variations.

Due to the limited sample sizes of the studies in this thesis and the literature review, it's difficult to claim if the composite biomarkers may generalize well to diverse populations, settings, or clinical trial protocols. As a result, the performance of composite biomarkers may vary across different trials and patient populations, which highlights the need to validate their effectiveness across different contexts.

Implications for clinicians

The benefits of using of MHEALTH technologies and ML to provide a clinical prediction include efficiency, consistency, accessibility, and data-driven insights. As these technologies do not experience fatigue or inter-rater variability, they can ensure more consistent and less variable clinical outcomes. The collection and analysis of diverse data sources, including patient-reported outcomes, physiological measurements, and behavioral data can enable a more comprehensive and faster understanding of disease status, disease activity, and treatment response. These biomarkers can potentially help clinicians refine or redefine how they view disease beyond traditional siloed disease-specific definitions. Further, the automated processing of large volumes of data could enable fast predictions, which would save valuable time for clinicians.

Despite their promise, it's important to note that composite biomarkers should not be considered as a replacement for traditional clinical assessments. Traditional clinical assessments, which typically involve a comprehensive evaluation of a patient's medical history, physical examination, and laboratory tests, are crucial in providing an accurate diagnosis and monitoring of disease activity. Further, they can infer an understanding of subjective and contextual factors that may not be easily captured in the medical datasets. ML rely on understanding the patterns within a training data, which may not represent all possible scenarios, and less likely to represent rare or complex cases. The critical thinking of clinicians may allow them to adapt their knowledge to diagnose challenging or atypical conditions. While MHEALTH biomarkers has shown promise for clinical assessment, this thesis argues that it is essential to view ML as a tool to augment human expertise rather than a complete replacement.

The objective of a remotely monitored clinical trial should be to develop a synergistic approach that leverages the strengths of traditional clinical assessments, MHEALTH devices, and ML. By harnessing the power of composite biomarkers alongside traditional clinical assessments, we can better quantify disease activity and provide more effective and personalized care to patients. This integrated approach has the potential to aid future developments in clinical research and contribute to significant advancements in healthcare.

Implications for other CNS disorders

Developing MHEALTH biomarkers for MDD, PD, FSHD, and hospitalized infants carries several potential implications for the development and application of MHEALTH biomarkers for other CNS disorders. The protocols and methodologies for the data collection and MHEALTH biomarker

development and application can potentially be transferred and applied to other areas such as bipolar disorder, Amyotrophic Lateral Sclerosis, and Alzheimer's disease. This cross-fertilization of methodologies can accelerate the progress of biomarker research in these related conditions. It could allow researchers and clinicians to identify similarities and differences in symptom severity and treatment responses across various conditions. Similar physiological and behavioural patterns may exist across different conditions, and using the same biomarker to monitor both populations may facilitate comparative analysis between different clinical populations. For example, the social activity biomarker to identify depressive episodes among MDD and bipolar patients. This enhances the generalizability of the research findings and allows for broader application and transferability of knowledge across a wider range of clinical populations.

Impact on future clinical trials

By identifying the optimal sensors, features, and data collection periods for the development of composite biomarkers, future clinical trials can be more efficient, less time-consuming, and less costly, which in turn can alleviate the study burden for both patients and clinicians. reducing the feature space and the amount of data required also reduces the need for more complex ML algorithms that may potentially limit interpretability and therefore adoption. More specifically, feature selection techniques can help remove noise and irrelevant data, improving the accuracy of the analysis and the interpretability of the final biomarker. **Parts 2 to 4** of the thesis employed various feature selection approaches to identify the most relevant features for analysis. This is crucial for informing future clinical trials about the specific features and corresponding sensors that are essential for achieving their research objectives. Additionally, in **Parts 2 and 3**, the studies described determined the amount of data necessary to develop a reliable composite biomarker. These findings emphasize the significance of data curation and its role in obtaining a dependable and informative composite biomarker.

Ethical implications

The ethical governance of MHEALTH biomarkers is a crucial aspect to consider in their integration into clinical trials. Clinicians and healthcare providers tend to exhibit higher levels of trust in ML-derived biomarkers that are explainable and transparent in their decision-making process. Understanding how each feature or input influences the final predictions of the biomarker can be important for its adoption. While deep learning models have shown remarkable prediction accuracy in various domains, they often lack interpretability.[4,5] Unlike traditional ML models that can provide insights into the relationships between input features and predictions, deep learning models operate as black boxes, making it challenging to explain their decision-making process. This lack of interpretability raises concerns about the accountability and fairness of MHEALTH biomarkers.

When an inaccurate prediction is made by an MHEALTH biomarker, it raises questions about who should be held responsible for any harmful or fatal consequences. The lack of interpretability in ML models hinders the ability to understand and address potential biases, errors, or limitations of the biomarker's predictions.[4,5] It becomes essential to ensure that the use of MHEALTH biomarkers in clinical trials follows rigorous ethical guidelines, including transparency, accountability, and mechanisms for addressing potential harms or errors. The integration of MHEALTH biomarkers in clinical practice requires a balance between the benefits they offer and the ethical consequences they entail. While high prediction accuracy is desirable, it should be accompanied by interpretability and transparency to ensure the fair and responsible use of these biomarkers. Ethical governance frameworks that emphasize explainability and accountability can help address concerns related to potential biases, errors, or unintended consequences associated with MHEALTH biomarkers.

Privacy implications

The integration of MHEALTH biomarkers in clinical trials brings forth significant privacy concerns and implications. The utilization of MHEALTH biomarkers in clinical trials entails the collection of an unprecedented amount of personal information about study participants.[6] In this thesis, the MHEALTH technologies used were the study participants' smartphones and third-party wearable devices. It is important to acknowledge that these technologies, although widely available, are not specifically designed as medical devices, which limits the clinician's control over their functionalities. One[6] aspect of concern is the level of control that individuals, including the study participants and device developers, have over these devices. Since these technologies are owned and operated by the participants themselves, the clinician or researcher may have limited ability to regulate or monitor their usage. This lack of control introduces potential vulnerabilities in terms of data security and privacy.[7] Unauthorized access to such sensitive information can have severe consequences, including identity theft, discrimination, or exposure of personal health details.[7] Aggregated and de-identified data, if mishandled or inadequately protected, can still carry privacy risks when re-identified or combined with other datasets. This highlights the importance of robust data anonymization and de-identification techniques to safeguard the privacy of study participants.

To mitigate these privacy concerns and potential harms, it is essential to implement stringent privacy protection measures. This includes obtaining informed consent from participants, ensuring secure data transmission and storage, and adhering to relevant privacy regulations and guidelines. Additionally, transparent communication with participants about data usage, anonymization practices, and the purpose of data collection can foster trust and promote participant engagement. By prioritizing privacy protection and adhering to best practices, clinicians can strike a balance between leveraging the benefits of MHEALTH biomarkers and safeguarding the privacy of study participants.

Conclusion

The development and application of composite biomarkers using MHEALTH devices and ML holds significant promise for clinical research. These biomarkers can integrate diverse data sources and provide a more comprehensive understanding of disease status, symptom severity, and treatment effects. The use of MHEALTH devices and ML in clinical trials presents opportunities for real-time data collection, disease symptom monitoring under free-living conditions, and more accurate and timely detection of treatment effects. However, there are challenges and considerations that need to be addressed. These include ensuring the clinical validity and reliability of these novel biomarkers, by addressing optimized and standard data collection protocols, and maintaining ethical and privacy governance in the integration of MHEALTH technologies in clinical trials. Further, the adoption and acceptance of MHEALTH biomarkers by clinicians and healthcare providers depend on factors such as interpretability and explainability. Explainable biomarkers that provide insights into how features effect the biomarker predictions can enhance trust and facilitate their integration into clinical (research) practice. Overall, these discussions highlight the potential of MHEALTH devices and ML in complementing clinical research. While there are challenges to overcome, the advancements in this field offer exciting opportunities for advancing the field of CNS research.

REFERENCES

1 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

2 Huml RA, Perez DP. FSHD: The Most Common Type of Muscular Dystrophy? In: *Muscular Dystrophy*. Springer International Publishing; 2015:9-19. doi:10.1007/978-3-319-17362-7_3

3 Mul K, Vincenten SCC, Voermans NC, et al. Adding quantitative muscle MRI to the FSHD clinical trial toolbox. *Neurology*. 2017;89(20):2057-2065. doi:10.1212/WNL.0000000000004647

4 Katuwal GJ, Chen R. Machine Learning Model Interpretability for Precision Medicine. Published online 2016. http://arxiv.org/abs/1610.09045

5 Carvalho D V., Pereira EM, Cardoso JS. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics (Basel)*. 2019;8(8):832. doi:10.3390/electronics8080832

6 Arora S, Yttri J, Nilse W. Privacy and Security in Mobile Health (MHEALTH) Research. *Alcohol Res*. 2014;36(1):143-151.

7 Nurgalieva L, O'Callaghan D, Doherty G. Security and Privacy of MHEALTH Applications: A Scoping Review. IEEE *Access*. 2020;8:104247-104268. doi:10.1109/ACCESS.2020.2999934