AHNJILI ZHUPARRIS

DEVELOPMENT OF MACHINE LEARNING – DERIVED MHEALTH COMPOSITE BIOMARKERS FOR TRIAL@HOME CLINICAL TRIALS

DEVELOPMENT OF MACHINE LEARNING DERIVED MHEALTH COMPOSITE BIOMARKERS FOR TRIAL@HOME CLINICAL TRIALS

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof.dr.ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op donderdag 13 juni 2024
klokke 13.45 uur

door
Ahnjili ZhuParris
geboren te New York, Verenigde Staten
in 1990

Cover illustration: Peter van Dijk
Design: Caroline de Lint, Den Haag (caro@delint.nl)

The printing of this thesis was financially supported by Centre for Human Drug Research.

PROMOTERS

Prof. Dr. G. J. Groeneveld

Prof. Dr. Ir. W. Kraaij *(Leiden Institute of Advanced Computer Science (LIACS), Leiden University)*

CO-PROMOTERS

Dr. Ir. R.J. Doll (*Leiden University*)

DOCTORATE COMMITTEE

Prof. Dr. A. F. Cohen

Prof. Dr. A. Brouwer (*Radboud University*)

Dr. M. Baratchi (*Leiden Institute of Advanced Computer Science (LIACS), Leiden University*)

Prof. Dr. M. Hoogendoorn (*Vrije Universiteit Amsterdam*)

PART I

INTRODUCTION

CHAPTER 1

Introduction

Development of novel biomarkers

Clinical biomarkers serve a critical role in diagnosing diseases, monitoring disease progression, measuring drug effects, and predicting treatment outcomes.[1] As our understanding of biology and diseases continue to evolve, there is a growing demand for the development of novel biomarkers that offer more precise, in-depth, and timely understanding of the disease and provide early detection and quantification of drug effects. To meet this need, researchers are increasingly turning towards novel technologies that enable the development of innovative biomarkers. This goal is not without hurdles. Challenges such as data collection, standardization, validation, and regulatory considerations need to be carefully addressed. Additionally, the translation of these biomarkers from research setting to clinical practice requires robust evidence of their clinical utility and reliability.

The primary objective of this thesis is to address the development and validation of innovative biomarkers by harnessing the data of mobile health (MHEALTH) devices, such as smartphones, tablets, and wearable devices. These widely available and data-intensive technologies offer an unprecedented opportunity to capture diverse physiological and behavioral data outside the traditional clinical setting. To effectively utilize this wealth of information, Machine Learning (ML) techniques will be employed to transform the unstructured and multifaceted MHEALTH data into meaningful clinical biomarkers. This research aims to address the challenges, important factors, and potential benefits associated with the development and validation of MHEALTH biomarkers.

MHEALTH devices for clinical trials

Clinical trials play a crucial role in assessing the efficacy of new pharmacological treatments and are typically conducted by academic hospitals and Contract Research Organizations (CROS). Conventionally, data for observational and randomized clinical trials is collected during patients' visits to in-patient facilities like hospitals or clinical research units. This approach has several benefits, such as strict control over the study environment and standardized data collection. However, a limitation is that the data collected only represents a snapshot of the patient's health and disease activity, often in an isolated context. As a result, evidence gaps between visits are created, and clinicians' insight into patients' overall health may be limited.

To overcome the limitations of conventional clinical trials, MHEALTH devices like smartphones, wearables, and tablets offer a unique opportunity for continuous and longitudinal data collection from clinical trial participants under free-living conditions.[2–6] Mobile applications (apps) installed on smartphones and tablets can be utilized to actively collect self-reported outcomes from patients through electronic diaries.[7] Simultaneously, apps can passively collect data from various sensors such as accelerometers, cameras, gyroscopes, microphones, and phone logs, providing an additional source of valuable physical and behavioral data.[8–10] Wearables support continuous tracking of physiological responses or physical activity, such as heart rate or steps, enable characterization of intra- and inter-individual variability in disease activity and quantification of drug response.[11–14] This approach of collecting data from multiple sensors acknowledges that a patient's experience of their disease is a consequence of multiple neurobiological processes, and therefore is expressed as a diverse array of symptoms simultaneously.

The use of MHEALTH devices in clinical trials has sharply increased since the global adoption of the smartphone. Between 2012 to 2022, the term 'MHEALTH' was incorporated in 1605 clinical studies posted on clinicaltrials.gov. Only 15 studies used the term between 2000 to 2011.[15] MHEALTH biomarkers have been shown to be effective in monitoring disease activity and estimating symptom severity for a wide range of diseases such as mood disorders,[16–21] neurodegenerative disorders,[22–24] and cardiovascular diseases.[25] The benefits of MHEALTH devices in clinical trials are twofold. First, real-world data collected under free-living conditions, which is data collected outside of controlled clinical trial settings, can be used to

generate novel hypotheses or insights into the most effective treatments. This can help to provide the ecological validity of findings produced by well-controlled clinical trials. Second, the use of MHEALTH devices for clinical trials may also be cost-effective due to the emerging concept of Bring Your Own Device (BYOD).[26,27] By leveraging participants' own devices for data collection, costs are reduced for clinical trials as study specific hardware does not need to be purchased, distributed, or maintained. The burden for participants is also reduced as they can use hardware that they are already familiar with and can have access to in their daily lives.

Despite these advantages, integrating MHEALTH devices into clinical trials presents its own challenges. The most significant issues include ensuring tolerability and usability of the MHEALTH devices by patients and clinicians and developing, validating, and interpreting the biomarkers given the lack of control under free-living conditions.[5] Unlike controlled clinical settings, free-living conditions offer minimal control over the environment in which data is collected. Participants may also engage in various activities and encounter unpredictable situations that can influence data quality and consistency. Factors such as variations in daily routines, social interactions, and environmental exposures can introduce variability and noise into the collected data. The accuracy and reliability of the collected data can be affected by factors such as user engagement, device performance, and data synchronization. Ensuring data quality requires clear patient instructions, participant compliance, and regular monitoring to address any issues that may arise. When collecting data in free-living conditions, there is a greater risk of breaching participants' privacy. The use of MHEALTH devices, such as smartphones and wearable devices, often involves capturing personal information and sensitive data. Safeguarding privacy becomes crucial to ensure participants' trust and compliance. Implementing robust data encryption, secure data storage, and strict privacy policies are essential to mitigate privacy risks. The datasets generated by these devices are often complex, large, and subject to influence by external factors such as differences in devices, lifestyles, weather, and location. ML provides a potential solution for processing these large and heterogeneous datasets into biomarkers that can aid the understanding and prediction of complex clinical outcomes.

ML and traditional statistical learning methods both play important roles in the analysis and interpretation of clinical trial data. While both share a common objective of extracting meaningful insights and informing decision-making, they have distinct approaches and applications.[28] Traditional statistical learning methods typically focus on hypothesis testing, parameter estimation, and model interpretability and inference and therefore are classically used to test the significance of individual covariates or predictors, estimating effect sizes, and calculating sample sizes.[29] As traditional statistical learning methods are typically designed to answer specific research questions or test predefined hypotheses, their primary focus is on estimating the effects of individual covariates or predictors rather than generating accurate predictions for new, unseen data. These methods may lack the ability to generalize well to different populations, settings, or contexts, as they are often tailored to the specific characteristics of the analyzed dataset. With time-honored techniques such as ANOVA, t-tests, linear and logistic regression, and survival analysis deeply rooted in the field of clinical trials, the continued utilization of traditional statistical learning remains pivotal in advancing medical research and improving patient outcomes.[29,30] However, their limitations can hinder their effectiveness in analyzing complex and diverse clinical trial data, where flexibility and adaptability may be required.

Conversely, ML is primarily focused on developing data-driven statistical models that are both generalizable and predictive in nature.[28,31,32] As a result, ML is often considered more 'data-hungry' compared to statistical learning due to its reliance on large and diverse datasets. Generalizability is a desirable characteristic of biomarkers as it indicates their ability to perform well in diverse scenarios. Generalizable and predictive biomarkers derived from ML techniques can be applied across different patient populations, settings, and clinical trial protocols. A key step in the ML pipeline is the use of cross-validation. By employing cross-validation, clinicians can obtain a reliable estimate of how well the ML model is likely to perform

on unseen data sourced from a similar population or setting. This assessment of predictive accuracy is crucial in determining whether the developed model can generalize its findings beyond the specific dataset used for training. This versatility allows for the broader utilization of biomarkers in various healthcare contexts, increasing their potential impact and value.

A ML model has the potential to build a representative composite biomarker by integrating and capturing complex relationships among different features, which would lead to a more comprehensive and informative representation of the underlying biological or pharmacological processes. However, while the complexity of the biomarker can increase its predictive accuracy, it may limit its interpretability. ML offers a wide range of model types, such as decision trees, neural networks, ensemble methods, transfer learning, and unsupervised learning methods that can be adapted to different types of data and objectives, allowing for more flexible and adaptable modelling approaches.[28,33] Many ML algorithms, particularly deep learning models, can automatically learn and extract features directly from the data, eliminating the need for manual feature engineering. The automation of the identification of relevant features and patterns in the data, reduces the need for manual feature selection and engineering. This can streamline the biomarker development process and improve the efficiency of clinical trial analyses. In addition, unsupervised learning algorithms, which can identify patterns in data without being explicitly told what to look for, can be useful for exploratory data analysis or for discovering hidden patterns or subgroups within data that may not be immediately apparent.[34] In conclusion, ML's data-driven approach, flexibility in model selection, automated feature extraction, and ability to identify hidden patterns offer significant advantages over traditional statistical learning methods in the development of biomarkers for clinical trials. Its reliance on large and diverse datasets may make it more data-hungry, but this enables the creation of generalizable and predictive models. By streamlining the biomarker development process and improving the efficiency of clinical trial analyses, ML has the potential to greatly impact clinical research and contribute to improved patient outcomes.

Clinical validation of composite MHEALTH biomarkers

Composite MHEALTH biomarkers can offer several benefits to both clinicians and patients. By consolidating multiple clinical features into a single composite digital biomarker, this biomarker can be used to predict clinical outcomes, serving as a complement rather than a replacement for multiple clinical endpoints. The resulting composite biomarkers have the potential for inference and prediction, contributing to the discovery of generalizable and robust evidence to guide clinical studies. This thesis proposes that there are three beneficial applications for composite biomarkers. Firstly, composite biomarkers may be more sensitive to subtle changes or treatment effects that may not be evident when assessing individual biomarkers independently. Secondly, by combining multiple biomarkers, this can help mitigate the measurement variability that are inherent in an individual biomarker. The aggregated biomarker can provide a more stable representation of the underlying phenomenon. Lastly, a composite biomarker may provide a more holistic evaluation of disease activity. A composite biomarker provides a more comprehensive and multi-faceted assessment, and therefore may capture a broader spectrum of treatment effects. However, to determine if these composite digital biomarkers have utility in clinical research, they must be clinically validated.[35] The following section addresses the validation criteria considered to evaluate if a biomarker is suitable for clinical adoption.

Validation of novel composite biomarkers before incorporating them into clinical trials is crucial. To validate these biomarkers, Kruizinga et al. have proposed five criteria, which we have adopted along with an optional criterion of Interpretability and Explainability.[35] The first criterion, *Classifying Patients and Healthy Controls*, focuses on accurately distinguishing between patients and healthy individuals to identify disease-specific biomarkers. The second criterion, *Correlation with Gold Standard or Disease Metrics*, involves establishing the validity of the biomarker and its ability to accurately reflect disease activity by correlating it with the

gold standards. The third criterion, *Detecting Changes in Disease Activity or Treatment Effects,* refers to detecting changes in disease activity over time, which is crucial for monitoring disease progression or response to treatment. The fourth criterion, *Tolerability and Usability*, is particularly important for MHEALTH devices that may be worn continuously or for extended periods. The device should not cause discomfort or irritation and should be easy to use. If tolerability and usability of the device are poor, the missing or poor-quality data collected will negatively impact the development of the biomarker. The fifth criterion, *Repeatability and Variability*, refers to the device producing consistent measurements under different conditions and over multiple time points. Finally, the optional criterion, *Interpretability and Explainability*, refers to the ability of the composite biomarker to provide clear and understandable explanations for its predictions. This is important for building trust in the biomarker and its ability to inform clinical decision-making.

Research objectives and structure of this thesis

The overall research question of this thesis is *How can MHEALTH devices and ML algorithms be used to develop composite biomarkers for clinical applications*? To address this question, we have outlined a series of research questions that will explore different aspects of the development and clinical validation of these biomarkers. These research questions will be addressed in their respective chapters, culminating in a discussion of the general findings and recommendations for future research in this field.

Parts 2 to 4 will use clinical trial data collected using Centre for Human Drug Research (CHDR)'s Trial@Home platform. The Trial@Home platform aims to investigate alternative approaches for collecting clinical trial data in non-traditional clinical settings. Serving as a comprehensive solution, Trial@Home offers end-to-end services, encompassing trial design, execution, and data analytics. By integrating smartphones, tablets, and wearables (such as smartwatches, smart scales, and sleep mats) into clinical trials, participants can experience reduced visit frequency while enabling more convenient and representative data collection. This innovative approach captures participants' real-world experiences in their daily lives, providing valuable insights under free-living conditions. Through the use of ML, the collected data is transformed into novel and validated digital biomarkers. The following chapters provide more insight into the type of data collected during these trials, and how the data was transformed into validated biomarkers for clinical applications.

Part 1 (Introduction) asks *What is the motivation behind creating composite MHEALTH biomarkers for clinical applications and how are they currently being developed?* This part addresses the challenges and limitations of using MHEALTH devices and ML for developing and validating composite biomarkers in clinical trials. **Chapter 1** provides a brief overview of concept, reasoning, and importance of using ML in clinical trials that use MHEALTH devices. **Chapter 2** offers a literature review of existing published studies that have used similar techniques to derive composite biomarkers. Given the rise and breadth of ML applications in clinical trials, we sought to identify both the generic and best practices of developing these ML applications. However, given the lack of consistent reporting in these studies, the literature review does not provide a complete or detailed overview. On the contrary, the literature review presents a set of recommended reporting practices aimed at enhancing the transparency and reproducibility of the methods utilized.

Part 2 (Classification of Diagnosis) asks *How can MHEALTH devices and ML be utilized to create composite biomarkers for the classification of diagnoses?* This part addresses how different types of MHEALTH devices compare in terms of their usability, tolerability, and data quality for developing composite biomarkers. Further, it examines the methods required for developing accurate and clinically relevant biomarkers for the classification of disease diagnoses using MHEALTH data and ML. **Chapter 3** use the Trial@Home platform to classify the remotely monitored behavioural activity of Facioscapulohumeral Muscular Dystrophy (FSHD) patients respectively from Healthy Controls. To assess the feasibility of piloting a Trial@Home study, these publications also report the data completion rate and patient experience of the Trial@Home app to reflect the tolerability and usability of the devices.

Part 3 (Estimation of Symptom Severity) asks *How can MHEALTH devices and ML be utilized to create composite biomarkers for the estimation of symptom severity?* This part investigates the effectiveness of the developed composite biomarkers in estimating the severity of disease symptoms in patients compared to traditional methods. **Chapter 4** and **5** use regression algorithms and the Trial@Home platform to estimate the symptom severity of the FSHD and Major Depressive Disorder (MDD) patients. In addition to estimating the symptom severity, we evaluated how varying time windows used to train the models can affect the repeatability and variability of their predicted outcomes. **Chapter 6** and **7** focus on developing ML models that can automatically quantify the number of coughs and cries using a smartphone microphone respectively. While these activities cannot be used as diagnostic tools themselves, they serve as relevant and informative proxies for disease activity.

Part 4 (Detection of Treatment Effects) asks *Can the use of MHEALTH devices and ML algorithms enable the detection of treatment effects in clinical trials and provide* insights *into the efficacy of pharmacological treatments*? To address this question, **Chapter 8** explore if a composite tapping biomarker can detect treatment effects and to estimate symptom severity among Parkinson's Disease patients respectively. The underlying motivation for this investigation lies in examining whether the same tapping biomarker can serve the dual purpose of monitoring both treatment effects and symptom severity in alignment with the gold standard, thus unveiling new possibilities for comprehensive biomarker applications.

Chapter 9, the discussion, reflects on the methodologies and analyses in **Parts 2 to 4** and addresses the motivations, factors, and limitations that contribute to the development and adoption of MHEALTH composite biomarkers for the purposes of diagnosis classification, symptom severity estimation, and treatment effects detection. Given the potential impacts of MHEALTH biomarkers, the discussion reflects on the practical and ethical implications of MHEALTH biomarkers for clinicians, other Central Nervous System (CNS) disorders, and future clinical trials.

Condensed structure of the thesis

Given the criteria for evaluating the clinical validity of candidate composite biomarkers, this thesis consists of 5 parts. Part 1 provides the theoretical and historical framework for the development of these biomarkers. Part 2, 3, and 4 focus on clinical trials that use ML to classify a clinical diagnosis, to estimate symptom severity, and to detect treatment effects respectively. In each of these sections, we provide a detailed account of our approach to the proposed clinical validation. **Chapter 9** discusses the general findings of this thesis and addresses general recommendations for developing future biomarkers that use MHEALTH devices and ML.

REFERENCES

1 A. F. Cohen, J. Burggraaf, J. M. A. van Gerven, M. Moerland, and G. J. Groeneveld, 'The Use of Biomarkers in Human Pharmacology (Phase I) Studies,' *Annu Rev Pharmacol Toxicol*, vol. 55, no. 1, pp. 55–74, Jan. 2015, doi: 10.1146/annurev-pharmtox-011613-135918.

2 D. G. Stuijt, V. Exadaktylos, A. D. Bins, J. J. Bosch, and M. G. H. van Oijen, 'Potential Role of Smartphone-Based Passive Sensing in Remote Monitoring of Patients With Cancer,' *JCO Clin Cancer Inform*, no. 6, Sep. 2022, doi: 10.1200/CCI.22.00079.

3 M. M. E. Vogel, S. E. Combs, and K. A. Kessel, 'MHEALTH and Application Technology Supporting Clinical Trials: Today's Limitations and Future Perspective of smartRCTs,' *Front Oncol*, vol. 7, Mar. 2017, doi: 10.3389/fonc.2017.00037.

4 S. P. Rowland, J. E. Fitzgerald, T. Holme, J. Powell, and A. McGregor, 'What is the clinical value of MHEALTH for patients?,' *NPJ Digit Med*, vol. 3, no. 1, p. 4, Jan. 2020, doi: 10.1038/s41746-019-0206-x.

5 A. Kakkar, P. Sarma, and B. Medhi, 'MHEALTH technologies in clinical trials: Opportunities and challenges,' *Indian J Pharmacol*, vol. 50, no. 3, p. 105, 2018, doi: 10.4103/ijp.IJP_391_18.

6 Q. Pham, D. Wiljer, and J. A. Cafazzo, 'Beyond the Randomized Controlled Trial: A Review of Alternatives in MHEALTH Clinical Trial Methods,' *JMIR MHEALTH Uhealth*, vol. 4, no. 3, p. e107, Sep. 2016, doi: 10.2196/MHEALTH.5720.

7 S. J. Coons, S. Eremenco, J. J. Lundy, P. O'Donohoe, H. O'Gorman, and W. Malizia, 'Capturing Patient-Reported Outcome (PRO) Data Electronically: The Past, Present, and Promise of ePRO Measurement in Clinical Trials,' *The Patient - Patient-Centered Outcomes Research*, vol. 8, no. 4, pp. 301–309, Aug. 2015, doi: 10.1007/s40271-014-0090-z.

8 R. Au, H. Lin, and V. B. Kolachalama, 'Tele-Trials, Remote Monitoring, and Trial Technology for Alzheimer's Disease Clinical Trials,' in *Alzheimer's Disease Drug Development*, Cambridge University Press, 2022, pp. 292–300. doi: 10.1017/9781108975759.026.

9 Z. W. Adams, E. A. McClure, K. M. Gray, C. K. Danielson, F. A. Treiber, and K. J. Ruggiero, 'Mobile devices for the remote acquisition of physiological and behavioral biomarkers in psychiatric clinical research,' *J Psychiatr Res*, vol. 85, pp. 1–14, Feb. 2017, doi: 10.1016/j.jpsychires.2016.10.019.

10 A. Zhuparris *et al.*, 'Smartphone and Wearable Sensors for the Estimation of Facioscapulohumeral Muscular Dystrophy Disease Severity: Cross-sectional Study,' *JMIR Form Res*, vol. 7, p. e41178, Mar. 2023, doi: 10.2196/41178.

11 G. Cosoli, A. Poli, S. Spinsante, and L. Scalise, 'The importance of physiological data variability in wearable devices for digital health applications,' *ACTA IMEKO*, vol. 11, no. 2, p. 1, May 2022, doi: 10.21014/acta_imeko.v11i2.1135.

12 S. Carreiro *et al.*, 'Real-Time Mobile Detection of Drug Use with Wearable Biosensors: A Pilot Study,' *Journal of Medical Toxicology*, vol. 11, no. 1, pp. 73–79, Mar. 2015, doi: 10.1007/s13181-014-0439-7.

13 A. K. Yetisen, J. L. Martinez-Hurtado, B. Ünal, A. Khademhosseini, and H. Butt, 'Wearables in Medicine,' *Advanced Materials*, vol. 30, no. 33, p. 1706910, Aug. 2018, doi: 10.1002/adma.201706910.

14 S. Bian, B. Zhu, G. Rong, and M. Sawan, 'Towards wearable and implantable continuous drug monitoring: A review,' *J Pharm Anal*, vol. 11, no. 1, pp. 1–14, Feb. 2021, doi: 10.1016/j.jpha.2020.08.001.

15 NIH U.S. National Library of Medicine, 'Information on Clinical Trials and Human Research Studies,' Apr. 01, 2023. https://www.clinicaltrials.gov/ct2/results?cond=&term=MHEALTH&type=&rslt=&age_v=&gndr=&intr=&titles=&outc=&spons=&lead=&id=&cntry=&state=&city=&dist=&locn=&rsub=&strd_s=01%2F01%2F2012&strd_e=12%2F31%2F2022&prcd_s=&prcd_e=&sfpd_s=&sfpd_e=&rfpd_s=&rfpd_e=&lupd_s=&lupd_e=&sort= (accessed Apr. 05, 2023).

16 E. Brietzke, E. R. Hawken, M. Idzikowski, J. Pong, S. H. Kennedy, and C. N. Soares, 'Integrating digital phenotyping in clinical characterization of individuals with mood disorders,' *Neurosci Biobehav Rev*, vol. 104, no. July, pp. 223–230, Sep. 2019, doi: 10.1016/j.neubiorev.2019.07.009.

17 A. Grünerbl *et al.*, 'Smartphone-based recognition of states and state changes in bipolar disorder patients,' IEEE *J Biomed Health Inform*, vol. 19, no. 1, pp. 140–148, Jan. 2015, doi: 10.1109/JBHI.2014.2343154.

18 S. Stanislaus *et al.*, 'Smartphone-based activity measurements in patients with newly diagnosed bipolar disorder, unaffected relatives and control individuals,' *Int J Bipolar Disord*, vol. 8, no. 1, p. 32, Dec. 2020, doi: 10.1186/s40345-020-00195-0.

19 M. Faurholt-Jepsen *et al.*, 'Differences in mood instability in patients with bipolar disorder type I and II: a smartphone-based study,' *Int J Bipolar Disord*, vol. 7, no. 1, 2019, doi: 10.1186/s40345-019-0141-4.

20 J.-Y. Liu, K.-K. Xu, G.-L. Zhu, Q.-Q. Zhang, and X.-M. Li, 'Effects of smartphone-based interventions and monitoring on bipolar disorder: A systematic review and meta-analysis,' *World J Psychiatry*, vol. 10, no. 11, pp. 272–285, 2020, doi: 10.5498/wjp.v10.i11.272.

21 M. Faurholt-Jepsen, M. Frost, E. M. Christensen, J. E. Bardram, M. Vinberg, and L. V. Kessing, 'The effect of smartphone-based monitoring on illness activity in bipolar disorder: The MONARCA II randomized controlled single-blinded trial,' *Psychol Med*, vol. 50, no. 5, pp. 838–848, 2020, doi: 10.1017/S0033291719000710.

22 L. C. Kourtis, O. B. Regele, J. M. Wright, and G. B. Jones, 'Digital biomarkers for Alzheimer's disease: the mobile/wearable devices opportunity,' *NPJ Digit Med*, vol. 2, no. 1, pp. 1–9, 2019, doi: 10.1038/s41746-019-0084-2.

23 R. Bhidayasiri and Z. Mari, 'Digital phenotyping in Parkinson's disease: Empowering neurologists for measurement-based care,' *Parkinsonism and Related Disorders*, vol. 80. Elsevier Ltd, pp. 35–40, Nov. 01, 2020. doi: 10.1016/j.parkreldis.2020.08.038.

24 P. Kassavetis *et al.*, 'Developing a Tool for Remote Digital Assessment of Parkinson's Disease,' *Mov Disord Clin Pract*, vol. 3, no. 1, pp. 59–64, Jan. 2016, doi: 10.1002/mdc3.12239.

25 J. X. Teo *et al.*, 'Digital phenotyping by consumer wearables identifies sleep-associated markers of cardiovascular disease risk and biological aging,' *bioRxiv*, 2019, doi: 10.1101/527077.

26 L. Pugliese, O. Crowley, M. Woodriff, V. Lam, J. Sohn, and S. Bradley, 'Feasibility of the 'Bring Your Own Device' Model in Clinical Research: Results from a Randomized Controlled Pilot Study of a Mobile Patient Engagement Tool,' *Cureus*, Mar. 2016, doi: 10.7759/cureus.535.

27 C. Demanuele *et al.*, 'Considerations for Conducting Bring Your Own 'Device' (BYOD) Clinical Studies,' *Digit Biomark*, vol. 6, no. 2, pp. 47–60, Jul. 2022, doi: 10.1159/000525080.

28 I. El Naqa and M. J. Murphy, 'What Is Machine Learning?,' in *Machine Learning in Radiation Oncology*, Springer International Publishing, 2015, pp. 3–11. doi: 10.1007/978-3-319-18305-3_1.

29 R. J. Calin-Jageman, 'Better Inference in Neuroscience: Test Less, Estimate More,' *The Journal of Neuroscience*, vol. 42, no. 45, pp. 8427–8431, Nov. 2022, doi: 10.1523/JNEUROSCI.1133-22.2022.

30 R. Iniesta, D. Stahl, and P. Mcguffin, 'Machine learning, statistical learning and the future of biological research in psychiatry,' 2016, doi: 10.1017/S0033291716001367.

31 T. Hastie, R. Tibshirani, and J. Friedman, 'Statistics The Elements of Statistical Learning,' *The Mathematical Intelligencer*, vol. 27, no. 2, pp. 83–85, 2009.

32 T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York City: Springer Science & Business Media, 2013.

33 D. Bzdok, N. Altman, and M. Krzywinski, 'Statistics versus machine learning,' *Nat Methods*, vol. 15, no. 4, pp. 233–234, Apr. 2018, doi: 10.1038/nmeth.4642.

34 Z. Ghahramani, 'Unsupervised Learning,' 2004, pp. 72–112. doi: 10.1007/978-3-540-28650-9_5.

35 M. D. Kruizinga *et al.*, 'Development of Novel, Value-Based, Digital Endpoints for Clinical Trials: A Structured Approach Toward Fit-for-Purpose Validation,' *Pharmacol Rev*, vol. 72, no. 4, pp. 899–909, Oct. 2020, doi: 10.1124/pr.120.000028.

CHAPTER 2

Machine learning techniques for developing remotely monitored central nervous system biomarkers using wearable sensors: a narrative literature review

Ahnjili ZhuParris,[1,2,3] Annika A. de Goede,[1] Iris E. Yocarini,[2] Wessel Kraaij,[2,4] Geert Jan Groeneveld,[1,2] and Robert-Jan Doll [1]

Sensors. 2023;23(11):5243. doi:10.3390/s23115243

1 Centre for Human Drug Research (CHDR), Leiden, NL

2 Leiden Institute of Advanced Computer Science (LIACS), Leiden, NL

3 Leiden University Medical Center (LUMC), Leiden, NL

4 The Netherlands Organisation for Applied Scientific Research (TNO), Den Haag, NL

Abstract

Background: Central nervous system (CNS) disorders benefit from ongoing monitoring to assess disease progression and treatment efficacy. Mobile health (MHEALTH) technologies offer a means for the remote and continuous symptom monitoring of patients. Machine Learning (ML) techniques can process and engineer MHEALTH data into a precise and multidimensional biomarker of disease activity. **Objective:** This narrative literature review aims to provide an overview of the current landscape of biomarker development using MHEALTH technologies and ML. Additionally, it proposes recommendations to ensure the accuracy, reliability, and interpretability of these biomarkers. **Methods:** This review extracted relevant publications from databases such as PubMed, IEEE, and CTTI. The ML methods employed across the selected publications were then extracted, aggregated, and reviewed. **Results:** This review synthesized and presented the diverse approaches of 66 publications that address creating MHEALTH-based biomarkers using ML. The reviewed publications provide a foundation for effective biomarker development and offer recommendations for creating representative, reproducible, and interpretable biomarkers for future clinical trials. **Conclusion:** MHEALTH-based and ML-derived biomarkers have great potential for the remote monitoring of CNS disorders. However, further research and standardization of study designs are needed to advance this field. With continued innovation, MHEALTH-based biomarkers hold promise for improving the monitoring of CNS disorders.

Introduction

MOTIVATION

Disorders that are affected by the Central Nervous System (CNS), such as Parkinson's Disease (PD) and Alzheimer's Disease (AD), have a significant impact on the quality of life of patients. These disorders are often progressive and chronic, making long-term monitoring essential for assessing disease progression and treatment effects. However, the current methods for monitoring disease activity are often limited by accessibility, cost, and patient compliance.[1,2] Limited accessibility to clinics or disease monitoring devices may hinder the regular and consistent monitoring of a patient's condition, especially for patients living in remote areas or for those who have mobility limitations. Clinical trials incur costs related to personnel, infrastructure, and equipment. A qualified healthcare team, including clinical raters, physicians, and nurses, contributes to personnel costs through salaries, training, and administrative support. Trials involving specialized equipment for measuring biomarkers can significantly impact the budget due to costs associated with procurement, maintenance, calibration, and upgrades. Furthermore, infrastructure costs may increase as suitable facilities are required for data collection during patient visits and equipment storage. Patient compliance poses challenges for disease monitoring, as some methods require patients to adhere to strict protocols, collect data at specific time intervals, or perform certain tasks that can be challenging for patients to execute. Low or no compliance can lead to incomplete or unreliable monitoring results, which in turn can hinder the reliability of the assessments. Given these limitations, there is a growing interest in exploring alternative approaches to monitoring CNS disorders that can overcome these challenges. The increasing adoption of smartphones and wearables among patients and researchers offers a promising avenue for remote monitoring.

Patient-generated data from smartphones, wearables, and other remote monitoring devices can potentially complement or supplement clinical visits by providing data during evidence gaps between visits. As

the promise of mobile Health (MHEALTH) technologies is to provide more sensitive, ecologically valid, and frequent measures of disease activity, the data collected may enable the development and validation of novel biomarkers. The development of novel 'digital biomarkers' using data collected from electronic Health (EHEALTH) and MHEALTH device sensors (such as accelerometers, GPS, and microphones) offers a scalable opportunity for the continuous collection of data regarding behavioral and physiological activity under free-living conditions. Previous clinical studies have demonstrated the benefits of smartphone and wearable sensors to monitor and estimate symptom severity associated with a wide range of diseases and disorders, including cardiovascular diseases,[3] mood disorders,[4] and neurodegenerative disorders.[5,6] These sensors can capture a range of physiological and behavioral data, including movement, heart rate, sleep, and cognitive function, providing a wealth of information that can be used to develop biomarkers for CNS disorders in particular. These longitudinal and unobtrusive measurements are highly valuable for clinical research, providing a scalable opportunity for measuring behavioral and physiological activity in real-time. However, these approaches may carry potential pitfalls as the data sourced from these devices can be large, complex, and highly variable in terms of availability, quality, and synchronicity, which can therefore complicate analysis and interpretation.[7,8] Machine Learning (ML) may provide a solution to processing heterogenous and large datasets, identifying meaningful patterns within the datasets, and predicting complex clinical outcomes from the data. However, the complexities involved in developing biomarkers using these new technologies need to be addressed. While these tools can aid the discovery of novel and important digital biomarkers, the lack of standardization, validation, and transparency of the ML pipelines used can pose challenges for clinical, scientific, and regulatory committees.

WHAT IS MACHINE LEARNING

In clinical research, one of the primary objectives is to understand the relationship between a set of observable variables (features) and one or more outcomes. Building a statistical model that captures the relationship between these variables and the corresponding outputs facilitates the attainment of this understanding.[9] Once this model is built, it can be used to predict the value of an output based on the features.

ML is a powerful tool for clinical research as it can be used to build statistical models. A ML model consists of a set of tunable parameters and a ML algorithm that enables the generation of outputs based on given inputs and selected parameters. Although ML algorithms are fundamentally statistical learning algorithms, ML and traditional statistical learning algorithms can differ in their objectives. Traditional statistical learning aims to create a statistical model that represents causal inference from a sample, while ML aims to build generalizable predictive models that can be used to make accurate predictions on previously unseen data.[10,11] However, it is essential to recognize that while ML models can identify relationships between variables and outcomes, they may not necessarily identify a causal link between them. This is because even though these models may achieve good performances, it is crucial to ensure that their predictions are based on relevant features rather than spurious correlations. This enables the researchers to gain meaningful insights from ML models while also being aware of their inherent limitations.

While ML is not a substitute for the clinical evaluation of patients, it can provide valuable insights into a patient's clinical profile. ML can help to identify relevant features that clinicians may not have considered, leading to better diagnosis, treatment, and patient outcomes. Additionally, ML can help to avoid common pitfalls observed in clinical decision making by removing bias, reducing human error, and improving the accuracy of predictions.[12–15] As the volume of data generated for clinical trials and outside clinical settings continues to grow, ML's support in processing data and informing the decision-making process becomes necessary. ML can help to uncover insights from large and complex datasets that would be difficult or impossible to identify manually.

To develop an effective ML model, it is necessary to follow a rigorous and standardized procedure. This is where ML pipelines come in. Table 1

showcases an exemplary ML pipeline, which serves as a systematic framework for automating and standardizing the model generation process. The pipeline encompasses multiple stages, as defined by the authors, to ensure an organized and efficient approach to model development. First, defining the study objective guides the subsequent stages and ensures the final model meets the desired goals. Second, raw data must be preprocessed to remove errors, inconsistencies, missing data, or outliers. Third, feature extraction and selection identify quantifiable characteristics of the data relevant to the study objective and extracts them for use in the ML model. Fourth, ML algorithms are applied to learn patterns and relationships between features, with optimal configurations identified through iterative processes until desired performance metrics are achieved. Finally, the model is validated against a new dataset that is not used in training to ensure generalizability. Effective reporting and assessment of ML procedures must be established to ensure transparency, reliability, and reproducibility.

OBJECTIVES

The objective of this narrative literature review is to provide an overview of the ML practices used in studies that use MHEALTH technologies and ML to develop novel biomarkers for clinical trials. In this review, each component of the ML pipeline has a dedicated section. Based on the results obtained from the review process, each ML component section provides a comprehensive analysis and discussion of the most common and notable practices. These sections delve into the motivations behind these practices, their limitations, and their overall impact on the ML pipeline. This review will not provide precise recommendations for best practices, as much of the research in this area is new and quickly evolving. Rather, the recommendation section discusses the approaches for standardization and validation procedures that are necessary for the development of ML biomarkers to ensure the effectiveness and acceptance of these biomarkers by clinical, scientific, and regulatory committees.

Methods

INFORMATION SOURCES AND SEARCH STRATEGY

Given the wide range of study designs and clinical populations that use smartphones and wearables to collect data, we used the Joanna Briggs Institute (JBI) guidelines to develop a search strategy.[16] Based on an initial limited search of online databases for clinical trials that report using MHEALTH devices and ML, we developed a custom keyword strategy and performed an in-depth search in PubMed, IEEE Xplore, and CTTI (Table 2). The search terms for the CNS disorder terms were based on the National Library of Medicine's CNS MeSH descriptor data.[17] The relevant papers were selected based on the title and abstract. Finally, other literature review studies that explore the same questions were reviewed; the references cited by these studies were then identified and reviewed if they met our criteria. The date range for the search was between 1 January 2012 and 31 December 2022. The search was conducted on 7 January 2023.

INCLUSION CRITERIA

The authors adopted the Population, Intervention, Comparator, Outcomes, Study type (PICOS) framework to define the inclusion and exclusion criteria (Table 3).[13] The studies included were restricted to those involving participants diagnosed with CNS disorders who were remotely monitored under free-living conditions. The intervention and device criteria were limited to passive data collected from smartphones and other non-invasive remote monitoring sensors, whereas data collected using active engagement from participants, such as disposable blood tests or small scales, were excluded. As we chose to focus on ML pipelines, we selected studies in which a statistical model was used to analyze a dataset and could potentially be used to generate future predictions using an independent dataset. Therefore, traditional statistical models such as linear or logistic regression were included, but statistical models such as ANOVA and correlation analyses were not included. Further, as the focus

was on the development and validation of ML models, we did not include studies that did not report on model performance.

DATA EXTRACTION

Two authors conducted the data extraction following the inclusion criteria, and the results were reviewed by the remaining authors. Data relating to the database source, title, DOI, publication year, trial setting or scenario, objective, devices used, data collection period, number of participants, inclusion of healthy controls, data processing steps, feature engineering, feature selection, machine learning models used, hyperparameters and hyperparameter optimization, model performance, and validation procedure were extracted. The comprehensive data extraction and review conducted by the authors encompassed various essential aspects of the studies, ensuring a thorough analysis of the database source, trial details, data processing steps, machine learning models, and validation procedures.

Results

STUDY SELECTION

Our initial keyword search revealed a total of 2310 articles that utilized digital phenotyping devices, such as smartphones and wearables, in a clinical study and applied ML techniques. After screening the titles and abstracts based on our predefined criteria, we narrowed down the articles to 66 studies, which were used for our analysis. Figure 1 provides an overview of the complete selection process.

STUDY CHARACTERISTICS

For each of the 66 studies, we extracted information about the clinical population and the ML pipeline that was used to develop the digital biomarkers. We found that only half of the studies included healthy controls (N = 34). As seen in Figure 2, Parkinson's disease (PD) (N = 27) was the most prevalent disorder identified in our search, followed by Bipolar Disorder (BD) (N = 11), and Unipolar Depression or Major Depressive Disorder (MDD) (N = 9). The sample size of the selected studies was heterogenous, ranging from 7 to 6221 participants (Figure 3). Overall, our review provides a comprehensive overview of the characteristics of studies that have utilized MHEALTH devices and ML techniques, which can help inform future research in this field. In the following sections, we addressed how the selected studies approached the construction of their ML pipelines.

Missing and Outlier Data

Missing and outlier data are commonly encountered problems for remote sensing clinical trials. Missing data can be the result of device charging frequency, device robustness, and participant compliance.[18] Outliers can be the result of sensor or device dysfunction or malfunction, incorrect data entry, and incorrect classifications.[19] Data preprocessing, which refers to the dropping or manipulation of data, is required for identifying and removing redundant or irrelevant data and for cleaning the data prior to analysis. Without preprocessing, learning from an imperfect dataset can influence the prediction accuracy of the models.[20] In this section, we address how the selected studies preprocessed their raw data by treating their missing data and outliers, and the limitations of doing so.

HANDLING OF MISSING DATA

Missing data can be Missing Completely at Random (MDD), Missing at Random (MDD), and Missing Not at Random (MNAR).[21] MDD assumes that each observation has the same probability of being included or being missed; therefore, there is no difference in the characteristics between participants or observations without missing data and those with missing data. For example, data may be missed due to the battery of the smartphone running out. MDD assumes that missing data may have systematic differences between the missing and non-missing data; however, the cause of the missing data can be explained by the non-missing data. For example, a smartphone may have more missing values when the smartphone

battery is low. If the battery percentage is known during the data acquisition, researchers can verify the probability of acquiring missing data depending on the battery percentage. MNAR assumes that missing data are caused by unknown reasons. For example, smartphone sensors may be gradually worn down, which therefore creates more missing data over time. The type of missing data present in the dataset influences whether a researcher should ignore, exclude, or impute the missing data.

Among the selected studies, we found that only 21 of the studies reported the quantity of missing data acquired. Only 29 studies reported how they handled their missing data. We found that complete-case analysis and imputation were the most popular. We identified 14 studies that report using complete-case analysis.[22–36] Complete-case analysis (otherwise known as listwise deletion) is the deletion of an observation involving one or multiple elements of missing data.[26,37,38] While complete-case analysis is the simplest approach to handle missing data, it does reduce the sample size and statistical power of the analysis [39] and can potentially lead to bias if the data are not MDD.[40] Imputation is the statistical process of replacing missing data with substituted inferred values.[41] We identified studies that imputed their missing data using linear interpolation,[29,42,43] forward filling,[44–1,45] zeros, median, means, and the most frequent value in the column.[24,46] The advantage of imputation is that it enables researchers to use all observations in the dataset. However, the inclusion of imputed values can lead to a false impression of the number of complete cases and reduce the variance in the dataset.[47–49]

IDENTIFICATION OF OUTLIERS

Aggarwal's Data mining: the textbook states that it is the subjective definition of the researcher that defines an outlier.[50] In cases where the outlier data were discussed in the selected studies, we found that researchers customized their definition of outliers by either defining a range of acceptable values [32] or by defining a threshold based on the mean and standard deviation.[51–53] Visual inspection by the researchers or the optimization of different threshold mechanisms can both be used to define the boundaries of normal or outlier data.[54,55] Maleki et al. defined outliers as observations that were most likely the result of measurement errors.[36] In terms of the handling of outliers, we only identified six studies that explicitly stated that outliers were excluded.[26,30,51–53,56]

Feature Engineering

FEATURE SCALING

Feature scaling is used to normalize the ranges of the features in a dataset.[57] Several feature engineering techniques and ML models (such as Principal Component Analysis and Linear Regression) calculate the distances between two observations. If one feature has a broader range of values compared to the other features, the calculated distances will be heavily influenced by this feature.[58] Therefore, the ranges of all the features should be normalized or standardized so that each feature is appropriately and proportionally considered with respect to the estimated distances.[57] Feature normalization is a common scaling method for rescaling the features into a bounding range using the minimum and maximum values, for example, between 0 and 1. Normalization is an ideal approach when the distribution of the data is not Gaussian, as normalization preserves the original distribution of the data. However, normalization uses minimum and maximum values to define ranges. This makes the method sensitive to outliers.[57,59] Alternatively, feature standardization, also known as z-score normalization, is a method for rescaling the data to fit a standard normalized distribution by using the mean and standard deviation and does not define a bounding range. Consequently, the standardization approach is not sensitive to outliers as it has no bounding range.[57,59] Normalization, log-transformation, and standardization have been reported in a small selection of the selected studies.[26,27,36,60,61]

EXPERT FEATURE ENGINEERING

Feature engineering is the process of constructing (new) features from the raw data or existing features while maintaining the original patterns and information in the data.[62] The newly engineered features can be added to or replace features in the original dataset. Engineering of the features

can speed up the model performance, improve learning accuracy, and ease the interpretability of the model. The latter is particularly important for clinical trials.[63] Features can be engineered manually by relying on domain-knowledge or automatically by using statistical models, such as Principal Component Analysis (PCA) and Deep Learning (DL).[62–64] All features aim to increase the separability between the classes or signals, which in turn reduces noise in the dataset. While expert engineered features are easy to interpret and explain and have been widely used in the development of digital biomarkers, these features are typically task- or population-dependent. Due to intra-class variability, some clinically relevant characteristics may be exhibited differently by different individuals (such as different symptom profiles among patients with the same diagnosis). Furthermore, expert engineered features may not be sufficient for representing the most important characteristics of complex patterns and can be time-consuming to acquire, especially when handling large-scale datasets.[65,66] As clinical data has expanded in terms of diversity, availability, and complexity, the aforementioned techniques may be insufficient for developing generic features. In the following sections, we address the notable and generic procedures used to perform feature engineering.

SIGNAL PROCESSING

To monitor changes in the physical activity of study participants using time series data collected from wearable sensors, signal processing is necessary to detect, clean, and analyze the components of interest. The feature extraction technique used is influenced by the sensor type, study objectives, and signal quality. Typically, signal features are extracted from the frequency, time, or cepstrum domain.[67] Frequency domain features show the prominence of a signal within a given frequency, whereas time-domain features show the changes in the signal of time. Cepstrum domain features represent the rate of change in the different frequency bands. The analysis of the frequency, time, or cepstrum domain features is not mutually exclusive. We identified studies that use both time- and frequency-based features for the estimation of gait speed,[68] speech-tasks,[69] seizure detection,[70] tremor detection,[71] and FOG detection.[72] In particular, Tougui et al. built 138 voice related features extracted from the cepstral, frequency, and time domains.[24] In sum, time series data collected from wearable sensors can be used to monitor the physical activity of study participants, but signal processing is necessary to extract meaningful features. Different feature extraction techniques can be used depending on the sensor type, signal quality, and study objectives. The analysis of these features is not mutually exclusive, and studies that use multiple domains for different clinical applications have been identified.

PRINCIPAL COMPONENT ANALYSIS

A common linear dimensionality reduction technique for feature engineering and selection is Principal Component Analysis (PCA).[28,73,74] PCA is used to sufficiently explain a high-dimensional dataset through a few principal components and, therefore, to reduce a high-dimensional dataset to one of fewer dimensions.[75] To this purpose, PCA converts a set of correlated features into a set of uncorrelated features by utilizing orthogonal transformation.[75] The principal components enable a reduction in the feature space by creating a linear combination of the original features, which consequently reduces the storage space and reduces the learning time. Therefore, the periodic components within a concurrent time series dataset can be isolated using PCA, which can subsequently be used to identify any underlying patterns within the dataset. It is important to note that PCA assumes that the data are normally distributed and is sensitive to feature variance.[75,76] Consequently, features with larger ranges will dominate features with smaller ranges. To make the variables comparable, transformation of the data prior to PCA is required.[75,76] Of the studies selected, PCA was used to engineer and select features from times series data sourced from waist-worn triaxial accelerometers and wearable activity trackers.[28,73,74] However, the limitations of PCA are its sensitivity to missing data and outliers and the limited interpretation of the original features. Hence, this observation highlights the need for thorough data preprocessing prior to using PCA.

CLUSTERING

A clustering algorithm is a common feature engineering method that assigns similar observations to a single cluster and assigns dissimilar observations to another.[77] While PCA compresses the features into principal components, clustering compresses the individual observations into clusters. The grouping of similar observations can improve the model's ability to discriminate between classes.[78] Clustering algorithms, more specifically DBSCAN and K-means clustering, have been deployed in smartphone GPS systems and Wi-Fi-network sensors to extract meaningful location features such as frequented location clusters,[79] location patterns,[80] and mobility patterns.[81] These studies demonstrate that clustering algorithms are a powerful method for reducing the number of observations into a smaller number of artificial variables that account for the variance within the dataset.

DEEP LEARNING

The performance of ML models can be limited by the development of manual and arbitrary features, and this potential obstacle can be overcome by DL algorithms. DL algorithms eliminate the need for manual feature engineering, as the DL layers can translate the data into more compact and intermediate abstractions of the data, which in turn can be used as features to predict the final output.[82] While DL can reduce the need for manual data preprocessing and feature extraction, which can potentially improve the generalizability and robustness of a model, the interpretation of the DL model is difficult, as the abstracted features may not be explainable by clinicians. However, it is important to note that the discriminative power of the DL-derived abstractions is strongly influenced by the architecture of the DL algorithm, which is also dependent on the trial-and-error process.[59] Due to DL's representation learning, DL is data-hungry, and therefore requires more data than other ML algorithms.[83,84] For clinical trial data, because of technological limitations and small sample sizes, there may not be enough data to train a sufficiently representative DL model.[76,83]

Four studies used DL to engineer features using time series data.[23,85–87] These models were used to extract gait features from accelerometer data [85,87] and tremor characteristics from IMU data.[23,86] However, it should be noted that the DL models do not always outperform the 'shallow learning' models, as shown in a study by Juen et al. in which smartphone accelerometers were used to predict natural walking speed and distance during a six-minute walk test.[85]

Feature Selection

In recent decades, high-dimensional clinical datasets have relied on feature selection.[88] Feature selection is the process of selecting a subset of the most informative features that will be processed by the ML algorithm.[89] Reducing the features for analysis has both computational and practical benefits. Selecting features can limit storage requirements, increase the algorithm processing speed, increase the interpretability of a model, and improve model performance.

OVERFITTING AND UNDERFITTING

Overfitting and underfitting are common pitfalls for ML models. Overfitting refers to when a ML model fits too well to its training dataset and is unable to generalize its patterns to unseen data. This problem can occur when the training dataset is small and not representative of the overall potential data distribution. Additionally, if the training dataset contains many outliers, the ML model may also fit the outlier data. Underfitting occurs when the trained ML model is too simple; therefore, it cannot identify the relationship between the features and the outputs. Underfitted models will perform poorly for both the train and validation datasets. To address overfitting, reducing the number of features considered by the model or updating the model architecture to include fewer features can be effective.[90] Underfitting can be improved by adding more features considered by the model or by updating the model architecture to increase the complexity of the feature space.[90]

Feature selection identifies the most important features in the dataset and eliminates the irrelevant ones, which thereby reduces noise. However, it is important to strike a balance, as strict feature selection may remove important signals from the data. Therefore, selecting the optimal set of features is important for preventing over- and underfitting. In the following sections, we will elaborate on the three general methods of feature selection that are suitable for ML models.[75]

FILTER METHODS

Filter methods are used during preprocessing prior to training the ML model. Filtering involves removing features based on domain knowledge, missing data, low variance, or correlation.[89,91,92] As filter methods are independent of any model that is to be used in later steps, they are typically faster to implement and reduce the need for repeating feature selection for different ML models. In our selected studies, we found five studies that used Analysis of Variance (ANOVA), Pearson's Correlation, or Spearman's Correlation to identify features that were statistically significant predictors of the outcomes.[24,93–96] p-value based feature selection, while commonly used in clinical studies, is not always suitable for training a ML model. The use of p-values to identify statistically significant features was a popular approach that relied on the belief that insignificant features were not informative. However, important features can be missed when sample sizes are small. Furthermore, p-values can be biased towards low values due to the increased risk of type 1 errors during multiple comparisons, which in turn increases the probability of random variables being included into the final statistical model.[97,98] Additionally, p-value based feature selection methods may be based on certain assumptions that may not be applicable to ML models, such as assuming that the distribution of scores for the groups among the independent variables are the same.[99]

We wanted to highlight one filtering method identified in our selected studies: Relief.[100] Relief is a feature selection technique that also ranks features and selects only the top-scoring features; however, it is notably sensitive to feature interactions.[101,102] Yaman et al. first obtained 177 speech-related features and used Relief to select 66 most predictive vocal biomarkers for the classification of PD.[103] Rodriguez-Molinero used Relief to select frequency features that were subsequently used to predict gait disturbances among PD patients.[104] Overall, Relief has demonstrated its effectiveness in selecting relevant features in various studies related to the prediction of PD using high-dimensional clinical datasets.

EMBEDDED METHODS

The embedded method is a feature selection technique integrated into the ML algorithm itself and is commonly seen in penalized regression.[105] Penalized regression algorithms aim to learn the optimal coefficients for each feature by minimizing its loss function. Regularization (also known as penalization) limits the learning process of the model by increasing the penalty of the loss function.[106] The two common penalized regression methods, identified in the selected studies, are LASSO (also known as L1 penalization) (N = 9) [22,24,29,33,42,95,100,101,107,108] and Ridge (L2 penalization) (N = 2).[109,110] An advantage of LASSO is that it eliminates non-informative features by reducing their coefficients to zero. The first limitation of LASSO is that, if the number of features f is greater than the number of observations o, LASSO will select a maximum of o predictors as non-zeros, regardless of the relevance of other features. The second limitation is that LASSO also suffers from collinearity; hence, if two or more variables are highly correlated, then LASSO will randomly select one feature and penalize the other correlated features. A disadvantage of Ridge is that it only reduces the weights of the non-informative features by reducing their coefficients towards zero, but it never reduces the number of variables. Therefore, all predictors are included in the final model. However, because of this approach, Ridge protects ML models from overfitting.[111]

WRAPPER METHODS

Wrapper methods rely on a stand-alone model to select features, but the performance of the selected features is reflected in the performance of the trained model.[112] The wrapper method algorithms tend to be greedy

search algorithms that aim to select the optimal feature subset by iteratively selecting the features based on ML performance. As the wrapper method is an iterative process and the model must be evaluated on each feature subset combination, this method is computationally expensive. Wrapper-based feature selection can be completed by ranking the features in terms of relative importance using a ML model (such as decision trees or random forests).[88,101,113] We identified a handful of feature ranking methods that include two stepwise regression techniques: Forward Selection and Backwards Elimination,[29,36,52,114–116] as well as Recursive Feature Selection (RFE).[30,117] Forward selection starts the modelling process with zero features and adds a new feature to the model incrementally, each time testing for statistical significance. Backwards elimination starts the modelling process with all features and incrementally removes each feature to evaluate its relative importance in predicting the model output.[97,118] RFE fits a model, ranks the features, and removes the least informative features and continues to remove features until a predefined number of features is met.[64,119,120] Senturk et al. illustrated that RFE-based feature selection increased the prediction accuracy of ANN, CART, and SVM when using vocal data to classify a PD diagnosis.[121]

Machine learning algorithms

ML algorithms build a statistical model based on a training dataset, which can subsequently be used to make predictions about a new, unseen dataset. ML algorithms have been used in a wide variety of clinical trial applications, such as the classification of a diagnoses, classification of physical or mental state (such as a seizure or mood), and the estimation of symptom severity. Within the realm of clinical research, ML algorithms can be broadly divided into two learning paradigms: supervised and unsupervised learning.[122] In this section, we will discuss the model objectives of supervised and unsupervised learning and the specific ML models used to achieve these model objectives.

Supervised ML algorithms use labeled data to map the patterns within a dataset to a known label, while unsupervised ML algorithms do not.[123] Rather, the unsupervised ML algorithms learn the structure present within a dataset without relying on annotations. Supervised learning can be used to automate the labelling process, detect disease cases, or predict clinical outcomes (such as treatment outcomes). There are scenarios when experts or participants can provide labelled data; however, it can become labor-intensive or time-consuming to label every observation. For example, a supervised learning algorithm trained to classify human sounds can be used to automatically annotate and quantify hours of coughs [124] and instances of crying.[125] These algorithms can also be used to differentiate between clinical populations and control participants [95] to identify known clinical population subtypes [23] or classify a clinical event (such as a seizure or tremor).[126] The majority of our selected studies (N = 38) used a clinician to provide the label data. Some studies (N = 22) used a combination of a clinician and self-reported label data, and six studies solely relied on self-reported assessments. Unsupervised ML algorithms can be used to investigate the similarities and differences within a dataset without human intervention. This makes it the ideal solution for exploratory data analysis, subgroup phenotype identification, and anomaly detection. Among digital phenotyping studies, unsupervised learning has been used to identify location patterns [81] and classify sleep disturbance subtypes using wrist-worn accelerometer data.[127]

It is important to recognize that unsupervised and supervised methods are not mutually exclusive, and they can be effectively combined. For instance, unsupervised methods can be employed to extract a meaningful latent representation of the input data. Subsequently, these latent vectors, along with the original inputs, can be used as inputs for a supervised model. This type of approach is commonly observed when applying techniques such as PCA, clustering, or other dimensionality reduction methods.[29,73,74,128] By combining unsupervised and supervised methods, valuable information can be extracted from the data and used to enhance the performance and interpretability of the overall model.

In clinical research, supervised ML algorithms have been used to classify class labels or estimate scores. Classification algorithms learn to map a new observation to a predefined class label. These algorithms can be used to classify patient populations and patient population subtypes and identify clinical events. Regression algorithms learn to map an observation to a continuous output. These algorithms are commonly used to estimate symptom severity,[129] quantify physical activity, and forecast future events.[130] Among the selected papers that were focused on the classification of a diagnosis or state, the four most common algorithms were Random Forest, Support Vector Machine, Logistic Regression, and k-Nearest Neighbors (Figure 4). Some additional classification algorithm families identified were Naïve Bayes, Ensemble-based methods (including Decision Trees, Bagging, and Gradient Boosting), and Neural Networks (such as Convolutional, Artificial, and Recurring Neural Networks). The three most common algorithms for the regression focused papers were Linear Regression (including linear mixed effects models), Support Vector Machine, and k-Nearest Neighbors (Figure 4). We found that most studies only considered or reported a single ML algorithm (N = 32). Additionally, 29 of the studies considered or reported two to five ML algorithms, and the remaining 5 studies considered six or more. The following section provides an overview of the most widely used machine learning models, their properties, advantages, and disadvantages. In addition, we discuss some notable off-the-shelf ML approaches and some custom-built ML methods such as transfer learning, multi-task learning, and generalized and personalized models.

TREE-BASED MODELS

A Decision Tree (DT) is a supervised non-parametric algorithm that is used for both classification and regression. A DT algorithm has a hierarchical structure in which each node represents a test of a feature, each branch represents the result of that test, and each leaf represents the class label or class distribution.[131,132] A Random Forest (RF) algorithm is a supervised ensemble learning algorithm consisting of multiple DTs that aims to predict a class or value.[133] Ensemble learning algorithms use multiple ML algorithms to obtain a prediction.[134] Tree-based models have several benefits. As each tree is only based on a subset of features and data and because they make no assumptions about the relationship between the features and distribution, they are not sensitive to collinearity between features, can ignore missing data, and are less susceptible to overfitting (for multiple trees), making the model more generalizable.[135] Another advantage of RF and DT models is that they can support linear and nonlinear relationships between the dependent and independent variables.[136] Further, as the design of the RF models can be interpreted in terms of feature importance and proximity plots, the interpretability of the RF model is feasible. However, a limitation of using tree-based models is that small changes in the data can lead to drastically different models. Additionally, the more complicated a tree-based model becomes, the less explainable a model becomes. However, pruning the trees can help to reduce the complexity of the model.

According to the selected studies, RF is a versatile and powerful model used for classification and regression tasks across multiple datatypes and populations. RF models have been used for the classification of diagnoses among PD patients,[107,110] Multiple Sclerosis,[34,118] and BD and unipolar depressed patients.[45,61] It is also a popular classification model for the classification of states or episodes, such as the detection of flares among Rheumatoid Arthritis or Axial Spondylarthritis patients [32] and tremor detection among PD patients,[137] to quantify physical activity among cerebral palsy patients [138] and detect the moods of BD patients.[69,139] RF regression algorithms have also been used to predict anxiety deterioration among patients who suffer with anxiety.[140]

SUPPORT VECTOR MACHINES

A Support Vector Machine (SVM) is a supervised algorithm that is used for classification and regression tasks. The objective of a SVM is to identify the optimal hyperplane based on the individual observations, also known as the support vectors. For SVM regression, the optimal hyperplane

represents the minimal distance between the hyperplane and the support vectors. Whereas for SVM classification, the objective is to find the hyperplane that represents the maximum distance between two classes.[141] The hyperplanes can separate the classes in either a linear or non-linear fashion.[136] Given that SVM are influenced by the support vectors closest to the hyperplanes, SVM are less influenced by outliers, making them more suitable for extreme case binary classification. The performance of a SVM can be relatively poor when the classes are overlapping or do not have clear decision boundaries. This makes SVM less appealing for classification tasks as inter class similarity is low. SVM are computationally demanding models as they compute the distance between each support vector; hence, SVM do not scale well for large datasets.[142]

SVM classifiers have been used to classify clinical populations (e.g., facial nerve palsy and their control participants).[143] SVM classifiers have also been used to classify events or states, such as detecting gait among PD patients [104] and classifying seizures among epileptic children.[144] We identified studies that used SVM regression to estimate motor fluctuations and gait speed among PD and Multiple Sclerosis patients, respectively.[74,145]

K-NEAREST NEIGHBORS

A k-Nearest Neighbor (K-NN) algorithm is a non-parametric supervised learning approach that can be used for multi-class classification and regression tasks. Classification K-NN algorithms determine class membership by the plurality vote of its nearest neighbors. They can estimate the continuous value of an output by calculating the average value of its nearest neighbors.[136] Given this, the quality of predictions is not only dependent on the amount of data but also on the density of the data (the number of points per unit). K-NN is simple to implement, intuitive to understand, and robust to noisy training data. However, the disadvantage is that K-NN is computationally slow when it is faced with large multi-dimensional datasets. Further, K-NN does not work well with imbalanced datasets, as under- or over-represented datapoints will influence the classification.[146]

The most popular application for K-NN algorithms is for wearable-based time series data. K-NN classification models have been used to classify PD and healthy controls,[24] classify tremor severity,[147] predict acute exacerbations of chronic obstructive pulmonary disease (AECOPD),[44] and identify mood stability among BD and MDD patients.[33,69,148] Using wearable data, K-NN regression models have been used to predict the deterioration of symptoms associated with anxiety disorder.[140]

NAÏVE BAYES

A Naïve Bayes (NB) classifier is a supervised multi-class classification algorithm. NB classifiers calculate the class conditional probability—the probability that a datapoint belongs to a given class in the data.[141,149] NB classifiers are computational efficient algorithms; thus, they are suitable for real-time predictions, scale well for larger datasets, and can handle missing values. A limitation of NB is that it assumes that all features are conditionally independent; hence, it is recommended that collinear features are removed in advance. Another limitation is that when new feature-observation pairs do not resemble the data in the training data, the NB assigns a probability of zero to that observation. This approach is particularly harsh, especially when dealing with a smaller dataset. Hence, the training data should represent the entire population.

As NB classifiers help form classification models, we found that NB classifiers have been used for the classification of tremors or for freezing gait among PD patients,[52] as well as to classify flares among Rheumatoid Arthritis and Axial Spondylarthritis patients [32] and classify bipolar episodes and mood stability among BD and MDD patients.[33,69,148]

LINEAR AND LOGISTIC REGRESSION

A Linear Regression model is a supervised regression model that predicts a continuous output. It finds the optimal hyperplane that minimizes the sum of squared difference between the true data points and the hyperplane. A Logistic Regression model is a supervised classification model that can be used for binomial, multinominal, and ordinal classification

tasks. Logistic Regression classifies observations by examining the outcome variables on the extreme ends and determines a logistic line that divides two or more classes.[136] Linear and Logistic Regression are popular in algorithms as they are easy to implement, efficient to train, and easy to interpret. However, a limitation of both models is that they make multiple assumptions, e.g., that a solution is linear, the input residuals are normally distributed, and that all features are mutually independent.[150] Multicollinearity, the correlation between multiple features, and outliers will inflate the standard error of the model and may undermine the significance of significant features.[151] Further, outliers that deviate from the expected range of the data can skew the extreme bounds of the probability, making both algorithms sensitive to outliers in the dataset.[150]

Linear Regression has been used to quantify tremors among Essential Tremor (ET) patients [116] and to estimate motor-related symptom severity among PD patients.[31,93] It has also been used to forecast convergence between body sides for Hemiparetic patients.[130] Logistic Regression was a popular approach for classifying PD diagnosis,[107,110] Post-Traumatic Stress Disorder,[109] and distinguishing fallers and non-fallers.[152] Logistic Regression has been used to classify drug effects, such as predicting the pre- and post-medication states among PD patients.[22]

NEURAL NETWORKS

Neural Networks (NN), also known as Artificial Neural Networks (ANN), can be used for unsupervised and supervised classification and regression tasks.[153] NN consists of a collection of artificial neurons (or nodes). Each artificial neuron receives, processes, and sends the signal to the artificial neuron connected to it. The neurons are aggregated into multiple layers, and each layer performs different transformations on the signal. The signal first travels from the input layer into the output layer while possibly traversing multiple hidden layers in between. NN offer several advantages, such as the ability to detect complex non-linear relationships between features and outcomes and work with missing data, while it also requires less preprocessing of the data and offers the availability of multiple training algorithms. However, the disadvantages of NN include increased computational burden, reduced explainability and interpretability (as NN are 'black box' in nature), and the fact that NN are prone to overfitting.[154] However, it is important to highlight the growing number of studies that specifically explore explainable deep learning approaches for biomarker discovery and development. Studies utilizing methodologies such as LIME (LIME Tabular Explainer), SHAP (SHAPley Additive exPlanations), and other visual inspections of feature distribution and importance have aided clinicians in understanding the model mechanisms. These approaches also provide patient-specific insights by describing the importance of each feature, which may, in turn, facilitate personalized treatment opportunities.[90,155–157]

The most popular applications for neural networks were for the classification of a diagnosis or classification of a state or event. The most popular application is the detection of tremors among PD patients.[23,52,86,137,158] NN have been used to classify unipolar and bipolar depressed patients based on motor activity,[45,159] estimate depression severity,[159] forecast seizures,[160] and classify a treatment response using keyboard patterns among PD patients.[161]

TRANSFER LEARNING

Transfer learning (also known as domain adaption) refers to the act of deriving the representations of a previously trained ML model to extract meaningful features from another dataset for an inter-related task.[162] One applicable scenario is the training of a supervised ML model on data collected in a controlled setting (such as in a lab or clinic). The performance of the model may suffer when applied to a dataset collected under free-living conditions. Rather than developing a new model trained solely on a free-living condition dataset, transfer learning can use patterns learned from the controlled setting dataset to improve the learning of the patterns from the free-living conditions dataset.

Transfer learning can also be a valuable technique for enhancing the utilization of limited or rare data.[163] One practical application is to employ

pretraining on abundant control data and subsequently finetune the model on the specific population of interest to improve the model's performance.[163–165] This approach not only optimizes the efficiency of utilizing scarce data but also facilitates model personalization. By adapting a pretrained model to individual characteristics or preferences, it becomes possible to create personalized models that better cater to unique needs or circumstances. Transfer learning thus offers a powerful means to leverage existing knowledge and make the most of available data resources, enhancing both the efficiency and personalization of biomarkers.

Given its application, transfer learning reduces the amount of labeled data and computational resources required to train new ML models,[162] thus making this method advantageous when the sensor modalities, sensor placements, and populations differ between studies. While we only identified two studies that applied transfer learning to estimate PD disease severity using movement sensor data,[166,167] we predict that the application of transfer learning will enable future researchers to overcome the challenges of a limited dataset and develop more sensitive and effective ML models.

MULTI-TASK LEARNING

Multi-task learning (MTL) enables the learning of multiple tasks simultaneously.[168] Learning the commonalities and differences between multiple tasks can improve both the learning efficiency and the prediction accuracy of the ML models.[168] A traditional single-task ML model can have a performance ceiling effect, given the limitations of the dataset size and the model's ability to learn meaningful representations. MTL uses all available data across multiple datasets and can learn to develop generalized models that are applicable to multiple tasks. To use MTL, there should be some degree of information shared between or across all tasks. The correlation allows MTL to exploit the underlying shared information or principles within tasks. Sometimes MTL models can perform worse than single-task models because of 'negative transfers'. This occurs when different tasks share no mutual information or if the information of tasks are contradictory.[169] MTL models have been used to simultaneously model data sourced from two separate sources or to model multiple outcomes.[170,171] For example, Lu ET AL. explored the use of MTL to jointly model data collected from two different smartphone platforms (iPhone and Android) to jointly predict two different types of depression assessments (QIDS and a DSM-5 survey).[79] They illustrated that the classification accuracy of the MTL approach outperformed the single-task learning approach by 48%; thus, the classification model benefited from learning from observations sourced from multiple devices.

GENERALIZED VERSUS PERSONALIZED

ML algorithms can be trained on population data or individual subject data. Generalized models, which are trained on population data, are fed data from all participants for the purpose of general knowledge learning. Conversely, personalized models are trained on an individual's data and take into consideration individual factors such as biological or lifestyle-related variations.[172] We have adopted these terms from Kahdemi et al.'s study, in which they developed generalized and personalized models for sleep-wake prediction.[173] The heterogenous nature of each population or individual can be a potential hinderance for generalizable models. A single individual's deviation from the 'norm' may be viewed as a source of 'noise' in a generalized model. For example, patients with mood disorders such as MDD and BD have large inter-individual symptom variability. Abdullah ET AL., reliably predicted the social rhythms of BD patients with personalized models using smartphone activity data.[30] Cho et al. compared the mood prediction accuracy of personalized and generalized models based on the circadian rhythms of MDD and BD participants.[38] Their studies illustrated that their personalized model predictions were, on average, 24% more accurate than the generalized models. These studies lay the groundwork for developing personalized models that are more sensitive to individual differences.

MODEL HYPERPARAMETERS

The process of building an effective ML model consists of two main steps: selecting the appropriate ML algorithm and optimizing the model performance by tuning its parameters. Each model consists of two types of parameters:

- The parameters that are initialized and continuously updated throughout the learning process (e.g., the weights of neurons of a neural networks).
- The hyperparameters that must be set prior to the learning process as they define the model architecture (e.g., the regularization parameters of a Linear Regression model, and the learning rates of a neural network).[174]

Every combination of the selected hyperparameters will have a direct influence on the performance of the learned model. For example, as the number of trees in a RF increases, the more features tend to be selected by the model, which may not always be relevant for the development of biomarkers.[175] Similarly, the number of layers, number of neurons per layer, activation functions, and the regularization techniques used for NN can each influence the model performance.[176] While most ML algorithms come with default values for the hyperparameters, these may not be optimal for the dataset at hand, and even tuned hypermeters are at risk of being non-optimal for a different dataset. The process of selecting the optimal hyperparameter configurations is known as hyperparameter tuning.[177]

To identify the optimal hyperparameters for a model, researchers must define the hyperparameter space and the hyperparameter search strategy. When defining the hyperparameter space, the distribution of the hyperparameter ranges can be either uniform or logarithmic. The uniform distribution assigns equal probability to all hyperparameter values within a manually defined range. The log-uniform distribution samples hyperparameter values uniformly between the logarithmic transformations of the lower and upper thresholds. We argue that log-uniform distribution is particularly useful when exploring values that vary over several orders of magnitude. Consider the example of tuning a linear regression model with the hyperparameter alpha, which determines the strength of regularization. To efficiently explore a wide range of alpha values, such as between 0.001 and 10, the log-uniform distribution allows for an evenly distributed search space over different orders of magnitude. Log-uniform distribution can be used for the initial exploration of a large range of hyperparameter values. The range can then be narrowed down to explore with a uniform-distribution to determine the optimal hyperparameters for the respective models.

The manual tuning of hyperparameters is impractical due to the large number of available hyperparameters, hyperparameter configurations, and time-consuming model evaluations. Automated tuning approaches are preferred, and there are a wide variety of approaches available, including GridSearch, RandomSearch, and Bayesian Optimization.[177] GridSearch uses brute force to test a finite combination of hyperparameters to identify the optimal hyperparameter configuration.[178] This approach can suffer from the effects of dimensionality, as more potential hyperparameter configurations can be time-consuming and computationally expensive. An alternative to GridSearch is RandomSearch. RandomSearch only samples a subset of all possible hyperparameter configurations within a specific time or computational budget.[179] While RandomSearch only relies on a subsample of configurations, it has been shown to outperform the GridSearch method.[179] As GridSearch and RandomSearch do not consider previous performance evaluations for their hyperparameter optimization strategy, they are inefficient in exploring the hyperparameter search space. Bayesian Optimization, which uses Bayes Theorem, is a powerful approach. It considers previous hyperparameter evaluations to choose which hyperparameters to evaluate next and disregards potential hyperparameter combinations that are deemed irrelevant.[178] This approach reduces the time and computations required for hyperparameter tuning. The benefit of using these more automated

approaches to hyperparameter tuning is three-fold. First, it reduces the time effort required to optimize a ML model. Next, the performance of the ML models is improved as the hyperparameters explore different optimal model configurations for different datasets. Finally, when the hyperparameters and their ranges (together also referred to as the hyperparameter space) and the hyperparameter tuning methods are reported, the models and the findings become reproducible.[180] When similar hyperparameter tuning processes can be used for different ML algorithms for different datasets, researchers can then identify the optimal ML model.

Among the selected studies, 25 discussed which hyperparameters were considered for their models,[23,24,34,43,44,46,53,69,73,86,87,94,95,107–110,114,138,158,159,181–184] of which one stated they used the default hyperparameters of the models.[69] Only nine studies discussed how they selected or optimized their hyperparameters. We identified four studies that stated GridSearch was used for the hyperparameter tuning.[36,46,95,110] We did not identify any studies that used RandomSearch or Bayesian Optimization. The limited reporting of hyperparameters and the hyperparameter tuning process poses a problem for the transparency, reproducibility, and comparison of ML models.

Model evaluation

Assessing a ML model's performance is an essential component for determining the usability and reliability of the model. Depending on the objective of the research, it is often necessary to try to compare the performance of multiple ML models to identify the optimal model.[185,186] In ML, the terms metric and measure are often used interchangeably, but they do have slightly different meanings. A metric is a function used to evaluate the performance of a model, while a measure is a numerical summary of the performance of a model obtained using one or more metrics. It is best practice to use multiple metrics and model performance visualizations for the model evaluation, as a model may perform well for one evaluation metric and poorly for another. Using multiple evaluation metrics ensures that the model is operating optimally and correctly. The following sections provide more details about the performance metrics used for classification and regression models. Table 4 provides an overview of the most common performance metrics used in the selected studies, their respective calculations, and their clinical interpretations.

CLASSIFICATION MEASURES

Classification models have discrete outcomes; thus, a metric must reflect how often an observation belongs to the correct label or class.[187] There are three categories of classification measures: Threshold Metrics, Ranking Metrics, and Error Metrics. Threshold Metrics (such as accuracy and F1 score) quantify the prediction errors of the classification model as a ratio or rate. Ranking Metrics (such as the Receiver Operating Characteristics (ROC) and Area Under the Curve (AUC)) focus on evaluating classification models based on how effective they can discern separate classes. Error Metrics (such as Root Mean Square Error) quantify the uncertainty of the classification model's predictions. While the Threshold and Ranking Metrics are focused on correct and incorrect predictions, the Error Metrics quantify the proportion of classification errors.

As ML models are increasingly being used to perform high-impact tasks pertaining to clinical assessments, an evaluation metric must be selected based on what the stakeholders find to be important regarding the model prediction, which can make the selection of the model metrics challenging. As seen in Table 4, accuracy, sensitivity, specificity, and precision are calculated based on four test results. The True Positive (TP) and True Negative (TN) indicate the presence or absence of a diagnostic or characteristic. The False Positive (FP) and False Negative (FN) indicate the opposite of the true condition.

Binary classification models typically involve a decision threshold hyperparameter that determines how the model assigns labels based on the predicted probabilities. The default threshold is typically 0.5, meaning that if the predicted probability is greater than 0.5, the positive label is assigned, and vice versa. However, it is important to note that this

threshold can be adjusted to accommodate specific needs or domain considerations. To evaluate the performance of binary classification models across different decision thresholds, the ROC curve is commonly used. The ROC curve provides an overview of the model's performance by illustrating the trade-off between TP and FP rates at various threshold values. ROC can aid the assessment of the model's performance across a range of decision thresholds and enable the selection of the threshold that aligns with a specific objective.

It is worth noting that many classification metrics, including accuracy, precision, recall, and F1 score, assume binary labels. However, when dealing with multiclass classification problems, another approach is to use one-vs-rest or one-vs-one strategies, wherein the problem is decomposed into multiple binary classification tasks. The performance of the model on each task can then be evaluated using the binary classification metrics, and the results can be aggregated or averaged to provide an overall assessment of the model's performance on the multiclass problem.

Class imbalance can be an obstacle for assessing model performance. In particular, accuracy, AUC, ROC, may be sensitive to such imbalances.[188] Hence, when facing class imbalance, there are two approaches to consider: one can choose a metric that accounts for class imbalance or one can choose to balance the classes. Metrics such as balanced accuracy, F1-score, or Matthews Correlation Coefficient (MCC) are common metrics for handling class imbalance, as identified by 15 studies.[23,24,29,36,44,60,61,107,108,110,114,140,159,161,189] Balanced accuracy represents the mean of the sensitivity and specificity, while the F1-score represents the mean of the precision and recall.[190] The MCC measures the correlation coefficient of the binary and even multiclass classes. Therefore, the MCC score is high only if the classification model correctly predicts both the positive and negative predictions.[190,191]

The other approach to handling class imbalances is adjusting the class distribution using oversampling or undersampling. We identified eight studies that used random over/under sampling or SMOTE.[29,44–46,61,95,109,192] Oversampling techniques duplicate the samples of the minority class, while undersampling removes samples of the majority class. However, these techniques also have their disadvantages, as the duplication of multiple samples can lead to overfitting of a model, while undersampling reduces the diverse representation of the majority class. Thus, we would specifically recommend using the Synthetic Minority Oversampling Technique (SMOTE) with Tomek Links or Edited Nearest Neighbor (ENN)—two undersampling techniques.[193,194] SMOTE is first applied to create an artificial minority class to minimize the class imbalance. Next, Tomek Links or ENN can be used to remove samples that are close to the boundaries between the classes, which would further separate the classes.[193,194]

REGRESSION MEASURES

As regression models generate predictions on a continuous scale, the objective is to estimate how close the predictions were to the true values.[195] Among the studies selected, we found that regression models used Distance Metrics and Error Metrics to estimate the strength of the association or the distance between the predicted values and the true values.[29,42,87,93,96,128,152] We would like to emphasize that these metrics are used to compare the performance of the composite biomarkers rather than the performance of the individual features. The most common Distance Metrics were the correlation (also known as R) and the percentage of the variance explained (R2). Both were used to assess the strength of the association between the predicted and true values.[196] There is no rule of thumb for interpreting the strength of R2. While an R2 closer to 1 can be obtained in clinical trials, a low R2 can still be useful with respect to trends in the data. We would like to address two points of caution when using the R2.[185,187] First, it is not always suitable to compare R2 across different datasets, as different clinical populations are likely to differ in their feature variance. Second, the R2 will increase with the number of features. To compensate for this, one may use the adjusted R2 to account for the number of features.[197,198]

The Error Metrics included the Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).[133] The MAE measures the average absolute difference between the true and predicted values. The MAE is easy to interpret and robust to outliers. The absolute difference accounts for negative differences. The MSE squares the error instead of providing the absolute error, which gives more weight to the bigger errors. The MSE is sensitive to outliers and not easy to interpret, as the results will not have the same unit as the output. However, the RMSE provides an estimation of the error in the same units as the output while maintaining the properties of the MSE.[199]

Model validation

In ML, model validation refers to the process of evaluating the generalizability of a trained model on an unseen dataset. Selecting the most appropriate model validation approach depends on the size and characteristics of the datasets. Three datasets are required for model validation: the training, test, and validation datasets. In most cases, the validation dataset can be a subset of the original dataset; however, this can lead to data leakage, which could produce overly optimistic results. Another approach is to create a validation dataset from an independent (but comparable) dataset, which ensures an unbiased and independent evaluation of the ML model. However, a limitation is that the performance evaluation may reflect high variance due to the limited size of the dataset.[200] Moreover, it is crucial to highlight that a participant should only be present in a single dataset, such as the training dataset, and should not simultaneously appear in other datasets such as the testing or validation datasets. When a participant's observations are distributed across multiple datasets, data leakage can occur, compromising the accuracy estimation and its applicability to new participants.[183] As a result, cross-validation on the observation level rather than the participant level is methodologically flawed. Unfortunately, this is a common issue and needs to be accounted for in future studies.[201]

Cross-validation is a popular validation method that uses resampling to train, test, and validate a model using different subsets of the data. The training dataset is used to train the ML model to learn the patterns within a dataset. The validation dataset is used to tune the hyperparameters of the model based on the performance of the ML model trained on the training dataset. The test dataset provides an unbiased estimate of the performance of the final ML model after training and validation. In the scenario when both validation and test datasets are used, the test datasets are only used to assess the model once (via hold-out validation) or multiple times (via nested cross-validation). In general, datasets need to meet two main requirements. The datasets should not have shared or overlapping observations to ensure that data leakage does not lead to bias in the estimates, and all observations must be statistically independent.[202] When applying feature engineering or feature selection with cross-validation, any transformation or selection steps should be performed within each fold of the cross-validation to prevent biasing in the training of the prediction model with information from the test dataset.[203] The overall performance of the prediction models, obtained by averaging across each iteration of the cross-validation, evaluates the effectiveness of the combined feature reduction and learning methods in estimating the label for a given dataset.

Among the selected studies, we found that the most popular cross-validation methods were k-fold cross-validation (N = 27), Leave-One-Out cross-validation (N = 16), and custom validation (N = 8). Overall, 15 studies did not report the use a validation method. K-fold cross-validation randomly splits the datasets in 'k' folds; one-fold is used for testing and the remaining folds are used for training. This step is repeated until every unique fold has been used as the test dataset, and the overall performance is based on the average of the performance of each model in each fold.[204] Leave-one-out cross-validation is a specific type of k-fold cross-validation, wherein individual observations (or participants) are the test datasets, and the remaining cases are used for training. Leave-one-out cross validation prevents data leakage across datasets, as repeated

measurements of the same subjects can lead to the violation of independence assumption for ordinary cross-validation.[204–206]

We would like to highlight the advantages of the nested cross-validation approach. While nested cross-validation was the least popular approach, we would argue that nested cross-validation is a more robust approach for selecting and evaluating a ML model.[207] Currently, the model section without the nested cross-validation approach uses the same data to both tune the model hyperparameters and evaluate its performance. Therefore, information is 'leaked' between the training and validation of the model, which can lead to overfitting.[207] Nested cross-validation consists of an inner loop and an outer loop. The outer loop assesses the model performance, while the inner loop assesses the hyperparameter selection.[207] Each iteration of the outer loop is split into a different combination of training and test sets. The outer loop training set is used in the inner loop, which is further split into a training and validation dataset. The inner loop split is repeated over k-folds, and the best performing model across the k-folds is evaluated in the outer loop. This ensures that different data are used to optimize the models' hyperparameters and evaluate the model's performance. The final model performance represents the average and standard deviation of the model performance as selected by each of the outer loops. Without the standard deviation or confidence intervals, it is not possible to evaluate the spread or stability of the prediction error of the given models.[208,209]

It is important to highlight that cross-validation is only used to approximate the generalization error of the models built and not to build the final model that will be used for making predictions.[205,210] The average prediction error across the folds gives an expected error for a single model built on the single dataset. If the variance of the prediction error is too high, then the model is considered unstable. To select a single model, it is recommended that researchers rebuild the model using the full dataset.[211] If an external validation set is available, then this validation set can be used to evaluate and compare the single prediction error to that of the cross-validation prediction error.

Recommendations

In this recommendation section, we address the main issues consistently identified in the selected studies and how to amend these issues for future trials (see Figure 5 for a simplified overview of these recommendations). It is important to bear in mind the regulatory implications for developing ML-derived biomarkers. Within the European Union, AI medical systems and devices are considered high risk; therefore, they are subject to stringent reviews prior to being made available on the market.[212] These review requirements emphasize the importance of achieving high levels of performance, transparency, and minimal risk in ML-derived biomarker development.[213] High performance implies that the developed ML models must be accurate, robust, and capable of reliably and consistently predicting the target outcome variable. Furthermore, transparency in ML-derived biomarker development refers to the provision of clear and adequate information to the user, including appropriate human-readable measures to minimize risks associated with the use of the system. The development of ML-derived biomarkers must also aim to minimize risks and discriminatory outcomes, which can be achieved by training the ML model on high-quality datasets that are representative of the target population and by conducting adequate risk assessment checks.[214] These considerations are critical for ensuring the safe and effective use of ML-derived biomarkers in clinical practice.

INCLUSION OF HEALTHY CONTROLS

When conducting a study focused on disease classification or estimation, the inclusion of control data can serve several purposes. By comparing the data from individuals with the condition of that of the healthy controls, researchers can discern whether the observed differences are specific to the condition or a result of unrelated factors. Moreover, analyzing the performance of a model on control subjects can shed light on the biomarker's effectiveness and reliability. By evaluating how well the model distinguishes between healthy controls and patients with the

condition, researchers can gain a better understanding of its predictive capabilities. This evaluation can provide insights into potential false positives or false negatives that may occur when using the model in real-world settings.

It is worth noting that, when including control data, the control data should be appropriately matched with the patient population data. Having age- and gender-matched control subjects can help minimize confounding variables, improving the accuracy of the analysis. This matching process allows researchers to draw more robust conclusions about the relationship between the identified features or patterns and the disease activity while also reducing the potential impact of demographic factors on the results.

The finding that only half of the studies included healthy controls is significant as it highlights a potential gap or limitation in the existing body of research. Without the inclusion of controls, it becomes challenging to attribute identified features or patterns solely to the CNS disorder or the severity of the condition. Further, if the dataset only contains a relatively homogeneous population, it calls the reliability and predictive capabilities of the models into question. We encourage future researchers to include control subjects in their studies, as it would improve the strength of their biomarkers and the validity of their findings.

DATA QUALITY AND PREPROCESSING

The remote monitoring of clinical trials can generate large and complex datasets that include longitudinal data from multiple subjects and data sourced from multiple sensors, resulting in a multi-dimensional data structure. To this point, we recommend using the WHO MHEALTH Technical Evidence Review Groups' MHEALTH evidence and evidence reporting and assessment (MERA) 16-item checklist to provide transparency on which MHEALTH invention was used, where, and how it was implemented to support the reproducibility of the MHEALTH data collection.[215] To ensure the quality and reliability of the data, it is important to assess the quality of the data. This assessment includes examining the data for missing and outlier data and understanding how these factors might affect the generalizability and reproducibility of the ML model. While most studies provide detailed information on patient populations, the devices used, and the data collected, they often underreport information related to data quality and preprocessing steps. Therefore, it is important to provide sufficient details on the methods used to preprocess the data, including the quantity of missing and outlier data and the strategies employed to handle such data. This information can ensure that the data collection and preprocessing process can be reproduced, which, in turn, can enhance the credibility and generalizability of the ML model.

FEATURE ENGINEERING AND SELECTION

There is a wide variety of manual or automated techniques used for engineering and selecting features to feed a model. ML models perform best when feature engineering and selection are leveraged to formulate potentially clinically relevant features from existing data. In addition, the performance of the ML model can be optimized, and the computational time can be reduced when the redundancy across the features is reduced. While only selecting the most informative features can remove noise (therefore reducing the likelihood of overfitting), selecting too few features may reduce the strength of the (combined) signal in the dataset, making the ML model vulnerable to underfitting. Feature engineering and selection can be guided by domain expertise and/or automated statistical models, where multiple features are evaluated by their importance in predicting the outcome. While automated feature engineering techniques, such as clustering, PCA, and DL, can be used to extract a reduced set of representative features, this risks a potential decline in interpretability, which may limit its clinical application.

MODEL CONFIGURATION AND OPTIMIZATION

When selecting the ML models, there are several factors that should be considered, such as model objectives, model types, model hyperparameters, and model evaluation. Poor design choices and lenient

hyperparameter tuning and validation in these steps can lead to poor model performance. We recommend that researchers carefully consider each step of building their ML pipeline by comparing multiple ML algorithms, using automated methods for assessing multiple hyperparameter configurations, and using nested cross validation to both optimize and validate the ML models.

MODEL VALIDATION

We would recommend using a minimum of three datasets to validate a ML model and train, validate, and test a dataset. At no point should the test set be used for the model configuration, which includes the data transformation, feature engineering, and selection, or the tuning of the hyperparameters. The test dataset could either be a subset of the original data (with no overlapping subjects or observations) or a separate external dataset. The use of an external dataset is ideal as this ensures that there is no influence of bias during the data collection period and that there is no data leakage between the datasets. If an external dataset is not available or if the dataset is not sufficiently large, we recommend nested cross-validation. This resampling method supports model hyperparameter tuning and performance evaluation without the risk of data leakage across the dataset.

It is crucial to report the evaluation metric results for each dataset. In the case of cross-validation reporting, we recommend that researchers report the distribution of the performance measures (e.g., the mean and standard deviation or median and 95% confidence interval) across the folds to show the average and variability of the performance of the models. As cross-validation evaluates the prediction error across multiple ML models, we would also recommend reporting the performance of the final model selected. This is achieved by re-training a ML model on the full dataset and evaluating the performance on an external dataset.[207,210] This would give insight into how well the model would perform under different circumstances. We also highly recommend using multiple evaluation metrics for assessing the model's performance. Seeing as a model might excel for one metric and fail for another, this underscores the need for comprehensive evaluation. Employing multiple metrics ensures optimal operation and reduces the likelihood of blind spots.

Once the final model has been trained, there are three approaches to choose from to apply the model to a new target dataset. The first approach is to test the model 'as-is', implying that the ready-made model can be used in its original state without modifications.[216] In the second scenario, the train data and the target data may have different characteristics, which may lead to a distribution shift. The type of distribution shift between the two datasets can occur for many reasons, including different MHEALTH devices used for data collection, environmental noise, and sampling bias.[217] When this occurs, transfer learning can be used to fine-tune the ready-made model and update its weights to better suit the target dataset.[216] In the third scenario, the target dataset may have different requirements than the original training dataset.[216] As a result, the decision boundary of the classification model can be altered, such as optimizing the model for a sensitivity of 90% instead of accuracy. Whether testing the model as-is, employing transfer learning, or adjusting the decision boundary, these strategies offer flexibility in adapting the model to different settings and improving its performance for validation purposes.

MODEL REPRODUCIBILITY AND INTERPRETABILITY

Equally important as the model performance are the ML models' reproducibility and interpretability. Reproducibility is a core component for ensuring that a ML model can be validated and reused by clinical researchers. Technical reproducibility involves using the same computational procedures to produce consistent model outcomes. Statistical reproducibility ensures that the model demonstrates similar statistical performance across different subsets of data. Conceptual reproducibility refers to achieving consistent results under new conditions, such as data collected from different settings.[216] Transparency regarding data quality, feature engineering and selection methods, the hyperparameters considered and selected, and the model validation protocol can help ease the

ability of the scientific community to recreate the work in the published literature. Best practices for reproducibility include publishing the code on GitHub or by publishing FAIR metadata.[211,218,219]

Given the potential clinical application of ML models, prior to modeling, researchers should determine the model's interpretability requirement. While ML models provide researchers with what was predicted, interpretability requires that the model can explain why it made the prediction.[185] Interpretability enables us to understand the causal relationships between the data and the ML model's predictions. There are two situations in which the interpretability of a model is required: when an inaccurate prediction can have severe or even fatal consequences for the patients (such as a misclassified diagnosis [220]) and when the interpretability can be used to identify novel relationships between clinical factors and the predicted outcome (such as factors influencing treatment outcomes [221]). There can be two situations in which interpretability is not required: situations in which incorrect predictions do not have severe consequences (such as counting the number of coughs [222]) or situations in which the ML model has been sufficiently validated in real clinical applications, even if the predictions are not perfect.[223] While black box models may offer more accurate predictions than an interpretable model, they only provide limited insight into how the predictions were made. Therefore, both interpretable and black box models have their respective merits.

There are two broad approaches towards achieving interpretability.[224] One approach is to use easy-to-interpret models, such as Linear or Logistic Regression, where the coefficients of the features can provide insight into the features' associations with the predicted outcome. The other approach is to use explanation methods for explaining complex or black box models, such as SHAPley Additive exPlanations plots (SHAP), Local Interpretable Model-agnostic Explanations (LIME), or Anchors.[224] We recommend that researchers report whether their final selected model was an interpretable model or a black box.[225] If it was interpretable, we recommend discussing what interpretations can be derived from the models.

Conclusions

The rise and breadth of ML applications in clinical trials highlight the increasing reliance and importance of ML in the development of novel biomarkers.[226] While the advances in ML applications have demonstrated great potential for innovative biomarker development, the process of its development is not well documented, which, in turn, limits the reproducibility of these findings. This review has illustrated the steps taken to translate raw data from MHEALTH technologies into meaningful clinical biomarkers using ML. Given the lack of consistent reporting in the ML methods, the present review cannot provide a complete or detailed picture of the notable and generic practices. However, the authors have provided an overview of the status quo of the development and translation of ML-derived biomarkers in MHEALTH-focused clinical trials. The recommended checklist provided in the review could serve as a foundation for the design of future ML-derived biomarkers in conventional ML practices. By encouraging consistent and transparent reporting, researchers can accelerate the integration of novel biomarkers derived from MHEALTH sensors and ML pipelines into future clinical trials.

REFERENCES

1 Au, R.; Lin, H.; Kolachalama, V.B. Tele-Trials, Remote Monitoring, and Trial Technology for Alzheimer's Disease Clinical Trials. In Alzheimer's Disease Drug Development; Cambridge University Press: Cambridge, UK, 2022; pp. 292–300.

2 Inan, O.T.; Tenaerts, P.; Prindiville, S.A.; Reynolds, H.R.; Dizon, D.S.; Cooper-Arnold, K.; Turakhia, M.; Pletcher, M.J.; Preston, K.L.; Krumholz, H.M.; ET AL. Digitizing clinical trials. NPJ Digit. Med. 2020, 3, 101.

3 Teo, J.X.; Davila, S.; Yang, C.; Hii, A.A.; Pua, C.J.; Yap, J.; Tan, S.Y.; Sahlén, A.; Chin, C.W.-L.; Teh, B.T.; ET AL. Digital phenotyping by consumer wearables identifies sleep-associated markers of cardiovascular disease risk and biological aging. bioRxiv 2019.

4 Brietzke, E.; Hawken, E.R.; Idzikowski, M.; Pong, J.; Kennedy, S.H.; Soares, C.N. Integrating digital phenotyping in clinical characterization of individuals with mood disorders. Neurosci. Biobehav. Rev. 2019, 104, 223–230.

5 Kourtis, L.C.; Regele, O.B.; Wright, J.M.; Jones, G.B. Digital biomarkers for Alzheimer's disease: The mobile/wearable devices opportunity. NPJ Digit. Med. 2019, 2, 9.

6 Bhidayasiri, R.; Mari, Z. Digital phenotyping in Parkinson's disease: Empowering neurologists for measurement-based care. Park. Relat. Disord. 2020, 80, 35–40.

7 Prosperi, M.; Min, J.S.; Bian, J.; Modave, F. Big data hurdles in precision medicine and precision public health. BMC Med. Inform. Decis. Mak. 2018, 18, 139.

8 Torres-Sospedra, J.; Ometov, A. Data from Smartphones and Wearables. Data 2021, 6, 45.

9 García-Santıllán, A.; del Flóres-Serrano, S.; López-Morales, J.S.; Rios-Alvarez, L.R. Factors Associated that Explain Anxiety toward Mathematics on Undergraduate Students. (An Empirical Study in Tierra Blanca Veracruz-México). Mediterr. J. Soc. Sci. 2014, 5.

10 Iniesta, R.; Stahl, D.; Mcguffin, P. Machine learning, statistical learning and the future of biological research in psychiatry. Psychol. Med. 2016, 46, 2455–2465.

11 Rajula, H.S.R.; Verlato, G.; Manchia, M.; Antonucci, N.; Fanos, V. Comparison of Conventional Statistical Methods with Machine Learning in Medicine: Diagnosis, Drug Development, and Treatment. Medicina 2020, 56, 455.

12 Getz, K.A.; Rafael, A.C. Trial watch: Trends in clinical trial design complexity. Nat. Rev. Drug. Discov. 2017, 16, 307.

13 Getz, K.A.; Stergiopoulos, S.; Marlborough, M.; Whitehill, J.; Curran, M.; Kaitin, K.I. Quantifying the Magnitude and Cost of Collecting Extraneous Protocol Data. Am. J. Ther. 2015, 22, 117–124.

14 Getz, K.A.; Wenger, J.; Campo, R.A.; Seguine, E.S.; Kaitin, K.I. Assessing the Impact of Protocol Design Changes on Clinical Trial Performance. Am. J. Ther. 2008, 15, 450–457.

15 Globe Newswire. Rising Protocol Design Complexity Is Driving Rapid Growth in Clinical Trial Data Volume, According to Tufts Center for the Study of Drug Development. Available online: https://www.globenewswire.com/news-release/2021/01/12 /2157143/0/en/Rising-Protocol-Design-Complexity-Is-Driving-Rapid-Growth-in-Clinical-Trial-Data-Volume-According-toTufts-Center-for-the-Study-of-Drug-Development.html (accessed on 12 January 2021).

16 Santos, W.M.D.; Secoli, S.R.; de Araújo Püschel, V.A. The Joanna Briggs Institute approach for systematic reviews. Rev. Lat. Am. Enferm. 2018, 26, e3074.

17 Central Nervous System Diseases—MeSH—NCBI. 2023. Available online: https://www.ncbi.nlm.nih.gov/mesh?Db=mesh& Cmd=DetailsSearch&Term=%22Central+Nervous+System+Diseases%22%5BMeSH+Terms%5D (accessed on 5 January 2023).

18 Martinez, G.J.; Mattingly, S.M.; Mirjafari, S.; Nepal, S.K.; Campbell, A.T.; Dey, A.K.; Striegel, A.D. On the Quality of Real-world Wearable Data in a Longitudinal Study of Information Workers. In Proceedings of the 2020 IEEE International Conference on Pervasive Computing and Communications Workshops, PerCom Workshops 2020, Austin, TX, USA, 23–27 March 2020.

19 Ruiz Blázquez, R.R.; Muñoz-Organero, M. Using Multivariate Outliers from Smartphone Sensor Data to Detect Physical Barriers While Walking in Urban Areas. Technologies 2020, 8, 58.

20 Poulos, J.; Valle, R. Missing Data Imputation for Supervised Learning. Appl. Artif. Intell. 2018, 32, 186–196.

21 Schafer, J.L.; Graham, J.W. Missing data: Our view of the state of the art. Psychol. Methods 2002, 7, 147–177.

22 Evers, L.J.; Raykov, Y.P.; Krijthe, J.H.; de Lima, A.L.S.; Badawy, R.; Claes, K.; Heskes, T.M.; Little, M.A.; Meinders, M.J.; Bloem, B.R. Real-life gait performance as a digital biomarker for motor fluctuations: The Parkinson@Home validation study. J. Med. Internet Res. 2020, 22, e19068.

23 Papadopoulos, A.; Kyritsis, K.; Klingelhoefer, L.; Bostanjopoulou, S.; Chaudhuri, K.R.; Delopoulos, A. Detecting Parkinsonian Tremor from IMU Data Collected In-The-Wild using Deep Multiple-Instance Learning. IEEE J. Biomed. Health Inform. 2019, 24, 2559–2569.

24 Tougui, I.; Jilbab, A.; El Mhamdi, J. Analysis of smartphone recordings in time, frequency, and cepstral domains to classify Parkinson's disease. Healthc. Inform. Res. 2020, 26, 274–283.

25 Meyerhoff, J.; Liu, T.; Kording, K.P.; Ungar, L.H.; Kaiser, S.M.; Karr, C.J.; Mohr, D.C. Evaluation of Changes in Depression, Anxiety, and Social Anxiety Using Smartphone Sensor Features: Longitudinal Cohort Study. J. Med. Internet Res. 2021, 23, e22844.

26 Dinesh, K.; Snyder, C.W.; Xiong, M.; Tarolli, C.G.; Sharma, S.; Dorsey, E.R.; Sharma, G.; Adams, J.L. A Longitudinal Wearable Sensor Study in Huntington's Disease. J. Huntingt. Dis. 2020, 9, 69–81.

27 Cho, C.-H.; Lee, T.; Lee, H.-J. Mood Prediction of Patients with Mood Disorders by Machine Learning Using Passive Digital Phenotypes Based on the Circadian Rhythm: Prospective Observational Cohort Study. 2019. Available online: https://www.ncbi. nlm.nih.gov/pmc/articles/PMC6492069/ (accessed on 5 January 2023).

28 Tanaka, T.; Kokubo, K.; Iwasa, K.; Sawa, K.; Yamada, N.; Komori, M. Intraday activity levels may better reflect the differences between major depressive disorder and bipolar disorder than average daily activity levels. Front. Psychol. 2018, 9, 2314.

29 Palmius, N.; Tsanas, A.; Saunders, K.E.A.; Bilderbeck, A.C.; Geddes, J.R.; Goodwin, G.M.; De Vos, M. Detecting bipolar depression from geographic location data. IEEE Trans. Biomed. Eng. 2017, 64, 1761–1771.

30 Abdullah, S.; Matthews, M.; Frank, E.; Doherty, G.; Gay, G.; Choudhury, T. Automatic detection of social rhythms in bipolar disorder. J. Am. Med. Inform. Assoc. 2016, 23, 538–543.

31 Ramsperger, R.; Meckler, S.; Heger, T.; van Uem, J.; Hucker, S.; Braatz, U.; Graessner, H.; Berg, D.; Manoli, Y.; Serrano, J.A.; ET AL. Continuous leg dyskinesia assessment in Parkinson's disease -clinical validity and ecological effect. Park. Relat. Disord. 2016, 26, 41–46.

32 Gossec, L.; Guyard, F.; Leroy, D.; Lafargue, T.; Seiler, M.; Jacquemin, C.; Molto, A.; Sellam, J.; Foltz, V.; Gandjbakhch, F.; ET AL. Detection of Flares by Decrease in Physical Activity, Collected Using Wearable Activity Trackers in Rheumatoid Arthritis or Axial Spondyloarthritis: An Application of Machine Learning Analyses in Rheumatology. Arthritis Care Res. 2019, 71, 1336–1343.

33 Bai, R.; Xiao, L.; Guo, Y.; Zhu, X.; Li, N.; Wang, Y.; Chen, Q.; Feng, L.; Wang, Y.; Yu, X.; ET AL. Tracking and monitoring mood stability of patients with major depressive disorder by machine learning models using passive digital data: Prospective naturalistic multicenter study. JMIR MHEALTH Uhealth 2021, 9, e24365.

34 Schwab, P.; Karlen, W. A Deep Learning Approach to Diagnosing Multiple Sclerosis from Smartphone Data. IEEE J. Biomed. Health Inform. 2021, 25, 1284–1291.

35 Aghanavesi, S. Smartphone-Based Parkinson's Disease Symptom Assessment. Licentiate Dissertation, Dalarna University, Falun, Sweden, 2017.

36 Maleki, G.; Zhuparris, A.; Koopmans, I.; Doll, R.J.; Voet, N.; Cohen, A.; van Brummelen, E.; Groeneveld, G.J.; De Maeyer, J. Objective Monitoring of Facioscapulohumeral Dystrophy During Clinical Trials Using a Smartphone App and Wearables: Observational Study. JMIR Form. Res. 2022, 6, e31775.

37 Twose, J.; Licitra, G.; McConchie, H.; Lam, K.H.; Killestein, J. Early-warning signals for disease activity in patients diagnosed with multiple sclerosis based on keystroke dynamics. Chaos 2020, 30, 113133.

38 Cho, C.H.; Lee, T.; Kim, M.G.; In, H.P.; Kim, L.; Lee, H.J. Mood prediction of patients with mood disorders by machine learning using passive digital phenotypes based on the circadian rhythm: Prospective observational cohort study. J. Med. Internet Res.2019, 21, e11029.

39 Little, R.J.A.; Rubin, D.B. Complete-Case and Available-Case Analysis, Including Weighting Methods; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 2014; pp. 41–58.

40 Demissie, S.; LaValley, M.P.; Horton, N.J.; Glynn, R.J.; Cupples, L.A. Bias due to missing exposure

data using complete-case analysis in the proportional hazards regression model. Stat. Med. 2003, 22, 545–557.
41 Enders, C.K.; London, N.Y. Applied Missing Data Analysis; Guilford Press: New York, NY, USA, 2010.
42 Zhang, Y.; Folarin, A.A. Predicting Depressive Symptom Severity Through Individuals' Nearby Bluetooth Device Count Data Collected by Mobile Phones: Preliminary Longitudinal Study. Available online: https://www.ncbi.nlm.nih.gov/pmc/articles/ PMC8367113/ (accessed on 5 January 2023).
43 Creagh, A.P.; Dondelinger, F.; Lipsmeier, F.; Lindemann, M.; De Vos, M. Longitudinal Trend Monitoring of Multiple Sclerosis Ambulation using Smartphones. IEEE Open J. Eng. Med. Biol. 2022, 3, 202–210.
44 Wu, C.-T.; Li, G.-H.; Huang, C.-T.; Cheng, Y.-C.; Chen, C.-H.; Chien, J.-Y.; Kuo, P.-H.; Kuo, L.-C.; Lai, F. Acute exacerbation of a chronic obstructive pulmonary disease prediction system using wearable device data, machine learning, and deep learning:Development and cohort study. JMIR mHealth Uhealth 2021, 9, e22591.
45 Jakobsen, P.; Garcia-Ceja, E.; Riegler, M.; Stabell, L.A.; Nordgreen, T.; Torresen, J.; Fasmer, O.B.; Oedegaard, K.J. Applying machine learning in motor activity time series of depressed bipolar and unipolar patients compared to healthy controls. PLoS ONE 2020, 15, e0231995.
46 Lekkas, D.; Jacobson, N.C. Using artificial intelligence and longitudinal location data to differentiate persons who develop posttraumatic stress disorder following childhood trauma. Sci. Rep. 2021, 11, 10303.
47 Richman, M.B.; Trafalis, T.B.; Adrianto, I. Missing data imputation through machine learning algorithms. In Artificial Intelligence Methods in the Environmental Sciences; Springer: Dordrecht, The Netherlands, 2009; pp. 153–169.
48 Jerez, J.M.; Molina, I.; García-Laencina, P.J.; Alba, E.; Ribelles, N.; Martín, M.; Franco, L. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artif. Intell. Med. 2010, 50, 105–115.
49 Lakshminarayan, K.; Harp, S.A.; Goldman, R.P.; Samad, T. Imputation of Missing Data Using Machine Learning Techniques.In KDD Proceedings 1996; AAAI Press: Palo Alto, CA, USA, 1996; Volume 96.
50 Aggarwal, C.C. Data Mining; Springer International Publishing: Cham, Switzerland, 2015.
51 Ledolter, J.; Kardon, R.H. Does Testing More Frequently Shorten the Time to Detect Disease Progression? Transl. Vis. Sci. Technol. 2017, 6, 1.
52 Bazgir, O.; Habibi, S.A.H.; Palma, L.; Pierleoni, P.; Nafees, S. A classification system for assessment and home monitoring of tremor in patients with Parkinson's disease. J. Med. Signals Sens. 2018, 8, 65–72.
53 Williamson, J.R.; Telfer, B.; Mullany, R.; Friedl, K.E. Detecting Parkinson's Disease from Wrist-Worn Accelerometry in the U.K. Biobank. Sensors 2021, 21, 2047.
54 Buda, T.S.; Khwaja, M.; Matic, A. Outliers in Smartphone Sensor Data Reveal Outliers in Daily Happiness. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 2021, 5, 1–19.
55 Buda, T.S.; Caglayan, B.; Assem, H. DeepAD: A generic framework based on deep learning for time series anomaly detection. In Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer: Berlin/Heidelberg, Germany, 2018; pp. 577–588.
56 Arora, S.; Venkataraman, V.; Zhan, A.; Donohue, S.; Biglan, K.; Dorsey, E.; Little, M. Detecting and monitoring the symptoms of Parkinson's disease using smartphones: A pilot study. Park. Relat. Disord. 2015, 21, 650–653.
57 Guyon, I.; Elisseeff, A. An Introduction to Feature Extraction. In Feature Extraction; Springer: Berlin/ Heidelberg, Germany, 2006; pp. 1–25.
58 Raju, V.N.G.; Lakshmi, K.P.; Jain, V.M.; Kalidindi, A.; Padma, V. Study the Influence of Normalization/ Transformation process on the Accuracy of Supervised Classification. In Proceedings of the 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 20–22 August 2022; pp. 729–735.
59 Dara, S.; Tumma, P. Feature Extraction by Using Deep Learning: A Survey. In Proceedings of the 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 29–31 March 2018; pp. 1795–1801.
60 Tizzano, G.R.; Spezialetti, M.; Rossi, S. A Deep Learning Approach for Mood Recognition from Wearable Data. In Proceedings of the IEEE Medical Measurements and Applications, MeMeA 2020—Conference Proceedings, Bari, Italy, 1 June–1 July 2020; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2020.
61 Garcia-Ceja, E.; Riegler, M.; Jakobsen, P.; Torresen, J.; Nordgreen, T.; Oedegaard, K.J.; Fasmer, O.B. Motor Activity Based Classification of Depression in Unipolar and Bipolar Patients. In Proceedings of the 2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS), Karlstad, Sweden, 18–21 June 2018; pp. 316–321.
62 Liu, H. Feature Engineering for Machine Learning and Data Analytics. In Feature Engineering for Machine Learning and Data Analytics; Taylor & Francis Group: Boca Raton, FL, USA, 2018.
63 Nargesian, F.; Samulowitz, H.; Khurana, U.; Khalil, E.B.; Turaga, D. Learning feature engineering for classification. IJCAI Int. Jt. Conf. Artif. Intell. 2017, 2529–2535.
64 Kuhn, M.; Johnson, K. Feature Engineering and Selection: A Practical Approach for Predictive Models; Chapman and Hall/CRC: Boca Raton, FL, USA, 2019.
65 Ronao, C.A.; Cho, S.-B. Human activity recognition with smartphone sensors using deep learning neural networks. Expert Syst. Appl. 2016, 59, 235–244.
66 Nweke, H.F.; Teh, Y.W.; Al-garadi, M.A.; Alo, U.R. Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges. Expert Syst. Appl. 2018, 105, 233–261.
67 Zdravevski, E.; Lameski, P.; Trajkovik, V.; Kulakov, A.; Chorbev, I.; Goleva, R.; Pombo, N.; Garcia, N. Improving Activity Recognition Accuracy in Ambient-Assisted Living Systems by Automated Feature Engineering. IEEE Access 2017, 5, 5262–5280.
68 McGinnis, R.S.; Mahadevan, N.; Moon, Y.; Seagers, K.; Sheth, N.; Wright, J.A., Jr.; Dicristofaro, S.; Silva, I.; Jortberg, E.; Ceruolo, M.; et al. A machine learning approach for gait speed estimation using skin-mounted wearable sensors: From healthy controls to individuals with multiple sclerosis. PLoS ONE 2017, 12, e0178366.
69 Maxhuni, A.; Muñoz-Meléndez, A.; Osmani, V.; Perez, H.; Mayora, O.; Morales, E.F. Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients. Pervasive Mob. Comput. 2016, 31, 50–66.
70 Yamakawa, T.; Miyajima, M.; Fujiwara, K.; Kano, M.; Suzuki, Y.; Watanabe, Y.; Watanabe, S.; Hoshida, T.; Inaji, M.; Maehara, T. Wearable epileptic seizure prediction system with machine-learning-based anomaly detection of heart rate variability. Sensors 2020, 20, 3987.
71 Fuchs, C.; Nobile, M.S.; Zamora, G.; Degeneffe, A.; Kubben, P.; Kaymak, U. Tremor assessment using smartphone sensor data and fuzzy reasoning. BMC Bioinform. 2021, 22, 57.
72 Aich, S.; Pradhan, P.M.; Park, J.; Sethi, N.; Vathsa, V.S.S.; Kim, H.C. A validation study of freezing of gait (FOG) detection and machine-learning-based FOG prediction using estimated gait characteristics with a wearable accelerometer. Sensors 2018, 18, 3287.
73 Rodríguez-Martín, D.; Samà, A.; Pérez-López, C.; Català, A.; Arostegui, J.M.M.; Cabestany, J.; Bayés, À.; Alcaine, S.; Mestre, B.; Prats, A.; et al. Home detection of freezing of gait using Support Vector Machines through a single waist-worn triaxial accelerometer. PLoS ONE 2017, 12, e0171764.
74 Supratak, A.; Datta, G.; Gafson, A.R.; Nicholas, R.; Guo, Y.; Matthews, P.M. Remote monitoring in the home validates clinical gait measures for multiple sclerosis. Front. Neurol. 2018, 9, 561.
75 Bro, R.; Smilde, A.K. Principal component analysis. Anal. Methods 2014, 6, 2812–2831.
76 Kim, J.; Lim, J. A Deep Neural Network-Based Method for Prediction of Dementia Using Big Data. Int. J. Environ. Res. Public Health 2021, 18, 5386.
77 Clustering. In Principles of Data Mining; Springer: London, UK, 2007; pp. 221–238.
78 Arabie, P.; Hubert, L.J. An Overview of Combinatorial Data Analysis. In Clustering and Classification; World Scientific: Singapore, 1996; pp. 5–63.
79 Lu, J.; Shang, C.; Yue, C.; Morillo, R.; Ware, S.; Kamath, J.; Bamis, A.; Russell, A.; Wang, B.; Bi, J. Joint Modeling of Heterogeneous Sensing Data for Depression Assessment via Multi-task Learning. In Proceedings of the Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies; Association for Computing Machinery: New York, NY, USA, 2018; Volume 2, pp. 1–21.
80 Sabatelli, M.; Osmani, V.; Mayora, O.; Gruenerbl, A.; Lukowicz, P. Correlation of significant places with self-reported state of bipolar disorder patients. In Proceedings of the 2014 4th International Conference on Wireless Mobile Communication and Healthcare-Transforming Healthcare through Innovations in Mobile and Wireless Technologies (MOBIHEALTH), Athens, Greece, 3–5 November 2014; pp. 116–119.
81 Faurholt-Jepsen, M.; Busk, J.; Vinberg, M.; Christensen, E.M.; HelgaÞórarinsdóttir; Frost, M.; Bardram, J.E.; Kessing, L.V. Daily mobility patterns

in patients with bipolar disorder and healthy individuals. J. Affect. Disord. 2021, 278, 413–422.
82 Miotto, R.; Wang, F.; Wang, S.; Jiang, X.; Dudley, J.T. Deep learning for healthcare: Review, opportunities and challenges. Brief. Bioinform. 2018, 19, 1236–1246.
83 Marx, V. The big challenges of big data. Nature 2013, 498, 255–260.
84 Li, Y.; Ding, L.; Gao, X. On the decision boundary of deep neural networks. arXiv 2018, arXiv:1808.05385.
85 Juen, J.; Cheng, Q.; Schatz, B. A Natural Walking Monitor for Pulmonary Patients Using Mobile Phones. IEEE J. Biomed. Health Inform. 2015, 19, 1399–1405.
86 Cole, B.T.; Roy, S.H.; De Luca, C.J.; Nawab, S.H. Dynamical learning and tracking of tremor and dyskinesia from wearable sensors. IEEE Trans. Neural Syst. Rehabil. Eng. 2014, 22, 982–991.
87 Peraza, L.R.; Kinnunen, K.M.; McNaney, R.; Craddock, I.J.; Whone, A.L.; Morgan, C.; Joules, R.; Wolz, R. An automatic gait analysis pipeline for wearable sensors: A pilot study in parkinson's disease. Sensors 2021, 21, 8286.
88 Saeys, Y.; Abeel, T.; Van De Peer, Y. Robust feature selection using ensemble feature selection techniques. In Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer: Berlin/Heidelberg, Germany, 2008; pp. 313–325.
89 Cai, J.; Luo, J.; Wang, S.; Yang, S. Feature selection in machine learning: A new perspective. Neurocomputing 2018, 300, 70–79.
90 Jabar, H.; Khan, R.Z. Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study). Comput. Sci. Commun. Instrum. Devices 2015, 70, 163–172.
91 Hall, M.A. Correlation-based Feature Selection for Machine Learning. Ph.D. Thesis, The University of Waikato, Hamilton, New Zealand, 1999.
92 Hall, M.A.; Smith, L.A. Feature selection for machine learning: Comparing a correlation-based filter approach to the wrapper. In Proceedings of the FLAIRS Conference 1999, Orlando, FL, USA, 1–5 May 1999; Volume 1999, pp. 235–239.
93 Galperin, I.; Hillel, I.; Del Din, S.; Bekkers, E.M.; Nieuwboer, A.; Abbruzzese, G.; Avanzino, L.; Nieuwhof, F.; Bloem, B.R.; Rochester, L.; ET AL. Associations between daily-living physical activity and laboratory-based assessments of motor severity in patients with falls and Parkinson's disease. Park. Relat. Disord. 2019, 62, 85–90.
94 Dong, C.; Ye, T.; Long, X.; Aarts, R.M.; van Dijk, J.P.; Shang, C.; Liao, X.; Chen, W.; Lai, W.; Chen, L.; ET AL. A Two-Layer Ensemble Method for Detecting Epileptic Seizures Using a Self-Annotation Bracelet with Motor Sensors. IEEE Trans. Instrum. Meas. 2022, 71, 4005013.
95 Creagh, A.P.; Simillion, C.; Bourke, A.K.; Scotland, A.; Lipsmeier, F.; Bernasconi, C.; van Beek, J.; Baker, M.; Gossens, C.; Lindemann, M.; ET AL. Smartphone-and Smartwatch-Based Remote Characterisation of Ambulation in Multiple Sclerosis during the Two-Minute Walk Test. IEEE J. Biomed. Health Inform. 2021, 25, 838–849.
96 Chen, O.Y.; Lipsmeier, F.; Phan, H.; Prince, J.; Taylor, K.I.; Gossens, C.; Lindemann, M.; de Vos, M. Building a Machine-Learning Framework to Remotely Assess Parkinson's Disease Using Smartphones. IEEE Trans. Biomed. Eng. 2020, 67, 3491–3500.
97 Steyerberg, E.W.; Eijkemans, M.J.C.; Habbema, J.D.F. Stepwise selection in small data sets: A simulation study of bias in logistic regression analysis. J. Clin. Epidemiol. 1999, 52, 935–942.
98 Austin, P.C.; Tu, J.V. Bootstrap Methods for Developing Predictive Models. Am. Stat. 2004, 58, 131–137.
99 Zimmerman, D.W. Power Functions of the Test and Mann-Whitney Test Under Violation of Parametric Assumptions. Percept. Mot. Skills 1985, 61, 467–470.
100 Urbanowicz, R.J.; Meeker, M.; la Cava, W.; Olson, R.S.; Moore, J.H. Relief-based feature selection: Introduction and review. J. Biomed. Inform. 2018, 85, 189–203.
101 Kira, K.; Rendell, L.A. A Practical Approach to Feature Selection; Elsevier: Amsterdam, The Netherlands, 1992; pp. 249–256.
102 Verma, N.K.; Salour, A. Feature selection. Stud. Syst. Decis. Control 2020, 256, 175–200.
103 Yaman, O.; Ertam, F.; Tuncer, T. Automated Parkinson's disease recognition based on statistical pooling method using acoustic features. Med. Hypotheses 2020, 135, 109483.
104 Rodriguez-Molinero, A.; Samà, A.; Pérez-Martínez, D.A.; López, C.P.; Romagosa, J.; Bayes, A.; Sanz, P.; Calopa, M.; Gálvez-Barrón, C.; De Mingo, E.; ET AL. Validation of a portable device for mapping motor and gait disturbances in Parkinson's disease. JMIR mHealth Uhealth 2015, 3, e9.
105 Chandrashekar, G.; Sahin, F. A survey on feature selection methods. Comput. Electr. Eng. 2014, 40, 16–28.
106 Goldsmith, J.; Bobb, J.; Crainiceanu, C.M.; Caffo, B.; Reich, D. Penalized functional regression. J. Comput. Graph. Stat. 2011, 20, 830–851.
107 Prince, J.; Andreotti, F.; De Vos, M. Multi-Source Ensemble Learning for the Remote Prediction of Parkinson's Disease in the Presence of Source-Wise Missing Data. IEEE Trans. Biomed. Eng. 2019, 66, 1402–1411.
108 Motin, M.A.; Pah, N.D.; Raghav, S.; Kumar, D.K. Parkinson's Disease Detection Using Smartphone Recorded Phonemes in Real World Conditions. IEEE Access 2022, 10, 97600–97609.
109 Cakmak, A.S.; Alday, E.A.P.; Da Poian, G.; Rad, A.B.; Metzler, T.J.; Neylan, T.C.; House, S.L.; Beaudoin, F.L.; An, X.; Stevens, J.S.; ET AL. Classification and Prediction of Post-Trauma Outcomes Related to PTSD Using Circadian Rhythm Changes Measured via Wrist-Worn Research Watch in a Large Longitudinal Cohort. IEEE J. Biomed. Health Inform. 2021, 25, 2866–2876.
110 Tracy, J.M.; Özkanca, Y.; Atkins, D.C.; Ghomi, R.H. Investigating voice as a biomarker: Deep phenotyping methods for early detection of Parkinson's disease. J. Biomed. Inform. 2020, 104, 103362.
111 Abdulhafedh, A. Comparison between Common Statistical Modeling Techniques Used in Research, Including: Discriminant Analysis vs Logistic Regression, Ridge Regression vs LASSO, and Decision Tree vs Random Forest. OAlib 2022, 9, 1–19. 112. Sánchez-Maroño, N.; Alonso-Betanzos, A.; Tombilla-Sanromán, M. Filter methods for feature selection—A comparative study. In Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer: Berlin/Heidelberg, Germany, 2007; pp. 178–187.
113 Porter, B.W.; Bareiss, R.; Holte, R.C. Concept learning and heuristic classification in weak-theory domains. Artif. Intell. 1990, 45, 229–263.
114 Wu, C.-T.; Wang, S.-M.; Su, Y.-E.; Hsieh, T.-T.; Chen, P.-C.; Cheng, Y.-C.; Tseng, T.-W.; Chang, W.-S.; Su, C.-S.; Kuo, L.-C.; ET AL. A Precision Health Service for Chronic Diseases: Development and Cohort Study Using Wearable Device, Machine Learning, and Deep Learning. IEEE J. Transl. Eng. Health Med. 2022, 10, 2700414.
115 de Lima, A.L.S.; Evers, L.J.; Hahn, T.; de Vries, N.M.; Daeschler, M.; Boroojerdi, B.; Terricabras, D.; Little, M.A.; Bloem, B.R.; Faber, M.J. Impact of motor fluctuations on real-life gait in Parkinson's patients. Gait Posture 2018, 62, 388–394.
116 Pulliam, C.; Eichenseer, S.; Goetz, C.; Waln, O.; Hunter, C.; Jankovic, J.; Vaillancourt, D.; Giuffrida, J.; Heldman, D. Continuous in-home monitoring of essential tremor. Park. Relat. Disord. 2014, 20, 37–40.
117 Goni, M.; Eickhoff, S.B.; Far, M.S.; Patil, K.R.; Dukart, J. Smartphone-Based Digital Biomarkers for Parkinson's Disease in a Remotely-Administered Setting. IEEE Access 2022, 10, 28361–28384.
118 Livingston, E.; Cao, J.; Dimick, J.B. Tread carefully with stepwise regression. Arch. Surg. 2010, 145, 1039–1040.
119 Li, F.; Yang, Y. Analysis of recursive feature elimination methods. In Proceedings of the the 28th ACM/SIGIR International Symposium on Information Retrieval 2005, Salvador, Brazil, 15–19 August 2005.
120 Kuhn, M.; Johnson, K.; Kuhn, M.; Johnson, K. An Introduction to Feature Selection. In Applied Predictive Modeling; Springer: New York, NY, USA, 2013; pp. 487–519.
121 Senturk, Z.K. Early diagnosis of Parkinson's disease using machine learning algorithms. Med. Hypotheses 2020, 138, 109603.
122 Zhang, X.D. Machine Learning. In A Matrix Algebra Approach to Artificial Intelligence; Springer: Singapore, 2020. 123. Russell, S.J.; Norvig, P. Artificial Intelligence: A Modern Approach, 4th ed.; Prentice Hall: Hoboken, NJ, USA, 2020.
124 Tinschert, P.; Rassouli, F.; Barata, F.; Steurer-Stey, C.; Fleisch, E.; Puhan, M.; Kowatsch, T.; Brutsche, M.H. Smartphone-Based Cough. Detection Predicts Asthma Control—Description of a Novel, Scalable Digital Biomarker; European Respiratory Society (ERS): Lausanne, Switzerland, 2020; p. 4569.
125 ZhuParris, A.; Kruizinga, M.D.; van Gent, M.; Dessing, E.; Exadaktylos, V.; Doll, R.J.; Stuurman, F.E.; Driessen, G.A.; Cohen, A.F. Development and Technical Validation of a Smartphone-Based Cry Detection Algorithm. Front. Pediatr. 2021, 9, 262.
126 Fatima, M.; Pasha, M. Survey of Machine Learning Algorithms for Disease Diagnostic. J. Intell. Learn. Syst. Appl. 2017, 9, 1–16.
127 Ensari, I.; Caceres, B.A.; Jackman, K.B.; Suero-Tejeda, N.; Shechter, A.; Odlum, M.L.; Bakken,

S. Digital phenotyping of sleep patterns among heterogenous samples of Latinx adults using unsupervised learning. Sleep. Med. 2021, 85, 211–220.
128 Ko, Y.-F.; Kuo, P.-H.; Wang, C.-F.; Chen, Y.-J.; Chuang, P.-C.; Li, S.-Z.; Chen, B.-W.; Yang, F.-C.; Lo, Y.-C.; Yang, Y.; ET AL. Quantification Analysis of Sleep Based on Smartwatch Sensors for Parkinson's Disease. Biosensors 2022, 12, 74.
129 Farhan, A.A.; Yue, C.; Morillo, R.; Ware, S.; Lu, J.; Bi, J.; Kamath, J.; Russell, A.; Bamis, A.; Wang, B. Behavior vs. introspection: Refining prediction of clinical depression via smartphone sensing data. In Proceedings of the 2016 IEEE Wireless Health (WH), Bethesda, MD, USA, 25–27 October 2016.
130 Derungs, A.; Schuster-Amft, C.; Amft, O. Longitudinal walking analysis in hemiparetic patients using wearable motion sensors: Is there convergence between body sides? Front. Bioeng. Biotechnol. 2018, 6, 57.
131 Freedman, D.A. Statistical Models. In Statistical Models: THeory and Practice; Cambridge University Press: Cambridge, UK, 2009. 132. Ahmed, S.T.; Basha, S.M.; Arumugam, S.R.; Kodabagi, M.M. Pattern Recognition: An Introduction, 1st ed.; MileStone Research Publications: Bengaluru, India, 2021.
133 Ruppert, D. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. J. Am. Stat. Assoc. 2004, 99, 567.
134 Opitz, D.; Maclin, R. Popular Ensemble Methods: An Empirical Study. J. Artif. Intell. Res. 1999, 11, 169–198. 135. Kosasi, S. Perancangan Prototipe Sistem Pemesanan Makanan dan Minuman Menggunakan Mobile Device. Indones. J. Netw. Secur. 2015, 1, 1–10.
136 Hastie, T.; Tibshirani, R.; Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction; Springer Science & Business Media: New York, NY, USA, 2013.
137 San-Segundo, R.; Zhang, A.; Cebulla, A.; Panev, S.; Tabor, G.; Stebbins, K.; Massa, R.E.; Whitford, A.; de la Torre, F.; Hodgins, J. Parkinson's disease tremor detection in the wild using wearable accelerometers. Sensors 2020, 20, 5817.
138 Ahmadi, M.N.; O'neil, M.E.; Baque, E.; Boyd, R.N.; Trost, S.G. Machine learning to quantify physical activity in children with cerebral palsy: Comparison of group, group-personalized, and fully-personalized activity classification models. Sensors 2020, 20, 3976.
139 Faurholt-Jepsen, M.; Busk, J.; Frost, M.; Vinberg, M.; Christensen, E.M.; Winther, O.; Bardram, J.E.; Kessing, L.V. Voice analysis as an objective state marker in bipolar disorder. Transl. Psychiatry 2016, 6, e856.
140 Jacobson, N.C.; Lekkas, D.; Huang, R.; Thomas, N. Deep learning paired with wearable passive sensing data predicts deterioration in anxiety disorder symptoms across 17–18 years. J. Affect. Disord. 2021, 282, 104–111.
141 Hastie, T.; Tibshirani, R.; Friedman, J. Statistics the Elements of Statistical Learning. Math. Intell. 2009, 27, 83–85.
142 Patle, A.; Chouhan, D.S. SVM kernel functions for classification. In Proceedings of the 2013 International Conference on Advances in Technology and Engineering, ICATE 2013, Mumbai, India, 23–25 January 2013.
143 Kim, H.S.; Kim, S.Y.; Kim, Y.H.; Park, K.S. A smartphone-based automatic diagnosis system for facial nerve palsy. Sensors 2015, 15, 26756–26768.
144 Luca, S.; Karsmakers, P.; Cuppens, K.; Croonenborghs, T.; Van de Vel, A.; Ceulemans, B.; Lagae, L.; Van Huffel, S.; Vanrumste, B. Detecting rare events using extreme value statistics applied to epileptic convulsions in children. Artif. Intell. Med. 2014, 60, 89–96.
145 Ghoraani, B.; Hssayeni, M.D.; Bruack, M.M.; Jimenez-Shahed, J. Multilevel Features for Sensor-Based Assessment of Motor Fluctuation in Parkinson's Disease Subjects. IEEE J. Biomed. Health Inform. 2020, 24, 1284–1295.
146 Kramer, O. K-Nearest Neighbors. In Dimensionality Reduction with Unsupervised Nearest Neighbors. Intelligent Systems Reference. Library; Springer: Berlin/Heidelberg, Germany, 2013; Volume 51.
147 Jeon, H.; Lee, W.; Park, H.; Lee, H.J.; Kim, S.K.; Kim, H.B.; Jeon, B.; Park, K.S. Automatic classification of tremor severity in Parkinson's disease using awearable device. Sensors 2017, 17, 2067.
148 Grunerbl, A.; Muaremi, A.; Osmani, V.; Bahle, G.; Ohler, S.; Troster, G.; Mayora, O.; Haring, C.; Lukowicz, P. Smartphone-based recognition of states and state changes in bipolar disorder patients. IEEE J. Biomed. Health Inform. 2015, 19, 140–148.
149 Pranckevičius, T.; Marcinkevičius, V. Comparison of Naive Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification. Balt. J. Mod. Comput. 2017, 5, 221–232.
150 Worster, A.; Fan, J.; Ismaila, A. Understanding linear and logistic regression analyses. Can. J. Emerg. Med. 2007, 9, 111–113.
151 Morrow-Howell, N. The M word: Multicollinearity in multiple regression. Soc. Work. Res. 1994, 18, 247–251.
152 Schwenk, M.; Hauer, K.; Zieschang, T.; Englert, S.; Mohler, J.; Najafi, B. Sensor-derived physical activity parameters can predict future falls in people with dementia. Gerontology 2014, 60, 483–492.
153 Lecun, Y.; Bengio, Y.; Hinton, G. Deep learning. Nature 2015, 521, 436–444.
154 Tu, J.V. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. J. Clin. Epidemiol. 1996, 49, 1225–1231.
155 Mudiyanselage, T.K.B.; Xiao, X.; Zhang, Y.; Pan, Y. Deep Fuzzy Neural Networks for Biomarker Selection for Accurate Cancer Detection. IEEE Trans. Fuzzy Syst. 2020, 28, 3219–3228.
156 Yagin, F.H.; Cicek, I.B.; Alkhateeb, A.; Yagin, B.; Colak, C.; Azzeh, M.; Akbulut, S. Explainable artificial intelligence model for identifying COVID-19 gene biomarkers. Comput. Biol. Med. 2023, 154, 106619.
157 Wang, Y.; Lucas, M.; Furst, J.; Fawzi, A.A.; Raicu, D. Explainable Deep Learning for Biomarker Classification of OCT Images. In Proceedings of the 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE), Cincinnati, OH, USA, 26–28 October 2020; pp. 204–210.
158 Fisher, J.M.; Hammerla, N.Y.; Ploetz, T.; Andras, P.; Rochester, L.; Walker, R.W. Unsupervised home monitoring of Parkinson's disease motor symptoms using body-worn accelerometers. Park. Relat. Disord. 2016, 33, 44–50.
159 Frogner, J.I.; Noori, F.M.; Halvorsen, P.; Hicks, S.A.; Garcia-Ceja, E.; Torresen, J.; Riegler, M.A. One-dimensional convolutional neural networks on motor activity measurements in detection of depression. In Proceedings of the HealthMedia 2019—Proceedings of the 4th International Workshop on Multimedia for Personal Health and Health Care, Co-Located with MM 2019, Nice, France, 21–25 October 2019; pp. 9–15.
160 Meisel, C.; el Atrache, R.; Jackson, M.; Schubach, S.; Ufongene, C.; Loddenkemper, T. Machine learning from wristband sensor data for wearable, noninvasive seizure forecasting. Epilepsia 2020, 61, 2653–2666.
161 Matarazzo, M.; Arroyo-Gallego, T.; Montero, P.; Puertas-Martín, V.; Butterworth, I.; Mendoza, C.S.; Ledesma-Carbayo, M.J.; Catalán, M.J.; Molina, J.A.; Bermejo-Pareja, F.; ET AL. Remote Monitoring of Treatment Response in Parkinson's Disease: The Habit of Typing on a Computer. Mov. Disord. 2019, 34, 1488–1495.
162 Weiss, K.; Khoshgoftaar, T.M.; Background, D.W. A survey of transfer learning. J. Big Data 2016, 3, 1345–1459.
163 Kamishima, T.; Hamasaki, M.; Akaho, S. TrBagg: A Simple Transfer Learning Method and its Application to Personalization in Collaborative Tagging. In Proceedings of the 2009 Ninth IEEE International Conference on Data Mining, Miami, FL, USA, 6–9 December 2009; pp. 219–228.
164 Fu, Z.; He, X.; Wang, E.; Huo, J.; Huang, J.; Wu, D. Personalized Human Activity Recognition Based on Integrated Wearable Sensor and Transfer Learning. Sensors 2021, 21, 885.
165 Chen, Y.; Qin, X.; Wang, J.; Yu, C.; Gao, W. FedHealth: A Federated Transfer Learning Framework for Wearable Healthcare. IEEE Intell. Syst. 2020, 35, 83–93.
166 Goschenhofer, J.; Pfister, F.M.J.; Yuksel, K.A.; Bischl, B.; Fietzek, U.; Thomas, J. Wearable-Based Parkinson's Disease Severity Monitoring Using Deep Learning. In Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer: Berlin/Heidelberg, Germany, 2020; Volume 11908 LNAI, pp. 400–415.
167 Hssayeni, M.D.; Jimenez-Shahed, J.; Burack, M.A.; Ghoraani, B. Ensemble deep model for continuous estimation of Unified Parkinson's Disease Rating Scale III. Biomed. Eng. Online 2021, 20, 1–20.
168 Zhang, Y.; Yang, Q. Special Topic: Machine Learning An overview of multi-task learning. Natl. Sci. Rev. 2018, 5, 30–43.
169 Lee, G.; Yang, E.; Hwang, S. Asymmetric multi-task learning based on task relatedness and loss. In Proceedings of the International Conference on Machine Learning 2016, New York, NY, USA, 19–24 June 2016; pp. 230–238.
170 Xin, W.; Bi, J.; Yu, S.; Sun, J.; Song, M. Multiplicative Multitask Feature Learning. J. Mach. Learn. Res. JMLR 2016, 17, 1–33.
171 Zhang, Z.; Jung, T.P.; Makeig, S.; Pi, Z.; Rao,

B.D. Spatiotemporal sparse Bayesian learning with applications to compressed sensing of multichannel physiological signals. IEEE Trans. Neural Syst. Rehabil. Eng. 2014, 22, 1186–1197.
172 Schneider, J.; Vlachos, M. Personalization of deep learning. In Data Science–Analytics and Applications: Proceedings of the 3rd International Data Science Conference–iDSC2020; Springer: Wiesbaden, Geramny, 2021; pp. 89–96.
173 Khademi, A.; El-Manzalawy, Y.; Buxton, O.M.; Honavar, V. Toward personalized sleep-wake prediction from actigraphy. In Proceedings of the 2018 IEEE EMBS International Conference on Biomedical and Health Informatics, BHI 2018, Vegas, NV, USA, 4–7 March 2018; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2018; pp. 414–417.
174 Kuhn, M.; Johnson, K. Applied Predictive Modeling; Springer: New York, NY, USA, 2013.
175 Pal, M. Random forest classifier for remote sensing classification. Int. J. Remote Sens. 2005, 26, 217–222.
176 Putin, E.; Mamoshina, P.; Aliper, A.; Korzinkin, M.; Moskalev, A.; Kolosov, A.; Ostrovskiy, A.; Cantor, C.; Vijg, J.; Zhavoronkov, A. Deep biomarkers of human aging: Application of deep neural networks to biomarker development. Aging 2016, 8, 1021–1033.
177 Yang, L.; Shami, A. On hyperparameter optimization of machine learning algorithms: Theory and practice. Neurocomputing 2020, 415, 295–316.
178 Waring, J.; Lindvall, C.; Umeton, R. Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. Artif. Intell. Med. 2020, 104, 101822.
179 Bergstra, J.; Ca, J.B.; Ca, Y.B. Random Search for Hyper-Parameter Optimization Yoshua Bengio. 2012. Available online: http://scikit-learn.sourceforge.net (accessed on 5 January 2023).
180 Beam, A.L.; Manrai, A.K.; Ghassemi, M. Challenges to the Reproducibility of Machine Learning Models in Health Care. JAMA 2020, 323, 305.
181 Ahlrichs, C.; Samà, A.; Lawo, M.; Cabestany, J.; Rodríguez-Martín, D.; Pérez-López, C.; Sweeney, D.; Quinlan, L.R.; Laighin, G.Ò.; Counihan, T.; ET AL. Detecting freezing of gait with a tri-axial accelerometer in Parkinson's disease patients. Med. Biol. Eng. Comput. 2016, 54, 223–233.
182 Rosenwein, T.; Dafna, E.; Tarasiuk, A.; Zigel, Y. Detection of Breathing Sounds during Sleep Using Non-Contact Audio Recordings; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2014.
183 Pérez-López, C.; Samà, A.; Rodríguez-Martín, D.; Moreno-Aróstegui, J.M.; Cabestany, J.; Bayes, A.; Mestre, B.; Alcaine, S.; Quispe, P.; Laighin, G.; ET AL. Dopaminergic-induced dyskinesia assessment based on a single belt-worn accelerometer. Artif. Intell. Med. 2016, 67, 47–56.
184 Bernad-Elazari, H.; Herman, T.; Mirelman, A.; Gazit, E.; Giladi, N.; Hausdorff, J.M. Objective characterization of daily living transitions in patients with Parkinson's disease using a single body-fixed sensor. J. Neurol. 2016, 263, 1544–1551.
185 Carvalho, D.V.; Pereira, E.M.; Cardoso, J.S. Machine Learning Interpretability: A Survey on Methods and Metrics. Electronics 2019, 8, 832.
186 Zhou, J.; Gandomi, A.H.; Chen, F.; Holzinger, A. Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics. Electronics 2021, 10, 593.
187 Hossin, M.; Sulaiman, M.N. A Review on Evaluation Metrics for Data Classification Evaluations. Int. J. Data Min. Knowl. Manag. Process 2015, 5, 1–11.
188 He, H.; Ma, Y. Imbalanced Learning; Wiley: Hoboken, NJ, USA, 2013.
189 Wan, S.; Liang, Y.; Zhang, Y.; Guizani, M. Deep Multi-Layer perceptron classifier for behavior analysis to estimate Parkinson's disease severity using smartphones. IEEE Access 2018, 6, 36825–36833.
190 Chicco, D.; Tötsch, N.; Jurman, G. The matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. BioData Min. 2021, 14, 1–22.
191 Jurman, G.; Riccadonna, S.; Furlanello, C. A Comparison of MCC and CEN Error Measures in Multi-Class Prediction. PLoS ONE 2012, 7, e41882.
192 Faurholt-Jepsen, M.; Busk, J.; HelgaÞórarinsdóttir; Frost, M.; Bardram, J.E.; Vinberg, M.; Kessing, L.V. Objective smartphone data as a potential diagnostic marker of bipolar disorder. Aust. N. Z. J. Psychiatry 2019, 53, 119–128.
193 Xu, Z.; Shen, D.; Nie, T.; Kou, Y. A hybrid sampling algorithm combining M-SMOTE and ENN based on Random forest for medical imbalanced data. J. Biomed. Inform. 2020, 107, 103465.
194 Zeng, M.; Zou, B.; Wei, F.; Liu, X.; Wang, L. Effective prediction of three common diseases by combining SMOTE with Tomek links technique for imbalanced medical data. In Proceedings of the 2016 IEEE International Conference of Online Analysis and Computing Science, ICOACS 2016, Chongqing, China, 28–29 May 2016; pp. 225–228.
195 Botchkarev, A. Performance Metrics (Error Measures) in Machine Learning Regression, Forecasting and Prognostics: Properties and Typology. Interdiscip. J. Inf. Knowl. Manag. 2018, 14, 45–76.
196 di Bucchianico, A. Coefficient of Determination. In Encyclopedia of Statistics in Quality and Reliability; Wiley: Hoboken, NJ, USA, 2007.
197 Piepho, H. A coefficient of determination (R2) for generalized linear mixed models. Biom. J. 2019, 61, 860–872.
198 Gelman, A.; Pardoe, I. Bayesian Measures of Explained Variance and Pooling in Multilevel (Hierarchical) Models. Technometrics 2006, 48, 241–251.
199 Hodson, T.O. Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not. Geosci. Model. Dev. 2022, 15, 5481–5487.
200 Mezzadri, G.; Laloë, T.; Mathy, F.; Reynaud-Bouret, P. Hold-out strategy for selecting learning models: Application to categorization subjected to presentation orders. J. Math. Psychol. 2022, 109, 102691.
201 Gholamiangonabadi, D.; Kiselov, N.; Grolinger, K. Deep Neural Networks for Human Activity Recognition with Wearable Sensors: Leave-One-Subject-Out Cross-Validation for Model Selection. IEEE Access 2020, 8, 133982–133994.
202 Little, M.A.; Varoquaux, G.; Saeb, S.; Lonini, L.; Jayaraman, A.; Mohr, D.C.; Kording, K.P. Using and understanding crossvalidation strategies. Perspectives on Saeb ET AL. Gigascience 2017, 6, 1–6.
203 Peterson, R.A.; Cavanaugh, J.E. Ordered quantile normalization: A semiparametric transformation built for the cross-validation era. J. Appl. Stat. 2020, 47, 2312–2327.
204 Zhang, Y.; Yang, Y. Cross-validation for selecting a model selection procedure. J. Econom. 2015, 187, 95–112.
205 Refaeilzadeh, P.; Tang, L.; Liu, H. Cross-Validation. In Encyclopedia of Database Systems; Springer: Berlin/Heidelberg, Germany, 2016; pp. 1–7.
206 Browne, M.W. Cross-validation methods. J. Math. Psychol. 2000, 44, 108–132.
207 Wainer, J.; Cawley, G. Nested cross-validation when selecting classifiers is overzealous for most practical applications. Expert Syst. Appl. 2021, 182, 115222.
208 Kohavi, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. 1995. Available online: http://robotics.stanford.edu/~ronnyk (accessed on 5 January 2023).
209 Vanwinckelen, G.; Blockeel, H. On estimating model accuracy with repeated cross-validation. In BeneLearn 2012: Proceedings of the 21st Belgian-Dutch Conference on Machine Learning; Benelearn 2012 Organization Committee: Ghent, Belgium, 2012; pp. 39–44.
210 Parvandeh, S.; Yeh, H.-W.; Paulus, M.P.; McKinney, B.A. Consensus Features Nested Cross-Validation. bioRxiv 2020.
211 Goble, C.; Cohen-Boulakia, S.; Soiland-Reyes, S.; Garijo, D.; Gil, Y.; Crusoe, M.; Peters, K.; Schober, D. Fair computational workflows. Data Intell. 2020, 2, 108–121.
212 Muehlematter, U.J.; Daniore, P.; Vokinger, K.N. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): A comparative analysis. Lancet Digit. Health 2021, 3, e195–e203.
213 Beckers, R.; Kwade, Z.; Zanca, F. The EU medical device regulation: Implications for artificial intelligence-based medical device software in medical physics. Phys. Med. 2021, 83, 1–8.
214 van Oirschot, J.; Ooms, G. Interpreting the EU Artificial Intelligence Act for the Health Sector; Health Action International: Amsterdam, The Netherlands, February 2022.
215 Agarwal, S.; LeFevre, A.; Lee, J.; L'engle, K.; Mehl, G.; Sinha, C.; Labrique, A. Guidelines for reporting of health interventions using mobile phones: Mobile health (MHEALTH) evidence reporting and assessment (mERA) checklist. BMJ 2016, 352, i1174.
216 Yang, J.; Soltan, A.A.S.; Clifton, D.A. Machine learning generalizability across healthcare settings: Insights from multi-site COVID-19 screening. NPJ Digit. Med. 2022, 5, 69.
217 Petersen, E.; Potdevin, Y.; Mohammadi, E.; Zidowitz, S.; Breyer, S.; Nowotka, D.; Henn, S.; Pechmann, L.; Leucker, M.; Rostalski, P.; ET AL. Responsible and Regulatory Conform Machine Learning for Medicine: A Survey of Challenges and Solutions. IEEE Access 2022, 10, 58375–58418.
218 FAIR Principles—GO FAIR. Available online: https://www.go-fair.org/fair-principles/ (accessed on 16 December 2021).
219 Fletcher, R.R.; Nakeshimana, A.; Olubeko, O. Addressing Fairness, Bias, and Appropriate Use

of Artificial Intelligence and Machine Learning in Global Health. Front. Artif. Intell. 2021, 3, 116.
220 Mei, J.; Desrosiers, C.; Frasnelli, J. Machine Learning for the Diagnosis of Parkinson's Disease: A Review of Literature. Front. Aging Neurosci. 2021, 13, 633752.
221 Chekroud, A.M.; Bondar, J.; Delgadillo, J.; Doherty, G.; Wasil, A.; Fokkema, M.; Cohen, Z.; Belgrave, D.; DeRubeis, R.; Iniesta, R.; ET AL. The promise of machine learning in predicting treatment outcomes in psychiatry. World Psychiatry 2021, 20, 154–170.
222 Kruizinga, M.D.; Zhuparris, A.; Dessing, E.; Krol, F.J.; Sprij, A.J.; Doll, R.; Stuurman, F.E.; Exadaktylos, V.; Driessen, G.J.A.; Cohen, A.F. Development and technical validation of a smartphone-based pediatric cough detection algorithm. Pediatr. Pulmonol. 2022, 57, 761–767.
223 Doshi-Velez, F.; Kim, B. Towards A Rigorous Science of Interpretable Machine Learning. arXiv 2017, arXiv:1702.08608.
224 Ignatiev, A. Towards Trustable Explainable AI. 2020. Available online: https://www.kaggle.com/uciml/zoo-animal-classification (accessed on 5 January 2023).
225 Walsh, I.; Fishman, D.; Garcia-Gasulla, D.; Titma, T.; Pollastri, G.; Capriotti, E.; Casadio, R.; Capella-Gutierrez, S.; Cirillo, D.; Del Conte, A.; ET AL. DOME: Recommendations for supervised machine learning validation in biology. Nat. Methods 2021, 18, 1122–1127.
226 Zippel, C.; Bohnet-Joschko, S. Rise of Clinical Studies in the Field of Machine Learning: A Review of Data Registered in ClinicalTrials.gov. Int. J. Environ. Res. Public Health 2021, 18, 5072.

TABLE 1 Representation of a standard machine learning pipeline.

Stage	Objective	Example
STUDY DESIGN	The ML pipeline is provided with a study objective in which the features and corresponding outputs are defined. The ML model aims to identify the associations between the features and outputs.	The study objective is to classify Parkinson's Disease patients and control groups using smartphone-based features.
DATA PREPROCESSING	Data preprocessing filters and transforms raw data to guarantee or enhance the ML training process.	To improve the model performance, one may identify and exclude any missing or outlier data.
FEATURE ENGINEERING AND SELECTION	Feature engineering uses raw data to create new features that are not readily available in the dataset. Feature selection selects the most relevant features for the model objective by removing redundant or noisy features. Together, the goal is to simplify and accelerate the computational process while also improving the model process. For deep learning methods, the concept of 'feature engineering' is typically embedded within the model architecture and training process, although substantial preprocessing steps may occur prior to that.	An interaction of two or more predictors (such as a ratio or product) or re-representation of a predictor are examples of feature engineering. Removing highly correlated or non-informative features are examples of feature selection. Note: The feature selection step can occur during model training
MODEL TRAINING AND VALIDATION	During training, the ML model(s) iterates through all the examples in the training dataset and optimizes the parameters of the mathematical function to minimize the prediction error. To evaluate the performance of the trained ML model, the predictions of an unseen test set are compared with a known ground truth label.	Cross-validation can be used to optimize and evaluate model performance. Classification models may be evaluated based on their prediction accuracy, sensitivity, and specificity, while regression models may be evaluated using variance explained (R2) and Mean Absolute Error.

TABLE 2 An overview of the keyword strategy used for this study.

Domain	Search String
TECHNOLOGY	((‘smartphone’[tiab] OR ‘wearable’[tiab] OR ‘remote + monitoring’[tiab] OR ‘home + monitoring’[tiab] OR ‘mobile + sensors’[tiab] OR ‘mobile + montoring’[tiab] OR ‘behavioral + sensing’[tiab] OR ‘geolocation’[tiab] OR ‘mHealth’[tiab] OR ‘passive + monitoring’[tiab] OR ‘digital + phenotype’[tiab] OR ‘digital + phenotyping’[tiab] OR ‘digital + biomarker’[tiab])
ANALYSIS	AND (‘machine + learning’[tiab] OR ‘deep + learning’[tiab] OR ‘random + forest’[tiab] OR ‘neural + network’[tiab] OR ‘time + series’[tiab] OR ‘regression’[tiab] OR ‘SVM’[tiab] OR ‘knn’[tiab] OR ‘dynamics + model’[tiab] OR ‘decision + tree’[tiab] OR ‘discriminant + analysis’[tiab] OR ‘feature + engineering’[tiab] OR ‘feature + selection’[tiab] OR ‘data + mining’[tiab] OR ‘model’[tiab] OR ‘classification’[tiab] OR ‘diagnostic’[tiab] OR ‘prognostic’[tiab] OR ‘symptom + severity’[tiab] OR ‘prediction’[tiab] OR ‘monitoring’[tiab])
POPULATION	AND (‘disease’[tiab] OR ‘disorder’[tiab] OR ‘diagnosis’[tiab] OR ‘prognosis’ OR ‘alzheimer’[tiab] OR ‘parkinson’[tiab] OR ‘Huntington’[tiab] OR ‘neurodegenerative’[tiab] OR ‘degenerative’ OR ‘tremor’[tiab] OR ‘bipolar’[tiab] OR ‘depression’[tiab] OR ‘manic’[tiab] OR ‘anxiety’[tiab] OR ‘vocal + biomarker’[tiab] OR ‘amyotrophic + lateral + sclerosis’[tiab] OR ‘central + nervous + system’[tiab] OR ‘symptom’[tiab] OR ‘psychosis’[tiab] OR ‘stroke’[tiab] OR ‘muscular dystrophy’[tiab] OR ‘Facioscapulohumeral Dystrophy’[tiab] OR ‘autoimmune’[tiab] OR ‘seizure’[tiab] OR ‘multiple + sclerosis’[tiab])
DATE	AND (‘2012/01/01’[PDAT]:’2022/12/31’[PDAT])
LANGUAGE	AND (English[lang])
EXCLUSION CRITERIA	NOT(‘animals’[tiab] OR ‘implant’[tiab] OR ‘hospital’[tiab] OR ‘caregiver’[tiab] OR ‘telemedicine’[tiab] OR ‘telerehabilitation’[tiab] OR ‘smartphone + addiction’[tiab] OR ‘nursing’[tiab] OR’screening’[tiab] OR ‘recruitment’[tiab] OR ‘diabetes’[tiab] OR ‘malaria’[tiab] OR ‘self-care’[tiab] OR ‘self-management’[tiab] OR ‘self-help’[tiab])
ARTICLE TYPE	AND (clinicalstudy[Filter] OR clinicaltrial[Filter] OR clinicaltrialphasei[Filter] OR clinicaltrialphaseii[Filter] OR clinicaltrialphaseiii[Filter] OR clinicaltrialphaseiv[Filter] OR controlledclinicaltrial[Filter] OR meta-analysis[Filter] observationalstudy[Filter] OR randomizedcontrolledtrial[Filter] OR systematicreview[Filter])

TABLE 3 Table of the inclusion and exclusion criteria used for study selection.

Category	Criteria
POPULATION	The study must be initiated by a research organization and not by the participants. The participants must have a clinical diagnosis that is affected by the CNS. Hence, studies that collected data from participants with no clinically confirmed diagnosis were not considered.
INTERVENTION	The study must include the use of smartphone or non-invasive wearables to remotely monitor and quantify passive biomarkers under free-living conditions.
COMPARATOR	A ground truth comparator for digital phenotyping such as clinical assessment, medical records, or self-reported outcomes.
OUTCOMES	A ML model that is used to classify a clinical label (such as a diagnosis, or clinical event), estimate symptom severity, or to detect treatment effects.
STUDY TYPE	The paper must be about a human-centered observational study (cohort or longitudinal) where the data were collected outside the clinic, lab, or hospital (free-living conditions). Hence, studies that use smartphones or wearables as a form of intervention or as screening tools are not of interest. The study must show if the ML models had ecological validity by validating the models using free-living data. The study has to have been written or translated into English and published within the last 10 years (2012 onwards).

TABLE 4 Clinical interpretations of common ML performance metrics.

Term	Equation	Objective
ACCURACY	$\frac{TP}{TP+TN}$	Out of all the predictions, how many predictions were correctly identified as positive or negative?
PRECISION	$\frac{2\times Precision\times Recall}{Precision+Recall}$	How many predictions were correctly labeled as patients out of all correctly classified patients and misclassified healthy controls?
SPECIFICITY	$\frac{1}{N}\sum Actual - Predicted$	How many predictions were correctly labeled as healthy controls out of all healthy controls? In other words, of all healthy controls, who were correctly identified as such?
RECALL/ SENSITIVITY	$\sqrt{\frac{\sum_{i=1}^{N}(Predicted_i - Actual_i)^2}{N}}$	Of all the patients, who were correctly classified/identified as such?
F1-SCORE	$\frac{2\times Precision\times Recall}{Precision+Recall}$	How many predictions were correctly labeled as patients (recall) and what was the accuracy with regards to correctly predicted patients (precision)?
MEAN SQUARE ERROR	$\frac{1}{N}\sum Actual - Predicted$	What is the absolute difference between the true scores and the predicted scores?
ROOT MEAN SQUARE ERROR	$\sqrt{\frac{\sum_{i=1}^{N}(Predicted_i - Actual_i)^2}{N}}$	What is the average difference between the true and the predicted scores (in the same unit of the true scores)?
R2	$1 - \frac{RSS}{TSS}$	What fraction of the variance in the data is captured by the model?

True Positive = TP, True Negative = TN, False Positives = FP, False Negatives = FN, Sum of Squares of Residuals = RSS, Total Sum of Squares = TSS, Number of Observations = N

FIGURE 1 Flow diagram illustrating the paper selection process for this review.

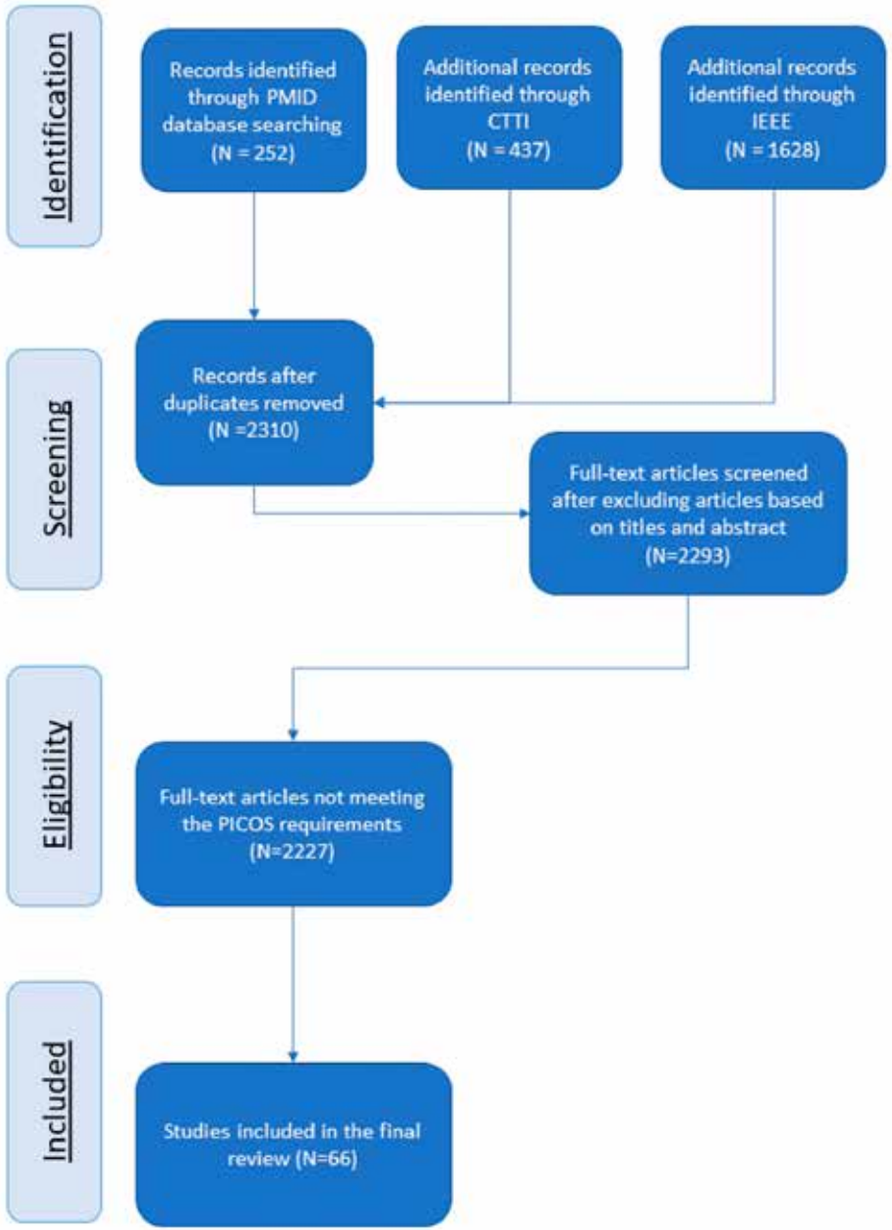


FIGURE 2 Clinical populations and the use of healthy controls in the selected studies.

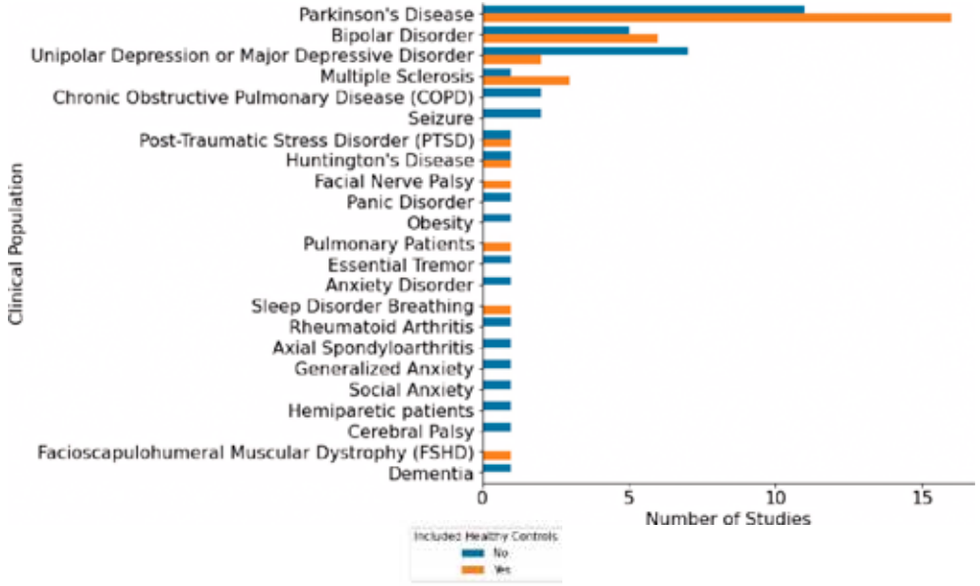


FIGURE 3 Sample sizes of clinical populations included in selected studies, with x-axis (sample size) presented on a logarithmic scale.

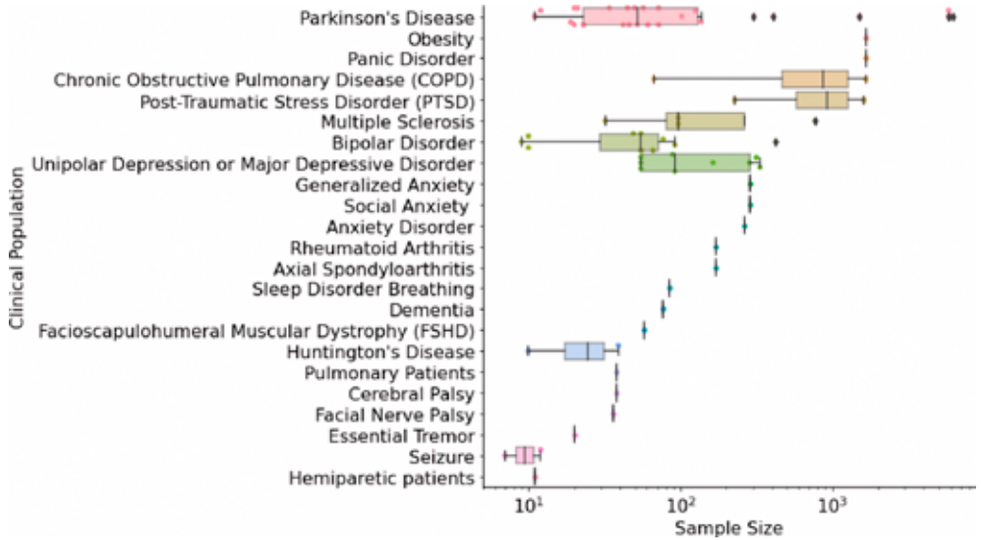


FIGURE 4 Machine learning algorithms and their respective objectives in the selected studies.

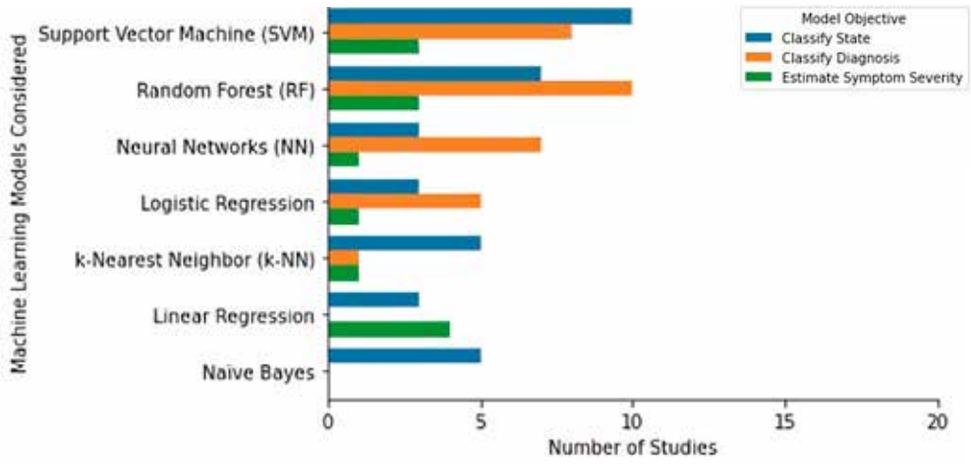
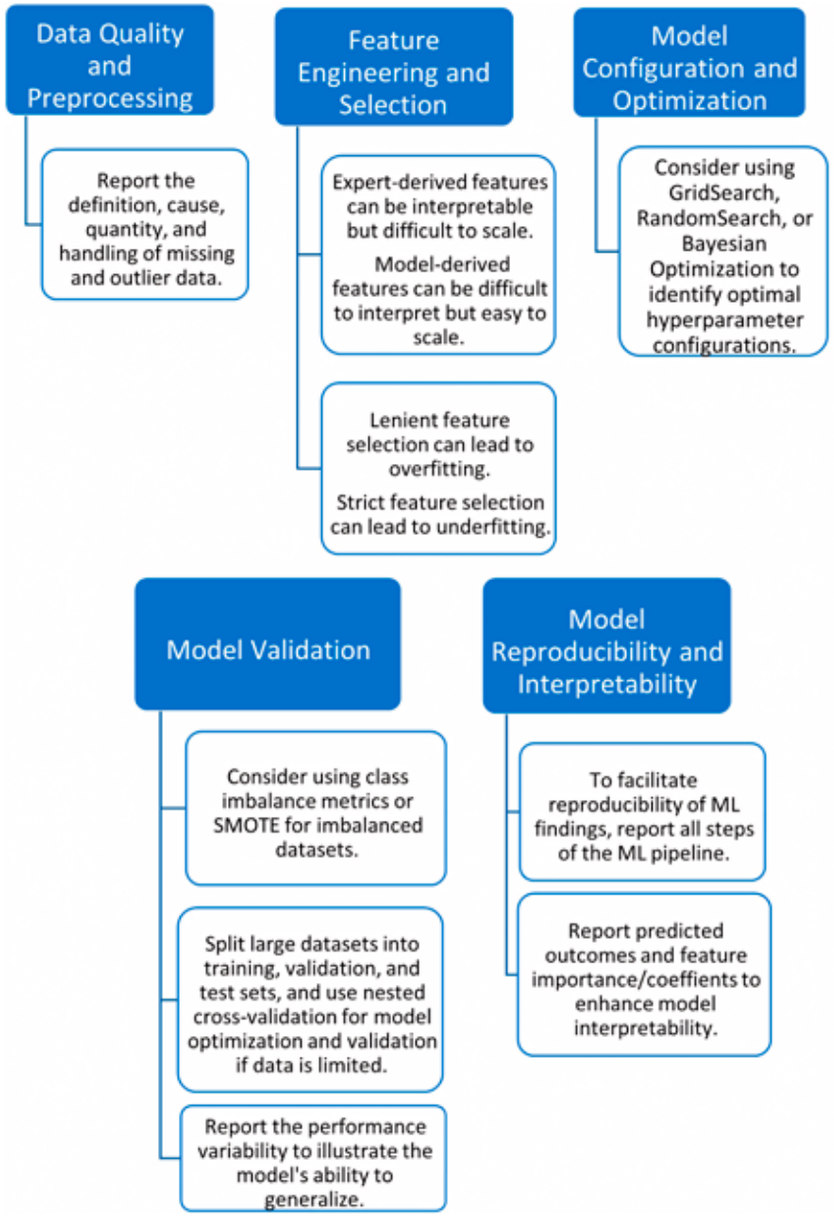


FIGURE 5 General recommendations for building an effective and reproducible ML pipeline.



PART II

CLASSIFICATION OF DIAGNOSIS

CHAPTER 3

Objective monitoring of facioscapulohumeral dystrophy during clinical trials using a smartphone app and wearables: observational study

Ghobad Maleki,[1,2*] BSc; Ahnjili Zhuparris,[1,2*] MSc; Ingrid Koopmans,[1] MSc; Robert J Doll[1], PhD; Nicoline Voet,[3,4] MD, PhD; Adam Cohen,[1] MD, PhD; Emilie van Brummelen[1], PhD; Geert Jan Groeneveld,[1,2] MD, PhD; Joris De Maeyer,[5] PhD **these authors contributed equally*

JMIR Form Res. 2022;6:1-13. doi:10.2196/31775

1 Centre for Human Drug Research, Leiden, Netherlands
2 Leiden University Medical Center, Leiden, Netherlands
3 Radboud University Medical Center, Nijmegen, Netherlands
4 Klimmendaal, Arnhem, Netherlands
5 Facio Therapies, Leiden, Netherlands

Abstract

Background: Facioscapulohumeral dystrophy (FSHD) is a progressive muscle dystrophy disorder leading to significant disability. Currently, FSHD symptom severity is assessed by clinical assessments such as the FSHD clinical score and the Timed Up-and-Go test. These assessments are limited in their ability to capture changes continuously and the full impact of the disease on patients' quality of life. Real-world data related to physical activity, sleep, and social behavior could potentially provide additional insight into the impact of the disease and might be useful in assessing treatment effects on aspects that are important contributors to the functioning and well-being of patients with FSHD. **Objective:** This study investigated the feasibility of using smartphones and wearables to capture symptoms related to FSHD based on a continuous collection of multiple features, such as the number of steps, sleep, and app use. We also identified features that can be used to differentiate between patients with FSHD and non-FSHD controls. **Methods:** In this exploratory noninterventional study, 58 participants (N=38, 66%, patients with FSHD and N=20, 34%, non-FSHD controls) were monitored using a smartphone monitoring app for 6 weeks. On the first and last day of the study period, clinicians assessed the participants' FSHD clinical score and Timed Up-and-Go test time. Participants installed the app on their Android smartphones, were given a smartwatch, and were instructed to measure their weight and blood pressure on a weekly basis using a scale and blood pressure monitor. The user experience and perceived burden of the app on participants' smartphones were assessed at 6 weeks using a questionnaire. With the data collected, we sought to identify the behavioral features that were most salient in distinguishing the 2 groups (patients with FSHD and non-FSHD controls) and the optimal time window to perform the classification. **Results:** Overall, the participants stated that the app was well tolerated, but 67% (39/58) noticed a difference in battery life using all 6 weeks of data, we classified patients with FSHD and non-FSHD controls with 93% accuracy, 100% sensitivity, and 80% specificity. We found that the optimal time window for the classification is the first day of data collection and the first week of data collection, which yielded an accuracy, sensitivity, and specificity of 95.8%, 100%, and 94.4%, respectively. Features relating to smartphone acceleration, app use, location, physical activity, sleep, and call behavior were the most salient features for the classification. **Conclusions:** Remotely monitored data collection allowed for the collection of daily activity data in patients with FSHD and non-FSHD controls for 6 weeks. We demonstrated the initial ability to detect differences in features in patients with FSHD and non-FSHD controls using smartphones and wearables, mainly based on data related to physical and social activity.

Introduction

BACKGROUND

A recent Dutch population study on facioscapulohumeral dystrophy (FSHD) estimated that approximately 2000 people in the Netherlands and approximately 800,000 people worldwide are living with FSHD.[1] Often, early symptoms include difficulty whistling, smiling, and closing the eyelids while asleep. Weakening of the facial muscles is generally followed by scapular winging. This abnormal positioning of the shoulder bone impairs the movement of the shoulders and arms. Further weakening of the muscles is commonly observed in the upper arms and may progress to the hip girdle and lower legs in severe cases. Less visible symptoms of FSHD are chronic pain and fatigue.[2] In addition to the physical symptoms the diagnosis of FSHD comes with an emotional and social burden. The highly variable and unpredictable progression of the disease can have a strong impact on the quality of life [3,4]: 90% of the affected individuals have visible symptoms by the age of 20 years and 1 in 5 patients with FSHD eventually becomes wheelchair dependent.[5]

No therapy is currently available that stops the progression of FSHD.[6-9] Patients thus must rely on symptomatic treatment such as medical devices or surgical intervention.[2] The development of novel treatment options to delay or halt disease progression is currently under investigation. However, measuring the effect of such new treatments is complicated because disease progression is slow and no objective surrogate end points, predictive for clinical benefit, have been established. App-based technologies may help to monitor FSHD symptom progression more closely and evaluate potential treatment effects on a continuous basis.

Currently, FSHD symptom severity is assessed by clinical scoring of symptoms such as the FSHD clinical score or mobility performance tests such as the Timed Up-and-Go test (TUG) and Reachable Workspace assessment.[10-12] These clinical severity and functional scores have several drawbacks. Scores change very slowly over time,[13] are assessed in a clinic at 1 specific moment, and do not cover the implications of the disease on social and physical activity during daily life. The progressive muscle weakness characterizing FSHD leads to massive changes in the way people live their lives, affecting how they get around, how they complete daily activities, and whether they can work or care for children. Therefore, assessing disease severity may be improved by not only measuring muscle function but also evaluating social and physical activity data. This study aimed to address this by first classifying disease using a smartphone app and wearables to continuously remotely monitor features relating to biometric, physical, and social activities of patients with FSHD in comparison with those of non-FSHD controls. Subsequently, we performed a second analysis in which we aimed to assess disease severity. This analysis will be described in a different paper.

OBJECTIVES

We investigated the feasibility of remotely monitoring multiple features such as step count, sleep, app use, and location tracking in patients with FSHD and non-FSHD controls. First, we evaluated the participants' tolerability of these devices. We then characterized the patients with FSHD and non-FSHD controls in terms of composites of social, physical, and biometric activities. We sought to:

1. Distinguish patients with FSHD from non-FSHD controls using a classification machine learning model and determine the minimum monitoring window needed to perform the classification
2. Identify which of the remotely monitored features were most salient in differentiating between the 2 groups.

Methods

STUDY OVERVIEW

We conducted a cross-sectional, noninterventional study in patients with FSHD and non-FSHD controls. A total of 58 participants (N=38, 66%, patients with genetically confirmed FSHD and N=20, 34%, non-FSHD

controls) were included in this study at the Centre for Human Drug Research (CHDR) in Leiden, The Netherlands, between April 2019 and October 2019. Patients were recruited from The Netherlands and Belgium.

ETHICS APPROVAL

This study was performed in compliance with International Council for Harmonisation Good Clinical Practice and approved by the Stichting Beoordeling Ethiek Biomedisch Onderzoek Medical Ethics Committee (Assen, The Netherlands; CCMO number NL69288.056.19) according to Wet medisch-wetenschappelijk onderzoek met mensen (Dutch law on medical-scientific research with humans).

PATIENT POPULATION

To represent the clinical FSHD spectrum based on symptom severity and age, up to 40 patients with FSHD (and 20 control participants) were deemed sufficient. As this study was exploratory, sample size was not based on power calculations. Eligible patients with FSHD were aged >16 years, had genetically confirmed FSHD (FSHD1 or FSHD2), were symptomatic as demonstrated by the FSHD clinical score of >0 and had an Android phone that they used as their main phone or were willing to use one for the duration of the study period. Patients with any comorbidity, expected to affect the measurements, were excluded. Eligible control participants were included using the same inclusion and exclusion criteria that were used to recruit the patients, except they did not have a diagnosis or symptoms of FSHD.

DATA COLLECTION

CLINICAL ASSESSMENTS On the first and last days of the study period, the FSHD clinical score assessment was performed in the group consisting of patients with FSHD, whereas the TUG was performed in both groups. On day 42 in both groups the user experience was assessed and the perceived burden questionnaire (Multimedia Appendix 1) administered. The FSHD clinical score is a standardized clinical score that quantifies muscle weakness by combining the functional evaluations of the 6 muscle groups affected in FSHD. The scale is divided into 6 independent sections that assess the strength and the functionality of facial muscles, scapular girdle muscles, upper limb muscles, distal leg muscles, pelvic girdle muscles, and abdominal muscles.[11] The TUG assesses mobility and balance by measuring the time it takes for a participant to stand up from a seated position in a chair, walk 3 meters, turn around, walk back 3 meters, and sit down again.[12] The user experience and perceived burden questionnaire was developed by the CHDR to measure the impact of remote monitoring of apps on smartphone performance. The questions are based on the overall experience of CHDR with mobile apps.

REMOTE MONITORING PLATFORM All participants were remotely monitored using the CHDR Monitoring Remotely (CHDR MORE) platform for 42 days. CHDR MORE is a highly customizable platform that allows remote monitoring of participants using smartphones and wearables. The infrastructure used includes an Android app to collect data from smartphone sensors and a connection to the Withings Health (Withings) web-based platform to collect wearable data. All collected features are described in Table 1.

SMARTWATCH, SMART SCALE, AND BLOOD PRESSURE MONITOR In total, three commercially available Withings devices were used: (1) heart rate, step count, and sleep patterns were assessed by the Withings Steel HR smartwatch; (2) weight, BMI, and skeletal muscle mass were assessed by the Withings Body+ scale; and (3) systolic blood pressure and diastolic blood pressure were assessed by the Withings blood pressure monitor. Data from the Withings devices were collected on the phone using Bluetooth and sent to the Withings storage servers before being transferred to a CHDR server. Participants were instructed to wear the Withings Steel HR smartwatch continuously for the duration of the study, and they measured their weight and blood pressure themselves weekly using the Withings Body+ scale and Withings blood pressure monitor, respectively.

PRIVACY The data collection as part of this study may raise privacy and data safety concerns. Therefore, during development of the CHDR MORE app, we addressed these concerns by building in several measures to maximize privacy for all participants. First, all data sources such as SMS text messaging logs, phone calls, and microphone activation only report summative outcomes. These sources cannot send the content of messages or whole recordings to the CHDR servers. In addition, location data only report relative location instead of absolute GPS coordinates. Furthermore, all calculations such as human voice detection are performed on the Android phone itself and removed afterward and all personal data are coded and safely stored on certified CHDR servers.

STATISTICAL ANALYSIS

DATA PREPROCESSING The data preprocessing and analysis pipelines were developed using Python (version 3.6.0; Python Software Foundation). The Python library scikit-learn was used for the feature extraction and the development of the machine learning models.[14] All data were manually and visually inspected for missing data and outlier data. The identified outliers (eg, traveling 10,000 kilometers in a single day) were subsequently removed from the analysis. Missing or excluded data points were not imputed.

FEATURE EXTRACTION As disease progression in FSHD is gradual, the FSHD clinical scores and TUG scores were expected to remain stable during the 6-week period. The daily features were therefore averaged across a defined time window (for more information see p. 95: *Identification of optimal time window*). Table 1 provides a simplified overview of the features that were extracted from the CHDR MORE app and Withings sensors.

FEATURE SELECTION Before fitting the classification models to the data set, features were excluded using manual and automated feature selection. The authors (AZ, RJD, AC, EVB, GJG, and JDM) of this paper manually excluded features based on the degree of missing data and the clinical relevance of the feature (eg, time spent on home and house apps were deemed clinically irrelevant). For the automated feature selection, variance inflation factor and stepwise regression were used to exclude multi-collinear features or features that did not provide additive information, respectively.

CLASSIFICATION MODELS We used 4 categories of data sets for the classification of patients with FSHD and non-FSHD controls. These categories include the composite data (all features), social data (smartphone features relating to social location, social and communication app use, and phone calls), physical activity data (smartwatch features), and biometric data (scale and blood pressure monitor features). We compared the performance of the logistic regression, random forest, and support vector machine classification models (Multimedia Appendix 2 [15-22]). The performance of these classification models was evaluated by the accuracy, sensitivity, specificity, and Matthews correlation coefficient (MCC). A grid search was performed to find the optimal hyperparameters (the parameters that determine the model's structure) that would yield the highest sensitivity and specificity for each model. Furthermore, we performed a 5-fold stratified cross-validation. Cross-validation is a resampling method used to evaluate the prediction performance of the classification models. The data were divided into 5 equal subsets, with the same FSHD-to–non-FSHD ratio within each subset; the model was trained on 4 (80%) partitions of the data and tested on 1 (20%) partition. This procedure was repeated 5 times, with each partition serving as a test set once. The performance of each model validation was then averaged.

IDENTIFICATION OF OPTIMAL TIME WINDOW In total, 6 weeks of data were collected for this study. As continuous and periodic data collection for long periods of time can be expensive and increase the risk of data loss, we investigated the minimum time window needed for reliable classification. First, we used an incrementally increasing time window to train the classification model, starting from day 1 and adding 1 day until we

included all 42 days of data. We examined which time window would yield the highest overall accuracy, sensitivity, and specificity. We compared the performances of 3 classification algorithms (least absolute shrinkage and selection operator [LASSO]-penalized logistic regression, random forest, and support vector machine) to classify patients with FSHD and non-FSHD controls using the incremental time windows. Second, we used the optimal time window to train the classification model and evaluated how stable the classification performance would be for the remaining 5 weeks of data. Here, we evaluated the stability of the algorithm based on the generalization error of the trained classification model.[23]

Results

DATA COLLECTED

In total, 58 participants (N=38, 66%, patients with FSHD and N=20, 34%, non-FSHD controls) participated in the study. We did not meet our goal of 40 patients because of difficulties in recruiting patients in an acceptable time span. The female-to-male ratio was the same in both populations; however, the median age of the control participants without FSHD was lower than that of their counterparts with FSHD. Table 2 illustrates the demographic and disease characteristics of the participants enrolled in this study. The FSHD clinical scores and TUG scores remained relatively stable during the 6-week period (with a maximum intraparticipant change of 1 point for the FSHD score and 0.63 seconds for the TUG score).

PERCEIVED BURDEN AND DATA LOSS

As shown in Figure 1, overall, 3% (2/58) of the participants found the app on their phone to be annoying. Furthermore, 67% (39/58) of the participants agreed that there was a noticeable difference in battery life, 43% (25/58) agreed that the constant presence of the app was noticeable on their smartphone, 28% (16/58) rated the constant visible notification as annoying, and 26% (15/58) of the participants noted a difference in the speed of their smartphone.

Data completeness is defined as having incoming data for each day of the clinical trial, except for the blood pressure and scale data, for which completeness is defined as having incoming data each week. As phone and SMS text messaging data are activity triggered and are aperiodic, it is not possible to know whether data were missing. Table 3 provides an overview of data completeness for the CHDR MORE app, Withings watch, Withings scale, and Withings blood pressure monitor and their respective sensors.

FEATURE SELECTION

Several features were manually excluded before modeling. Because of the number of participants missing body composition data, we excluded all the body composition data with the exception of weight. Furthermore, we excluded SMS text message use features and app categories that were only used by only 5% (3/58) of the participants.

IDENTIFICATION OF OPTIMAL TIME WINDOW AND CLASSIFICATION PERFORMANCE

Using all 6 weeks of data, the optimal classification model (LASSO-penalized logistic regression) achieved 93% accuracy, 100% sensitivity, 80% specificity, and 85% MCC. This classification model identified 15 features that were relevant for differentiating between patients with FSHD and non-FSHD controls. Specifically, features such as app use, weight, location, physical activity, and sleep were important for differentiating between the 2 populations (Figure 2). Table 4 shows the predictive features and their positive or negative associations with the classification label. The predictive features indicate that the participants in the group consisting of patients with FSHD were less likely to engage in moderate physical activity and spend time on recreational apps such as entertainment apps, music and audio apps, video players and editing apps, and games. The predictive features also showed that the participants in the group consisting of patients with FSHD were more likely to spend more time at home and health locations than their non-FSHD counterparts.

Table 5 provides a summary of the number of selected features and the respective performance metric for each of the data sets fitted to the 6-week LASSO-penalized logistic regression model. The table illustrates that the composite data set model outperformed the models fitted to the social, physical activity, and biometric data sets. The MCC is used to select the best model because it corrects for class imbalances. The scores of the individual data sets are included to give an overview of their performance on their own. The MCC values of the social activity, physical activity, and biometric logistic regression models were 52%, 38%, and −21%, respectively.

As for identifying the optimal time window for accurately classifying the patients with FSHD and non-FSHD controls, we found that training the random forest on the data collected on the first day and the data collected during the first week yielded an accuracy, sensitivity, specificity, and MCC of 95.8%, 100%, 94.4%, and 93.8% (Figure 3). This approach outperformed the classification models that were trained on all 6 weeks of data. We also trained classification models on the first week's data and fitted the data from subsequent weeks to assess the stability of the classification performance over time (Figure 4). We found that the random forest achieved the best overall performance, with a mean accuracy, sensitivity, specificity, and MCC of 95% (SD 0.9%), 97.6% (SD 3.6%), 94.1% (SD 0.9%), and 93.6% (SD 0.1%), respectively. Figure 5 provides a SHAPley additive explanations plot that illustrates the magnitude and direction of the effect of a feature on a prediction. Of the 20 selected features, the top 5 (25%) most important features for the classification were mean kilometers traveled, 95% maximum distance from home, total kilometers traveled, 95% highest heart rate, and intense activity duration. For each of these features, the participants in the group consisting of patients with FSHD had lower scores than the non-FSHD controls.

Discussion

PRINCIPAL FINDINGS

We investigated the feasibility of monitoring and characterizing the physical, social, and biometric features of patients with FSHD and non-FSHD controls using remote monitoring technologies. The use of the remote monitoring platform was well tolerated by all participants. Next, we found that a minimum of 1 day of data and a maximum of 1 week of data can be used to reliably classify the 2 populations. In fact, an FSHD classification model trained on data from a shorter time window outperformed a classification model trained on data from the entire 6-week period. Furthermore, we illustrated that a classification model trained on the first week's data yielded stable and reliable classification predictions across the remaining 5-week period.

Most (37/58, 64%) of the participants tolerated the CHDR MORE app constantly running on their smartphone (Figure 1). Of the 58 participants, only 2 (3%) stated that the app was annoying. However, the results show that some of the participants agreed that there was a noticeable difference in smartphone speed performance (14/58, 25%), stability (8/58, 14%), and overall battery life (39/58, 67%). Therefore, the presence of the app was noticeable for some (25/58, 43%) of the participants. The decrease in smartphone performance (ie, speed, stability, and battery performance) was likely due to the continuous sampling of the sensors. As this was the first study in this specific patient group with this platform, all smartphone sensors were frequently sampled to capture all possible features. With the collected data in this study, we identified the features that are useful in differentiating between patients with FSHD and non-FSHD controls. In future studies, noncontributing raw data such as data from the accelerometer and gyroscope (both sampled at 5 Hz) can be turned off to reduce the burden on the battery performance and overall user experience. We do not know for certain whether, and how, the noticeability of the app affects participants' behavior. Of the 58 participants, 6 (10%) stated that they noticed a change in smartphone use for themselves, which may

mean that they changed their behavior. Therefore, participants will know that they are participating in a study and that they are being constantly monitored even if the app is perfectly optimized. As a result, some sort of change in behavior can be expected.

As for the user experience and perceived burden questionnaire, we designed a questionnaire based on our own experiences with smartphone use and the predicted effects of the CHDR MORE app on smartphones. This questionnaire was not validated in any other study. At the time of designing the study, there were no validated and published smartphone app questionnaires that would fit our purpose. For example, the mHealth App Usability Questionnaire[24] focuses more on active smartphone apps, where there is interaction between the app and the participants. The CHDR MORE app is a passive app, requiring almost no interaction between the app and the user. Therefore, the questions should be more focused on the indirect effects of the app, such as more frequent crashes in other apps, subjective loss of snappiness of the operating system, or issues with battery performance. Although our questionnaire is not validated, it was considered the best way to accurately capture the perceived impact of the CHDR MORE app on smartphone use.

Feature selection is one of the most important processes for building a classification model. The inclusion of irrelevant features can confound the interpretability of the model because potentially predictive features would be excluded and therefore seem to be irrelevant. For example, because the patients with FSHD had more text-related activity than the non-FSHD controls, the SMS text messaging features were selected as important classification features. Given that the SMS text messaging features were not deemed clinically relevant because only 55% (21/38) of the patients with FSHD and 50% (10/20) of the non-FSHD controls actively sent outgoing SMS text messages and the majority of the SMS text messages were exchanged with unknown contacts, we excluded the SMS text messages as a feature. As a result, features that were initially not selected by the model for inclusion, such as sleep, were now deemed important features. The SMS text messaging features masked the relevance of other potentially predictive features. The features that researchers manually choose to include or exclude will influence the interpretability and stability of the model. It should be noted that although SMS text messaging features were excluded, features regarding instant messaging app use were included.

Our classification models allowed for the identification of a stable set of features that were distinctive of FSHD symptomology. We believe that identifying which remotely monitored features are relevant to FSHD can be a first step toward continuous monitoring of symptom severity and disease progression. For example, our classification model identified sleep as a relevant feature for classifying patients with FSHD. Other studies have found that patients with FSHD typically experience sleep anomalies because of anxiety, respiratory muscle dysfunction, and pain.[25-27] This illustrates that the CHDR MORE platform is sensitive enough to detect and monitor sleep anomalies among individuals with FSHD outside of the clinic. Furthermore, location-related features were relevant for differentiating between the 2 populations. In this study, the patients with FSHD spent more time at home, in areas with public transportation, or at health locations than the healthy participants. Patients with FSHD face a range of physical challenges because of the functional deterioration in the affected muscular regions. Consequently, patients with FSHD may become more home bound and more reliant on public transportation for travel, as well as require more visits to their physicians. In conclusion, the CHDR MORE platforms provide data that can be used to show differences in the daily lives of patients with FSHD and controls without FSHD.

We demonstrated that there is a trade-off among the classification accuracy, the number of sensor measurements, and the duration of the monitoring period. Previous studies have demonstrated that using data from multiple sensors improves the detection of mental and physical health status compared with using data from a single sensor.[28-30] We illustrated that social activity, physical activity, and biometric data alone are insufficient for the accurate classification of FSHD. Rather, the inclusion of data from the smartphone, smartwatch, and scale improves the

performance of the FSHD classification algorithm. Although the modeling of multi-sensor data can be advantageous, it can lead to several practical limitations. The inclusion of more features can increase the model's complexity and thus limit the model's explainability. Furthermore, the inclusion of more sensors and a longer monitoring period can be more expensive, potentially limiting the number of participants enrolled in a study, and increase the risk of data loss. Future studies will need to weigh the advantages and disadvantages of integrating smartphones, smartwatches, scales, and monitoring period into their remotely monitored FSHD clinical trials.

Despite the good performance of our model, this study includes some limitations. The patients with FSHD and non-FSHD controls were comparable except for the age demographic. The median age of the non-FSHD controls was approximately 13 years less than that of the patients with FSHD. Generally, the older the person, the less they tend to use their smartphone and, in particular, the less they tend to use communication and social apps.[31] When characterizing patients with FSHD and non-FSHD controls based on active smartphone use, the model may be biased because of the difference in age. However, as seen in the results, only 1 feature of active smartphone use—time spent on recreational apps—was included in the final model for the characterization of patients with FSHD, which may limit the impact of this difference. The other features used in the composite model consist of either physical activity features collected passively from the smartphone or biometric data collected from the Withings devices. Therefore, we believe that the impact of these contaminated data on the performance of our model is low.

The objective of our study was to capture continuous sensor data. However, these data can only be considered reliable when participants carry their smartphone and have it turned on all the time. During this study, all participants were instructed to do so. However, data captured when the participant was not carrying their smartphone could not be distinguished from data captured when the participant was carrying the smartphone. Therefore, all instances in which the smartphone is not carried or turned on result in unrepresentative data. These data get mixed in the real data because these moments cannot be filtered out of the data with full certainty, resulting in unreliable data. Of note, there is no easy solution to this problem. It would be difficult to continuously check whether the participants are carrying their smartphone using the built-in sensors. However, adherence to this requirement is an important aspect in remote data collection, emphasizing the need for clear instructions on this adherence aspect to participants during training sessions before study start.

The level of data loss from the Withings scale indicates that improvement is needed to gather reliable scale data (Table 3). Data loss occurred for both the patients with FSHD and the non-FSHD controls, indicating that the loss of data was unlikely related to any of the FSHD symptoms. Although clear instructions were given at the beginning of the study and all participants received a manual with the same instructions, we believe that the data loss was caused by improper use of the scale by the participants. The weight measurement consisted of two parts: measurement of weight and measurement of body composition.

Weight was determined first, followed by a blinking notification on the display during the measurement of body composition. This might have given the impression to the participant that the measurement had been completed, causing them to interrupt the second part of the measurement, resulting in an incomplete measurement. For future studies, we recommend incorporating a live training at the beginning of the study on the correct use of the scale.

Efficient clinical testing of any FSHD intervention or of any drug targeted at improving function of patients with FSHD or delaying disease progression requires the availability of clinical biomarkers that ideally change relatively rapidly over time; correlate with, and allow for, prediction of progression of the existing clinical severity and functional scores; and allow for identification of fast progressors. Using data collected in a home setting might provide a more comprehensive picture of the evolution of a patient's overall condition over time. This study is a first step in the development and validation process of using data collected by a

specific remote monitoring platform for use in patients with FSHD. The features described in this paper may be useful in further evaluating the impact of the disease and monitoring disease progression in patients with FSHD in the future.[13] More extensive data from longitudinal studies are needed to further define how social, physical, and biometric data collected remotely can be used to monitor symptoms.

Conclusions

To conclude, this study illustrates that the collection of smartphone data and wearable data is acceptable to patients with FSHD and non-FSHD controls and can be used to differentiate between the 2 populations. We showed that remotely monitored end points can capture behavioral differences between patients and controls. Further longitudinal studies are warranted to study the potential of using a remote monitoring system for detecting FSHD symptom severity and possible drug effects.

REFERENCES

1 Deenen JC, Arnts H, van der Maarel SM, Padberg GW, Verschuuren JJ, Bakker E, et al. Population-based incidence and prevalence of facioscapulohumeral dystrophy. Neurology 2014 Aug 13;83(12):1056-1059. [doi: 10.1212/wnl.0000000000000797]

2 Voet N, Bleijenberg G, Hendriks J, de Groot I, Padberg G, van Engelen B, et al. Both aerobic exercise and cognitive-behavioral therapy reduce chronic fatigue in FSHD: an RCT. Neurology 2014 Oct 22;83(21):1914-1922. [doi: 10.1212/wnl.0000000000001008]

3 Padua L, Aprile I, Frusciante R, Iannaccone E, Rossi M, Renna R, et al. Quality of life and pain in patients with facioscapulohumeral muscular dystrophy. Muscle Nerve 2009 Aug 16;40(2):200-205. [doi: 10.1002/mus.21308] [Medline: 19609906]

4 Tawil R, McDermott MP, Mendell JR, Kissel J, Griggs RC. Facioscapulohumeral muscular dystrophy (FSHD): design of natural history study and results of baseline testing. FSH-DY Group. Neurology 1994 Mar 01;44(3 Pt 1):442-446. [doi: 10.1212/wnl.44.3_part_1.442] [Medline: 8145913]

5 Wang LH, Tawil R. Facioscapulohumeral dystrophy. Curr Neurol Neurosci Rep 2016 Jul 23;16(7):66. [doi: 10.1007/s11910-016-0667-0] [Medline: 27215221]

6 Kissel JT, McDermott MP, Mendell JR, King WM, Pandya S, Griggs RC, FSH-DY Group. Randomized, double-blind, placebo-controlled trial of albuterol in facioscapulohumeral dystrophy. Neurology 2001 Oct 23;57(8):1434-1440. [doi: 10.1212/wnl.57.8.1434] [Medline: 11673585]

7 Kley R, Tarnopolsky M, Vorgerd M. Creatine for treating muscle disorders. Cochrane Database Systematic Rev 2004(2):CD004760. [doi: 10.1002/14651858.CD004760]

8 Tawil R, McDermott MP, Pandya S, King W, Kissel J, Mendell JR, et al. A pilot trial of prednisone in facioscapulohumeral muscular dystrophy. FSH-DY Group. Neurology 1997 Jan 01;48(1):46-49. [doi: 10.1212/wnl.48.1.46] [Medline: 9008492]

9 Le Gall L, Sidlauskaite E, Mariot V, Dumonceaux J. Therapeutic strategies targeting DUX4 in FSHD. J Clin Med 2020 Sep 07;9(9):2886 [FREE Full text] [doi: 10.3390/jcm9092886] [Medline: 32906621]

10 Validation of the upper extremity reachable work space (RWS) as a clinical outcome assessment (COA) for FSHD clinical trials. Muscular Dystrophy Association. URL: https://www.mdaconference.org/node/918 [accessed 2021-02-20]

11 Lamperti C, Fabbri G, Vercelli L, D'Amico R, Frusciante R, Bonifazi E, et al. A standardized clinical evaluation of patients affected by facioscapulohumeral muscular dystrophy: the FSHD clinical score. Muscle Nerve 2010 Aug 18;42(2):213-217. [doi: 10.1002/mus.21671] [Medline: 20544930]

12 Podsiadlo D, Richardson S. J Am Geriatr Soc 1991 Feb;39(2):142-148. [doi: 10.1111/j.1532-5415.1991.tb01616.x] [Medline: 1991946]

13 Kruizinga MD, Stuurman FE, Exadaktylos V, Doll RJ, Stephenson DT, Groeneveld GJ, et al. Development of novel, value-based, digital endpoints for clinical trials: a structured approach toward fit-for-purpose validation. Pharmacol Rev 2020 Oct 21;72(4):899-909. [doi: 10.1124/pr.120.000028] [Medline: 32958524]

14 Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12:2825-2830 [FREE Full text] [doi: 10.5555/1953048.2078195]

15 Tolles J, Meurer WJ. Logistic regression: relating patient characteristics to outcomes. JAMA 2016 Aug 02;316(5):533-534. [doi: 10.1001/jama.2016.7653] [Medline: 27483067]

16 Czepiel SA. Regression models. In: Maximum Likelihood Estimation for Sample Surveys. Milton Park, Abingdon-on-Thames, Oxfordshire United Kingdom: Taylor & Francis; 2012.

17 Tibshirani R. Regression shrinkage and selection via the LASSO. J Royal Stat Soc: series B (Methodological) 2018 Dec 05;58(1):267-288. [doi: 10.1111/j.2517-6161.1996.tb02080.x]

18 Ho T. Proceedings of 3rd International Conference on Document Analysis and Recognition. In: Proceedings of 3rd International Conference on Document Analysis and Recognition. 1995 Presented at: Proceedings of 3rd InternationalConference on Document Analysis and Recognition; Aug 14-16, 1995; Montreal, QC, Canada URL: https://IEEEexplore.IEEE.org/document/598929/citations#citations [doi: 10.1109/icdar.1995.598929]

19 Muchlinski D, Siroky D, He J, Kocher M. Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data. Polit anal 2017 Jan 04;24(1):87-103. [doi: 10.1093/pan/mpv024]
20 Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995 Sep;20(3):273-297. [doi: 10.1007/bf00994018]
21 Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning Data Mining, Inference, and Prediction, Second Edition. Cham: Springer; 2009.
22 Suykens J, Vandewalle J. Least squares support vector machine classifiers. Neural Processing Letters 1999;9(3):293-300. [doi: 10.1023/A:1018628609742]
23 Bousquet O, Elisseeff A. Stability and generalization. J Mach Learn Res 2002;2(3):499-526 [FREE Full text]
24 Zhou L, Bao J, Setiawan IM, Saptono A, Parmanto B. The mHealth app usability questionnaire (MAUQ): development and validation study. JMIR Mhealth Uhealth 2019 Apr 11;7(4):e11500 [FREE Full text] [doi: 10.2196/11500] [Medline:30973342]
25 Leclair-Visonneau L, Magot A, Tremblay A, Bruneau X, Pereon Y. P.16.5 Anxiety is responsible for altered sleep quality in Facio-Scapulo-Humeral Muscular Dystrophy (FSHD). Neuromuscular Disorders 2013 Oct;23(9-10):823-824. [doi: 10.1016/j.nmd.2013.06.642] https://formative.jmir.org/2022/9/e31775 JMIR Form Res 2022 | vol. 6 | iss. 9 | e31775 | p. 12
26 Della Marca G, Frusciante R, Vollono C, Iannaccone E, Dittoni S, Losurdo A, et al. Pain and the alpha-sleep anomaly: a mechanism of sleep disruption in facioscapulohumeral muscular dystrophy. Pain Med 2013 Apr 01;14(4):487-497. [doi: 10.1111/pme.12054] [Medline: 23387524]
27 Runte M, Spiesshoefer J, Heidbreder A, Dreher M, Young P, Brix T, et al. Sleep-related breathing disorders in facioscapulohumeral dystrophy. Sleep Breath 2019 Apr 26;23(3):899-906. [doi: 10.1007/s11325-019-01843-1]
28 Garcia-Ceja E, Riegler M, Nordgreen T, Jakobsen P, Oedegaard KJ, Tørresen J. Mental health monitoring with multimodal sensing and machine learning: a survey. Pervasive Mobile Comput 2018 Dec;51:1-26. [doi: 10.1016/j.pmcj.2018.09.003]
29 Garcia-Ceja E, Galván-Tejada CE, Brena R. Multi-view stacking for activity recognition with sound and accelerometer data. Inf Fusion 2018 Mar;40:45-56. [doi: 10.1016/j.inffus.2017.06.004]
30 Soleymani M, Riegler M, Halvorsen P. Multimodal analysis of image search intent: intent recognition in image search from user behavior and visual content. In: Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval.2017 Presented at: ICMR '17: International Conference on Multimedia Retrieval; Jun 6 - 9, 2017; Bucharest Romania. [doi: 10.1145/3078971.3078995]
31 Andone I, Błaszkiewicz K, Eibes M, Trendafilov B, Montag C, Markowetz A. How age and gender affect smartphone usage. In: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct. 2016 Presented at: UbiComp '16: The 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing; Sep 12 - 16, 2016; Heidelberg Germany. [doi: 10.1145/2968219.2971451]

TABLE 1 Overview of all smartphone and wearable sensors used in this study and their respective extracted features.

Device and Sensor	Features
SMARTPHONE	
ACCELEROMETER	Maximum magnitude of the acceleration: 98%
APPS	Number of times an app is opened; amount of time app is open in foreground
GPS	Total kilometers traveled per day; average kilometers traveled per trip; 95% maximum distance from home
GOOGLE PLACES	Number of unique places visited; time spent at each unique location
CALLS	Number of outgoing, incoming, and missed calls; number of calls from known and unknown contacts
MICROPHONE	Percentage of time a human voice is present
WEARABLES (WITHINGS)	
WATCH STEP COUNT	Total step count; mean steps per minute; mean steps per hour; maximum steps per hour
WATCH HEART RATE	Heart rate: 5%, 50%, and 95% ranges and SD of heart rate percentage of time spent in resting heart rate
WATCH SLEEP	Awake as well as light and deep sleep duration (minutes); number of awake as well as light and deep sleep periods; time to fall asleep (minutes)
WATCH PHYSICAL ACTIVITY	Soft, moderate, and hard activity duration
BLOOD PRESSURE MONITOR	Systolic and diastolic blood pressure
SCALE	Weight (kg); muscle mass (kg); bone mass (kg); body fat (%); body water (%)

TABLE 2 Demographics of patients with facioscapulohumeral dystrophy (FSHD) and controls without FSHD (N=58).

Demographics	Patients with FSHD	Non-FSHD Controls
Sex, n(%)		
Female	23 (61)	11 (55)
Male	15 (39)	9 (45)
Age (years), mean (SD: range)	45 (14.5; 18-64)	33 (12; 23-69)
Weight (kg), mean (SD: range)	80 (16; 52-130)	78 (18; 56-129)
BMI (kg/m^2), mean (SD: range)	26 (4; 20-44)	25 (5; 19-35)
FSHD clinical score, mean (SD: range)	5 (3; 1-13)	0 (0; 0-0)
Timed Up-and-Go test (seconds), mean (SD: range)	8.8 (35; 5-15.81)	7.8(1.55; 6-12.09)

TABLE 3 Overview of data completeness. The data completeness shows what percentage of data was collected among the participants during the 42 days of the study; hence, in total, there should be 2436 daily instances and 232 weekly instances.

Sensor	Feature	Overall data completion N (%)			
		Patients with FSHD		Controls without FSHD	
		n (%)	N	n (%)	N
Microphone (smartphone)	Voice activation	1181 (74)	1596	688 (81.9)	840
Accelerometer (smartphone)	Phone Acceleration	1260 (78.95)	1596	656 (78)	840
Google Places (smartphone)	Places	1109 (69.49)	1596	616 (73.33)	840
GPS (smartphone)	Relative Location	1373 (86.03)	1596	785 (93.45)	840
App use (smartphone)	Use event aggregate	1404 (87.97)	1596	779 (92.74)	840
Withings blood pressure monitor	Blood pressure and heart rate	1452 (91.15)	1596	630 (75)	840
Withings scale	Body composition	173 (75.88)	228	88 (73.33)	120
	Weight	205 (89.91)	228	108 (90)	120
Withings watch	Activity duration	1505 (94.3)	1596	744 (88.57)	840
	Heart rate	1181 (74)	1596	588 (70)	840
	Step count	1491 (93.42)	1596	708 (84.29)	840
	Sleep Summary	1408 (88.22)	1596	685 (81.55)	840

TABLE 4 Selected features for classifying patients with facioscapulohumeral dystrophy and controls without facioscapulohumeral dystrophy based on the complete 6-week composite data set. Unstandardized estimated coefficients indicate the direction of the association between the feature and the classification label.

Feature category	Features	Unstandardized estimated coefficients
Activity	Moderate activity duration	-0.04
App	Time spent on recreational apps	-0.53
Body	Weight (kg)	-0.45
Location	Distance from home: 95%	0.85
Time spent at location	Travel location	1.00
	Home location	0.67
	Unknown location	0.53
	Health location	0.29
	Public location	-0.12
	Social location	-0.14
	Commercial location	-0.94
Sleep	Average total sleep duration	0.65
	Light sleep duration	-0.35
	Number of awake periods during a sleep session	-0.61
	Maximum total sleep duration	-0.69

TABLE 5 Summary of number of selected features and the respective performance metric for each of the data sets used to classify the patients with FSHD from the controls without facioscapulohumeral dystrophy.

Dataset	Number of selected Features	Accuracy (%)	Sensitivity (%)	Specificity (%)	Matthews Correlation Coefficient (%)
Composite	15	93	100	80	85
Biometric	5	57	89	0	-21
Social	10	79	90	60	52
Physical Activity	13	71	78	60	38

FIGURE 1 Feasibility and perceived burden of remote monitoring in patients with facioscapulohumeral dystrophy using smartphone-based technologies.

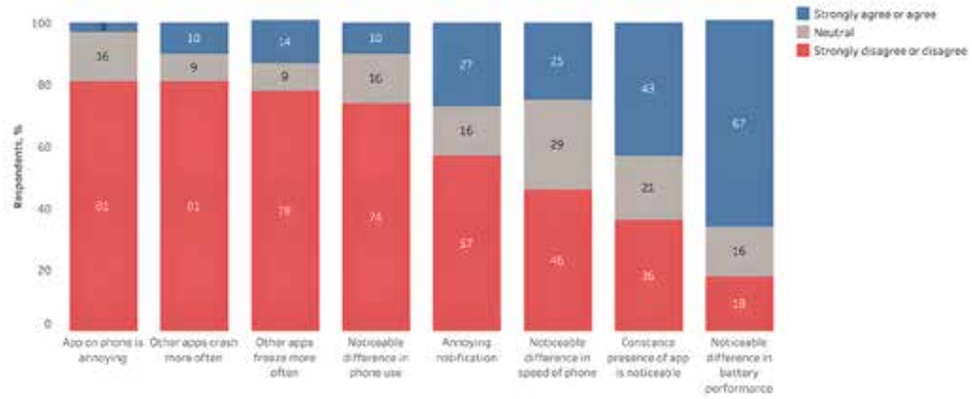


FIGURE 2 Selected features for classifying patients with facioscapulohumeral dystrophy and those without FSHD based on the composite data set using all 6 weeks of data and the least absolute shrinkage and selection operator–penalized logistic regression model. Unstandardized estimated coefficients indicate the direction of the association between the feature and the classification label.

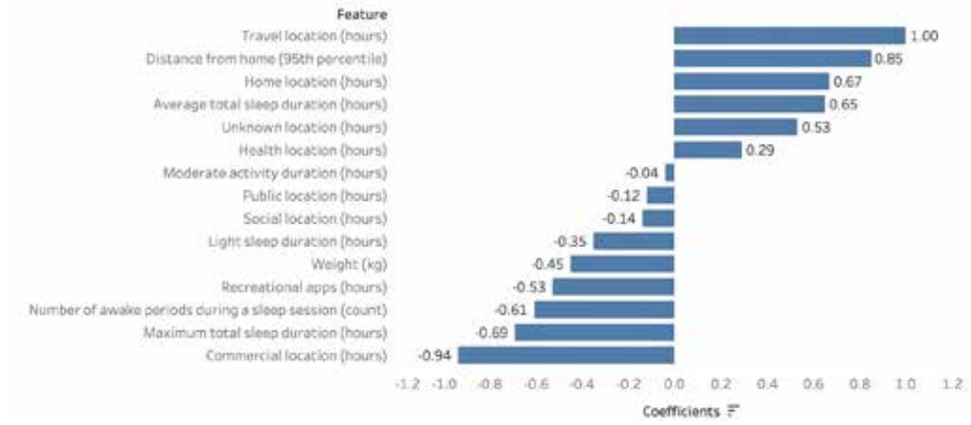


FIGURE 3 Performance of the incremental classification predictions for 3 classifiers (logistic regression, random forest, and support vector machine). The x-axis shows the time window for training the classification models starting from day 1 to day 42. The error bands represent the SD of the classification performance for the 5-fold cross-validation.

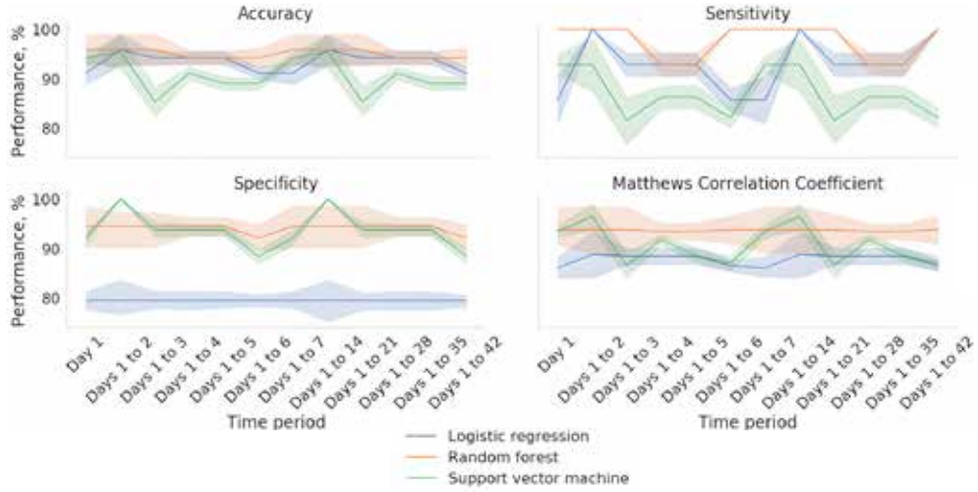


FIGURE 4 Performance of 3 classifiers (logistic regression, random forest, and support vector machine) trained on the week 1 data and used to predict the classification diagnosis of the subsequent weeks of data. The error bands represent the SD of the classification performance for the 5-fold cross-validation.

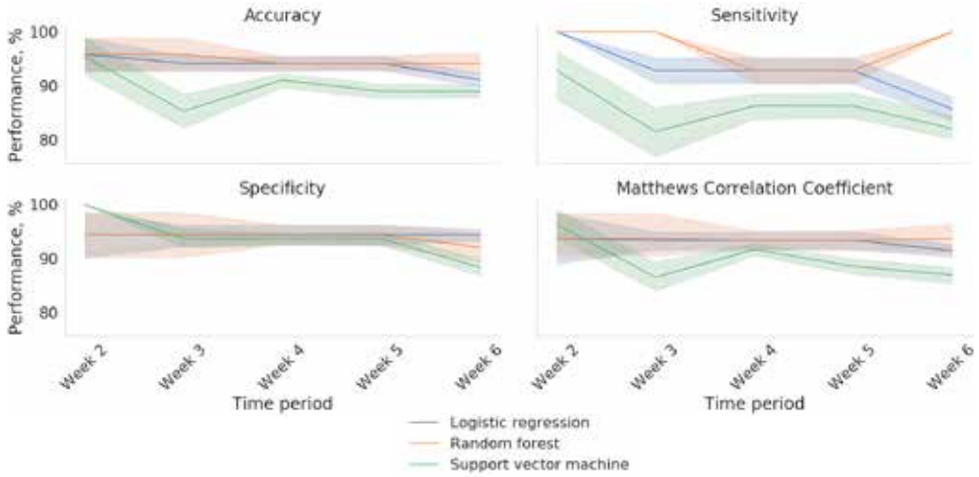
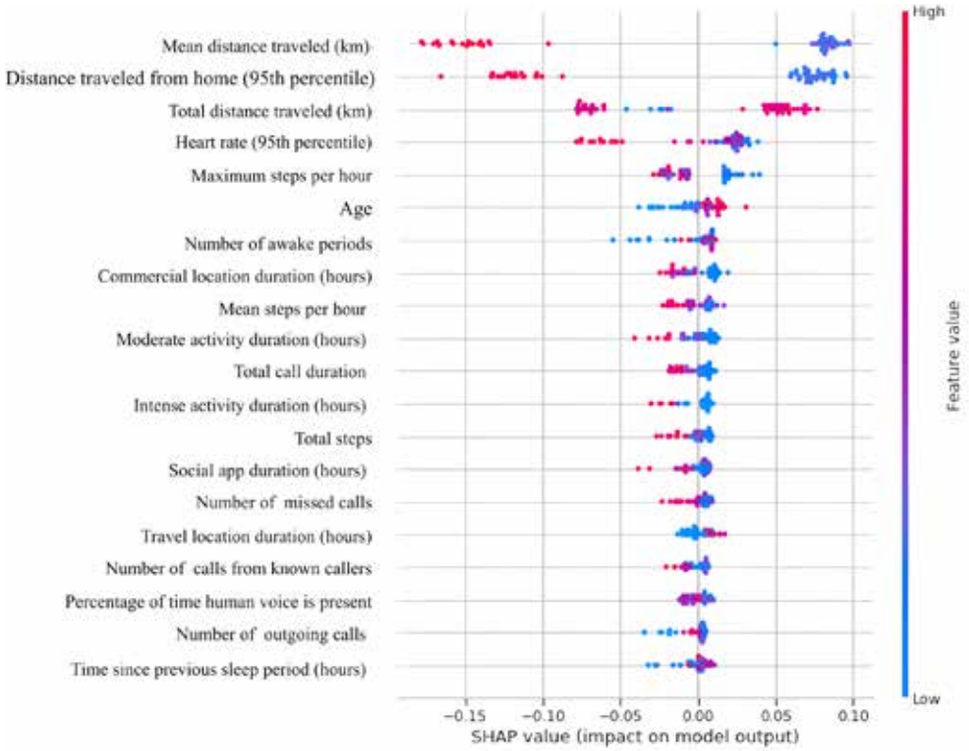


FIGURE 5 SHAPley additive explanations (SHAP) summary plot based on a random forest classifier that was trained on the week 1 data. The x-axis shows the feature importance, where features are ranked in descending order. The y-axis shows the SHAP value that illustrates the direction of the association between the feature and facioscapulohumeral dystrophy severity. The color scheme reflects the probability of a participant being classified as a patient with facioscapulohumeral dystrophy.



PART III

ESTIMATION OF SYMPTOM SEVERITY

CHAPTER 4

Smartphone and wearable sensors for the estimation of facioscapulohumeral muscular dystrophy disease severity: cross-sectional study

Ahnjili Zhuparris,[1] MSc; Ghobad Maleki,[1] BSc, MD; Ingrid Koopmans,[1] MSc; Robert J Doll,[1] PhD; Nicoline Voet,[2] PhD; Wessel Kraaij,[3] PhD, Prof Dr; Adam Cohen,[1] MD, PhD, Prof Dr; Emilie van Brummelen,[1] PhD; Joris H De Maeyer,[4] PhD; Geert Jan Groeneveld,[1] MD, PhD, Prof Dr

JMIR Form Res. 2023;7:e41178. doi:10.2196/41178

1 Centre for Human Drug Research (CHDR), Leiden, NL

2 Department of Rehabilitation, Rehabilitation Center Klimmendaal, Nijmegen, NL

3 Leiden Institute of Advanced Computer Science, Leiden University, Leiden, NL

4 Facio Therapies, Leiden, NL

Abstract

Background: Facioscapulohumeral muscular dystrophy (FSHD) is a progressive neuromuscular disease. Its slow and variable progression makes the development of new treatments highly dependent on validated biomarkers that can quantify disease progression and response to drug interventions. **Objective:** We aimed to build a tool that estimates FSHD clinical severity based on behavioral features captured using smartphone and remote sensor data. The adoption of remote monitoring tools, such as smartphones and wearables, would provide a novel opportunity for continuous, passive, and objective monitoring of FSHD symptom severity outside the clinic. **Methods:** In total, 38 genetically confirmed patients with FSHD were enrolled. The FSHD Clinical Score and the Timed Up and Go (TUG) test were used to assess FSHD symptom severity at days 0 and 42. Remote sensor data were collected using an Android smartphone, Withings Steel HR+, Body+, and BPM Connect+ for 6 continuous weeks. We created 2 single-task regression models that estimated the FSHD Clinical Score and TUG separately. Further, we built 1 multitask regression model that estimated the 2 clinical assessments simultaneously. Further, we assessed how an increasingly incremental time window affected the model performance. To do so, we trained the models on an incrementally increasing time window (from day 1 until day 14) and evaluated the predictions of the clinical severity on the remaining 4 weeks of data. **Results:** The single-task regression models achieved an R2 of 0.57 and 0.59 and a root-mean-square error (RMSE) of 2.09 and 1.66 when estimating FSHD Clinical Score and TUG, respectively. Time spent at a health-related location (such as a gym or hospital) and call duration were features that were predictive of both clinical assessments. The multitask model achieved an R2 of 0.66 and 0.81 and an RMSE of 1.97 and 1.61 for the FSHD Clinical Score and TUG, respectively, and therefore outperformed the single-task models in estimating clinical severity. The 3 most important features selected by the multitask model were light sleep duration, total steps per day, and mean steps per minute. Using an increasing time window (starting from day 1 to day 14) for the FSHD Clinical Score, TUG, and multitask estimation yielded an average R2 of 0.65, 0.79, and 0.76 and an average RMSE of 3.37, 2.05, and 4.37, respectively. **Conclusions:** We demonstrated that smartphone and remote sensor data could be used to estimate FSHD clinical severity and therefore complement the assessment of FSHD outside the clinic. In addition, our results illustrated that training the models on the first week of data allows for consistent and stable prediction of FSHD symptom severity. Longitudinal follow-up studies should be conducted to further validate the reliability and validity of the multitask model as a tool to monitor disease progression over a longer period.

Introduction

Facioscapulohumeral muscular dystrophy (FSHD) is a progressive neuromuscular disease characterized by the wasting of muscles in the face, upper body, and legs.[1] The onset and progression vary greatly between individuals.[2] Early symptoms include difficulties in smiling, whistling, and shutting of the eyelids during sleep. These symptoms are followed by impaired upper-arm movements and walking. A total of 20% of individuals with FSHD eventually become wheelchair bound.[2] Less visible FSHD symptoms include fatigue and chronic pain.[3] In addition to the physical burden, individuals with FSHD also experience emotional, social, and socioeconomic burdens.[4,5] As a result, patients report increased deterioration in quality of life as the disease progresses.[6]

Currently, there are no therapies or interventions that prevent the wasting of muscles in patients with FSHD.[7] Muscle-strengthening drugs have been shown to have limited effect on the disease progression.[8] As a result, patients with FSHD largely rely on symptomatic treatments (eg, analgesics, exercise, and cognitive therapy). The development of novel treatment options to delay or halt FSHD disease progression is currently under investigation.[9,10] However, measuring the effect of such new treatments is complicated, as disease progression is slow and no objective surrogate end points, predictive for clinical benefit, have been established.

Two common clinical assessments for assessing FSHD symptom severity are the FSHD Clinical Score and Timed Up and Go (TUG) test. The FSHD Clinical Score is composed of an evaluation of the extent of the muscle weakness among 6 regions of the body.[11] The TUG is a test used to assess functional mobility.[12] The test requires a participant to rise from a chair, walk 3 m forward, turn around, and return to the chair. These clinician-rated assessments provide a snapshot of the disease status and are primarily focused on muscular strength and function that are inherently subjective. Identifying novel objective biomarkers for monitoring disease progression could additionally provide clinically relevant insights and aid drug development. Novel digital end points for neuromuscular disease drug development have already demonstrated to be sensitive to differentiating patients from healthy volunteers and are strongly correlated with clinician assessments.[13-15] The widespread adoption of smartphones and wearables could provide new opportunities for objective and continuous monitoring of FSHD disease progression outside the laboratory.

This study was designed to identify smartphone-based and remote sensor–based features that could be used to assess FSHD disease severity. These features may enable the passive remote monitoring of disease progression and might potentially facilitate early detection of treatment effects on FSHD symptoms and the patient's quality of life. We hypothesized that the behavioral features captured by these remote monitoring devices would capture the daily physical and social burden that patients with FSHD experience. Although other neuromuscular disease studies with similar protocols have used machine learning to construct their digital end points, until now, different monitoring periods were arbitrarily selected by various researchers.[16,17] Here, we investigated how different time windows affect the model's performance to estimate one's symptom severity over time.[18,19] As these features can vary considerably over time, we assessed the stability and test-retest reliability of the first week of data to estimate FSHD disease severity for the remainder of the trial. In this paper, we describe the development of a novel tool based on smartphone and remote sensor data to provide remote estimation of FSHD disease severity.

Methods

OVERVIEW

This study is an extension of a previous longitudinal clinical study that investigated the feasibility of monitoring and characterizing patients with FSHD and healthy controls in terms of biometric, physical, and social activities using data sourced from smartphones and other remote monitoring devices. Therefore, additional information regarding the data collection and data quality has been previously published.[15]

PATIENTS

This was a noninterventional, cross-sectional study involving patients with FSHD. The study was performed between April and October 2019 in the Centre for Human Drug Research (CHDR) research unit in Leiden, the Netherlands. Table 1 provides an overview of the demographic distribution of the patients with FSHD enrolled in this study.

In total, 38 patients with genetically confirmed FSHD from the Netherlands and Belgium were included in the study. Eligible patients were 16 years or older, had genetically confirmed FSHD, and had an FSHD Clinical Score greater than zero. Patients had to be Android smartphone owners and willing to use either their own smartphone or an Android smartphone provided by CHDR for the duration of the study period. Patients with internal medical devices such as a pacemaker or deep brain stimulator were excluded from the study, as these could interfere with the Withings scale measurements.[20] Participants could not be pregnant or have a severe coexisting illness.

ETHICS APPROVAL

This study was approved by the Ethics Committee of BEBO, Assen, the Netherlands (NL69288.056.19) and was registered on ClinicalTrials.gov (NCT04999735). Before any study-related activities, written informed consent was obtained from the patients. Participants received monetary compensation for their time and effort during the trial.

To preserve the privacy of the patients, we deidentified the data and limited the amount of personally identified information collected from the smartphone and the connected devices. The location coordinates of the GPS or the cellular networks were collected as relative coordinates (GPS coordinates with respect to another predetermined location). For the calls and SMS text messaging, only metadata are stored (ie, no actual phone calls or text is being processed and stored). The call and SMS text messaging logs only store a partial phone number, making it impossible to identify the original phone numbers. As for the Withings devices, we created a unique email address (containing patient identifiers) for each patient to couple the Withings device with CHDR MORE, thus eliminating the need for using the patients' personal email.

INVESTIGATIONAL TECHNOLOGIES

Smartphone and remote sensor data were collected on the CHDR MORE platform. This customizable platform enables the collection, ingestion, and management of data sourced from monitoring devices. The CHDR MORE app was installed on the smartphone of each participant and allows for the unobtrusive collection of smartphone sensor data (sourced from the smartphone's accelerometer, gyroscope, magnetometer, GPS, light sensor, and microphone) as well as phone usage logs (eg, app usage, battery level, calls, and SMS text messages).

The smartphone sensor data provide insights into a participant's environment, such as location type and travel patterns (GPS), if human voices are present in the environment (microphone), and their physical activity (accelerometer and gyroscope). The phone usage logs give an indication of social activity (through social media and communication apps, calls, and SMS text messages) and smartphone usage (app usage). The app also collected Withings health data.

In this study, 3 Withings devices were used: Withings Steel HR smartwatch (monitors heart rate, sleep states, and a number of steps), Withings Body+ scale (monitors weight and body composition) and Withings BPM Connect (monitors heart rate, systolic blood pressure, and diastolic blood pressure). Together the Withings features reflect the daily physical activities of each of the participants.

This is the first study that aimed to monitor and estimate FSHD symptom severity using smartphone and wearable data. As this was an exploratory longitudinal study, specifically aimed to identify smartphone- and wearable-based features that were predictive of FSHD symptom severity, we did not identify any literature with a similar protocol. To identify these novel features, we decided to collect data from all available sensors and features from the CHDR MORE platform. As the symptoms of FSHD can

affect a patient's travel abilities,[21] physical activity, sleep,[11,22] and social lives,[23] we deemed these features relevant for estimating FSHD symptom severity.

DATA COLLECTION

Participants were monitored for 6 continuous weeks. On days 1 and 42, the clinical evaluations (FSHD Clinical Score and TUG) were performed. On day 1, the CHDR MORE and Withings Health Mate apps were installed on their smartphones. Participants were asked to use their smartphones as normal. Participants were asked to continuously wear their Withings Steel HR smartwatch and weigh themselves and take their blood pressure weekly.

DATA PREPROCESSING

Before modeling of the data, all sensor data were preprocessed and converted into features using Python (version 3.6.0) and the PySpark (version 3.0.1) library. The raw data were checked for missing values and outliers. Missing values were defined as the absence of data for a specific feature for each day, except for 2 types of measurements: the weekly measurements (eg, weight and blood pressure) and the data related to aperiodic activities (eg, phone calls or SMS text messages). Missing data were not imputed. Outliers were detected by manual visual inspection rather than automated statistical techniques, as our objective was to identify potential outliers that were a result of potential measurement errors rather than participants' behaviors. Measurement errors were deemed not relevant to our analysis, whereas outliers in behavior could still provide insights into a participant's symptom severity; therefore, sensitivity analysis was not conducted. Outliers would be subsequently excluded at the discretion of the authors (eg, removing overlapping sleep stages).

FEATURE EXTRACTION

All raw data were collected from the smartphone and Withings devices. The features were then aggregated per day, as the symptom severity exhibited on a given day is the focus of FSHD clinical evaluation. As there are no FSHD assessments that assess FSHD symptoms over a longer period, we did not explore other aggregation methods. Discrete features (eg, step count) were summed per day per participant. Continuous features (eg, heart rate) were averaged per day per participant. Table 2 provides an overview of how the features were aggregated based on the data type. Table 3 summarizes which features were extracted from the smartphone and Withings sensors. In addition, Table 3 shows the features that were provided from the MORE platform but were not included for the analysis either due to outliers, missing data, or because they were not of clinical interest.

FEATURE SELECTION

Before modeling, both expert-based manual and automated feature selections were performed. First, features were visually inspected by all authors. Excluded features were based on the number of available data points (eg, 9 participants did not have body composition data) and clinical relevance (eg, time spent on parenting apps was deemed clinically irrelevant). Next, two automated feature selection strategies were compared: (1) stepwise regression and (2) variance inflation factor (VIF). The stepwise regression strategy was an iterative process to select predictive variables that met a significance criterion ($P<.05$). Both forward and backward stepwise regression strategies were used. The VIF was calculated for all pairwise combinations of features to identify collinear features. Pairs of features having a VIF value greater than 10 were identified, and one of the features was subsequently removed for each of the pairs.[24] For comparison, we also fitted the model without any automated feature selection strategies. For each regression model, we applied each of the feature selection strategies.

STATISTICAL ANALYSIS

Python (version 3.6.0) was used for the data analysis and modeling in conjunction with the Pandas,[25] NumPy,[26] Matplotlib,[27] and Sklearn

packages.[28] Three regression models were created: 2 single-task regression models, 1 for each clinical assessment and 1 for each multitask regression model, simultaneously estimating both clinical assessments. For the multitask regression model, a dummy variable was included to denote either the FSHD Clinical Score or TUG.

For all models, linear regression, random forest regressor, and gradient boost regressor were used. A grid search was performed to optimize the hyperparameters for each model. For the Elastic Net linear regression model, we optimized the hyperparameters for the α (range 0-200) and L1 ratio (range 0.0-1.0). For the random forest and gradient boost regressors, we optimized the hyperparameters for the number of estimators (range 0-200), maximum depth (range 1-20), maximum features (range: auto, square root, log2), and maximum leaf nodes (range 2-20). In addition, we optimized the learning rate (range 0.0-1.0) for the gradient boost regressor.

Each model was validated using a group 5 outer-fold and 5 inner-fold nested cross-validation. By using group cross-validation, for each fold, we ensure that the participants in the validation are not also present in the training fold. While the data for all participants were used for the modeling, the cross-validation procedure was used for out-of-sample testing; hence, for each fold of the cross-validation procedure, only a subsample of participants' data were used. Further, the random forest and gradient boost regressor models only consider a subsample of participants and features per decision tree node. The elastic-net linear regression penalization would also reduce the potential features considered in the model. The cross-validation and models together would improve the generalizability and robustness of the models and therefore reduce the probability of spurious correlations.

We applied each of the feature selection strategies to each of the regression models and compared the results of each model. The model that provided the highest R2 (variance explained) and the lowest root-mean-square error (RMSE) was selected as the best-performing model. The R2 and the RMSE explain the variance and the error between the true clinical scores and the predicted scores of the regression models, respectively.

To assess how varying time window affects the model's estimation of symptom severity, we used an incrementally increasing time window to train the regression models, starting with day 1 and adding the following days until the first 2 weeks of data were included in the training set. To train, optimize, and assess each model's generalizability, we applied a 5-fold nested cross-validation model. To validate the performance of these models, we used the remaining 4 weeks of data as an external validation data set. To assess the stability of the trained models to yield consistent estimations of symptom severity, we trained the FSHD Clinical Score, TUG, and multitask models on the first week of data. We estimated the symptom severity for the subsequent weeks. We selected the first week, as each patient would have each day of the week represented in their data set.

In sum, we investigated 3 final models, 2 single-task models, and 1 multitask model. For each model, we considered 3 types of regression models (the linear regression, the random forest regressor, and the gradient boost regressor). For each model, we considered 3 feature selection strategies (no automated feature selection, stepwise regression, and VIF); hence, in total, we compared 27 models. Given that we are mainly interested in the comparison of the predictions of single-task and multitask models and the influence of the time windows on the predictions, we reported only the results of these models.

Results

No patients dropped out of the study. One patient was wheelchair-bound and therefore unable to perform the TUG. The FSHD Clinical Scores ranged between 1 and 13, with a median score of 5. The TUG times ranged between 5.5 seconds and 15.8 seconds, with a median time of 7.7 seconds. Before modeling, several features were manually excluded. Nine patients had

no body composition (eg, fat and muscle mass) data. As a result, the Withings body composition data (except weight) were excluded from the final analysis. We excluded SMS text message–related features as not all the patients used SMS text messaging (less than 30% of patients), and the SMS text message features were not deemed clinically relevant. Further, we excluded smartphone apps from the analysis that were used by less than 5% of the patients. We did not exclude any outliers as none of the data points were viewed as potential measurement errors. In a previous publication, we provided an overview of the proportion of observations that were missing per feature.[15]

The FSHD Clinical Score for 24 participants did not change over the 6 weeks. The scores of the remaining 14 participants changed by +1 or −1 point. The average difference between the day 1 and day 42 TUG scores was 0.38 seconds (95% CI 0.12-0.63). After reviewing the stability of the TUG and FSHD scores, we decided to use the averaged clinical assessment scores as the outcomes for all models. Subsequently, each feature was also averaged over the 6 weeks. These averaged features were used as inputs for the regression models.

Using all 6 weeks of data, we built a single-task model that used the CHDR MORE features to estimate the FSHD Clinical Score for each participant. Comparing the estimated scores and the true FSHD Clinical Score yielded an R2 of 0.57 and an RMSE of 2.09. This was achieved using VIF-selected features and Elastic Net–penalized linear regression. A total of 11 features were predictive of the FSHD Clinical Score, as seen in Figure 1. The features were related to app usage, blood pressure, location visits, and calling behaviors. Figure 2 (top) shows the estimated FSHD Clinical Score in relation to the actual FSHD Clinical Score.

Similarly, the comparison of the TUG single-task model estimated TUG and the actual TUG yielded an R2 of 0.59 and an RMSE of 1.66 (seconds) for each participant. This was achieved with forwarding selection stepwise regression and Elastic Net–penalized linear regression. In total, 13 feature were predictive of the TUG score (Figure 1). The feature categories related to age, app usage, calling behaviors, sleep, physical activity, and location visits were predictive of TUG. Figure 2 (bottom) illustrates the relationship between the predicted and actual TUG times.

The multitask model achieved an R2 of 0.74 and an RMSE of 1.89 for the FSHD Clinical Score and TUG prediction together. The same model achieved an R2 of 0.66 and an RMSE of 1.97 for the FSHD Clinical Score and an R2 of 0.81 and an RMSE of 1.61 for the TUG separately. The gradient boost regressor selected 50 predictive features. The relative feature importance is presented in Figure 3. The 5 most important features were light sleep duration, total steps per day, mean steps per minute, the number of times the social and communication apps were opened, and the number of incoming calls. Figure 4 illustrates the relationship between the predicted clinical scores and the actual clinical scores.

For each clinical score, we evaluated the effect of different monitoring periods on the estimation of symptom severity. The best performing FSHD Clinical Score single-task model, TUG single-task model, and multitask model yielded the highest R2 on day 3 (0.70), week 2 (0.86), and day 1 (0.86), and the lowest RMSE on day 3 (2.8), week 2 (1.9), and day 6 (3.4), respectively. As seen in Figure 5, although our analysis has identified windows that yielded the highest R2 and RMSE, we found that the mean (SD) of the R2 and RMSE for the FSHD Clinical Score single-task model, TUG single-task model, and multitask model was 0.65 (0.03) and 3.37 (0.19), 0.79 (0.05) and 2.05 (0.09), and 0.76 (0.08) and 4.37 (0.20), respectively. When evaluating the stability, the models trained on a week's worth of data were used to estimate the symptom severity for subsequent days. We found that the FSHD Clinical Score, TUG, and multitask models achieved median R2 (median RMSE) of 0.51 (3.66), 0.42 (2.44), and 0.72 (2.61), respectively (as seen in Figure 6).

Discussion

PRINCIPAL FINDINGS

We developed and compared 2 regression models to monitor and estimate FSHD symptom severity outside the clinic with remote sensor data

to estimate the FSHD Clinical Score and TUG for each participant. For the first type of model, both clinical assessment scores were separately estimated using 2 single-task regression models. For the second type of model, both clinical assessment scores were simultaneously estimated using a multitask regression model.

The 2 single-task models selected features that were uniquely predictive of each of the clinical scores. In addition, the models' selected features were found to be predictive for both scores (time spent at health locations and total call duration). Other studies have found that (a modified version of) the TUG significantly correlated to the FSHD Clinical Score,[12,29] indicating that these clinical scores share mutual information. Simultaneously estimating multiple tasks with shared features can improve the model performance.[30-32] This supports the notion that a multitask approach would improve the estimation of FSHD symptom severity.

Indeed, the multitask modeling of both the FSHD Clinical Score and the TUG outperformed the single-task models. Additionally, the multitask model identified features not selected as important by the single-task models (eg, sleep and the resting heart rate). The clinical assessments and their respective single-task models only captured a limited range of disease symptoms, which misses the opportunity to model other aspects of the disease (eg, sleep impairments [33,34] and arrhythmic abnormalities [35]). The multitask model, however, identified features representative of a broader range of FSHD symptoms. As shown in the SHAP (SHAPley Additive exPlanations) plot (Figure 3), participants with a higher mean step per minute, light sleep duration, soft activity duration, and total steps (indicated by the red feature value) had lower SHAP values. This indicates that participants with more physical activity and better sleep quality had a lower FSHD Clinical Score and TUG. Although the multitask model outperformed the single-task models, the multitask model required approximately twice as many features as the single-task models. Using fewer features could be considered beneficial as it reduces the number of sensors needed. Additionally, it eases the interpretation of the results. Therefore, there is a tradeoff between the performance of estimation of disease severity and the complexity of the data set and model. However, given that the multitask model showed an important improvement over the single-task models, we recommend using the multitask model for future estimation of the FSHD Clinical Score and TUG.

It is critical to determine how much data are needed to obtain reliable inferences without burdening the patients and the clinicians. Insufficient data can lead to inaccurate extrapolations, whereas excessive data can lead to wasted time and resources. This study investigated how long a patient needs to be monitored to estimate symptom severity reliably. Our results demonstrated that behaviors exhibited that based on our sample, the optimal time window (based on the highest R2 and lowest RMSE) varied for each task. The multitask model yielded the overall highest R2 based on a training data set of the first day. Although we identified that 5 days of data seem sufficient for training the multitask model, a longer or shorter time window would still provide consistent estimation of the symptom severity. However, our results also demonstrate that selecting any time window between days 1 and 14 would produce relatively stable results. Our results also demonstrated that training the multitask model on the first week of data allowed for constant and reliable estimations of symptom severity for the subsequent weeks. This further supports the notion that multitask should be used to estimate the clinical scores for longitudinal studies.

The agreement between the clinical scores and the remotely monitored features did not achieve 100% adherence. This may be due to the sensors being unable to capture specific aspects of the clinical score. For example, features captured by the remote monitoring system may not provide sufficient proxies for arm, scapular, and abdominal weaknesses (which the FSHD Clinical Score specifically addresses). Adding additional sensors and features could potentially allow for more complete modeling of FSHD. For example, an additional accelerometer could try to capture arm swings [36] or detect the (limited) shoulder range of motion.[37] Another explanation for the imperfect model fit is that the clinical scores have limited accuracy in capturing disease severity. There can be variation within

a specific clinical score, as patients with the same scores may exhibit different FSHD symptoms. For example, patients with scores between 2 and 4 may have impairments related to facial muscles and upper limbs, whereas others may be unable to walk on their heels.[11]

The clinical scores provide snapshots of muscular strength and function, whereas the remote monitoring approach provides a more continuous measure of (FSHD-related) social and physical activity. Additionally, the clinical scores were assessed at the clinic, whereas the sampling of the remotely monitored features occurred at home, and in daily practice. Altogether, these 2 clinical scores may not be the optimal clinical assessment strategies for fully assessing FSHD symptom severity. These are only 2 of several FSHD-related assessments that can be used in a clinical trial. The remotely monitored features may show different correlations with other FSHD-related assessments such as the Clinical Severity Scale for FSHD [38,39] and the Pittsburgh Sleep Quality Index.[39,40] Although the remotely monitored features may not correlate strongly with the 2 clinical scores, they still provide relevant insights into FSHD-related symptoms. Our multitask model could prove to be a promising tool for monitoring the FHSD severity based on patients' everyday activities outside the clinic.

Although the models cannot replace the TUG or FSHD Clinical Scores for estimating the disease severity, these models can potentially be used as a (complimentary) tool in clinical studies. When validated in longitudinal studies, given the continuous sampling of data from multiple sensors, this FSHD tool could potentially be used to track the symptom severity for long periods of time without patients having to visit a clinic. Previous studies have demonstrated that this approach of using smartphone-based models to quantify medication responses can be advantageous.[37,38] When implemented in a clinical trial, the FSHD tool might be evaluated as a tool to monitor drug effectiveness by tracking drug-induced changes in FSHD symptom severity.[41] Additionally, it might enable the identification of improvements in specific aspects of the disease severity (e.g., muscle function or sleep quality). Therefore, remote monitoring might aid clinicians' assessments of a patient's status during a clinical trial based on the review of the patient's in-clinic assessments and out-of-clinic daily activity.

We present an FSHD tool that estimates the FSHD Clinical Score and TUG using smartphone and remote sensor data. The conclusions drawn from this study are preliminary in view of the relatively small sample size and cross-sectional study nature. Given the short observation period, we did not expect changes in the patients' FSHD scores. As a result, we could not validate the use of the model to estimate changes in the FSHD severity over time. A trial where the FSHD clinical score is expected to change could help validate the FSHD tool's capacity to detect changes in FSHD symptom severity. Additionally, the FSHD tool could be improved by including more patients with FSHD and adding other remote sensors. All in all, the remote monitoring approach presented here could be a promising tool for monitoring FSHD severity outside the clinic environment.

Conclusions

We presented a smartphone-based and remote sensor–based FSHD tool that can estimate a patient's FSHD symptom severity. This is the first study to demonstrate how to monitor patients with FSHD remotely and subsequently model their FSHD Clinical Score and TUG simultaneously. The tool holds potential for monitoring disease progression and drug intervention effects outside the clinic, pending a longitudinal follow-up study to validate the capacity of the FSHD tool to detect changes in the disease severity score over time due to disease progression or drug intervention.

REFERENCES

1 Statland JM, McDermott MP, Heatwole C, Martens WB, Pandya S, van der Kooi E, et al. Reevaluating measures of disease progression in facioscapulohumeral muscular dystrophy. Neuromuscul Disord 2013;23(4):306-312 [doi:10.1016/j.nmd.2013.01.008] [Medline: 23406877]

2 Tawil R, Van Der Maarel SM. Facioscapulohumeral muscular dystrophy. Muscle Nerve 2006;34(1):1-15. [doi: 10.1002/mus.20522] [Medline: 16508966]

3 Statland JM, Tawil R. Risk of functional impairment in facioscapulohumeral muscular dystrophy. Muscle Nerve 2014;49(4):520-527. [doi: 10.1002/mus.23949] [Medline: 23873337]

4 Blokhuis AM, Deenen JCW, Voermans NC, van Engelen BGM, Kievit W, Groothuis JT. The socioeconomic burden of facioscapulohumeral muscular dystrophy. J Neurol 2021;268(12):4778-4788 [doi:10.1007/s00415-021-10591-w] [Medline: 34043041]

5 Hamel J, Johnson N, Tawil R, Martens WB, Dilek N, McDermott MP, et al. Patient-reported symptoms in facioscapulohumeral muscular dystrophy(PRISM-FSHD). Neurology 2019;93(12):e1180-e1192 [doi: 10.1212/WNL.0000000000008123] [Medline: 31409737]

6 Morís G, Wood L, FernáNdez-Torrón R, González Coraspe JA, Turner C, Hilton-Jones D, et al. Chronic pain has a strong impact on quality of life in facioscapulohumeral muscular dystrophy. Muscle Nerve 2018;57(3):380-387 [doi: 10.1002/mus.25991] [Medline: 29053898]

7 Tawil R, Mah JK, Baker S, Wagner KR, Ryan MM, Sydney Workshop Participants. Clinical practice considerations in facioscapulohumeral muscular dystrophy Sydney, Australia, 21 September 2015. Neuromuscul Disord 2015;26(7):462-471. [doi: 10.1016/j.nmd.2016.03.007] [Medline: 27185458]

8 Ramos VFML, Thaisetthawatkul P. A case of fascioscapulohumeral muscular dystrophy misdiagnosed as Becker's muscular dystrophy for 20 years. Age Ageing 2012;41(2):273-274. [doi: 10.1093/ageing/afr095] [Medline: 21795275]

9 Zhou L, Parmanto B. Reaching people with disabilities in underserved areas through digital interventions: systematic review. J Med Internet Res 2019;21(10):e12981 [doi: 10.2196/12981] [Medline: 31654569]

10 Churová V, Vyškovský R, Maršálová K, Kudláček D, Schwarz D. Anomaly detection algorithm for real-world data and evidence in clinical research: implementation, evaluation, and validation study. JMIR Med Inform 2021;9(5):e27172 [doi: 10.2196/27172] [Medline: 33851576]

11 Lamperti C, Fabbri G, Vercelli L, D'Amico R, Frusciante R, Bonifazi E, et al. A standardized clinical evaluation of patients affected by facioscapulohumeral muscular dystrophy: the FSHD clinical score. Muscle Nerve 2010;42(2):213-217. [doi: 10.1002/mus.21671] [Medline: 20544930]

12 Chan V, Hatch M, Kurillo G, Han J, Cadavid D. Development of an optimized timed up and go (OTUG) for measurement of changes in mobility impairment in facioscapulohumeral muscular dystrophy (FSHD) clinical trials (2228). Neurology 2020;94(15):2228.

13 Boukhvalova AK, Fan O, Weideman AM, Harris T, Kowalczyk E, Pham L, et al. Smartphone level test measures disability in several neurological domains for patients with multiple sclerosis. Front Neurol 2019;10:358 [doi: 10.3389/fneur.2019.00358] [Medline: 31191424]

14 Servais L, Camino E, Clement A, McDonald CM, Lukawy J, Lowes LP, et al. First regulatory qualification of a novel digital endpoint in Duchenne muscular dystrophy: a multi-stakeholder perspective on the impact for patients and for drug development in neuromuscular diseases. Digit Biomark 2021;5(2):183-190 [doi: 10.1159/000517411]. Medline: 34723071]

15 Maleki G, Zhuparris A, Koopmans I, Doll RJ, Voet N, Cohen A, et al. Objective monitoring of facioscapulohumeral dystrophy during clinical trials using a smartphone app and wearables: observational study. JMIR Form Res 2022;6(9):e31775 [doi: 10.2196/31775] [Medline: 36098990]

16 Jauhiainen M, Puustinen J, Mehrang S, Ruokolainen J, Holm A, Vehkaoja A, et al. Identification of motor symptoms related to Parkinson disease using motion-tracking sensors at home (KÄVELI): protocol for an observational case-control study. JMIR Res Protoc 2019;8(3):e12808 [doi: 10.2196/12808] [Medline: 30916665]

17. Jeannet PY, Aminian K, Bloetzer C, Najafi B, Paraschiv-Ionescu A. Continuous monitoring and quantification of multiple parameters of daily physical activity in ambulatory Duchenne muscular dystrophy patients. Eur J Paediatr Neurol 2011;15(1):40-47. [doi: 10.1016/j.ejpn.2010.07.002] [Medline: 20719551]

18 Zhong T, Zhuang Z, Dong X, Wong KH, Wong WT, Wang J, et al. Predicting antituberculosis drug-induced liver injury using an interpretable machine learning method: model development and validation study. JMIR Med Inform 2021;9(7):e29226 [doi: 10.2196/29226] [Medline: 34283036]

19 Chae SH, Kim Y, Lee K, Park H. Development and clinical evaluation of a web-based upper limb home rehabilitation system using a smartwatch and machine learning model for chronic stroke survivors: prospective comparative study. JMIR Mhealth Uhealth 2020;8(7):e17216 [doi: 10.2196/17216] [Medline: 32480361] 20. Body+ - Frequently asked questions about safety. Withings. URL: https://support.withings.com/hc/en-us/articles/218438708-Body-Frequently-asked-questions-about-safety [accessed 2021-02-24]

21 Faux-Nightingale A, Kulshrestha R, Emery N, Pandyan A, Willis T, Philp F. Upper limb rehabilitation in facioscapulohumeral muscular dystrophy: a patients' perspective. Arch Rehabil Res Clin Transl 2021;3(4):100157 [doi: 10.1016/j.arrct.2021.100157] [Medline: 34977539]

22 Mul K, Lassche S, Voermans NC, Padberg GW, Horlings CG, van Engelen BG. What's in a name? The clinical features of facioscapulohumeral muscular dystrophy. Pract Neurol 2016;16(3):201-207. [doi: 10.1136/practneurol-2015-001353] [Medline: 26862222]

23. Johnson NE, Quinn C, Eastwood E, Tawil R, Heatwole CR. Patient-identified disease burden in facioscapulohumeral muscular dystrophy. Muscle Nerve 2012;46(6):951-953 [doi: 10.1002/mus.23529] [Medline: 23225386]

24 Hair JFJ, Anderson RE, Tatham RL, Black WC. F. In: Multivariate Data Analysis, 3rd Ed. New York: Macmillan; 1995.

25 McKinney W. Data structures for statistical computing in python. 2010 Presented at: Proceedings of the 9th Python in Science Conference; June 28-July 3, 2010; Austin, Texas. [doi: 10.25080/majora-92bf1922-00a]

26 Oliphant TE. A Guide to NumPy. USA: Trelgol Publishing; 2006.

27 Hunter JD. Matplotlib: a 2D graphics environment. Comput Sci Eng 2007;9(3):90-95. [doi: 10.1109/mcse.2007.55]

28 Buitinck L, Louppe G, Blondel M, Pedregosa F, Muller AC, Grisel O, et al. API design for machine learning software: experiences from the scikit-learn project. ArXiv. Preprint posted online September 1, 2013 2013:108-122.

29 Huisinga J, Bruetsch A, Mccalley A, Currence M, Herbelin L, Jawdat O, et al. An instrumented timed up and go in facioscapulohumeral muscular dystrophy. Muscle Nerve 2018;57(3):503-506 [doi: 10.1002/mus.25955] [Medline: 28877559]

30 Yu G, Liu Y, Shen D. Graph-guided joint prediction of class label and clinical scores for the Alzheimer's disease. Brain Struct Funct 2016;221(7):3787-3801 [doi: 10.1007/s00429-015-1132-6] [Medline: 26476928]

31. Li Y, Tian X, Liu T, Tao D. Multi-task model and feature joint learning. 2015 Presented at: Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence; 25-31 July, 2015; Buenos Aires, Argentina.

32 Zhang D, Shen D, Alzheimer's Disease Neuroimaging Initiative. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. Neuroimage 2012;59(2):895-907 [doi:10.1016/j.neuroimage.2011.09.069] [Medline: 21992749]

33 Runte M, Spiesshoefer J, Heidbreder A, Dreher M, Young P, Brix T, et al. Sleep-related breathing disorders in facioscapulohumeral dystrophy. Sleep Breath 2019;23(3):899-906. [doi: 10.1007/s11325-019-01843-1] [Medline: 31025273]

34 Leclair-Visonneau L, Magot A, Tremblay A, Bruneau X, Pereon Y. Anxiety is responsible for altered sleep quality in facio-scapulo-humeral muscular dystrophy (FSHD). Neuromuscul Disord 2013;23(9-10):823-824. [doi:10.1016/j.nmd.2013.06.642]

35 Emre Cagliyan C, Gelincik's H, Celic AI, Filiz Koc A. Impaired autonomic and repolarization abnormalities are observed in patients with facioscapulohumeral dystrophy despite normal myocardial functions. J Neurol Neurosurg 2018;05(01):127-131. [doi: 10.19104/jnn.2018.42]

36 LeMoyne R, Tomycz N, Mastroianni T, McCandless C, Cozza M, Peduto D. Implementation of a smartphone wireless accelerometer platform for establishing deep brain stimulation treatment efficacy of essential tremor with machine learning. Annu Int Conf IEEE Eng Med

Biol Soc 2015;2015:6772-6775. [doi: 10.1109/EMBC.2015.7319948] [Medline: 26737848]
37 Boissy P, Diop-Fallou S, Lebel K, Bernier M, Balg F, Tousignant-Laflamme Y. Trueness and minimal detectable change of smartphone inclinometer measurements of shoulder range of motion. Telemed J E Health 2017;23(6):503-506. [doi: 10.1089/tmj.2016.0205] [Medline: 27911652]
38 Ricci E, Galluzzi G, Deidda G, Cacurri S, Colantoni L, Merico B, et al. Progress in the molecular diagnosis of facioscapulohumeral muscular dystrophy and correlation between the number of KpnI repeats at the 4q35 locus and clinicalphenotype. Ann Neurol 1999;45(6):751-757. [doi: 10.1002/1531-8249(199906)45:6<751::aid-ana9>3.0.co;2-m] [Medline: 10360767]
39 Della Marca G, Frusciante R, Vollono C, Dittoni S, Galluzzi G, Buccarella C, et al. Sleep quality in facioscapulohumeral muscular dystrophy. J Neurol Sci 2007;263(1-2):49-53. [doi: 10.1016/j.jns.2007.05.028] [Medline: 17597162]
40 Buysse DJ, Reynolds CF, Monk TH, Berman SR, Kupfer DJ. The Pittsburgh sleep quality index: a new instrument for psychiatric practice and research. Psychiatry Res 1989;28(2):193-213. [doi: 10.1016/0165-1781(89)90047-4] [Medline: 2748771]
41 Zhang H, Guo G, Song C, Xu C, Cheung K, Alexis J, et al. PDLens: smartphone knows drug effectiveness among Parkinson's via daily-life activity fusion. 2020 Presented at: MobiCom '20: The 26th Annual International Conference on Mobile Computing and Networking; 21-25 September, 2020; London, United Kingdom. [doi: 10.1145/3372224.3380889]

TABLE 1 An overview of characteristics of the FSHD participants (N=38).

Demographics	Values
GENDER, N	
Female	23
Male	15
RACE, N	
African American	-
Mixed	1
White	37
Age (years), mean (SD) (minimum, maximum)	44 (14.5) (18, 64)
Weight (kg), median (SD) (minimum, maximum)	79 (16) (52, 130)
BMI (kg/m2), median (SD) (minimum, maximum)	25 (4) (20, 44)
FSHD Clinical Score, median (SD) (minimum, maximum)	5 (3) (1, 13)
Timed Up and Go test (seconds), median (SD) (minimum, maximum)	7.7 (2.4) (5.5, 15.8)

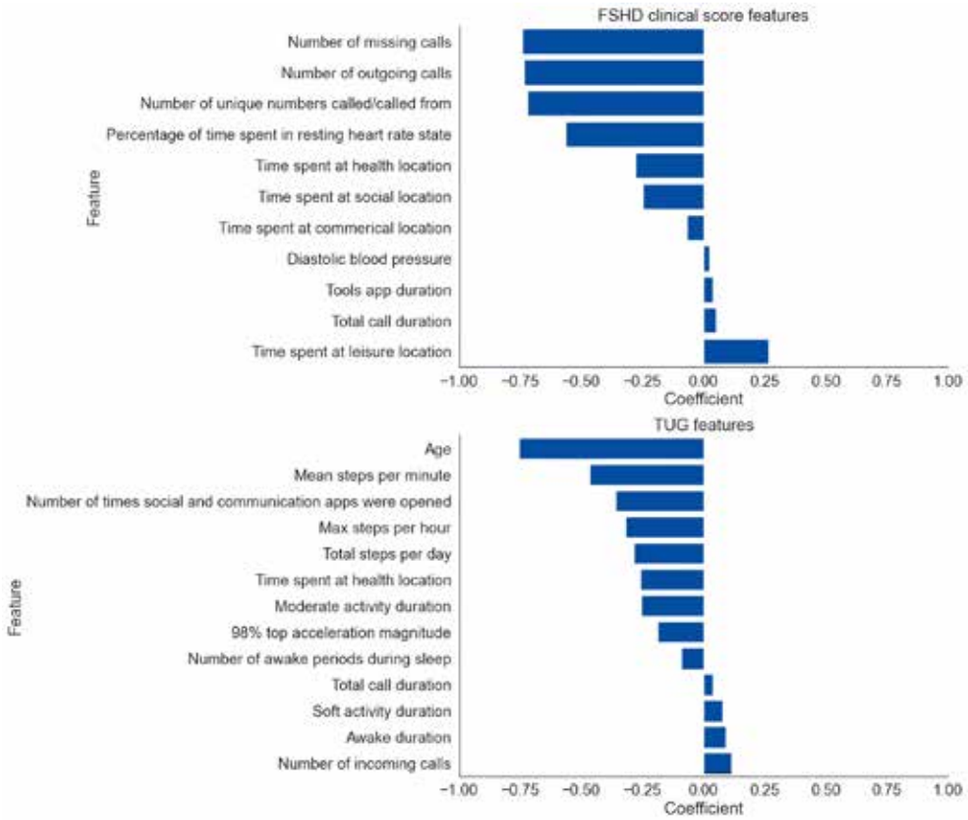
TABLE 2 A simplified summation of how the features were aggregated based on the data type.

Data Type	Time Unit	Example Feature	Aggregation Format	Example Aggregation
COUNT	Per day Per hour	Steps	Sum Mean Max	Total Steps Max Steps Per Hour Mean Steps Per Hour
CONTINUOUS DATA WITHIN A RANGE	Per day	Heart Rate	Min (5%) Median (50%) Max (95%)	Lowest 5% Heart Rate Median Heart Rate Maximum 95% Hr
DURATION	Per day	App Usage	Total Duration Mean Duration	Total Duration of Social Apps Opened Mean Duration of Social App Use Per Interaction
GPS COORDINATES	Per day	Location	Sum Max Mean	Total Distance Travelled Mean And Max Distance From Home

TABLE 3 An overview of the features provided from the MORE platform and the features that were subsequently aggregated per day (with the exception of the body measurements as that was measured once a week).

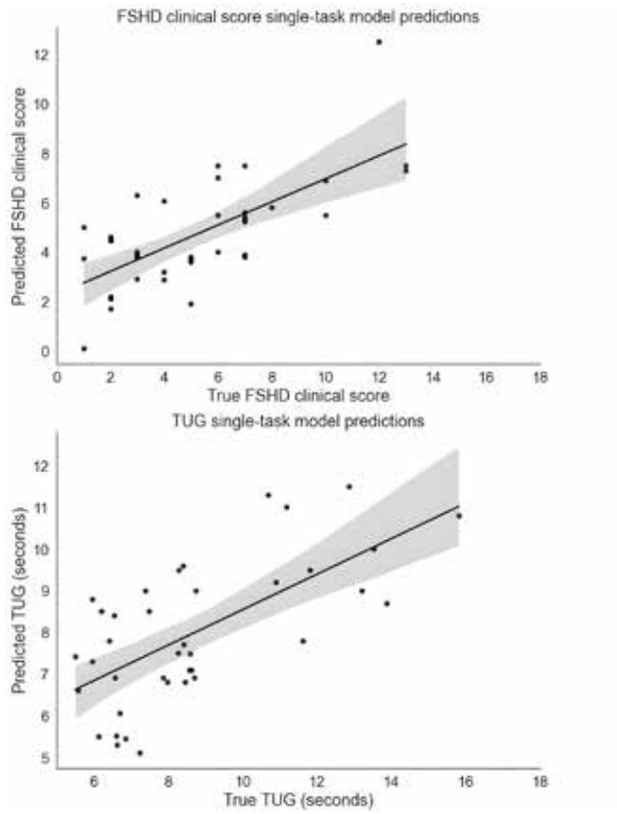
Category	MORE Features	Derived Features (Per day)	Excluded Features
DEMOGRAPHICS	Age Gender	Age Gender	
ACCELERATION	Acceleration magnitude Gyroscope Magnetometer	98% Acceleration magnitude	Mean Acceleration Magnitude
ACTIVITY	Steps Heart Rate Physical activity duration Calories	**STEPS:** Total steps, max steps per hour, mean steps per hour **HEART RATE:** 5%, 50% & 95% beats per minute (BPMs), standard deviation of BPMs, % time spent in resting state **PHYSICAL ACTIVITY:** soft, moderate and intense activity duration	Calories Distance Travelled Distance Per Step
APPS	APP CATEGORIES: Health & Fitness, Recreational, Communication & Social, Tools, Shopping	Duration Times Open	House & Home Libraries & Demo Reading Travel
BODY	Diastolic Blood Pressure Systolic Blood Pressure Heart Pulse (Bpm) Weight	Diastolic blood pressure Systolic blood pressure Heart pulse (bpm) Weight	Height (M) Fat mass (kg) Fat ratio (%) Hydration Muscle Mass
LOCATION	LOCATION CATEGORIES: Commercial, Health, Home, Leisure, Public, Social, Travel	Total duration at place Total distance travelled Total no of unique places visited Max distance from home Time spent commuting	
SOCIAL	Calls Voice	Number of calls Number of unique numbers Number of incoming, outgoing and missing calls Number of calls from known & unknown numbers Total duration of calls Average duration of calls % Time human voice is detected	Text messages (SMS)
SLEEP		Number of sleep sessions Total sleep duration Number of sleep phases (awake, light sleep and deep sleep) Duration of sleep phases (awake, light and deep sleep) Time between sleep sessions Time to fall asleep	

FIGURE 1 Linear regression coefficients for the features selected by the single-task FSHD Clinical Score and TUG models. Features with a coefficient of zero are not shown.



FSHD: facioscapulohumeral muscular dystrophy; TUG: Timed Up and Go.

FIGURE 2 True FSHD Clinical Scores and TUG times against the predicted scores using the respective FSHD Clinical Score and TUG regression models. The lines represent a regression line with a 95% CI band.



FSHD: facioscapulohumeral muscular dystrophy; TUG: Timed Up and Go.

FIGURE 3 SHAP (SHAPley Additive exPlanations) variable importance plot showing the feature importance of the top 20 most important features, in which the features are ranked in descending order. Each scatter point represents one prediction. The color of the scatter point reflects the value of the real data. If the actual value of the data point was high, then the color was red. If the value was low, then the color was blue. The SHAP value, as illustrated on the x-axis, shows the direction and magnitude of each feature's contribution toward predicting the facioscapulohumeral muscular dystrophy symptom severity.

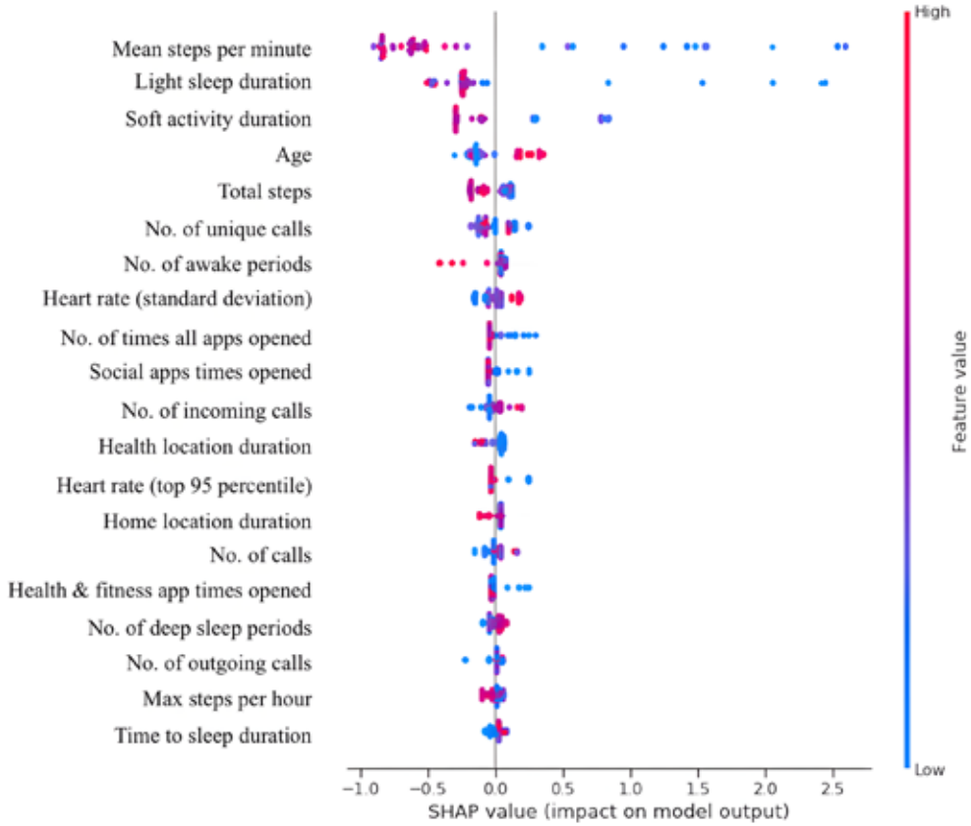
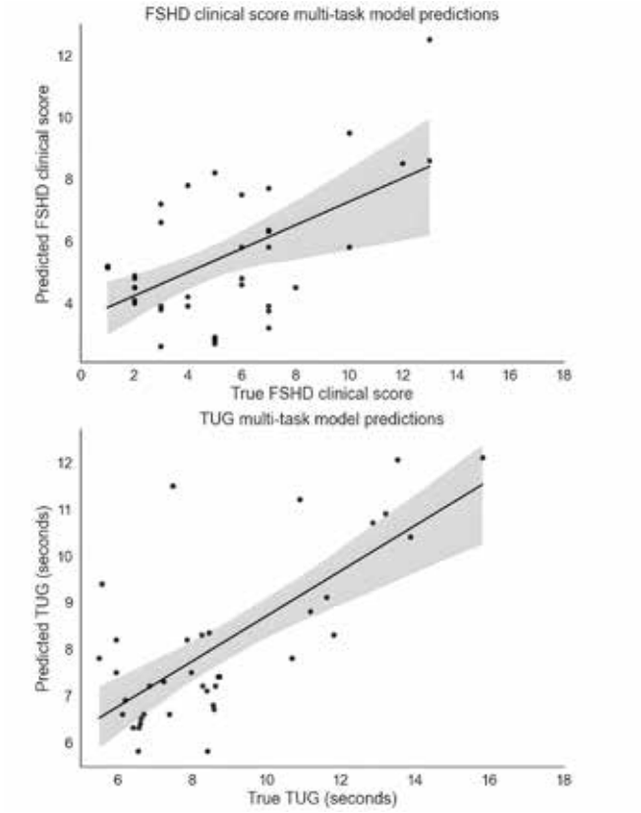
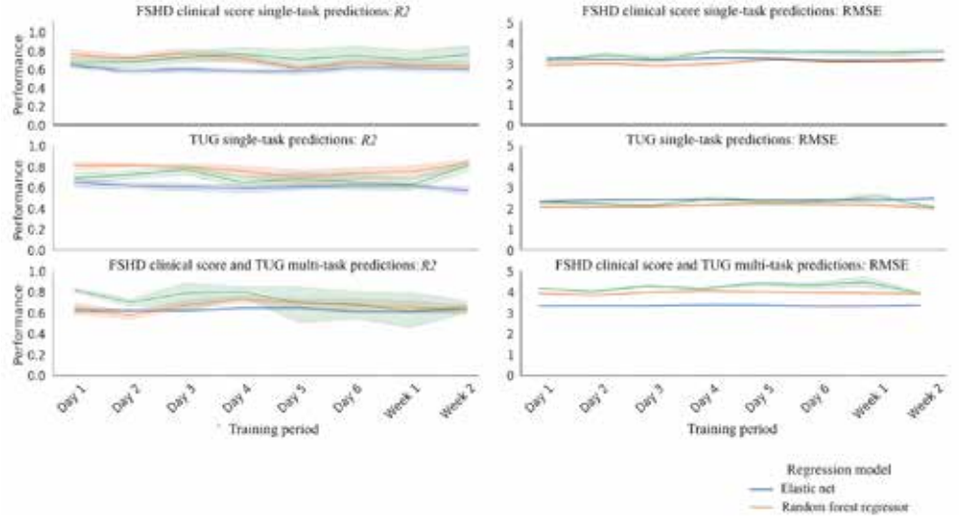


FIGURE 4 Scatterplot of the estimated FSHD Clinical Scores and TUG times in relation to the actual FSHD Clinical Scores and TUG using the multi-task learning regression model. The lines represent the regression lines with a 95% CI band.



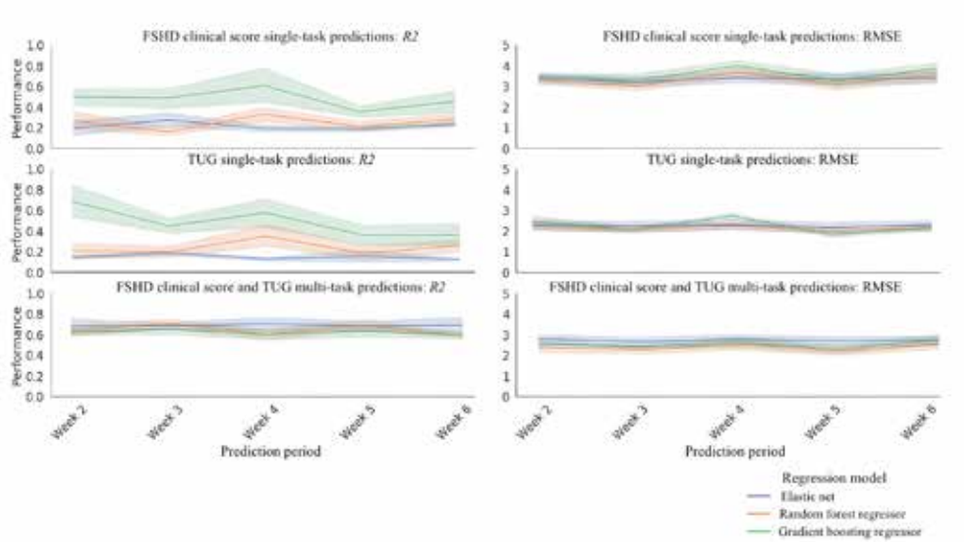
FSHD: facioscapulohumeral muscular dystrophy; TUG: Timed Up and Go.

FIGURE 5 Evaluating the performance of the single-task FSHD Clinical Score, TUG, and the multitask FSHD Clinical Score and TUG regression models trained on an incrementally increasing time window. The colored lines represent the 3 types of regression models trained on the data (Elastic Net, Random Forest Regressor, and Gradient Boosting Regressor). For each model and each incremental time window, the top and bottom plots show the R2 and RMSE, respectively. The lines represent the median performance, and the bands represent the 95% CI.



FSHD: facioscapulohumeral muscular dystrophy; RMSE: root mean square error; TUG: Timed Up and Go.

FIGURE 6 Evaluating the performance of the single-task FSHD Clinical Score, TUG, and the multitask FSHD Clinical Score and TUG regression models trained on the first week of data to estimate symptom severity for the subsequent weeks. The colored lines represent the 3 types of regression models trained on the data (Elastic Net, Random Forest Regressor, and Gradient Boosting Regressor). For each model and each week, the top and bottom plots show the R2 and RMSE respectively. The lines represent the median performance, and the bands represent the 95% CI.



FSHD: facioscapulohumeral muscular dystrophy; RMSE: root mean square error; TUG: Timed Up and Go.

CHAPTER 5

A smartphone- and wearable-based biomarker for the estimation of unipolar depression severity

Ahnjili Zhuparris,[1,2] Ghobad Maleki[1,2], Liesbeth van Londen,[5] Ingrid Koopmans,[1,2] Vincent Aalten,[1,3] Iris E. Yocarini,[4] Vasileios Exadaktylos,[1] Albert van Hemert,[2] Adam Cohen,[1,2] Pim Gal,[1,2] Robert-Jan Doll,[1] Geert Jan Groeneveld,[1,2] Gabriël Jacobs,[1,3] Wessel Kraaij[4]

Sci Rep 13, 18844 (2023). doi:10.1038/s41598-023-46075-2

1 Centre for Human Drug Research (CHDR), Leiden, NL
2 Leiden University Medical Centre (LUMC), Leiden, NL
3 Department of Psychiatry, Leiden University Medical Center (LUMC), Leiden, NL
4 Leiden Institute of Advanced Computer Science (LIACS), Leiden, NL
5 Transparant Centre for Mental Health Care, Leiden, NL

Abstract

Drug development for mood disorders can greatly benefit from the development of robust, reliable, and objective biomarkers. The incorporation of smartphones and wearable devices in clinical trials provides a unique opportunity to monitor behavior in a non-invasive manner. The objective of this study is to identify the correlations between remotely monitored self-reported assessments and objectively measured activities with depression severity assessments often applied in clinical trials. 30 unipolar depressed patients and 29 age- and gender-matched healthy controls were enrolled in this study. Each participant's daily physiological, physical, and social activity were monitored using a smartphone-based application (CHDR MORE) for 3 weeks continuously. Self-reported Depression Anxiety Stress Scale-21 (DASS-21) and Positive and Negative Affect Schedule (PANAS) were administered via smartphone weekly and daily respectively. The Structured Interview Guide for the Hamilton Depression Scale and Inventory of Depressive Symptomatology–Clinical Rated (SIGHD-IDSC) was administered in-clinic weekly. Nested cross-validated linear mixed-effects models were used to identify the correlation between the CHDR MORE features with the weekly in-clinic SIGHD-IDSC scores. The SIGHD-IDSC regression model demonstrated an explained variance (R2) of 0.80, and a Root Mean Square Error (RMSE) of ±15 points. The SIGHD-IDSC total scores were positively correlated with the DASS and mean steps-per-minute, and negatively correlated with the travel duration. Unobtrusive, remotely monitored behavior and self-reported outcomes are correlated with depression severity. While these features cannot replace the SIGHD-IDSC for estimating depression severity, it can serve as a complementary approach for assessing depression and drug effects outside the clinic.

Introduction

An ideal biomarker would serve as a dynamic indicator of disease activity. The biomarker should be capable of predicting changes in disease progression over time, regardless of the treatment intervention.[1,2] By leveraging advanced machine learning algorithms, researchers can integrate multiple objective biomarkers into composite biomarkers, enabling a more comprehensive and multifaceted understanding of disease activity and the impact of treatment interventions. Drug development for the treatment of depression is expected to benefit greatly from robust biomarkers that reflect the etiology, phenomenology, and treatment management of the disorder. Depression is not only associated with subjective symptoms such as sadness, despair, and anhedonia, but also with negative behavioral and neurovegetative effects such as decreased psychomotor activity and changes in appetite and sleep. A combination of objective physiological indicators and frequent subjective assessments can potentially be used as features to create a composite biomarker to estimate the presence or severity of depression, or even to quantify the effects of therapeutic interventions with drugs and/or psychotherapy.

The current gold standards for assessing depression severity and treatment effects, such as the Hamilton Depression Rating Scale (HAMD) and the Montgomery & Åsberg Depression Rating Scales (MADRS), are clinician-administered questionnaires.[3,4] As these questionnaires require an interview with a clinician, they are applied infrequently, and thus real-time behavioral assessments of depressed individuals cannot be obtained.[5] Further, retrospective self-reported appraisals can be compromised by recall bias and altered by socially desired reporting from patients.[6,7] By relying on the current gold standards for the assessment of depression severity, researchers routinely miss out on real-time and real-world behavioral patterns associated with depression, which may potentially attenuate treatment effects. To address such limitations, there is a demand for developing and validating methodologically sound biomarkers to quantify depression severity in real-time under free-living conditions.

Mobile health (MHEALTH) biomarkers are biomarkers derived from mobile health technologies, such as smartphones, wearables, and other portable devices that can be worn outside a controlled setting.[8] Emerging literature on depression and MHEALTH biomarkers supports the notion that smartphones and wearable devices can overcome the limitations of traditional depression rating scales. The sensors embedded in these devices (e.g., accelerometers, Global Positioning Systems (GPS), and microphones) provide real-time, unobtrusive, passively collected data relating to behavioral patterns exhibited under free-living conditions.[9–13] In turn, these data can offer insights into an individual's sleep rhythms,[14] social interactions,[15] and daily physical activities,[16] all of which can be useful for quantifying depression severity. While the existing body of evidence demonstrates that these digital MHEALTH biomarkers can be used to identify the presence of depressive symptoms or the estimation of daily mood, however, there are still three major critical gaps that remain to be understood. First, several studies in this field have relied on self-reported psychometric assessments, such as the Depression Anxiety and Stress Scale (DASS), the Positive and Negative Affect Schedule (PANAS), and Quick Inventory of Depressive Symptomatology (QIDS), for documenting depression severity.[17,18] To date, we have only identified two studies that correlated digital MHEALTH biomarkers sourced from smartphone and wearable data with clinician's assessment of depression among unipolar depressed patients.[19,20] Therefore, more evidence is required for corroborating the clinical validity of these remotely monitored biomarkers in depression clinical trials. Next, these studies rarely include age- or sex-matched non-depressed controls. Healthy controls can also present behaviors and symptoms observed among depressed patients.[21] Observing behaviors exhibited by both depressed and non-depressed controls enables the identification of behaviors specific to depressed patients. This allows for the discovery of new candidate drugs that target the core symptoms of unipolar depression. Lastly, determination of the optimal monitoring period and data resolution needed for developing depression biomarkers has been overlooked in previous studies. Depression is a highly variable and heterogenous disorder;[2] thus, an effective depression biomarker should consistently correspond with the heterogenous changes in depression over time. While the advances of remote sensing can provide researchers with fine-grain longitudinal datasets, it can be operationally and financially burdensome for patients and researchers to collect, store, and process such expansive and information-dense datasets. Therefore, evaluating how much data is required to identify the earliest, reliable, and minimally observable changes in the patients' clinical status is crucial. This evaluation is necessary to minimize the impact of data collection on both the patients and researchers.

The current study consisted of two research objectives. First, we investigated the correlation of clinical ratings of depression, among unipolar depressed patients and healthy controls, with remotely self-reported psychometric assessments and smartphone- and wearable-based features. Here, we defined features as individual measurable variables, such as average heart rate or total steps. Second, we examined how many data points are required to develop a reliable statistical model that can consistently estimate the longitudinal variability of depression. The primary objective allows for the identification of reliable and clinically relevant depression biomarkers that can be monitored continuously in real-world conditions. The secondary objective focuses on the validation of a minimum dataset required to maintain the accuracy, sensitivity, and specificity of the biomarkers. To achieve these objectives, we adopted linear mixed effects models to estimate the weekly Structured Interview Guide for the Hamilton Depression Scale and Inventory of Depressive Symptomatology (SIGHD-IDSC) clinician ratings using one, two, and three weeks of remotely collected data. Together, such correlated features can potentially represent a composite digital MHEALTH biomarker for monitoring depression severity in longitudinal clinical trials.

Methods

STUDY OVERVIEW

This was a cross-sectional, non-interventional pilot study conducted by Centre for Human Drug Research (CHDR) and Transparant Centre for Mental Health in Leiden, The Netherlands. The participants were monitored between March 2019 to March 2020. Prior to any assessments, patients provided written informed consent. The trial was approved by the Stichting Beoordeling Ethiek Biomedisch Onderzoek ethics committee, Assen, the Netherlands, and was conducted in accordance with the Declaration of Helsinki at the Centre for Human Drug Research, Leiden, the Netherlands.

PARTICIPANTS

Eligible patients and healthy controls were between the ages of 18-65 years old and had a Body Mass Index (BMI) between 18 to 30 kg/m.^{2} Patients and healthy controls with severe coexisting illnesses that might interfere with study adherence or pregnant were excluded. Patients and healthy controls were required to use their own Android smartphone (version 5.0 or higher) as the CHDR MORE app was only available on Android App Store. Due to the Apple operating systems restrictions, the iPhone user device logs could not be accessed by the app.

Eligible patients had either a diagnosis of Major Depressive Disorder (MDD) without psychotic features or Persistent Depressive Disorder (PDD) according to the DSM-IV (Diagnostic and Statistical Manual of Mental Disorders) or DSM-V. The diagnosis was provided by an attending general practitioner, psychologist, or psychiatrist and was confirmed with the Mini International Neuropsychiatric Interview (MINI) version 7.0. To be included in the study, each patient must have had a Structured Interview Version of Montgomery-Åsberg Depression Rating Scale (MADRS-SIGMA) score of more than 22 at screening. Further, the patients either received no antidepressant drug treatment at least 2 weeks prior to screening, or they were receiving an antidepressant drug treatment with a stable dose for at least 4 weeks prior to screening. Patients were excluded if they presented specific psychiatric co-morbidities (psychotic disorder, bipolar disorder, mental retardation, or cluster B personality disorders), presented a Columbia-Suicide Severity Rating Scale (C-SSRS) greater than 5, alterations of antidepressant drug (including dose) during the trial period or use of sedative medications within 2 weeks of the beginning of the clinical trial. This was confirmed by their general practitioner, psychologist, or psychiatrist.

Eligible healthy controls were included if they had no previous or current history (or family history) of psychiatric disorder or chronic co-morbidities. Healthy controls were age and sex-matched with the MDD and PDD patients.

Participants received monetary compensation for their time and effort. The reimbursement was determined by a schedule approved by the Ethics Committee and was based on the amount of time the participants had to spend participating in the study. This compensation was not linked to the quantity or quality of the data obtained.

CHDR MORE AND WITHINGS DEVICES

On Day 0 of the trial, the CHDR MORE,[23,24] Withings Healthmate,[25] and CHDR Promasys EPRO smartphone applications were installed on the participant's Android smartphones. The participants were also provided with a Withings Steel HR smartwatch. Training sessions were provided for the Withings devices and the Promasys EPRO application. All participants were monitored for 21 days continuously.

The CHDR MORE app enables the unobtrusive collection of data from multiple smartphone sensors (the accelerometer, gyroscope, Global Positioning System, and microphone) and the smartphone usage logs (app usage and calls). The Withings Healthmate app collects data from the Withings devices provided to the participants. The Steel HR smartwatch monitors the participants heart rate, sleep states, and step activity. The EPRO app prompted participants to fill in the Positive and Negative Affect Schedule (PANAS) twice daily and Depression, Anxiety and Stress Scale-21 (DASS-21) weekly. PANAS is a validated self-reported, brief and

easy to administer, 20-item questionnaire that assess positive and negative affect.[26] DASS-21 is a validated self-reported, 21-item measure of three negative emotional states: Depression, Anxiety and Stress.[27,28] More information about the apps and their respective sensors and features can be found in Supplementary Table 1.

CLINICAL ASSESSMENTS

The Structured Interview Guide for the Hamilton Depression Scale and Inventory of Depressive Symptomatology (SIGHD-IDSC) assessments were conducted weekly (Day 7, 14, and 21) for all participants in-person at CHDR by trained raters. The SIGHD-IDSC is a single and multi-faceted, and therefore efficient, assessment of depression. The SIGHD-IDSC interview is a combination of the 17-item Hamilton Depression Rating Scale (SIGH-D) and the 30-item Inventory of Depressive Symptomatology-Clinician Rated (IDS-C).[29,30] The SIGH-D assesses single symptoms on a continuous scale. It is a multidimensional scale that assesses a profile of factors relating to agitation, anxiety (psychic and somatic), guilt, libido, suicide, work, and interest.[31] However, the 17-item scale is still limited in terms of scope. Some symptoms which are often associated with depressed behaviors (such as hypersomnia, weight gain, and reactivity of mood) are not rated.[32] The IDS-C provides additional ratings relating to anxiety, anhedonia, mood, cognitive changes, and vegetative symptoms (relating to sleep, appetite, weight, and psychomotor changes).[32] Hence, we included the IDS-C as a complementary assessment to provide a broader assessment of depressive symptomatology. IDS-C has been shown to have a higher sensitivity to detect changes in depression severity, therefore deeming it more advantageous for monitoring changes in symptom severity, especially for depression-related drug trials.[33]

SIGHD-IDSC DIMENSIONS For this study, we investigated the correlation between the remotely monitored features with the total depression severity scores (SIGHD-IDSC) and the scores of individual symptom dimensions. Multiple approaches can be taken to transform the raw data, collected from smartphones and wearable devices, into clinically relevant features. As illustrated by Mohr et. al, raw sensor data can be converted in low-level features and high-level behavioral markers.[34] These features and behavioral markers can be used to identify a clinical state or disorder. Low-level features represent descriptive activities, such as time spent at home and total calls per day. High-level behavioral markers can reflect cognition (e.g., distractibility), behaviors (e.g., social avoidance), and emotions (e.g., depressed mood), which can be measured or estimated by the low-level features. For this study, we developed low-level features (e.g., total number of steps per day) that we correlated directly with the clinical state (i.e., depression severity) and to create high-level behavioral markers (e.g., mood) that could be correlated with the clinical state (as described in Supplementary Table 2).

In Table 1, we defined the high-level behavioral markers as SIGH-IDSC symptom dimensions. The categorizations were manually grouped based on their conceptual similarities. In total, the authors created 15 dimensions relating to Agitation, Anxiety (Psychic), Anxiety (Somatic), Guilt, Hypochondria, Interpersonal relationships, Mood, Retardation, Sex, Sleep, Somatic (General), Somatic (Gastrointestinal), Suicidal Ideation, Weight, and Work. In addition, the authors defined global dimensions as the total scores of SIGH-D, IDS-C, and SIGHD-IDSC (the SIGH-D and IDS-C combined) individually.

DATA PRE-PROCESSING All data were inspected and preprocessed using Python (version 3.6.0) and the Pyspark (version 3.0.1) library. Raw data were inspected for missing data, outliers, and normality by the authors AZ and RJD. Missing data were defined as the absence of data for periodic features on a given day or given week (e.g., weight, blood pressure, and the DASS). No missing data definition was provided for the aperiodic activities (e.g., phone calls) as there was no method to distinguish between missing data or no activity. As we used weekly aggregates for the modelling (for more information see p. 154: *Feature engineering*), missing values were not imputed. The advantage is that when missing data are

limited to a small number of observations, we can still achieve a comprehensive analysis with incomplete data without adjustment. The disadvantage is that if participants were missing several days of data within one week, then the weekly aggregate would be biased towards days containing data. Outliers were removed if they were deemed illogical and impossible (such as walking more than 70,000 steps per day). Log- or square root-transformation was applied if the distribution of the feature was not normally distributed.

FEATURE ENGINEERING

The features were provided by the Withings devices and CHDR MORE app at different sampling frequencies (varying from each interaction to every 10 minutes). Feature engineering is the process of selecting and transforming features from raw data to extract and identify the most informative set of features. These engineered features represent a summarized measure of the collected data. For this study, cumulative parameters, such as step count, were summated per day per subject. Averaged features, such as the heart rate (average beats per minute), which was provided every 10 minutes, were averaged per day per subject. Supplementary Table 1 illustrates how all the features were aggregated for each data type. The design of these features was based on available data provided by the smartphone and wearable devices, and on a previous published study that had a similar protocol.[35]

SIGHD-IDSC scores represent the depression severity over the last week. To create a dataset that is representative of activity over the last week, we transformed the daily activities into weekly averages. Hence, each patient and control had three data points, each point representing an average day in a single week. We have defined a 'week' as 6 days prior to the SIGHD-IDSC assessment and the day of the SIGHD-IDSC assessment.

FEATURE SELECTION

Feature selection is the process of identifying relevant features that can be used for model construction. The elimination of irrelevant features would increase the interpretability of the final statistical models.[36] Typically, domain knowledge plays a pivotal role in selecting the most relevant features. However, domain knowledge may not be sufficient when dealing with a multi-dimensional dataset. Hence, automatic feature selection techniques can be used to remove features that are highly correlated, exhibit low variance, or provide a limited amount of information about the dependent variable.[37,38] Prior to the feature selection, 61 features were provided by the CHDR MORE and EPRO platform (as seen in Supplementary Table 2). The number of features was reduced in a two-step approach. First, we used domain knowledge to eliminate features. We visually inspected features to remove features which exhibited a high degree of missing data (e.g., if the majority of subjects had missing values or had no data) or had limited clinical relevance (e.g., time spent on the 'comics' apps category was deemed irrelevant). Second, we used and compared three automated feature selection techniques: Correlation-based Feature Selection,[39] Variance Thresholding,[40] and Variance Thresholding in combination with Variance Inflation Factor (VIF).[41] Each feature selection technique was used to select a subset of relevant features (based on the weekly aggregated features) and these features were subsequently fitted to the regression models (see section Statistical Analysis).

Statistical analysis

ESTIMATION OF SIGHD-IDSC R (version 3.6.2) was used for statistical analysis. While the Pearson's correlations are typically employed to estimate the correlation coefficient between two outcome variables, correlation coefficients in longitudinal settings (with possible missing values) cannot be obtained with this approach. Hence, we used Linear Mixed-Effects models (LMM) to account for the between- and within-subject variation over time.

We compared the LMM from the lme4 R package[42,43] and the generalized linear mixed models with L1-penalization from the glmmLASSO R package.[44] The glmmLASSO models allow for further feature selection by reduc-

ing the weight of irrelevant features to zero.[45] As seen in Equation 1, each of the employed LMMs included a subject-specific random effect to account for the intra-subject correlations between the dependent and independent variables. All other variables were included as fixed effects. No interaction terms were included in the model as we already had more unique features than unique participants, adding more interaction terms would only increase the complexity of the model, as observations within participants may be autocorrelated. To assess if model assumptions were met, each model was visually inspected using quantile-quantile (Q-Q) plots.[46]

EQUATION 1 Depression severity linear mixed effects model. **Y** is the vector that represents the weekly depression scores. **X** is the fixed effects design matrix, which includes columns for the intercept and the features. **Z** is the random effects design matrix, which includes columns for the subject-specific random effects. **β** and **b** represent the vectors for the fixed effects and subject-specific random effects coefficients respectively. **ε** represents the vector of the Independent and Identically Distributed (I.I.D.) error terms.

$$Y = X\beta + Zb + \varepsilon$$

While a LMM of the SIGHD-IDSC total score would provide a broad assessment of depression severity, LMMs of the SIGHD-IDSC dimension scores would provide insights into an individual's depression symptom profile. In total, we developed 18 LMMs, one for each of the global dimension scores, SIGH-IDSC total score, SIGH-D total score, IDS-C total score, and one for each of the SIGH-IDSC symptom dimensions scores (as seen in Table 1). We did not develop a LMM for the insight dimension as there was no variation in this assessment during the study period and only one participant had a score of one (the remaining participants had a score of zero).

All LMMs were validated using a repeated nested stratified shuffle split 100 outer-fold (and 50 inner-fold) cross-validation. Cross-validation is a resampling method to assess the generalizability of a statistical model.[47] Nested cross-validation consists of having two non-overlapping cross-validation layers. The inner cross-validation loop optimizes the model configuration, and the outer cross-validation loop assesses the performance of the model generated in the inner loop.[48] In each outer loop, 80% of the data was used for model training, while the remaining 20% was used for model validation. For each loop, all features were standardized (by scaling to the unit variance after subtracting the mean), using the training data only. The 80% training data in the outer loop was used for the train and test split in the inner loop. By using stratification, each dataset split had the same distribution of patients and controls in each fold. This approach mitigates the risk of biased model evaluation due to class imbalance. The limitation of nested cross-validation is that the validation procedure generates a model for each outer-fold. For this study, we reported the average R2 and RMSE (Root Mean Square Error) of the 100 outer-fold models. The R2 represents the percentage of variance that is explained by the remotely monitored features. The RMSE represents the standard deviation of the error between the true depression severity scores from the predicted depression severity scores.

Training LMMs with 1,2, and 3 weeks of data

For the secondary objective, we examine the impact of the number of data points used to train the model would affect the model performance. To do so, we trained the regression models on the first week, the first two weeks, and three weeks of data. Here, we assume that an individual's week-to-week behavior is habitual and therefore one week of data would constitute a minimally sufficient dataset for model building. We adopted a weekly aggregation approach for each model, where the data were aggregated on a weekly basis. Specifically, for the week 1 model, we had one aggregated weekly observation per subject. As for the week 2 models, we expanded the observations to two aggregated weekly data points per subject. For the training of the LMMs, the dependent variable was the SIGHD-IDSC scores for each week. For the evaluation of the model for the hold-out dataset, the dependent variable was the SIGHD-IDSC for the third week of data (as shown in the Supplementary Figure 1). As shown in the Supplementary Figure 1, we validated the performance of the models using a hold-out validation dataset consisting of the third week of data.

To ensure that there was no data leakage between the training and validation datasets, we used 70% of the participants for the training dataset, and the remaining 30% for the validation dataset. The dataset was stratified based on the depression symptom severity to ensure that the population distribution was the same in each training and validation datasets. To assess the generalizability of the regression models, we applied 100 outer-fold (50 inner-fold) nested cross validation, with each of the inner-folds creating the optimal regression models based on the training datasets and outer-folds consisting of the third week validation dataset.

Discussion

PARTICIPANT CHARACTERISTICS 30 patients and 29 healthy controls were enrolled in the study. Data was collected between March 2019 to March 2020. Supplementary Table 3 provides an overview of the demographic characteristics of the enrolled patients and healthy controls. In total, 177 SIGHD-IDSC total scores were collected (3 weeks for all 30 patients and 29 healthy controls). The last healthy control was not included due to the COVID-19 lockdown.[49] The patients had a mean MADRS total score of 29 (and standard deviation of ± 3.5), and MADRS was not collected for the healthy controls as it was only used to screen the unipolar depressed patients. The patients had a mean SIGH-D total score of 14.5 (± 4.5) and a mean IDS-C total score of 30.5 (± 8.5). The healthy volunteers had a mean SIGH-D total score and IDS-C total score of 1 (± 2) and 1(± 3) respectively. Figure 1 illustrates the distribution of the SIGHD-IDS, SIGH-D, IDS-C, and SIGHD-IDSC symptom dimensions total scores for both the patients and healthy controls.

DATA QUALITY To assess the quality of our data, we examined the number of days, features, and participants with missing data. In Supplementary Table 4, we found that most of the missing data were from the sleep and location features, however the percentage of missing days were less than 5% of the days and related to 12% of the participants. In the case of the DASS, our expectation was to receive 4 responses per person, totaling 236 responses. However, we received only 196 responses, resulting in an 83% completion rate. Similarly, for the PANAS, we anticipated 42 responses per person, amounting to a total of 2478 responses. However, we obtained 1585 responses, indicating a completion rate of 66%. We found that 64% of the 61 features had no outliers, 29% of the features (concerning 15% of the participants) had one outlier, and the remaining 5% of the features (concerning 5% of the participants) had two outliers.

PERFORMANCE OF LMMS Among the different feature selection methods and LMMs used, the Variance Thresholding in combination with the LMM consistently yielded the highest R2 and lowest RMSE across all the dependent variables. Hence, we only reported the results of these Variance Thresholding LMM depression severity models. When including both the healthy controls and the patients, the SIGH-D, IDS-C, and SIGHD-IDSC LMMs achieved an R2 of 0.80, 0.80, and 0.73 and a scaled RMSE of 5.3, 9.9, and 15.1 respectively. Table 2 provides an overview of the performance of the 18 SIGHD-IDSC dimension LMMs. The LMMs with the highest R2 were the SIGHD-IDSC dimensions related to mood (0.72) and work (0.65). While the LMMs with the lowest R2 were the SIGHD-IDSC dimensions related to retardation (0.40) and hypochondria (0.40). Supplementary Table 1 highlights the advantages of including healthy controls in the LMMs. When examining the predictive performances separately for patients and healthy controls, it is observed that the R2 and RMSE are lower compared to when they are combined. However, it is important to note that the overall predictive performance may still be valuable in both cases.

CORRELATIONS For each of the LMMs, we identified the correlation coefficients and their significance between the remotely monitored features and the depression severity scores. As seen in Figure 2, there was a significantly positive correlation between the mean SIGH-D total score with the DASS-Anxiety and DASS-Stress ($p<.05$). Both the IDS-C and the SIGHD-IDSC total scores were significantly positively correlated with the DASS-

Depression, Anxiety, and Stress total scores and significantly negatively ($p<.05$) correlated with the mean steps-per-minute and time spent travelling. We found that the Depression, Anxiety, and Stress total scores (from the DASS) and location features were significantly correlated with 7 (Agitation, Anxiety (Psychic), Anxiety (Somatic), Guilt, Interpersonal, Mood and Sex) and 6 (Agitation, Anxiety (Psychic), Guilt, Hypochondriasis, Retardation, and Sex) of the mean SIGHD-IDSC symptom dimensions respectively.

TRAINING LMMS WITH 1,2, AND 3 WEEKS OF DATA Overall, we found that training the models on three weeks of data consistently yielded the highest R2 and the lowest RMSE for each of the SIGHD-IDSC global and symptom dimensions compared to the models trained on the first week and first two weeks of data with the exception one dimension, Agitation (as seen in Figure 3). For the Agitation dimension, the models trained on the first two weeks of data yielded the highest R2. The difference in R2 between the first week and the third weeks models was relatively marginal (a difference of 0.07) for the SIGHD-IDSC global dimension. However, the difference in the scaled RMSE between the two models was notable, with a difference of 0.13.

Discussion

In this pilot study, we provided a comprehensive assessment of the relationship between depression severity and subjective and objective features sourced from data collected by smartphone and wearable devices under free-living conditions. Our results illustrate that features related to self-reported depression, anxiety scores, stress scores, physical activity, and not social activities, were significantly correlated with depression severity. These features can collectively serve as a composite biomarker to estimate the gold standard in-clinic assessment, the SIGHD-IDSC.

DATA QUALITY The missing and outlier data only impacted a minority of the participant's data and did not lead to the exclusion of any weekly aggregated features used in the analysis (Supplementary Table 4). Given the low number of missing data and outliers, we did not observe any differences in data quality between the depressed patients and controls. While we could not identify any similar trials to compare data quality, we deem that our protocol led to the collection of a robust and reliable dataset. However, the aggregation of the data undermines the opportunity to identify potentially nuanced daily behaviors and higher order interactions between multiple features. For example, social and physical activity behavior most likely differs per location and between weekdays and weekends, but these daily interaction features are not reflected in the current dataset. The identification of higher order behavioral patterns or routines per location and per day could enrich the sensitivity of the composite biomarkers.

ESTIMATION OF THE SIGHD-IDSC Our findings indicate that a combination of remotely monitored self-reported and objective features can serve as a composite biomarker to estimate weekly depression severity. We found our approach was better suited for evaluating the global dimensions (SIGH-D, IDS-C, and SIGH-IDSC total scores), rather than the manually defined SIGHD-IDSC symptom dimensions, such as mood, weight, or sex (Table 2). The symptom dimension models were a moderate to strong representation of work, somatic (general), interpersonal, anxiety (psychic) and mood dimensions and a poor representation of the hypochondria and retardation dimensions. This illustrates that the features obtained correspond to some but not all the SIGHD-IDSC dimensions. One explanation for the limited agreement between the remotely monitored biomarkers and the SIGHD-IDSC dimensions is the comparison of objective measures with subjective assessments. For example, we compared objective sleep measurements (such as sleep duration, and the number of light and deep sleep periods) to the subjective interpretations of sleep quality by the patient or the clinician as reflected in the SIGHD-IDSC. Despite having several objective measures relating to sleep, we found that the sleep model captured less than half of the variance. Previous studies have

illustrated that objective sleep assessments are not strongly correlated with subjective reports of sleep.[50,51] Discrepancies between the objective and subjective measures of sleep could be influenced by several factors, such as mood at the time of awakening,[52] insomnia, negative bias, and impaired memory.[53] These findings highlight that those subjective experiences are not always represented by objective measures. Hence, in the context of clinical trials for depression, the identified relevant features are better suited for monitoring overall depression severity rather than monitoring specific depression symptoms.

INCLUSION OF HEALTHY CONTROLS The inclusion of health controls in the models provides several benefits. Firstly, by incorporating more participants, the number of observations available for analysis increases. This larger sample size enhances the statistical power of the LMMs, which leads to more reliable and robust predictions. Additionally, the inclusion of healthy controls introduces a broader range of depression severity scores, spanning from zero to minimal symptoms. In addition to enhancing the model's ability to capture the full spectrum of depression severity and improving its generalizability, the wider range of scores also allows for the inclusion of potential remission in depressed patients. As their scores move towards zero, the model can accurately capture the possibility of their condition improving and reaching a state of remission.

CORRELATION WITH THE SIGHD-IDSC DIMENSIONS Both the self-reported DASS and daily travel routines were consistently significantly correlated with the SIGH-D, IDS-C and SIGHD-IDSC global dimension total scores (Figure 2). More specifically, we found that depression, anxiety, and stress total scores were positively correlated with overall depression severity. In addition, participants with higher depression scores were more likely to walk faster, however, spent less time travelling. Our findings are supported by previous studies that found correlations between both smartphone-based self-reported assessments and location-based behaviors [16,54,55] with in-clinic depression rating scales.[13,56,57]

Notwithstanding, we have not identified any research that supports the notion that unipolar depressed patients have increased walking speeds, rather, the current literature suggests that depressed patients exhibit more motor disturbances and thus reduced walking speeds.[58] However, these inferences were based on instrumented gait assessments performed in controlled settings, and not based on real-world evidence. This implies that inferences regarding gait or other motor disturbances assessed in the clinic may not always correspond with behaviors outside the clinic. Together, our findings highlight the importance of collecting both self-reported subjective and objective behavioral features, such as DASS, gait and travel patterns, in depression drug trials as they represent a more holistic biomarker of depression. Further, behaviors characteristic to depression that were identified within a clinical setting may not correspond to behaviors exhibited outside a clinical setting.

NUMBER OF WEEKS OF DATA FOR TRAINING Our findings indicate that the models overall performed better when trained on three weeks of data, rather than one or two weeks (Figure 3). However, for the SIGHD-IDSC global dimensions, the difference in the variance explained between the first week and three weeks of data was marginal. While the inclusion of three weeks of data notably reduced the prediction error. Depending on the mechanism of action of any given antidepressant drug, therapeutic effects may only become evident after several weeks of treatment with, for example SSRIs, or may rapidly occur and then dissipate over a week or two as with the NMDAR antagonist ketamine.[59,60] It is therefore crucial to determine how long and how often patients need to be monitored to extract reliable and meaningful inferences from the data following an intervention. Collecting excessive data can be time-consuming and resource-demanding, however having insufficient data can undermine the accuracy of the extrapolations. Although the present study was of non-interventional nature, this suggests that a minimum of three weeks of data are required to create a representative dataset that would build an accurate model that represents depression severity in future

interventional trials. However, the trade-off between the number of weeks used for training and the model performance was marginal.

LIMITATIONS There are several limitations to our approach. Due to the small sample size, relatively short observation period, and the number of technical devices used (Android smartphone and Withings wearables), there is a limited understanding of what degree our findings are generalizable to other cohorts, technical devices, and clinical assessments. A follow-up study is needed to assess how well our findings can translate to other depressed patients whose data are collected in a different time period using different devices (such as an iPhone and Apple Watch). Further, given the limited agreement between the objective measures of sleep and the SIGHD-IDSC sleep dimension scores, a follow-up study may choose to incorporate both objective and subjective measures of sleep such as polysomnography and self-report questionnaires related to sleep to further improve the reliability of the features.

APPLICATION Based on our findings, remotely monitored features cannot substitute the clinical assessment of depression severity. However, our approach can potentially serve as a complementary tool to assess clinical symptoms of depression over time in free-living conditions, since a number of subjectively reported indicators of depression can be missed between assessments and/or may be subject to recall bias during interviews. Remotely monitored composite biomarkers therefore are strong candidates for filling-in and complementing the retrospective gaps that are typical of in-person clinical assessments. Hence our approach is expected to benefit drug development for mood disorders, since it could aid the monitoring and assessment of depression severity during clinical trials based on both in-clinic rater-based interviews and out-of-clinic activities and self-reported outcomes.

Conclusion

We presented a novel approach to monitoring depression severity among unipolar depressed patients using data sourced from smartphone and wearable devices. In this longitudinal non-interventional study, we collected a relatively robust dataset, consisting of a few missing data points and outliers. We identified the relevant smartphone- and wearables-based features that collectively create a biomarker that could estimate the SIGH-D, IDS-C and SIGHD-IDSC global and symptom dimension total scores. Together, these findings suggest that objective and subjective features captured by these remote monitoring devices can collectively serve as a composite biomarker to estimate depression severity under free-living conditions.

TABLE 1 Overview of the SIGHD IDS-C symptom and global dimensions and their associated SIGH-D and IDS-C questions.

SIGHD-IDSC symptom dimensions	SIGH-D	IDS-C
Agitation	09. Agitation	24. Psychomotor agitation
Anxiety (Psychic)	10. Anxiety (Psychological)	06. Mood (Irritable) 07. Mood (Anxious) 27. Panic/phobic symptoms
Anxiety (Somatic)	31. Anxiety (Somatic)	26. Sympathetic arousal
Guilt	02. Feelings of Guilt	
Hypochrondia	15. Hypochondriasis	
Insight	17. Insight	
INTERPERSONAL RELATIONSHIPS		29. Interpersonal sensitivity
Mood	01. Depressed mood (sad, hopeless, helpless, worthless)	05. Mood (Sad) 08. Reactivity of Mood 09. Mood variation 10. Quality of mood 16. Outlook (Self) 17. Outlook (Future)
Psychomotor retardation	08. Retardation; Psychomotor	23. Psychomotor slowing
Sexual function	14. Genital symptoms	22. Sexual interest
Sleep	04. Insomnia (Early) 05. Insomnia (Middle) 06. Insomnia (Late)	01. Sleep onset insomnia 02. Mid-nocturnal insomnia 03. Early morning insomnia 04. Hypersomnia
Somatic (General)	12. Somatic Symptoms General	20. Energy/Fatigability 25. Somatic complaints 30. Leaden paralysis / physical energy
Somatic (Gastrointestinal)	12. Somatic Symptoms (Gastrointestinal)	11. Appetite decreased 12. Appetite increased 28. Gastrointestinal
Suicidal Ideation	03. Suicide	18. Suicidal Ideation
Weight	16. Loss of Weight	13. Weight decreased 14. Weight increased
Activity/reward/hedonic tone	07. Work and Activities	15. Concentration/decision making 19. Involvement 21. Pleasure/enjoyment
Global dimensions	SIGH-D global score: Sum of all SIGH-D dimension scores	IDS-C: Sum of all IDS-C dimension scores
	SIGH-D IDS-C: Sum of SIGH-D and IDS-C	

TABLE 2 Performance of the Variance Thresholding and LMM to estimate the total scores of the SIGH-D, IDS-C, SIGHD-IDSC global dimensions, and each of the SIGHD IDS-Csymptom dimensions.

SIGHD-IDSC global and symptom dimensions	Marginal R2 Mean	Mean RMSE
SIGH-D	0.73 (±0.01)	5.30 (±0.17)
IDS-C	0.80 (±0.01)	9.90 (±0.32)
SIGHD-IDSC	0.80 (±0.01)	15.1 (±0.48)
AGITATION	0.47 (±0.01)	0.99 (±0.04)
ANXIETY (PSYCHIC)	0.63 (±0.01)	1.70 (±0.06)
ANXIETY (SOMATIC)	0.57 (±0.01)	1.16 (±0.06)
GUILT	0.57 (±0.02)	1.01 (±0.04)
HYPOCHRONDIA	0.40 (±0.02)	0.27 (±0.02)
INTERPERSONAL	0.60 (±0.01)	0.56 (±0.02)
MOOD	0.72 (±0.01)	3.04 (±0.10)
RETARDATION	0.40 (±0.02)	0.61 (±0.03)
SEX	0.45 (±0.02)	1.01 (±0.05)
SLEEP	0.47 (±0.02)	2.34 (±0.07)
SOMATIC (GENERAL)	0.62 (±0.02)	1.88 (±0.07)
SOMATIC (GASTROINTESTINAL)	0.43 (±0.02)	0.71 (±0.03)
SUICIDE	0.50 (±0.01)	0.32 (±0.02)
WEIGHT	0.43 (±0.01)	0.37 (±0.02)
WORK	0.65 (±0.01)	2.02 (±0.07)

FIGURE 1 A)Distribution of the SIGH-D, IDS-C, and SIGHD-IDSC global dimensions total scores for patients and healthy controls. (B) Distribution of the total scores of the SIGHD-IDSC symptom dimensions for patients and healthy controls. In both figures, red represents the healthy controls while blue represents the patients. The lower and upper box boundaries of the boxplots represent the 25th and 75th percentile range respectively. The line within the boxplot represents the median score. The black scatter plots represent the outliers. The width of the violinplot represents the population distribution of each of the scores.

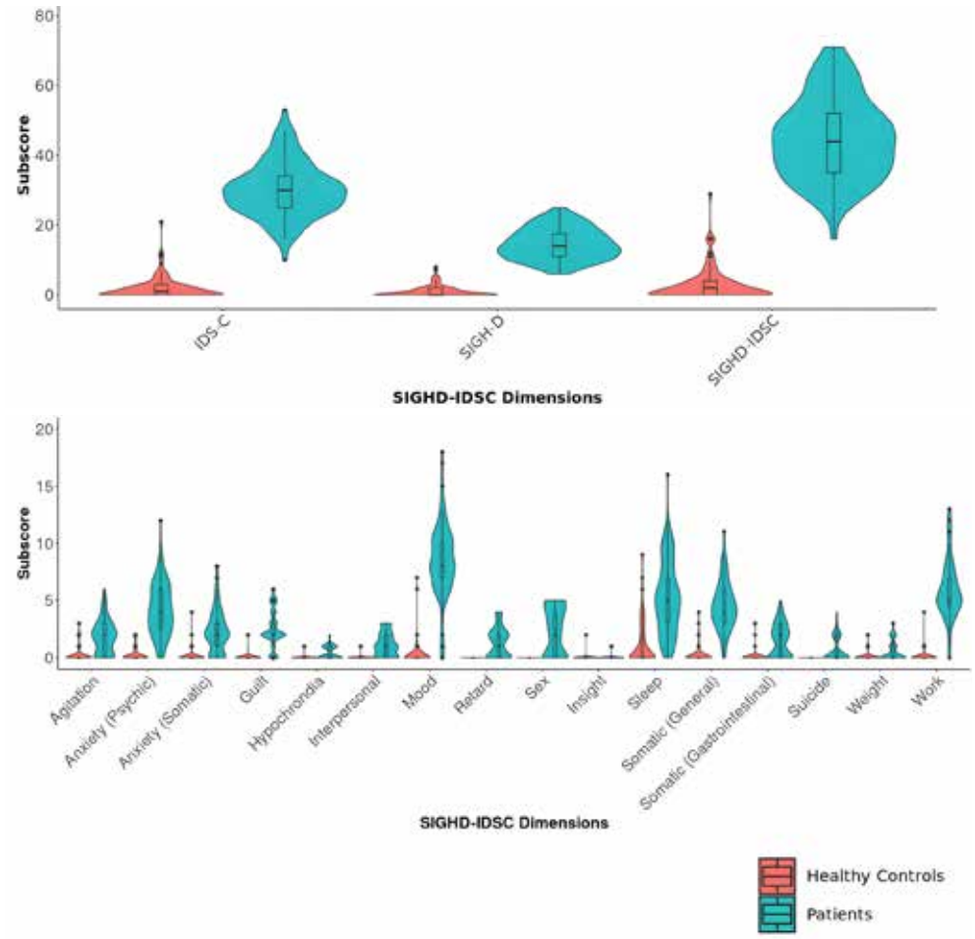


FIGURE 2 Overview of all significantly correlated features (p<0.05) for each of the (A) SIGHD-IDSC global and (B) symptom dimensions. The bars represent the correlation coefficients for each of the significant features. The color of the bars represents each of the SIGHD-IDSC global and symptom dimensions.

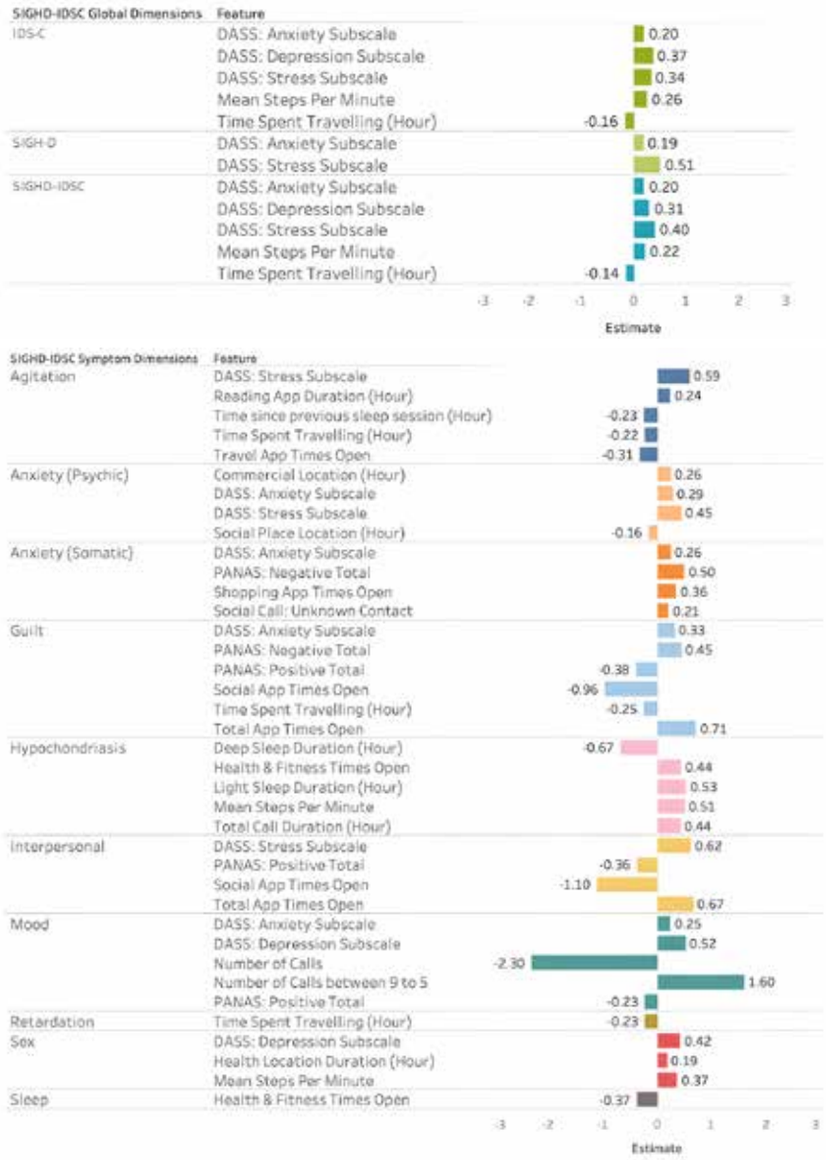
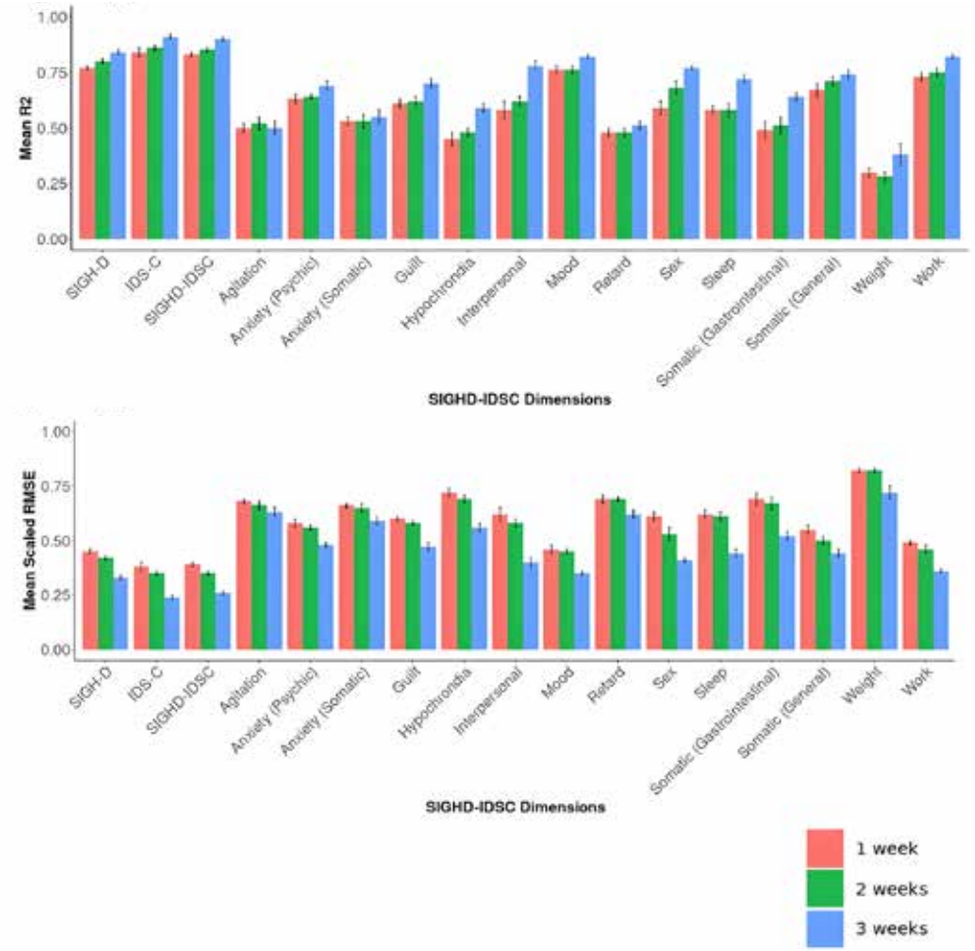


FIGURE 3 (A) and (B) represent the mean R2 and mean scaled RMSE for each of the SIGHD-IDSC global and symptom dimension LMMs. Each color represents the dataset used for training the models. The error bars represent the standard deviation across each of the 100 outer-fold predictions.



SUPPLEMENTARY TABLE 1 A summary of how the features were aggregated based on the data type.

Data Type	Time Unit	Example Feature	Aggregation Format	Example Aggregation
Count	Per day	Steps	Sum Mean Max	Total steps Max steps per hour Mean steps per hour
Continuous data within a range	Per day	Heart Rate	Min (5%) Median (50%) Max (95%)	Lowest 5% heart rate Median heart rate Maximum 95% heart rate
Duration	Per day	App Usage	Total Duration Mean Duration	Total duration of social apps opened Mean duration of social app opened per instance
GPS coordinates	Per day	Location	Sum Max Mean	Total distance travelled Mean and max distance from home

SUPPLEMENTARY TABLE 2 An overview of the CHDR MORE™ extracted features.

Category	MORE Features	Derived features	Excluded Features
DEMOGRAPHICS	Age; Gender		
ACCELERATION (SMARTPHONE)	Acceleration Magnitude Gyroscope Magnometer	98% Acceleration magnitude	Mean acceleration magnitude
ACTIVITY (SMARTPHONE)	Steps Heart Rate Physical activity duration Calories	**STEPS:** total steps, max steps per hour, mean steps per hour **HEART RATE:** 5%, 50% & 95% eats per minute (bpms), standard deviation of BPMs, % time spent in resting state **PHYSICAL ACTIVITY:** soft, moderate and intense activity duration	Calories Distance travelled Distance per step
APPS (SMARTPHONE)	**APP CATEGORIES** Communication & Social Health & Fitness, Recreational, Shopping, Tools, Travel	Duration Times open	House & Home App Libraries & Demo App Reading App All duration features
BODY (WITHINGS)	Diastolic blood pressure Systolic blood pressure Heart pulse (Bpm) Weight		Height (M) Fat mass (kg) Fat ratio (%) Hydration Muscle Mass
LOCATION (SMARTPHONE)	**LOCATION CATEGORIES** Commercial, Health, Home, Leisure, Public, Social, Travel	Total duration at place Total distance travelled Total no of unique places visited Max distance from home Time spent commuting	
SOCIAL (SMARTPHONE)	Calls Voice	Number of calls Number of unique numbers Number of incoming, outgoing and missing calls Number of calls from known and unknown numbers Total duration of calls Average duration of calls % Time human voice is detected	Text messages (SMS)
SLEEP (WITHINGS)		Number of sleep sessions Total sleep duration Number of sleep phases (awake, light sleep and deep sleep) Duration of sleep phases (awake, light and deep sleep) Time between sleep sessions Time to fall asleep	
EPRO (SMARTPHONE)	Self-assessments	Twice daily PANAS Weekly DASS-21	

SUPPLEMENTARY TABLE 3 An overview of demographic characteristics of the enrolled patients and healthy controls

Demographics	Descriptor	Patients	Healthy controls
GENDER	Female Male	24 6	25 4
RACE	African American or Black Asian Mixed Other White	2 2 4 1 21	1 3 0 1 25
AGE	Mean (STD) Min, Max]	35(13) 18, 64]	35(13) 20, 63]
BMI (KG/M^2)	Mean (STD) Min, Max]	24(3) 20, 31.5]	24(3) 18, 31]
MADRS	Mean (STD) Min, Max]	29 (4) 23, 38]	N/A
SIGH-D TOTAL	Mean (STD) Min, Max]	14.5(4.5) 6, 25]	1(2) 0, 8]
IDS-C TOTAL	Mean (STD) Min, Max]	30.5(8.5) 10, 62]	1(3) 0,21]
SIGH-IDSC TOTAL	Mean (STD) Min, Max]	45(12) 16, 71]	3(5) 0,29]

SUPPLEMENTARY TABLE 4 A summary table of the number of missing days or days containing excluded outliers and number of participants with missing or outlier days are shown. Features with no missing data or excluded outliers are not shown. For aperiodic features, it is not possible to differentiate between missing data and no data, thus missing data for these features are not represented.

Feature Category	Feature	Number of days with missing data	Number of participants with missing data days	Number of excluded outliers	Number of participants with excluded outliers
ACCELEROMETER	Acceleration Magnitude 98%	1.12	2	0	0
APPS	Times Open Shopping App	-	-	2	1
	Times Open Travel App	-	-	1	1
	Total Times App Open	-	-	1	1
CALLS	Calls from known contact	-	-	1	1
	Missed Calls	-	-	1	1
	Calls from unknown contact	-	-	1	1
LOCATION	Total unique places visited	4.49	7	1	1
	Total km travelled	3.93	5	1	1
	Time spent travelling	3.93	5	1	1
STEPS	All steps parameter	1.69	3	1	1
HEART RATE	All heart rate parameters	2.25	4	0	0
SLEEP	Light and deep sleep duration and count	3.93	7	0	0
	Longest sleep session	3.93	7	0	0
	Time since previous sleep session	3.93	7	1	1
	Time to fall asleep	3.93	7	1	1
	Total sleep duration	3.93	7	0	0

SUPPLEMENTARY TABLE 5 Comparison of R2 and RMSE Values for Patients and Healthy Controls Across SIGH Dimensions

Population	SIGHD-IDSC Dimensions	Marginal R2 Mean	Mean RMSE
PATIENTS ONLY	SIGH-D	0.64	6.54
	IDS-C	0.80	19.07
	SIGH-D IDSC	0.75	16.67
HEALTHY CONTROLS ONLY	SIGH-D	0.65	5.62
	IDS-C	0.71	6.53
	SIGH-D IDSC	0.70	7.13

REFERENCES

1 R. L. Holland, 'What makes a good biomarker?,' *Advances in Precision Medicine*, vol. 1, no. 1, p. 66, Mar. 2016, doi: 10.18063/APM.2016.01.007.

2 V. RACHAKONDA, T. H. PAN, and W. D. LE, 'Biomarkers of neurodegenerative disorders: How good are they?,' *Cell Research*, vol. 14, no. 5, pp. 349–358, Oct. 2004, doi: 10.1038/sj.cr.7290235.

3 S. Leucht, H. Fennema, R. R. Engel, M. Kaspers-Janssen, and A. Szegedi, 'Translating the HAM-D into the MADRS and vice versa with equipercentile linking,' *Journal of Affective Disorders*, vol. 226, pp. 326–331, Jan. 2018, doi: 10.1016/j.jad.2017.09.042.

4 A. M. Carneiro, F. Fernandes, and R. A. Moreno, 'Hamilton depression rating scale and montgomery-asberg depression rating scale in depressed and bipolar I patients: psychometric properties in a Brazilian sample,' 2011, doi: 10.1186/s12955-015-0235-3.

5 A. M. Mofsen, T. L. Rodebaugh, G. E. Nicol, C. A. Depp, J. P. Miller, and E. J. Lenze, 'When all else fails, listen to the patient: A viewpoint on the use of ecological momentary assessment in clinical trials,' *JMIR Mental Health*, vol. 6, no. 5. JMIR Publications Inc., p. e11845, May 01, 2019. doi: 10.2196/11845.

6 V. Zeigler-Hill and T. Shackelford, 'Encyclopedia of Personality and Individual Differences.'

7 J. Gorzelitz, P. E. Peppard, K. Malecki, K. Gennuso, F. J. Nieto, and L. Cadmus-Bertram, 'Predictors of discordance in self-report versus device-measured physical activity measurement,' *Annals of Epidemiology*, vol. 28, no. 7, pp. 427–431, Jul. 2018, doi: 10.1016/j.annepidem.2018.03.016.

8 A. ZhuParris, A. A. de Goede, I. E. Yocarini, W. Kraaij, G. J. Groeneveld, and R. J. Doll, 'Machine Learning Techniques for Developing Remotely Monitored Central Nervous System Biomarkers Using Wearable Sensors: A Narrative Literature Review,' *Sensors*, vol. 23, no. 11, p. 5243, May 2023, doi: 10.3390/s23115243.

9 D. Ben-Zeev, E. A. Scherer, R. Wang, and H. Xie, 'Next-Generation psychiatric assessment: using smartphone sensors to monitor behavior and mental health,' *Psychiatric Rehabilitation Journal*, vol. 38, no. 3, pp. 218–226, 2015, doi: 10.1037/prj0000130.

10 R. Wang *et al.*, 'Studentlife: Assessing mental health, academic performance and behavioral trends of college students using smartphones,' *UbiComp 2014 - Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 3–14, 2014, doi: 10.1145/2632048.2632054.

11 K. Ellis, S. Godbole, S. Marshall, G. Lanckriet, J. Staudenmayer, and J. Kerr, 'Identifying Active Travel Behaviors in Challenging Environments Using GPS, Accelerometers, and Machine Learning Algorithms,' *Frontiers in Public Health*, vol. 2, no. APR, p. 36, Apr. 2014, doi: 10.3389/fpubh.2014.00036.

12 S. Wan, L. Qi, X. Xu, C. Tong, and Z. Gu, 'Deep Learning Models for Real-time Human Activity Recognition with Smartphones,' *Mobile Networks and Applications*, vol. 25, no. 2, pp. 743–755, Apr. 2020, doi: 10.1007/s11036-019-01445-x.

13 J. Goltermann *et al.*, 'Smartphone-based self-reports of depressive symptoms using the remote monitoring application in psychiatry (ReMAP): Interformat validation study,' *JMIR Mental Health*, vol. 8, no. 1, p. e24333, Jan. 2021, doi: 10.2196/24333.

14 T. Aledavood *et al.*, 'Smartphone-Based Tracking of Sleep in Depression, Anxiety, and Psychotic Disorders,' 1920, doi: 10.1007/s11920-019-1043-y.

15 M. Boukhechba, A. R. Daros, K. Fua, P. I. Chow, B. A. Teachman, and L. E. Barnes, 'DemonicSalmon: Monitoring mental health and social interactions of college students using smartphones,' *Smart Health*, vol. 9–10, pp. 192–203, Dec. 2018, doi: 10.1016/j.smhl.2018.07.005.

16 S. Saeb, E. G. Lattie, K. P. Kording, and D. C. Mohr, 'Mobile phone detection of semantic location and its relationship to depression and anxiety,' *JMIR MHEALTH and uHealth*, vol. 5, no. 8, p. e112, 2017, doi: 10.2196/mhealth.7297.

17 J. Lu *et al.*, 'Joint Modeling of Heterogeneous Sensing Data for Depression Assessment via Multi-task Learning,' *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 1, pp. 1–21, 2018, doi: 10.1145/3191753.

18 N. C. Jacobson and Y. J. Chung, 'Passive sensing of prediction of moment-to-moment depressed mood among undergraduates with clinical levels of depression sample using smartphones,' *Sensors (Switzerland)*, vol. 20, no. 12, pp. 1–16, 2020, doi: 10.3390/s20123572.

19 M. L. Tønning, M. Faurholt-Jepsen, M. Frost, J. E. Bardram, and L. V. Kessing, 'Mood and Activity Measured Using Smartphones in Unipolar Depressive Disorder,' *Frontiers in Psychiatry*, vol. 12, no. July, pp. 1–12, 2021, doi: 10.3389/fpsyt.2021.701360.

20 P. Pedrelli *et al.*, 'Monitoring Changes in Depression Severity Using Wearable and Mobile Sensors,' *Frontiers in Psychiatry*, vol. 11, p. 1413, Dec. 2020, doi: 10.3389/fpsyt.2020.584711.

21 M. Zimmerman, I. Chelminski, and M. Posternak, 'A review of studies of the Hamilton Depression Rating Scale in healthy controls: Implications for the definition of remission in treatment studies of depression,' *Journal of Nervous and Mental Disease*, vol. 192, no. 9, pp. 595–601, 2004, doi: 10.1097/01.nmd.0000138226.22761.39.

22 S. De Vos, K. J. Wardenaar, E. H. Bos, E. C. Wit, and P. De Jonge, 'Decomposing the heterogeneity of depression at the person-, symptom-, and time-level: Latent variable models versus multimode principal component analysis,' *BMC Medical Research Methodology*, vol. 15, no. 1, pp. 1–10, 2015, doi: 10.1186/s12874-015-0080-4.

23 CHDR, 'Trial@home — CHDR,' 2022. https://chdr.nl/trialhome (accessed Aug. 16, 2022).

24 The Hyve, 'CHDR MORE® | The Hyve,' 2022. https://www.thehyve.nl/cases/chdr-more (accessed Aug. 16, 2022).

25 Withings, 'Health Tracker App | Fitness Tracker | Withings Health Mate,' 2022. https://www.withings.com/pt/en/health-mate (accessed Aug. 16, 2022).

26 D. Watson, L. A. Clark, and A. Tellegen, 'Development and validation of brief measures of positive and negative affect: The PANAS scales.,' *Journal of Personality and Social Psychology*, vol. 54, no. 6, pp. 1063–1070, 1988, doi: 10.1037//0022-3514.54.6.1063.

27 P. F. Lovibond and S. H. Lovibond, 'The structure of negative emotional states: Comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories,' *Behaviour Research and Therapy*, vol. 33, no. 3, pp. 335–343, Mar. 1995, doi: 10.1016/0005-7967(94)00075-U.

28 J. D. Henry and J. R. Crawford, 'The short-form version of the Depression Anxiety Stress Scales (DASS-21): Construct validity and normative data in a large non-clinical sample,' *British Journal of Clinical Psychology*, vol. 44, no. 2, pp. 227–239, Jun. 2005, doi: 10.1348/014466505X29657.

29 M. HAMILTON, 'A rating scale for depression,' *Journal of neurology, neurosurgery, and psychiatry*, vol. 23, no. 1, pp. 56–62, Feb. 1960, doi: 10.1136/jnnp.23.1.56.

30 A. J. Rush, C. M. Gullion, M. R. Basco, R. B. Jarrett, and M. H. Trivedi, 'The inventory of depressive symptomatology (IDS): Psychometric properties,' *Psychological Medicine*, vol. 26, no. 3, pp. 477–486, 1996, doi: 10.1017/s0033291700035558.

31 R. D. Gibbons, D. C. Clark, and D. J. Kupfer, 'Exactly what does the Hamilton depression rating scale measure?,' *Journal of Psychiatric Research*, vol. 27, no. 3, pp. 259–273, 1993, doi: 10.1016/0022-3956(93)90037-3.

32 A. John Rush, D. E. Giles, M. A. Schlesser, C. L. Fulton, J. Weissenburger, and C. Burns, 'The inventory for depressive symptomatology (IDS): Preliminary findings,' *Psychiatry Research*, vol. 18, no. 1, pp. 65–87, 1986, doi: 10.1016/0165-1781(86)90060-0.

33 E. Corruble, J. M. Legrand, C. Duret, G. Charles, and J. D. Guelfi, 'IDS-c and IDS-SR: Psychometric properties in depressed in-patients,' *Journal of Affective Disorders*, vol. 56, no. 2–3, pp. 95–101, Dec. 1999, doi: 10.1016/S0165-0327(99)00055-5.

34 D. C. Mohr, M. Zhang, and S. M. Schueller, 'Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning,' *Annual Review of Clinical Psychology*, vol. 13, no. 1, pp. 23–47, 2017, doi: 10.1146/annurev-clinpsy-032816-044949.

35 G. Maleki *et al.*, 'Objective Monitoring of Facioscapulohumeral Dystrophy During Clinical Trials Using a Smartphone App and Wearables: Observational Study,' *JMIR Formative Research*, vol. 6, pp. 1–13, 2022, doi: 10.2196/31775.

36 K. Kira and L. A. Rendell, *A practical approach to feature selection*. Elsevier, 1992, pp. 249–256. doi: 10.1016/b978-1-55860-247-2.50037-1.

37 G. Chandrashekar and F. Sahin, 'A survey on feature selection methods,' *Computers and Electrical Engineering*, vol. 40, no. 1, pp. 16–28, Jan. 2014, doi: 10.1016/j.compeleceng.2013.11.024.

38 I. Guyon and A. M. De, 'An Introduction to Variable and Feature Selection André Elisseeff,' 2003.

39 M. A. Hall, 'Correlation-based Feature Selection for Machine Learning,' 1999.

40 Guiseppe Bonaccorso, 'Machine Learning Algorithms - Giuseppe Bonaccorso - Google Books,' *Packt Publishing*, 2017. https://books.google.com.et/books?hl=en&lr=&id=_-ZDDwAAQBAJ&oi=fnd&pg=PP1&dq=machine+learning++algorithms+&ots=epfCF2Hx6K&sig=iKsf6fGqa8OrsjZayoR16LqVfgs&redir_esc=y#v=onepage&q=machine learning algorithms&f=false%0Ahttps://books.google.co.uk/books?hl=en (accessed Dec. 10, 2021).

41 J. Miles, 'Tolerance and Variance Inflation Factor,' in *Wiley StatsRef: Statistics Reference Online*, Wiley, 2014. doi: 10.1002/9781118445112.stat06593.

42 D. Bates *et al.*, 'Linear Mixed-Effects Models using 'Eigen' and S4.' CRAN, 2021.

43 A. Groll, 'Variable Selection for Generalized Linear Mixed Models byL1-Penalized Estimation.' CRAN, 2017.

44 A. Groll, 'glmmLASSO: Variable Selection for Generalized Linear Mixed Models by L1-Penalized Estimation.' 2022.

45 A. Groll and G. Tutz, 'Variable selection for generalized linear mixed models by L1-penalized estimation,' *Statistics and Computing*, vol. 24, no. 2. pp. 137–154, 2014. doi: 10.1007/s11222-012-9359-z.

46 J. D. Pleil, 'QQ-plots for assessing distributions of biomarker measurements and generating defensible summary statistics,' *Journal of Breath Research*, vol. 10, no. 3, p. 035001, Aug. 2016, doi: 10.1088/1752-7155/10/3/035001.

47 P. Refaeilzadeh, L. Tang, and H. Liu, 'Cross-Validation,' *Encyclopedia of Database Systems*, pp. 1–7, 2016, doi: 10.1007/978-1-4899-7993-3_565-2.

48 M. W. Browne, 'Cross-validation methods,' *Journal of Mathematical Psychology*, vol. 44, no. 1, pp. 108–132, Mar. 2000, doi: 10.1006/jmps.1999.1279.

49 Government of The Netherlands, 'New measures to stop spread of coronavirus in the Netherlands | News item | Government.nl,' 2020. https://www.government.nl/latest/news/2020/03/12/new-measures-to-stop-spread-of-coronavirus-in-the-netherlands (accessed Aug. 16, 2022).

50 M. Jackowska, S. Dockray, H. Hendrickx, and A. Steptoe, 'Psychosocial Factors and Sleep Efficiency,' *Psychosomatic Medicine*, vol. 73, no. 9, pp. 810–816, Nov. 2011, doi: 10.1097/PSY.0b013e3182359e77.

51 R. Armitage, M. Trivedi, R. Hoffmann, and A. J. Rush, 'Relationship between objective and subjective sleep measures in depressed patients and healthy controls,' *Depression and Anxiety*, vol. 5, no. 2, pp. 97–102, Jan. 1997, doi: 10.1002/(SICI)1520-6394(1997)5:2<97::AID-DA6>3.0.CO;2-2.

52 M. Baillet *et al.*, 'Mood Influences the Concordance of Subjective and Objective Measures of Sleep Duration in Older Adults,' *Frontiers in Aging Neuroscience*, vol. 08, no. JUN, p. 181, Jul. 2016, doi: 10.3389/fnagi.2016.00181.

53 E. A. Dinapoli *et al.*, 'Subjective-Objective Sleep Discrepancy in Older Adults with MCI and Subsyndromal Depression,' *Journal of Geriatric Psychiatry and Neurology*, vol. 30, no. 6, pp. 316–323, Nov. 2017, doi: 10.1177/0891988717731827.

54 S. Saeb, L. Lonini, A. Jayaraman, D. Mohr, and K. Kording, 'Voodoo Machine Learning for Clinical Predictions,' *bioRxiv*, p. 059774, 2016, doi: 10.1101/059774.

55 M. T. Masud, M. A. Mamun, K. Thapa, D. H. Lee, M. D. Griffiths, and S. H. Yang, 'Unobtrusive monitoring of behavior and movement patterns to detect clinical depression severity level via smartphone,' *Journal of Biomedical Informatics*, vol. 103, Mar. 2020, doi: 10.1016/j.jbi.2019.103371.

56 J. Torous *et al.*, 'Utilizing a Personal Smartphone Custom App to Assess the Patient Health Questionnaire-9 (PHQ-9) Depressive Symptoms in Patients With Major Depressive Disorder,' *JMIR Mental Health*, vol. 2, no. 1, p. e8, Mar. 2015, doi: 10.2196/mental.3889.

57 M. Faurholt-Jepsen *et al.*, 'Electronic monitoring of psychomotor activity as a supplementary objective measure of depression severity,' *Nordic Journal of Psychiatry*, vol. 69, no. 2, pp. 118–125, Feb. 2015, doi: 10.3109/08039488.2014.936501.

58 M. Belvederi Murri *et al.*, 'Instrumental assessment of balance and gait in depression: A systematic review,' *Psychiatry Research*, vol. 284, no. July 2019, p. 112687, 2020, doi: 10.1016/j.psychres.2019.112687.

59 M. B. Keller, 'The Long-Term Treatment of Depression,' *J Clin Psychiatry*, vol. 60, no. 17, pp. 41–45, 1999.

60 M. B. Keller, 'Depression: A long-term illness,' in *British Journal of Psychiatry*, Royal College of Psychiatrists, 1994, pp. 9–15. doi: 10.1192/s0007125000293239.

CHAPTER 6

Development and technical validation of a smartphone-based pediatric cough detection algorithm

Matthijs D. Kruizinga MD,[1,2,3*] Ahnjili Zhuparris,[1*] Eva Dessing MD,[1,2] Fas J. Krol MD,[1,3] Arwen J. Sprij MD,[2] Robert-Jan Doll PhD,[1] Frederik E. Stuurman PhD,[1] Vasileios Exadaktylos PhD,[1] Gertjan J. A. Driessen PhD,[2,4] Adam F. Cohen PhD[1,3]

**These authors are shared first authors*

Pediatr Pulmonol. 2022;57(3):761-767.doi:10.1002/ppul.25801

1 Centre for Human Drug Research, Leiden, NL

2 Juliana Children's Hospital, HAGA Teaching Hospital, The Hague, NL

3 Leiden University Medical Centre, Leiden, NL

4 Department of pediatrics, Maastricht University Medical Centre, Maastricht, NL

Abstract

Introduction: Coughing is a common symptom in pediatric lung disease and cough frequency has been shown to be correlated to disease activity in several conditions. Automated cough detection could provide a non-invasive digital biomarker for pediatric clinical trials or care. The aim of this study was to develop a smartphone-based algorithm that objectively and automatically counts cough sounds of children. **Methods:** The training set was composed of 3228 pediatric cough sounds and 480,780 non-cough sounds from various publicly available sources and continuous sound recordings of 7 patients admitted due to respiratory disease. A Gradient Boost Classifier was fitted on the training data, which was subsequently validated on recordings from 14 additional patients aged 0–14 admitted to the pediatric ward due to respiratory disease. The robustness of the algorithm was investigated by repeatedly classifying a recording with the smartphone-based algorithm during various conditions. **Results:** The final algorithm obtained an accuracy of 99.7%, sensitivity of 47.6%, specificity of 99.96%, positive predictive value of 82.2% and negative predictive value 99.8% in the validation dataset. The correlation coefficient between manual- and automated cough counts in the validation dataset was 0.97 ($p < .001$). The intra- and inter-device reliability of the algorithm was adequate, and the algorithm performed best at an unobstructed distance of 0.5–1 m from the audio source. **Conclusion:** This novel smartphone-based pediatric cough detection application can be used for longitudinal follow-up in clinical care or as digital endpoint in clinical trials

.

Introduction

Coughing is a physiological mechanism of the respiratory system to clear excessive secretions. It can be caused by various acute and chronic diseases, such as viral upper respiratory tract infections, bacterial infections, asthma, protracted bacterial bronchitis or tic cough, and is a common reason for parents to seek medical consultation for their children.[1,2] Several studies have shown that cough severity is correlated with disease activity in asthma and other pulmonary diseases,[3–6] making cough frequency an attractive candidate biomarker for respiratory disease severity. Although coughing is traditionally quantified via self- or parent-report in the form of questionnaires, technological advances allow for more sophisticated (semi-) automatic cough monitoring methods. Indeed, several commercial and academic entities have endeavoured to develop cough detection algorithms, with varying success.[7] The most notable and reliable examples are the Leicester Cough Monitor and the VitaloJak, which record sounds with a dedicated body-contact device and microphone, and subsequently use semi-automated counting methods.[8,9] Several completely automated cough counting algorithms have been developed, mostly for an adult population, but none have proceeded towards widespread availability.[7] A summary of the key principles of automatic cough detection and a thorough overview of cough counting technologies used in a clinical setting is provided by Hall et al. [10] A notable disadvantage of body-contact devices is that they are inconvenient in the field of pediatrics, especially in infants and toddlers. Additionally, pediatric cough sounds exhibit more variability across different ages due to the developing respiratory- and vocal system, which can make robust detection more challenging. [11] An ideal algorithm would require no manual input, be able to monitor from a distance, and be operational on low-cost consumer devices that are readily available, such as smartphones. To date, no such algorithm has been developed in the field of pediatrics. This study aimed to develop an algorithm that objectively and automatically counts cough sounds in children based on audio features collected via a smartphone application.

Materials and methods

ETHICS AND LOGISTICS

This study was conducted at the Centre for Human Drug Research (CHDR, Leiden, The Netherlands) and the Haga Teaching Hospital, Juliana Children's Hospital (The Hague, The Netherlands). Institutional review board approval was obtained (registration number: T19-080), and the study was conducted in compliance with the general data protection regulation. The algorithm was developed as part of the CHDR MORE® system, a remote monitoring clinical trial platform. Reporting was performed in accordance with EQUATOR guidelines.[12]

DATA COLLECTION

A comprehensive training dataset was obtained from multiple sources. First, audio was extracted from 91 publicly available videos on YouTube that contained coughing children with an estimated age between 0 and 16 years old. Furthermore, 334 non-coughing audio clips were gathered from YouTube, GitHub, and the British Broadcasting Corporation sound library. The non-coughing set contained various sounds that were expected to occur in real-life settings, such as talking, breathing, footsteps, cats, sirens, dogs barking, cars honking, snoring, glass breaking, and church clocks. Additionally, 21 children aged 0–16 and admitted due to pulmonary disease were included, after obtaining informed consent from parents, on the general ward of Juliana Children's Hospital. Children were recorded during a day or night during the admission with a G6 (Motorola) smartphone. The smartphone contains two microphones and runs on Android 8.0 Oreo. Data of the first 7 children (3 diagnosed with bronchiolitis, 2 diagnosed with pneumonia, 1 with viral wheezing and 1 with an upper respiratory infection, age range from 2 weeks to 15 years) were used to supplement the training dataset, with a maximum of the first 150 coughs per child to avoid overrepresentation of a single subject. Remaining cough sounds of the 7 children were discarded. Data from the other 14 subjects were used as validation dataset. All audio clips were manually annotated by an investigator using Audition software (Adobe). No filter was applied to remove 'silent' sections of the recording to ensure that the estimated accuracy reflects real-life conditions. As a result, the proportion of cough sounds in the validation dataset was 0.7%. The composition of the final training- and validation dataset are displayed in Table 1.

AUDIO FEATURE EXTRACTION AND SELECTION

Audio feature were extracted from all audio clips using the Open-SMILE software (version 2.3.0, audEERING).[13] The software converted all audio clips into 1582 features per epoch. Epoch length was fixed at 0.5 s since the average cough duration in the training dataset was 0.3 s. The extracted features included several audio domains, such as Mel-frequency cepstral coefficients and fundamental frequencies (F0) (Supporting Information Text S1). Using manual inspection, the most robust features across multiple conditions were selected (Supporting Information Text S2) and only these features were included in the final dataset used for algorithm development.

ALGORITHM DEVELOPMENT AND VALIDATION

For the cough detection algorithm, we compared the classification performance of two ensemble-based decision-tree classifiers: Random Forests and Gradient Boosting Machines. Both differ in their process to build learners (also known as 'trees'). Random Forests classifiers build multiple trees simultaneously, each tree learnings a random subsample of the data. This subsampling makes the final model more robust as it is less likely to be biased towards the training data. Gradient Boosting Machines classifiers build one tree at a time, and each new tree corrects the prediction error of the previous tree. Five fold cross-validation was used to select the optimal features and hyperparameters for the model. Given that the number of coughs and non-coughs are imbalanced, the optimal classifier was selected based on the highest overall Matthew's Correlation Coefficient (MCC). The MCC score provides a more informative and reliable evaluation of binary classifications compared to accuracy as MCC takes

into account the number of true and false positives and negatives when assessing classification performance. The selected model was then used to classify all 0.5-s epochs in the validation dataset. The sensitivity, specificity, MCC, positive predictive value (PPV), and negative predictive value (NPV) were calculated for the complete validation dataset and per subject.

INITIAL ROBUSTNESS TESTS

Limited robustness tests were conducted to ensure the algorithm performs comparably across a range of different conditions when applied as a smartphone application. First, a 27-min long audio-clip was generated which included coughing- and household sounds, as well as sections with silence. The clip was subsequently played repeatedly from a speaker, while a G6 smartphone (Motorola) with the CHDR MORE® application was placed in proximity. The application has incorporated openSMILE software and is able to calculate and transmit the generated audio features. The following conditions were tested: first, the intra-device variability was tested by repeating the assessment 7 times with the same device; second, the inter-device variability was tested by repeating the assessment 4 times with different devices of the same type; third, the effect of device distance (0.5, 1, and 4 m) from the audio source was assessed and finally accuracy was assessed when a small (plant and book) or large (loft bed) barrier was placed in front of the audio source and when television sounds were played in the background. Because the 0.5-s epochs from the original file and the output of the MORE® application could not be paired, cumulative cough count plots were generated and compared across conditions.

RESULTS

ALGORITHM TRAINING

The training set consisted of 3424 0.5-s cough epochs of various sources, as well as 431,622 0.5-s non-cough epochs. The final algorithm, fitted through a Gradient Boost Classifier, achieved an accuracy of 99.6%, MCC of 73.7%, sensitivity of 99.6% and specificity of 99.9% in the training set (Table 2). The most important audio features the algorithm relied on were derived from the mel frequency and loudness categories (Supporting Information Text S3).

ALGORITHM VALIDATION

For validation, 14 patients with respiratory disease aged 0–14 were recorded during a hospital admission. The median recording duration was 632 (interquartile range [IQR]: 477–775) minutes. In total, 4123 0.5-s epochs contained coughing. The median cough count per subject was 150 (IQR: 38–446). Table 2 displays the overall accuracy of the algorithm in the validation dataset. Overall sensitivity was 47.6% and specificity was 99.96%. Due to the relatively low frequency of cough counts in the dataset, the NPV and PPV in these real-world settings were 99.78% and 82.2%, respectively. The performance of the algorithm differed between subjects. Individual patient characteristics and classification accuracies are displayed in Table 3. The correlation coefficient between manual cough count and automated cough count was 0.97 ($p < .001$, Figure 1).

LIMITED ALGORITHM ROBUSTNESS TESTS

Repeated (N = 7) tests with the same device and show comparable performance during each iteration (Figure 2A), while the inter-device variability tests show some variability in cumulative cough count across devices (Figure 2B). The effect of the distance of the device to the audio source was assessed (Figure 2C) and demonstrated comparable accuracy for 0.5 and 1 m distance. The accuracy was lower when the distance of the monitoring device from the audio source was increased. Finally, the effect of a small- and large barrier was investigated, as well as the effect of ambient television sounds playing in the background (Figure 2D). During this test, it appeared that a small physical barrier did not impact algorithm performance, but a large physical barrier and background television sounds led to a lower cumulative cough count.

Discussion

The current manuscript described the development and initial validation of a novel cough detection algorithm in pediatrics. Publicly available audio recordings were combined with real-life recordings to fit an algorithm that had excellent classification capability in the training dataset. In the validation dataset, a sensitivity of 47.6% and specificity of 99.96% was obtained, which resulted in a PPV of 82.2% and an NPV of 99.8% in these real-world conditions. There was a strong correlation between manual cough count and automatic cough count. The accuracy of the algorithm in the validation set was confirmed by several robustness tests, which repeatedly showed a cumulative cough count that was roughly half of the true cough count across various conditions. The algorithm performed best when there was a relatively unobstructed maximum distance of 0.5–1 m from the audio source.

The current sensitivity is suboptimal but does not disqualify the algorithm, and we envision the current algorithm is already suitable for application in several settings. Algorithm-derived cough count could be incorporated as (secondary) digital endpoint in pediatric pulmonary disease trials. For this application, clinical validation of cough count as digital endpoints should be performed first, focusing on demonstrating a difference between patients and healthy children, correlation of the novel endpoint with traditional endpoints or patient reported outcomes, and sensitivity to change in disease activity.[14] In addition to clinical trials, applying this algorithm in clinical care is likely to be much more reliable than patient- or parent recall regarding cough frequency.[15,16] The strong correlation between manually- and automatically- counted coughs means the algorithm can discriminate children that cough excessively from children that do not and can uncover individual trends over time, e.g., to characterize clinical recovery after a hospital admission, or to assess the effect of treatment in excessively coughing patients with persistent bacterial bronchitis. This is further supported by the very high specificity of the algorithm that is maintained in all validation tests. For example, change in nocturnal cough frequency in the case of an asthma exacerbation could be identified reliably with the current algorithm, and subsequent treatment leading to a significant decrease in nocturnal coughing will also be detectable even with the current sensitivity. In the future, algorithm output could be combined with other non-invasive assessments known to be related to pulmonary disease activity, such as physical activity, heart rate and pulmonary function monitoring, as well as electronic patient reported outcome measures. Together, this could provide a holistic overview of multiple aspects of pulmonary disease severity and quality of life.[17]

Multiple research groups have developed cough detection algorithms in recent years. However, only one was developed specifically for a pediatric population.[18] Although this algorithm was not applied in a mobile device. Still, pediatric cough detection is theoretically more challenging due to changing vocal cord acoustics during various stages of development. In adults, the most widely reported cough detection devices are the Leicester cough monitor and the VitaloJak.[7] These methods have been validated in independent datasets and appear both sensitive (91%–99%) and specific (99%), but the use of dedicated microphones is less user-friendly in general, and the use contact-devices precludes their use in several age categories in pediatrics. Furthermore, the semi-automated counting method used by both devices remains laborious and requires training, which means that widespread use in large-scale clinical trials or in general care is not feasible. Other algorithms that count coughs automatically have reported sensitivities of 78%–99% and specificities of 92%–99%,[7,18–23] but only a few have been applied on a smart phone.[21,22,24] The one that most resembles the current study is a smartphone-based algorithm developed by Barata et al.,[21] who use a convolutional neural network to classify nocturnal sounds in adult asthmatics and obtained a sensitivity of 99.9% with a specificity of 91.5%.[21] In addition, other projects are often based on data obtained in tightly controlled environments and lack validation in independent or clinical datasets,[18,22–24] and may show a similar drop in accuracy during validation as was observed for the algorithm developed here. For example, the PulmoTrack® device, designed for

automatic clinic-based monitoring, showed a reduced sensitivity of 26% compared to human annotation during validation in a new cohort.[25]

A major advantage of the algorithm developed in this study is the conversion of raw audio into audio features on the smartphone before transmission to the study center, which ensures the privacy of participants. The automated classification is another advantage, allowing devices to analyze and transmit cough counts in real-time. This study focuses on detecting single coughs, which was the reason for using a 0.5 s epoch during algorithm development. In the future, aggregation of data into 'cough bouts' could add additional value in measuring the impact and severity of respiratory diseases.[26] For real-world application of the algorithm, we envision that parents could use a spare phone to run the algorithm and leave the phone close to their child. Additionally, miniaturization of current technology could lead to a dedicated clip-on device to attach to (the bed of) infants with respiratory illness. A limitation was the manual feature selection performed, which introduces a potentially subjective factor to the analysis. Furthermore, a laptop speaker was used during the initial robustness tests and using a higher quality speaker may have led to slightly different performance during these tests. However, we believe the device quality is sufficient for the purpose of testing repeatability and investigating the effects of differing conditions. During this study, a single smartphone type (Motorola G6) was used, and the observed performance may vary when other devices are used.[27] Another potential problem would arise when the sensitivity of the algorithm would be highly dependent on the underlying disease that is studied, although there is no evidence of this in the validation dataset, such factors need to be studied further during clinical validation for which we can supply the algorithm to other interested academic groups. The current algorithm is developed as a one-size-fits-all solution that can classify coughs of all pediatric patient groups and ages and that only used sound features as input variables. Although the current accuracy appears sufficient to include as digital biomarker in the applications mentioned above, the accuracy of future algorithms could improve significantly with the cost of added complexity. First, accuracy could improve by addition of additional covariates such as age, sex, and diagnosis, although this would require some user input before use. Second, the exponential increase in processing power of mobile devices could allow for the development of personalized models in the future, which would both be trained, validated, and deployed on the participants' own smartphones. A personalized classification model that is tuned to the cough characteristics of an individual could potentially be much more accurate, considering the intra-individual variability in cough sounds is assumed to be smaller compared to inter-individual variability. Future studies could also aim to quantify cough intensity, as this characteristic may have greater impact on quality of life than cough frequency.

CONCLUSION

This novel smartphone-based cough detection application is one of the first of its kind and able to count coughs in pediatric patients with a sensitivity of 47%, specificity of 99.96%, PPV of 82% and NPV of 99.8%. Although the observed sensitivity in the intended use must be improved in the future, the current algorithm may be reliable enough for longitudinal monitoring in the context of clinical trials- or care, which will be evaluated during a clinical validation process.

TABLE 1 Composition of training and validation datasets

	Training dataset				Validation dataset
	YouTube (91 clips)	Various sources (334 clips)	Hospital (7 children)	Total	Hospital (14 children)
Cough sounds (n)	2229	–	999	3228	4123
Noncough sounds (n)	9702	39,456	431,622	480,780	100,522
Total (n)	11,931	39,456	432,621	484,008	104,645
Cough proportion (%)[1]	18.5%	0%	0.2%	0.7%	0.4%
Mean cough duration (s)	0.3	–	0.3	0.3	0.3

1. Proportion of 0.5-s epochs that contain cough sounds.

TABLE 2 Performance of the final algorithm

	Training dataset	Validation dataset
Parameter	Mean (SD) performance[1]	Overall performance
Accuracy	99.61% (±0.13%)	99.74%
MCC	73.67% (±0.16%)	62.40%
Sensitivity	99.62% (±0.13%)	47.56%
Specificity	99.89% (±0.09%)	99.96%
PPV	99.65% (±0.08%)	82.16%
NPV	99.82% (±0.02%)	99.78%

1. Mean (SD) performance of fivefold cross-validation.
MCC, Matthew's Correlation Coefficient; NPV, negative predictive value; PPV, positive predictive value.

TABLE 3 Performance of the final algorithm among individual subjects

Subject (#)	Age	Diagnosis	Recording duration (min)	Manual Count (N)	Algorithm count (N)	Sens.	Spec.	MCC
1	14 years	Pneumonia	4	22	7	32%	100%	55%
2	4 years	Wheezing	717	63	49	73%	100%	73%
3	5 years	Pneumonia	237	29	21	72%	100%	85%
4	1.5 years	Pneumonia	609	16	6	19%	100%	31%
5	6 weeks	Bronchiolitis	727	85	70	58%	100%	63%
6	3 years	Pneumonia	792	454	344	69%	100%	79%
7	9 weeks	Bronchiolitis	967	895	436	34%	100%	69%
8	4 years	Pneumonia/ wheezing	497	29	17	52%	100%	88%
9	11 years	Asthma	598	171	98	56%	100%	73%
10	5 weeks	Bronchiolitis	873	1038	516	37%	100%	53%
11	2 years	Pneumonia	434	474	355	70%	100%	81%
12	3 years	Pneumonia	470	420	256	54%	100%	68%
13	13 weeks	Bronchiolitis	654	128	45	34%	100%	57%
14	4 years	Pneumonia	791	299	166	40%	100%	53%

FIGURE 1 Correlation manual- and automatic cough count in validation dataset. Pearson correlation between manually counted coughs and automatically detected coughs. Each dot represents an individual subject in the validation dataset..

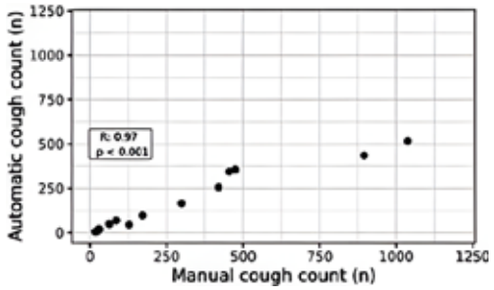
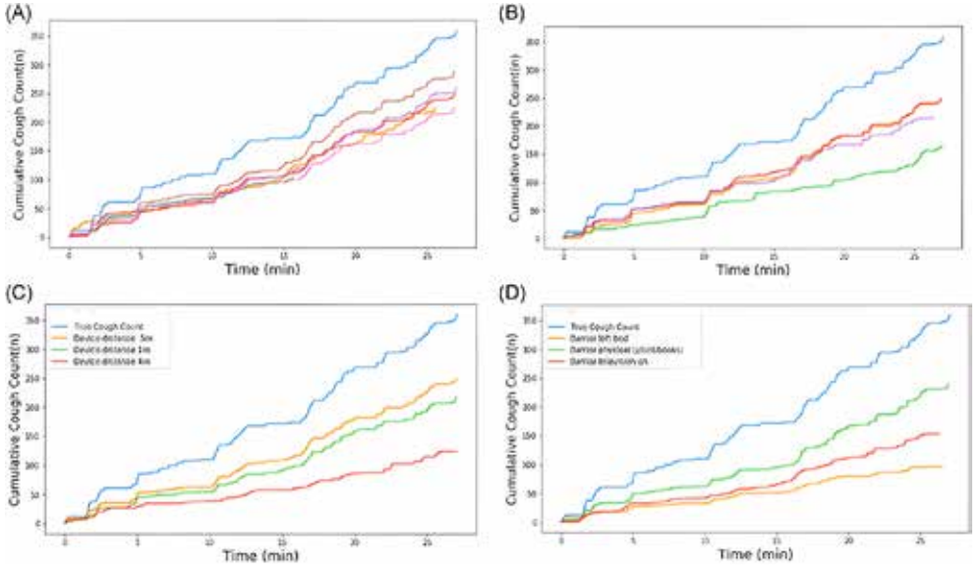


FIGURE 2 Performance of the algorithm under varying circumstances. (A) Intra-device repeatability. Each individual line represents a different session with the same device. (B) Inter-device repeatability. Each individual line represents a different session with a different device of the same type. (C) Influence of device distance from the audio source. (D) Influence of physical barrier or ambient background noise. In each of the panels, the light-blue line is the reference from the audio file.



SUPPLEMENTARY TEXT S1

OPENSMILE AUDIO features

OpenSMILE generated features from each 0.5 second epoch in the following domains: for each domain, the following statistics were derived by the openSMILE software:

Feature group	Description
Fundamental frequency (F0)	Pitch
Jitter and shimmer	Voice quality
Mel-frequency cepstrum (coefficients)	Power spectrum
Line spectral frequencies	Frequencies
Loudness	Sum of auditory spectrum. (Intensity & approximate loudness)
Voicing	Probability of voicing

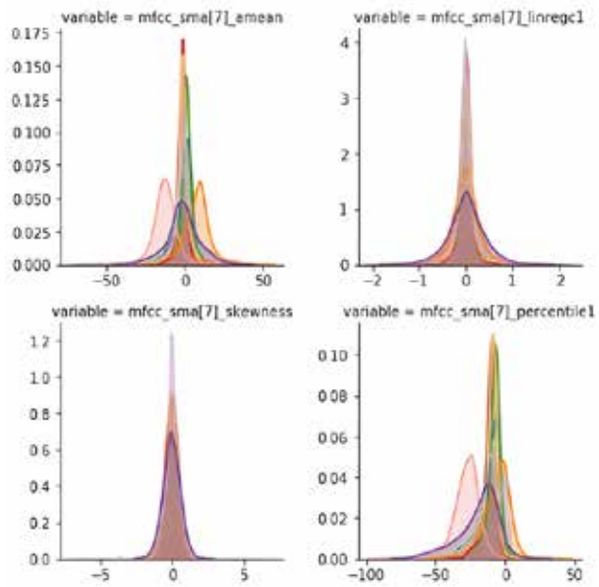
Statistics obtained from each feature during each 5-second epoch
Arithmetic mean
Quartiles and IQR ranges (1-2, 1-3, 2-3)
Skewness and kurtosis
Linear regression slope, offset and approximation error
Relative position of minimum and maximum
Percentile 1%, percentile 99% and range
Standard deviation
Percentage of frames above 75/90% of range

SUPPLEMENTARY TEXT S2

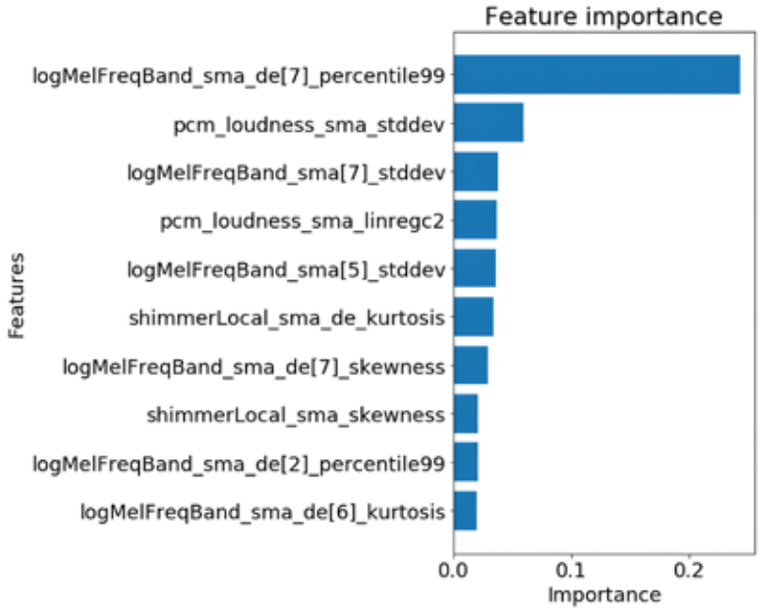
OpenSMILE feature selection

Feature selection was performed using the audio file generated during the robustness tests. The file was played back through a laptop speaker (B&O PLAY, incorporated in HP Pavilion 15-CK094ND) during differing ambient conditions (see paragraph Initial robustness tests in Materials & Methods), once more through a dedicated speaker (Luxman L-114A amplifier, Dali 6006SE speaker), and finally also processed using openSMILE software on a personal computer. Considering the data was derived from the exact same audio file, the frequency distribution of features should be identical during all conditions (see ***Supplementary Figure S2a*** below). However, this was not the case for all features, particularly those that were derived from the extremes of each feature (e.g. Percentile 1% percentile 99%). Therefore, distribution plots were judged visually by the authors and each feature that demonstrated a clear difference in means or standard deviations across conditions was excluded from the final dataset. Manual selection was preferred over statistical methods to compare distributions, as the large size of the dataset meant that statistical tests such as the Kolmogorov-Smirnov would have too much statistical power and irrelevant deviations would be flagged as significant difference.

SUPPLEMENTARY FIGURE S2A Example of distribution plots of each feature used during the feature selection process. Each color represents a different condition. Of the displayed features, the top right (mfcc_sma[7]_linregc1) and bottom left (mfcc_sma[7]_skewness) features were included in the final datasets.



SUPPLEMENTARY FIGURE S3 Feature importance plot of the final algorithm. On the y-axis, the 10 most important features derived from the openSMILE software are displayed. The bars and the x-axis represent the relative importance of each feature.



REFERENCES

1 Kantar A. Phenotypic presentation of chronic cough in children. J Thorac Dis. 2017;9:907-913.
2 Goldsobel AB, Chipps BE. Cough in the pediatric population. J. Pediatr. [Internet] Mosby, Inc. 156, 2010:352-358. doi:10.1016/j.jpeds.2009.12.004
3 Theodore AC, Tseng CH, Li N, Elashoff RM, Tashkin DP. Correlation of cough with disease activity and treatment with cyclophosphamide in scleroderma interstitial lung disease: findings from the scleroderma lung study. Chest. 2012;142:614-621.
4 Sato R, Handa T, Matsumoto H, Kubo T, Hirai T. Clinical significance of self-reported cough intensity and frequency in patients with interstitial lung disease: A cross-sectional study. BMC Pulm. Med. 2019;19:1-10.
5 Li AM, Tsang TWT, Chan DFY, et al. Cough frequency in children with mild asthma correlates with sputum neutrophil count. Thorax. 2006;61:747-750.
6 Van Der Giessen L, Loeve M, De Jongste J, Hop W, Tiddens H. Nocturnal cough in children with stable cystic fibrosis. Pediatr Pulmonol. 2009;44:859-865.
7 Cho PSP, Birring SS, Fletcher HV, Turner RD. Methods of cough assessment. J. Allergy Clin. Immunol. Pract. [Internet] Elsevier Inc. 7,2019:1715-1723. doi:10.1016/j.jaip.2019.01.049
8 Birring SS, Fleming T, Matos S, Raj AA, Evans DH, Pavord ID. The leicester cough monitor: preliminary validation of an automated cough detection system in chronic cough. Eur Respir J. 2008;31:1013-1018.
9 McGuinness K, Holt K, Dockry R, Smith J. P159 Validation of the VitaloJAK 24 Hour Ambulatory Cough Monitor. Thorax [Internet]. 67. BMJ Publishing Group Ltd; 2012:A131-A131 Available from.https://thorax.bmj.com/content/67/Suppl_2/A131.1
10 Hall JI, Lozano M, Estrada-Petrocelli L, Birring S, Turner R. The present and future of cough counting tools. J Thorac Dis. 2020;12: 5207-5223.
11 Chang AB. Pediatric cough: children are not miniature adults. Lung United States. 2010;188(Suppl):S33-S40.
12 Luo W, Phung D, Tran T, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J Med Internet Res. 2016;18:1-10.
13 Eyben F, Schuller B. OpenSMILE:). ACM SIGMultimedia Rec. 2015;6: 4-13.
14 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.
15 Morey MJ, Cheng AC, McCallum GB, Chang AB. Accuracy of cough reporting by carers of Indigenous children. J Paediatr Child Health. 2013;49:49-E203.
16 Chang AB, Newman RG, Carlin JB, Phelan PD, Robertson CF. Subjective scoring of cough in children: parent-completed vs child-completed diary cards vs an objective method. Eur Respir J. 1998;11:462-466.
17 Kruizinga MD, Stuurman FE, Groeneveld GJ, Cohen AF The Future of Clinical Trial Design: The Transition from Hard Endpoints to Value-Based Endpoints. 2019;371–397. Available from: http://link.springer.com/10.1007/164_2019_302
18 Amrulloh YA, Abeyratne UR, Swarnkar V, Triasih R, Setyati A. Automatic cough segmentation from non-contact sound recordings in pediatric wards. Biomed. Signal Process. Control [Internet] Elsevier Ltd. 2015;21:126-136. doi:10.1016/j.bspc.2015.05.001
19 Coyle M, Keenan D, Henderson L, et al. Evaluation of an ambulatory system for the quantification of cough frequency in patients with chronic obstructive pulmonary disease. Cough. 2005;1:3.
20 Vizel E, Yigla M, Goryachev Y, et al. Validation of an ambulatory cough detection and counting application using voluntary cough under different conditions. Cough. 2010;6:1-8.
21 Barata F, Tinschert P, Rassouli F, et al. Automatic recognition, segmentation, and sex assignment of nocturnal asthmatic coughs and cough epochs in smartphone audio recordings: observational field study. J Med Internet Res. 2020;22(7):e18082.
22 Monge-Alvarez J, Hoyos-Barcelo C, Lesso P, Casaseca-De-La- Higuera P. Robust detection of audio-cough events using local hu moments. IEEE J. Biomed. Heal. Informatics. 2019;23:184-196.
23 Pramono RXA, Imtiaz SA, Rodriguez-Villegas E. A cough-based algorithm for automatic diagnosis of pertussis. PLoS One. 2016;11:1-20.
24 Hoyos-Barceló C, Monge-Álvarez J, Pervez Z, San-José-Revuelta LM, Casaseca-de-la-Higuera P. Efficient computation of image moments for robust cough detection using smartphones. Comput Biol Med. 2018;100:176-185.
25 Turner RD, Bothamley GH. How to count coughs? Counting by ear, the effect of visual data and the evaluation of an automated cough monitor. Respir. Med. [Internet] Elsevier Ltd. 2014;108:1808-1815. doi:10.1016/j.rmed.2014.10.003
26 Chung KF, Bolser D, Davenport P, Fontana G, Morice A, Widdicombe J. Semantics and types of cough. Pulm. Pharmacol. Ther. [Internet] Elsevier Ltd. 2009;22:139-142. doi:10.1016/j.pupt.2008.12.008
27 Barata F, Kipfer K, Weber M, Tinschert P, Fleisch E, Kowatsch T Towards device-agnostic mobile cough detection with convolutional neural networks. 2019 IEEE Int. Conf. Healthc. Informatics, ICHI 2019 IEEE; 2019; 1–11.

CHAPTER 7

Development and technical validation of a smartphone-based cry detection algorithm

Ahnjili ZhuParris,[1] Matthijs D. Kruizinga,[1,2,3] Max van Gent,[1,2] Eva Dessing,[1,2] Vasileios Exadaktylos,[1] Robert Jan Doll,[1] Frederik E. Stuurman,[1,3] Gertjan A. Driessen [2,4] and Adam F. Cohen[1,3]*

Front Pediatr. 2021;9:262. doi:10.3389/fped.2021.651356

1 Centre for Human Drug Research, Leiden, Netherlands
2 Juliana Children's Hospital, Haga Teaching Hospital, The Hague, Netherlands
3 Leiden University Medical Centre, Leiden, Netherlands,
4 Department of Pediatrics, Maastricht University Medical Centre, Maastricht, Netherlands

Abstract

Introduction: The duration and frequency of crying of an infant can be indicative of its health. Manual tracking and labeling of crying is laborious, subjective, and sometimes inaccurate. The aim of this study was to develop and technically validate a smartphone-based algorithm able to automatically detect crying. **Methods:** For the development of the algorithm a training dataset containing 897 5-s clips of crying infants and 1,263 clips of non-crying infants and common domestic sounds was assembled from various online sources. OpenSMILE software was used to extract 1,591 audio features per audio clip. A random forest classifying algorithm was fitted to identify crying from non-crying in each audio clip. For the validation of the algorithm, an independent dataset consisting of real-life recordings of 15 infants was used. A 29-min audio clip was analyzed repeatedly and under differing circumstances to determine the intra- and inter- device repeatability and robustness of the algorithm. **Results:** The algorithm obtained an accuracy of 94% in the training dataset and 99% in the validation dataset. The sensitivity in the validation dataset was 83%, with a specificity of 99% and a positive- and negative predictive value of 75 and 100%, respectively. Reliability of the algorithm appeared to be robust within- and across devices, and the performance was robust to distance from the sound source and barriers between the sound source and the microphone. **Conclusion:** The algorithm was accurate in detecting cry duration and was robust to various changes in ambient settings.

Introduction

Crying is a primary indicator of decreased infant well-being.[1] Besides the normal crying-behavior that is natural for every infant, a change in cry duration, intensity or pitch can be a symptom of illness.[2] Cry duration has been used as a biomarker for diagnostic and follow-up purposes for a wide range of clinical conditions of infancy, such as gastroesophageal reflux and cow milk allergy.[3,4] However, traditional methods to record cry behavior, such as parent- or nurse- reported cry duration, are subjective and vulnerable to observer bias.[5] On the other hand, more objective manual annotating of audio recordings is labor intensive and may be subject to privacy-concerns by parents. An objective, automated and unobtrusive method to quantify crying behavior in an at-home and clinical setting may improve the diagnostic process in excessively crying infants, allow for objective determination of treatment effects by physicians, and enable researchers to include objectively determined cry duration as digital biomarker in clinical trials. Therefore, a classification algorithm is necessary for the automatic recognition of cries in audio files. Given the importance for researchers to study the relationship between an infant's crying patterns and their health, automatic detection and quantification of infant cries from an audio signal is an essential step in remote baby monitoring applications.[6]

Automatic cry detection has been reported in the form of remote baby monitors for non-intrusive clinical assessments of infants in hospital settings,[6–9] and several researchers have shown that classification of cry- and non-cry-sounds is possible with machine-learning algorithms.[10–12] However, most algorithms lack validation in a completely independent dataset, which is crucial to predict performance in new- and real-world settings, while data regarding intra- and inter-device variability and other factors that may influence repeatability is lacking as well.[10,13,14] Finally, algorithms are often developed for use on personal computers or dedicated devices. Usability of an algorithm would be increased if it were available on low-cost consumer-devices such as smartphones, which are readily available

in most households and are easy to operate. Furthermore, smartphones have adequate processing power to analyse and transmit data continuously for monitoring in real-time. The aim of this study was to develop and validate a smartphone-based cry-detection algorithm that is accurate, reliable, and robust to changes in ambient conditions.

Materials and methods

LOCATION AND ETHICS

This was a prospective study conducted by the Center for Human Drug Research (CHDR) and Juliana Children's Hospital. The study protocol was submitted to the Medical Ethics Committee Zuidwest Holland (ID 19-003, Leiden, Netherlands), who judged the protocol did not fall under the purview of the Dutch Law for Research with Human Subjects (WMO). The study was conducted in compliance with the General data protection regulation (GDPR). The algorithm was developed and reported in accordance with EQUATOR guidelines.[15]

ALGORITHM DEVELOPMENT

TRAINING DATASET A training dataset was obtained from various online sources (Supplementary Table 2) and consisted of both crying- and non-crying sounds. Non-crying sounds consisted of common real-life sounds and included talking, breathing, footsteps, cats, sirens, dogs barking, cars honking, snoring, glass breaking, and ringing of church clocks. Furthermore, non-crying infant sounds (hiccoughs, wailing, yelling, babbling, gurgles, and squeaking), as well as adult crying sounds, were included in the training dataset. All sounds were played back through a loudspeaker and processed into non-overlapping 5-s epochs on a G5 (Motorola, Chicago, IL, USA) or G6 (Motorola, Chicago, IL, USA) smartphones and. A total of 1,591 audio features (Supplementary Text 3) were extracted from each 5-s epoch with openSMILE (version 2.3.0, audEERING, Gilching, Germany) [16] on the smartphone. Each 5-s epoch was manually annotated as crying or non-crying by a single investigator. A 5-s epoch was selected because the median cry duration (without a silent break) in the training dataset was 4s.

ALGORITHM TRAINING

To prevent overfitting of the algorithm on non-robust audio features provide by the software, manual feature selection was performed to exclude features that exhibited different distributions when analyzed under different conditions (Supplementary Text 3). Feature selection was performed using the audio file generated during the robustness tests. The file was played back through a laptop speaker during differing ambient conditions with (see paragraph Robustness-tests in section Materials and Methods), a dedicated speaker, and processed to openSMILE features with the CHDR MORE® application. Additionally, the raw file was processed using openSMILE software on a personal computer. Considering the data was derived from the exact same audio file, the distribution of features should be identical during all conditions (Supplementary Text 3). However, this was not the case for all features, particularly those that were derived from the extremes of each feature (e.g., Percentile 1% percentile 99%). Therefore, distribution plots were judged visually by the authors and each feature that demonstrated a clear difference in means or standard deviations across conditions was excluded from the final dataset. After selection, 980 features audio features remained in the dataset. Two discriminative classifiers Random Forest and Logistic Regression[17–20] and one generative classifier (Naïve Bayes) were considered for the classification of crying and non-crying sounds. For each classifier, a 5-fold cross-validated grid-search to select the best combination of features and hyperparameters was performed to minimize the error estimates in the final model. The primary objective of the model was to identify crying and therefore, hyper-parameters that optimized for sensitivity were prioritized. This was followed by 5-fold cross-validation to robustly estimate the model performance and generalization of the model. The classifier with the highest Matthew's Correlation Coefficient (MCC) was chosen as the final model and subjected to algorithm validation.

ALGORITHM VALIDATION

DATA COLLECTION An independent validation dataset was obtained from two sources. First, audio recordings were made in an at-home setting of 4 babies aged 0–6 months using the G5 or G6 smartphones. Second, audio recordings were made with the G5 or G6 smartphones of 11 babies aged 0–6 months admitted to the pediatric ward due to various reasons. Audio recordings were made after obtaining informed consent from both parents and were stripped of medical- and personal information prior to analysis.

PERFORMANCE ANALYSIS Each 5-s epoch in the recordings was annotated as crying- and non-crying by one annotator. In the case of doubt on how to classify an epoch, two additional annotators were included, and a choice was made via blinded majority voting. The developed algorithm was used to classify each epoch, and annotations and classifications were compared to calculate the accuracy, MCC, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) in the complete dataset and in the hospital- and home datasets separately.

POST-PROCESSING OF CRY EPOCHS INTO NOVEL BIOMARKERS Some infants are reported to cry often, but with short intervals in between. Only counting the number of epochs that contain crying for such infants could result in an underestimation of the burden for infants and parents. As such, the duration of 'cry sequences' (periods during which an infant is crying either continuously or occasionally) is an important additional feature. To calculate this, post-processing of detected cries was performed to calculate the number and duration of cry sequences as separate candidate biomarkers. A cry sequence was defined by the authors with a start criterion (at least six 5-s epochs containing crying within 1 min) and a stop criterion (no crying detected for 5 min). Individual timelines were constructed for true- and predicted cry sequences to determine the reliability of the algorithm for this novel biomarker.

ROBUSTNESS TESTS A series of robustness tests was conducted to ensure that the developed algorithm was robust to varying conditions when used with a smartphone with the final application (CHDR MORE®) installed, which is how the algorithm would be deployed in practice. A 29-min-long clip containing 16.7 min of crying was played from a speaker with a smartphone with the CHDR MORE® application in proximity. This application, developed in-house, has incorporated openSMILE technology and is able to extract and transmit audio features. The following conditions were tested during this phase of the study: intra-device variability (N = 10), inter-device variability (N = 10), distance from audio source (0.5, 1, 2, and 4 m) and by placing the phone behind several barriers and in the presence of background tv sounds. For intra-device variability, a single phone was used 10 times to determine repeatability within a single device. For inter-device variability, 10 different devices of the same type (G6) were used to determine the repeatability across devices. Because it was not technically possible to pair the application output with the raw audio features of the original recording, cumulative cry count plots were construed for each condition and compared with cumulative cries in the original recording. A schematic overview of the analysis steps is displayed in Supplementary Figure 1.

Results

ALGORITHM TRAINING

The training set consisted of 897 5-s audio clips, as well as 1,263 non-crying 5-s clips. Of the three methods applied to develop the algorithm, the Random Forest method achieved the highest accuracy and MCC with 93.8 and 87.3%, respectively (Table 1). The 10 most important audio features for the algorithm were derived from Mel Frequency cepstral coefficients, Mel frequency bands and Voicing Probability. A variable importance plot of the 10 most important features included in the final algorithm is displayed in Supplementary Figure 4.

ALGORITHM VALIDATION

The 15 infants [mean age: 2 months (SD 1.9)] created a total of 150 min (1,805 5-s epochs) of crying and 4,372 min (52,464 5-s epochs) of non-crying. The median cry duration of the infants recorded at home was shorter (1.4 min, IQR 0.58–2.6) compared to children recorded during their admission to the hospital (5.8 min, IQR 2.2–16.7). Performance of the algorithm in the independent validation dataset is displayed in Table 1. Overall accuracy was 98.7%, but sensitivity was lower (83.2%) compared to the performance in the training dataset. Due to the relatively low crying incidence compared to non-crying incidence, the specificity of 99.2% led to a PPV of 75.2%. Supplementary Figure 5 displays individual timelines for each infant, displaying the epochs where crying- and misclassifications were present. After post-processing of cry epochs into cry sequences, the median number of cry sequences per infant in the validation dataset was 3 (IQR 1–3), for a total of 39 cry sequences. The median difference between true and predicted cry sequences was 1 (IQR 0.25–1). Furthermore, the median difference between true and predicted cry sequences duration was 6 min (IQR 2–15 min, Table 2). Individual timelines and concordance between true and predicted cry sequences are displayed in Figure 1.

ALGORITHM ROBUSTNESS

To ensure the algorithm and smartphone application performs sufficiently for the intended use, multiple tests were conducted to test robustness with the resulting smartphone application. Figure 2A shows the estimated repeatability of the algorithm by repeatedly classifying the same recording with the same device. Figure 2B shows the cumulative cry count of 8 different devices of the same type, which gives an indication of repeatability. The distance from the audio source, up to 4 meters, did not appear to impact the accuracy of the algorithm (Figure 2C). Finally, blocking the audio signal by placing the phone behind several physical barriers in front of the audio source demonstrated comparable accuracy across conditions (Figure 2D). Creating additional background noise generated by a television appeared to slightly decrease the specificity of the algorithm, as the final cry count according to the algorithm was higher compared to the true number of cries in the audio file.

Discussion

This paper describes the development and validation of a smartphone-based cry detection algorithm in infants. A random forest classifier had the highest accuracy in the training dataset and achieved a 98.7% accuracy in an independent validation set. Although the sensitivity of 83.2% was slightly lower compared to the estimated accuracy in the training dataset, the individual classification timelines show that this should not lead to unreliable estimation of cry duration. The fact that most misclassifications occurred directly before or after crying indicates that such misclassifications may be due to cry-like fussing, which are difficult to classify for both the algorithm and the human annotators. Post-processing of the detected cry epochs into cry sequences decreased the mismatch and resulted in excellent performance for each individual infant.

The observed accuracy of the algorithm is comparable to others described in the literature, although there is large variation in reported accuracy. Traditional machine learning classifiers and neural network-based classifiers have been used for infant cry analysis and classification.[21] We found that several studies that explored the use of minimum, maximum, mean, standard deviation and the variance of MFCCs and other audio features to differentiate normal, hypo-acoustic and asphyxia types using the Chillanto database.[6] Support Vector Machines (SVM) are among the most popular infant classification algorithms and routinely outperform neural network classifiers.[22,23] Furthermore, Osmani et al. have illustrated that boosted and bagging trees outperform SVM cry classification.[24] Additionally, sensitivities between 35 and 90% with specificities between 96 and 98% have been reported using a convoluted neural network approach.[10,14] Ferreti et al. and Severini et al. also used a neural network approach and achieved a reported precision of 87 and 80%,

respectively.[11,12] However, algorithms often lack validation in an independent dataset as, and real-life performance in new and challenging environments will most likely be lower. Our algorithm has several advantages compared to other approaches that have been described in the past. Most importantly, the algorithm was validated on independent and real-life data obtained from two settings where the application could be used in the future. Validation invariably leads to a drop in accuracy compared to the performance of the training data but gives reassurance regarding the generalizability of the algorithm in new settings that were not included during training. Furthermore, the algorithm can be deployed on all Android smartphones and no additional equipment is needed for acquiring the acoustic features. Although it is possible to implement complex deep learning algorithms on portable devices, we demonstrated that a shallow learning algorithm such as a random forest achieves good classifying capability. This means that audio processing and classification can be performed on the device in real-time with the MORE® application, and thus, precludes direct transmission of audio to a central location with inherent preservation of privacy. Finally, the manual feature selection that was performed should lead to further generalizability of the algorithm in new condition, since the observed variability in the excluded audio features would most likely result in a drop in accuracy in challenging acoustic environments. While automated feature selection methods could have been used, automated feature selection requires a static definition of similarity between distributions within features. This is not a straightforward task. Given the nature of the features, we chose to manually exclude features that presented a clearly different distribution from the rest of the features.

All in all, the performance of the algorithm in combination with the mentioned advantages indicate reliability of the algorithm and may be preferable over manual tracking of cry duration through a diary in several situations. Although the literature regarding sources of inaccuracy in cry monitoring via a diary is sparse, several factors make manual tracking through a diary a subjective assessment.[5] Observer bias can cause parents to overestimate the true duration of crying, and placebo-effects may cause parents to underestimate true cry duration after an intervention.[25] Additionally, parents may underreport nocturnal cry duration when they sleep through short cry sequences during the night. Current tracking of cry duration in clinical settings is performed by nurses, who have other clinical duties as well, possibly making the quality of the cry diary dependent on the number of patients under their care. While the consequences of all these factors are not easy to quantify, the combination of these sources of inaccuracy leads to the conclusion that objective and automated cry-monitoring could significantly improve the reliability of objective follow-up of cry duration in both clinical trials and -care. Still, parental report of cry duration and cry behavior will remain an important component of follow-up.

A technical limitation of any Android application, including the MORE® application, is that continuous recording can be interrupted by other smartphone applications apps that also access the microphone, like phone calls. However, using a dedicated smartphone for the purpose of cry monitoring will diminish this limitation. Only Motorola G5/G6 phones were used during each phase of algorithm development and validation. Although performance on other smartphones is uncertain, the approach used in this paper could easily be replicated to adapt the algorithm to other devices and obtain a similar accuracy. In the future, incorporation of covariates such as age, sex or location in the model may improve classifying capability even further, and further stratification could allow to discriminate different types of crying. In this manner cries from asphyxiated infants,[26] pre-term infants,[27] or infants with respiratory distress syndrome could be differentiated from healthy infants.[13] One potential technical limitation of our approach is the use of loudspeakers to create the training dataset. An ideal training dataset would include smartphone-based audio recordings of multiple subjects under different conditions over a long period of time. We found the most appropriate alternative was to re-record open-sourced cry corpus using smartphone. While the playback could have potentially hindered the quality of the openSMILE features

and thus the classification, it resulted in excellent classification performance of the home and hospital recordings. Hence the impact of the quality of the loudspeaker-based dataset was deemed acceptable. A follow-up study that uses an original smartphone-based cry corpus could potentially improve the accuracy of the classification algorithm. The start- and stop criteria used to determine the beginning and end of a cry sequence are a new proposal that was not previously described in the literature. However, the criteria appear reasonable and individual timeline figures demonstrated that this post-processing step was able to generate a solid high-level overview of individual cry behavior. Still, alternative criteria could obtain similar accuracy and may be explored in the future.

The developed algorithm already provides an excellent overview of the cry behavior of infants and preliminary tests of the robustness of the resulting algorithm show inter- and intra-device repeatability and reliability up to 4 m from the audio source. The algorithm can replace current methods to track cry behavior, such as cry diaries, in clinical and at-home settings. However, more research is needed before implementing the cry duration and the amount of cry sequences as digital endpoint in trials. Clinical validation of cry duration and cry sequence count as digital biomarker in a patient population is necessary, and should focus on establishing new normative values for objectively determined cry sequence duration and count, the difference between patients and healthy controls, correlation with disease-severity and sensitivity to change after an intervention.[28]

Conclusion

The proposed smartphone-based algorithm is accurate, robust to various conditions and has the potential to improve clinical follow-up of cry behavior and clinical trials investigating interventions to enhance infant well-being.

TABLE 1 Performance of the final algorithm

Training Dataset		Validation Dataset		
Parameter	**Performance [Mean (SD)]***	**Hospital Subjects (N=11)(%)**	**Home Subjects (N=4) (%)**	**All Subjects (N=15)(%)**
Accuracy	93.8% (±1%)	98.5	99.7	98.7
MCC	87.3% (±2.2%)	75.5	98.6	78.4
Sensitivity	93.8% (±1.1%)	80.6	97.5	83.2
Specificity	94.8% (±1.1%)	99.1	100	99.2
PPV	-	72.2	100	75.2
NPV	-	99.4	99.6	99.5

TABLE 2 Individual Algorithm Performance

Characteristics			Cry Epochs					Cry Sessions			
Subject	**Duration (min)**	**Annotated count (n)**	**Algorithm count (n)**	**Sensitivity (%)**	**Specificity (%)**	**PPV (%)**	**NPV (%)**	**Annotated cry sequence count (n)**	**Algorithm cry sequence count (n)**	**Annotated cry sequence duration**	**Algorithm cry sequence duration**
Hospital Dataset											
1	764	145	120	80	99.5	66.2	99.7	3	5	37	59
2	610	65	43	90.7	99.6	60	99.9	3	3	19	21
3	245	12	11	90.9	99.9	83.3	99.9	1	1	5	6
4	648	52	20	80	99.5	30.7	99.5	3	3	17	25
5	540	17	12	91.7	99.9	64.7	99.9	1	1	7	8
6	317	721	711	82.3	95.6	81.1	95.9	7	7	117	122
7	16.5	26	24	87.5	97.1	80.7	98.2	1	1	6	8
8	441	200	148	66.5	98.2	52.5	98.9	7	8	55	72
9	77.5	70	80	75	98.8	85.7	97.7	3	3	18.5	26
10	365	99	79	62	98.8	49.5	99.2	3	3	22	36
11	452	320	290	87.9	98.7	79.7	99.2	6	7	64	80
Home Dataset											
12	36	38	40	95	100	100	99.5	1	1	2.8	2.4
13	13	7	7	100	100	100	100	0	0	0	0
14	2	25	25	100	100	100	100	0	0	0	0
15	1	8	8	100	100	100	100	0	0	0	0

FIGURE 1 True and predicted cry sequence per infant

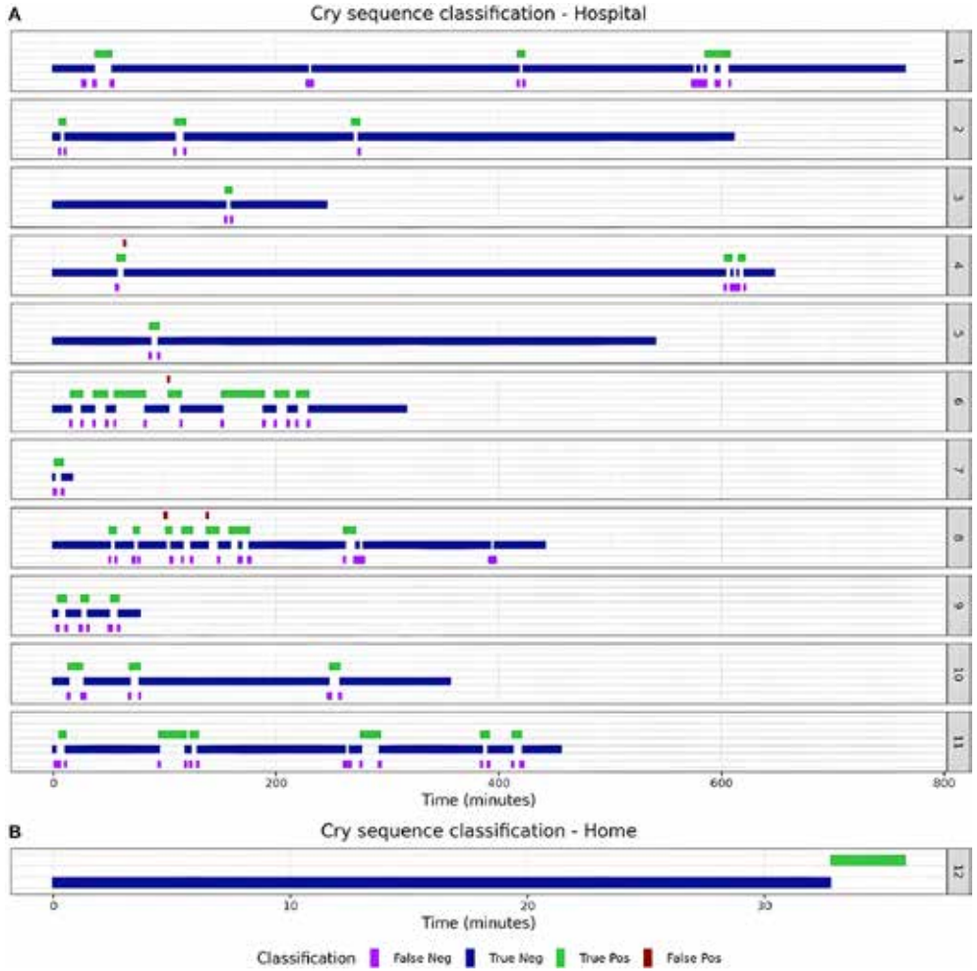
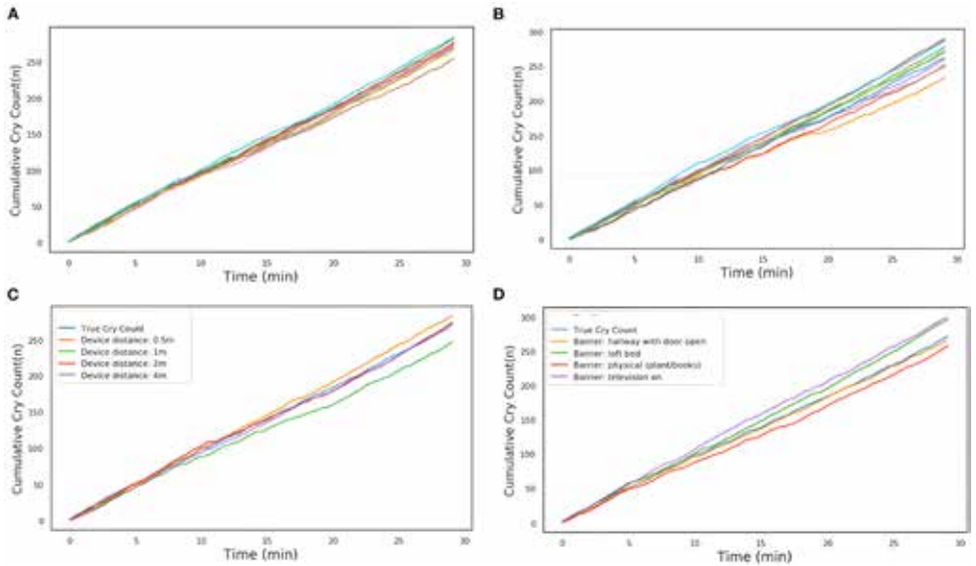
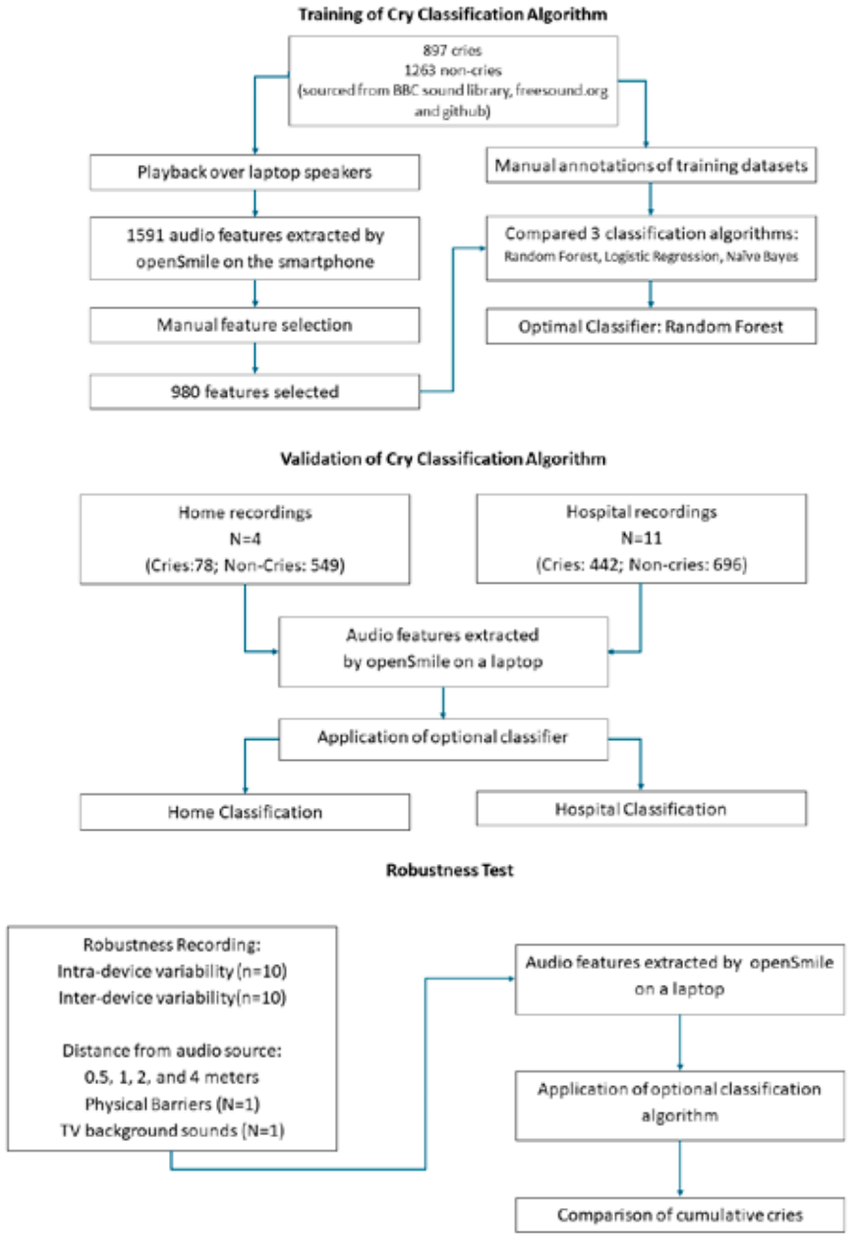


FIGURE 2 Cumulative cry count during robustness tests. (A) Intra-device repeatability. Each individual line is a different run with the same phone. (B) Inter-device repeatability. Each individual line is a run with a different phone of the same type. (C) Influence of device distance from the audio source. (D) Influence of physical barrier or ambient background noise. In each of the panels, the light-blue line is the reference from the audio file.



SUPPLEMENTARY FIGURE S1 Schematic overview of analysis steps



SUPPLEMENTARY TABLE 2 Audio sources

Dataset	Source	Crying (5-second epochs)	Non-crying** (5-send epochs)
Training Datasets	Github repository: *** https://github.com/giulbia/baby_cry_detection	146	216
	Freesound.org: *** https://freesound.org/search/?q=infant+cry	102	15
	British Broadcasting Company sound library: *** https://sound-effects.bbcrewind.co.uk/	207	336
Home Validation Dataset	Home Recordings	78	549
Hospital Validation Dataset	Hospital recordings	350	594
	Merged epochs*	92	102
Total		975	1812

** Merged crying sounds with additional background noise. ** The non-crying sound included common baby sounds (babies hiccoughing, gurling, babbling and yelling), common human sounds (breathing, coughing, talking), general indoor sounds (doors closing, footsteps and vacuuming) and general outdoor sounds (birds, thunder, sirens). *** This is a labelled collection of environmental audio recordings. The audio recordings have been extracted from public field recordings.*

SUPPLEMENTARY TEXT S3 Audio features and feature selection.

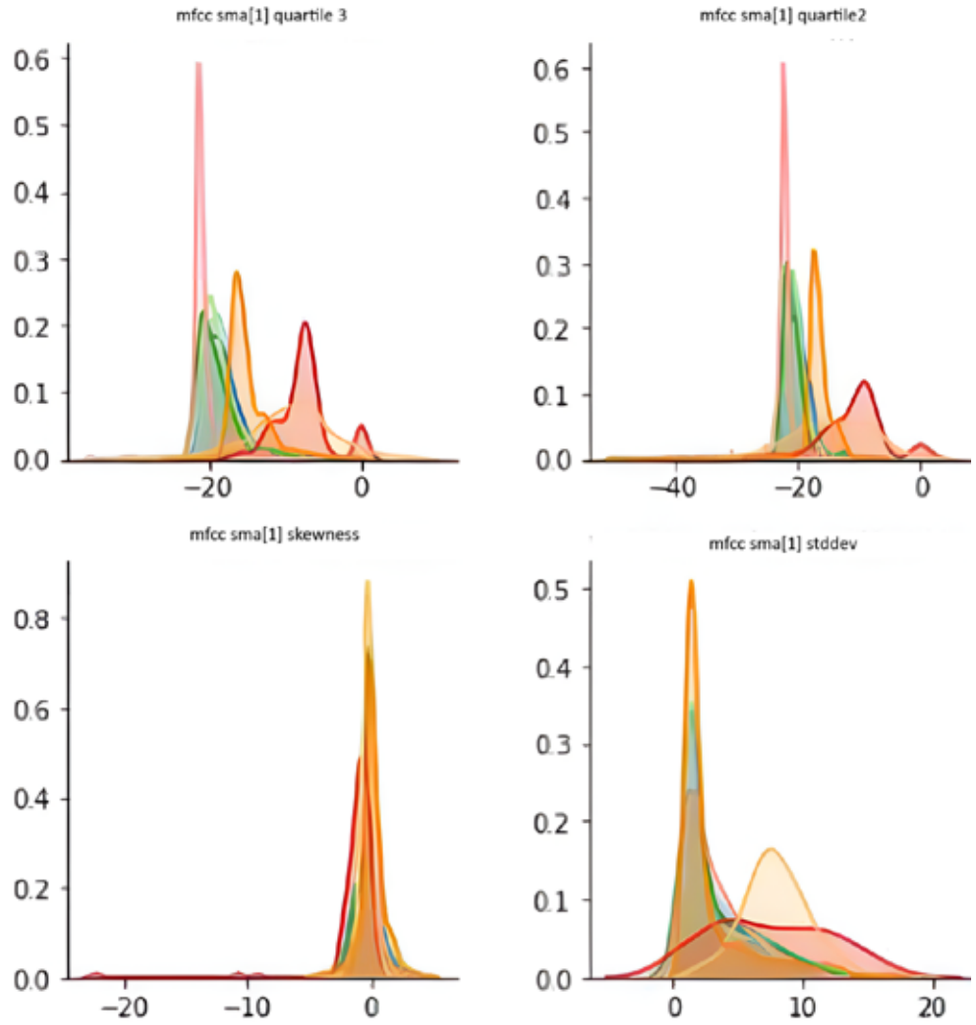
OpenSMILE generated features from each 5 second epoch in the following domains:

Feature group	Description
Fundamental frequency (F0)	Pitch
Jitter and shimmer	Voice quality
Mel-frequency cepstrum (coefficients)	Power spectrum
Line spectral frequencies	Frequencies
Loudness	Sum of auditory spectrum (Intensity & approximate loudness)
Voicing	Probability of voicing

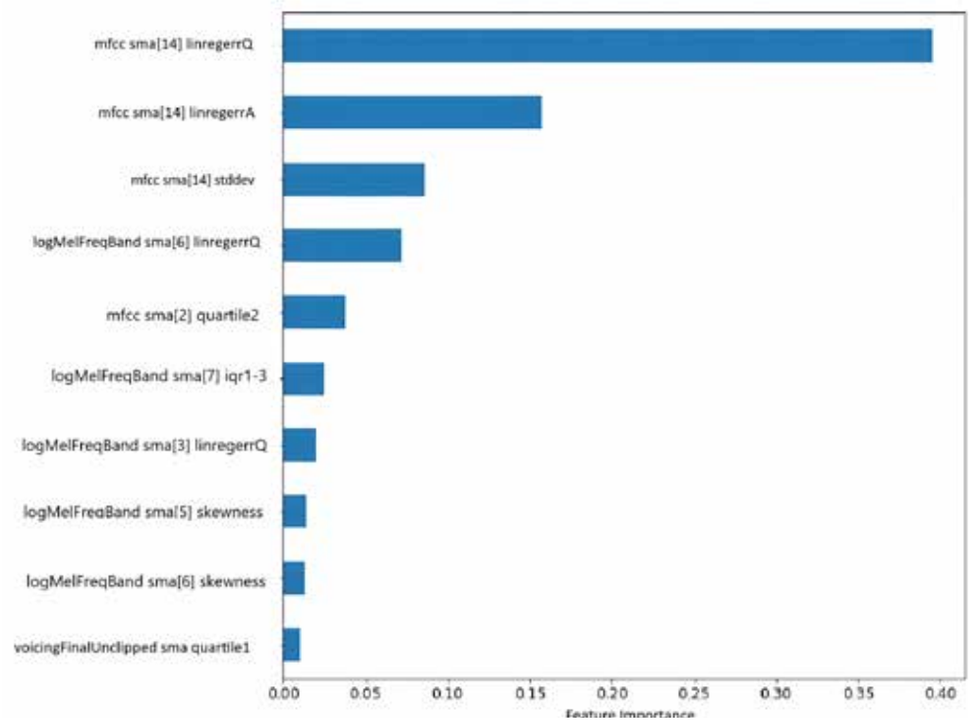
For each domain, the following statistics were derived by the openSMILE software:

Statistics obtained from each feature during each 5-second epoch
Arithmetic mean
Quartiles and IQR ranges (1-2, 1-3, 2-3)
Skewness and kurtosis
Linear regression slope, offset and approximation error
Relative position of minimum and maximum
Percentile 1%, percentile 99% and range
Standard deviation
Percentage of frames above 75/90% of range

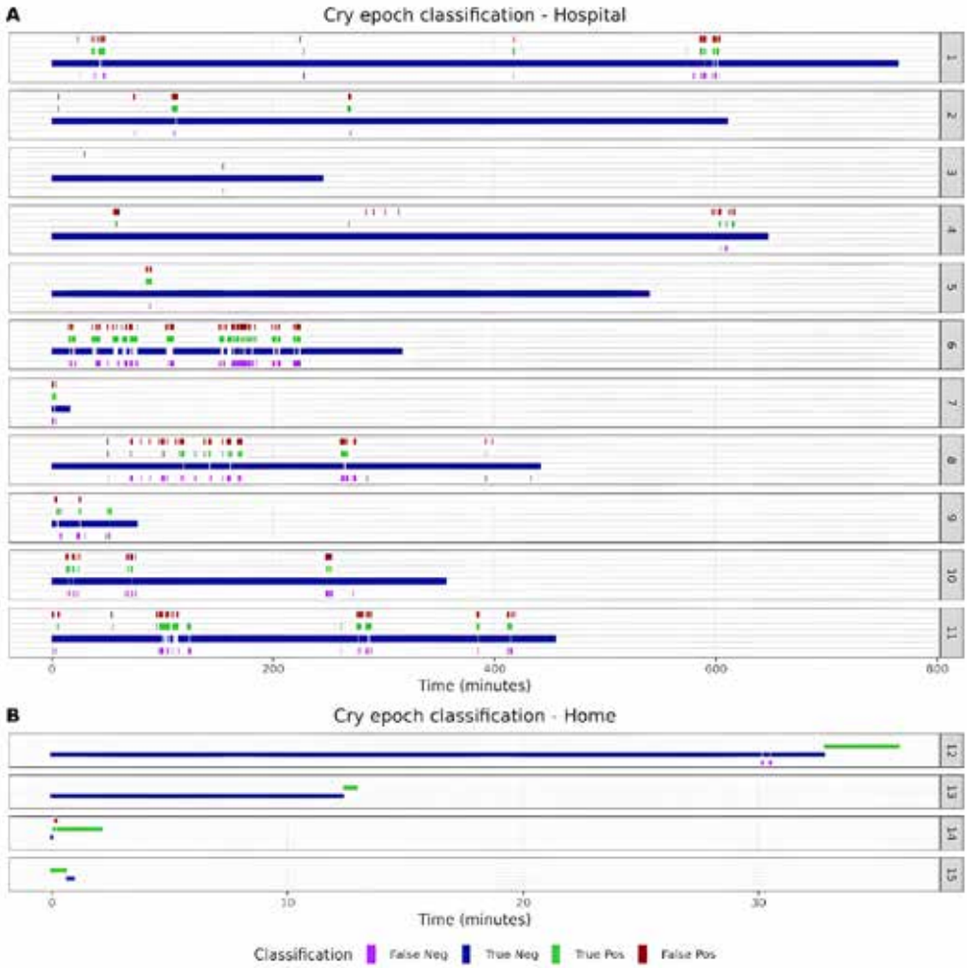
SUPPLEMENTARY FIGURE S3A Example of distribution plots of each feature used during the feature selection process. Each color represents a different condition. Of the displayed features, only the bottom left feature (mfcc_sma[1] skewness) was included in the final dataset.



SUPPLEMENTARY FIGURE S4 Variable importance Feature importance plot of the final algorithm. On the y-axis, the 10 most important features derived from the openSMILE software are displayed. The bars and the x-axis represent the relative importance of each feature.



SUPPLEMENTARY FIGURE S5 True and predicted crying epochs per infant



REFERENCES

1 Wolke D, Bilgin A, Samara M. Systematic review and meta-analysis: fussing and crying durations and prevalence of colic in infants. J Pediatr. (2017) 185:55–61.e4. doi: 10.1016/j.jpeds.2017.02.020

2 Freedman SB, Al-Harthy N, Thull-Freedman J. The crying infant: diagnostic testing and frequency of serious underlying disease. Pediatrics. (2009) 123:841–8. doi: 10.1542/peds.2008-0113

3 Moore DJ, Siang-Kuo Tao B, Lines DR, Hirte C, Heddle ML, Davidson GP. Double-blind placebo-controlled trial of omeprazole in irritable infants with gastroesophageal reflux. J Pediatr. (2003) 143:219–23. doi: 10.1067/S0022-3476(03)00207-5

4 Lucassen PLBJ, Assendelft WJJ, Gubbels JW, Van Eijk JTM, Douwes AC. Infantile colic: crying time reduction with a whey hydrolysate: a double-blind, randomized, placebo-controlled trial. Pediatrics. (2000) 106:1349–54. doi: 10.1542/peds.106.6.1349

5 Barr RG, Kramer MS, Boisjoly C, McVey-White L, Pless IB. Parental diary of infant cry and fuss behaviour. Arch Dis Child. (1988) 63:380– 7. doi: 10.1136/adc.63.4.380

6 Jeyaraman S, Muthusamy H, Khairunizam W, Jeyaraman S,Nadarajaw T, Yaacob S, et al. A review: survey on automatic infant cry analysis and classification. Health Technol (Berl). (2018) 8:20–9. doi: 10.1007/s12553-018-0243-5

7 Saraswathy J, Hariharan M, Yaacob S, Khairunizam W. Automatic classification of infant cry: a review. In: 2012 International Conference on Biomedical Engineering (ICoBE 2012), Vol. 8. Penang (2012). p. 543–8. doi: 10.1109/ICoBE.2012.6179077

8 LaGasse LL, Neal AR, Lester BM. Assessment of infant cry: acoustic cry analysis and parental perception. Ment Retard Dev Disabil Res Rev. (2005) 11:83–93. doi: 10.1002/mrdd.20050

9 Ntalampiras S. Audio pattern recognition of baby crying sound events. AES J Audio Eng Soc. (2015) 63:358–69.

10 Lavner Y, Cohen R, Ruinskiy D, Ijzerman H. Baby cry detection in domestic environment using deep learning. 2016 IEEE Int Conf Sci Electr Eng ICSEE 2016. (2017) doi: 10.1109/ICSEE.2016.7806117

11 Ferretti D, Severini M, Principi E, Cenci A, Squartini S. Infant cry detection in adverse acoustic environments by using deep neural networks. Eur Signal Process Conf. (2018) 2018-Septe:992–6. doi: 10.23919/EUSIPCO.2018.8553135

12 Severini M, Ferretti D, Principi E, Squartini S. Automatic detection of cry sounds in neonatal intensive care units by using deep learning and acoustic scene simulation.IEEE Frontiers in Pediatrics | www.frontiersin.org 7 April 2021 | Volume 9 | Article 6513567:51982–93. doi: 10.1109/ACCESS.2019.2911427

13 Salehian Matikolaie F, Tadj C. On the use of long-term features in a newborn cry diagnostic system. Biomed Signal Process Control. (2020) 59:101889. doi: 10.1016/j.bspc.2020.101889

14 Choi S, Yun S, Ahn B. Implementation of automated baby monitoring: CCBeBe. Sustain. (2020) 12:2513. doi: 10.3390/su12062513

15 Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J Med Internet Res. (2016) 18:1–10. doi: 10.2196/jmir.5870

16 Eyben F, Schuller B. openSMILE. ACM SIGMultimedia Rec. (2015) 6:4–13. doi: 10.1145/2729095.2729097

17 Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Statistics SS, editor. New York City, NY: Springer Science & Business Media (2013). p. 536.

18 Pranckevi ˇcius T, Marcinkevi ˇcius V. Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. Balt J Mod Comput. (2017) 5:221– 32. doi: 10.22364/bjmc.2017.5.2.05

19 Muchlinski D, Siroky D, He J, Kocher M. Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data. Polit Anal. (2016) 24:87–103. doi: 10.1093/pan/mpv024

20 Czepiel SA. Maximum Likelihood Estimation of Logistic Regression Models: Theory and Implementation. Class Notes (2012). p. 1–23. Available online at: https://papers3://publication/uuid/4E1E1B7E-9CAC-4570-8949-E96B51D9C91DEstimation%of%Logistic%Regre

21 Ji C, Mudiyanselage TB, Gao Y, Pan Y. A review of infant cry analysis and classification [Internet]. Vol. 8, Eurasip Journal on Audio, Speech, and Music Processing. Springer Science and Business Media Deutschland GmbH (2021). p. 1–17.

22 Joshi G, Dandvate C, Tiwari H, Mundhare A. Prediction of probability of crying of a child and system formation for cry detection and financial viability of the system. Proc - 2017 Int Conf Vision, Image Signal Process ICVISP 2017. (2017) 2017-November:134–41. doi: 10.1109/ICVISP.2017.33

23 Felipe GZ, Aguiat RL, Costa YMG, Silla CN, Brahnam S, Nanni L, et al. Identification of infants' cry motivation using spectrograms. Int Conf Syst Signals, Image Process. (2019)2019-June:181–6. doi: 10.1109/IWSSIP.2019.8787318

24 Osmani A, Hamidi M, Chibani A. Machine learning approach for infant cry interpretation. Proc - Int Conf Tools with Artif Intell ICTAI. (2018) 2017-November:182–6. doi: 10.1109/ICTAI.2017.00038

25 Berseth CL, Johnston WH, Stolz SI, Harris CL, Mitmesser SH. Clinical response to 2 commonly used switch formulas occurs within 1 day. Clin Pediatr (Phila). (2009) 48:58–65.

26 Ji C, Xiao X, Basodi S, Pan Y. Deep learning for asphyxiated infant cry classification based on acoustic features and weighted prosodic features. In: Proceedings – 2019 IEEE International Congress on Cybermatics: 12th IEEE International Conference on Internet of Things, 15th IEEE International Conference on Green Computing and Communications, 12th IEEE International Conference on Cyber, Physical and Social Computing and 5th IEEE International Conference on Smart Data, iThings/GreenCom/CPSCom/SmartData 2019. (2019). p. 1233–40.

27 Orlandi S, Reyes Garcia CA, Bandini A, Donzelli G, Manfredi C. Application of pattern recognition techniques to the classification of full-term and preterm infant cry. J Voice. (2016) 30:656–63. doi: 10.1016/j.jvoice.2015.08.007

28 Kruizinga MD, Stuurman FE, Exadaktylos V, Doll RJ, Stephenson DT, Groeneveld GJ, et al. Development of novel, value-based, digital endpoints for clinical trials: a structured approach toward fit-for-purpose validation. Pharmacol Rev. (2020) 72:899–909. doi: 10.1124/pharmrev.120.000028

PART IV

DETECTION OF TREATMENT EFFECTS

CHAPTER 8

Treatment detection and movement disorder society-unified Parkinson's disease rating scale, part III estimation using finger tapping tasks

Ahnjili ZhuParris, MSc,[1,2,3] Eva Thijssen, MSc,[1,2] Willem O. Elzinga, MSc,[1]
Soma Makai-Bölöni, MSc,[1,2] Wessel Kraaij, PhD,[3]
Geert J. Groeneveld, MD, PhD[1,2] and Robert J. Doll, PhD[1]

Movement disorders. 2023. doi:10.1002/mds.29520

1 Centre for Human Drug Research (CHDR), Leiden, NL
2 Leiden University Medical Centre (LUMC), Leiden, NL
3 Leiden Institute of Advanced Computer Science (LIACS), Leiden, NL

Abstract

The validation of objective and easy-to-implement biomarkers that can monitor the effects of fast-acting drugs among Parkinson's disease (PD) patients would benefit antiparkinsonian drug development. We developed composite biomarkers to detect levodopa/carbidopa effects and to estimate PD symptom severity. For this development, we trained machine learning algorithms to select the optimal combination of finger tapping task features to predict treatment effects and disease severity. Data were collected during a placebo-controlled, crossover study with 20 PD patients. The alternate index and middle finger tapping (IMFT), alternative index finger tapping (IFT), and thumb–index finger tapping (TIFT) tasks and the Movement Disorder Society-Unified Parkinson's Disease Rating Scale (MDS-UPDRS) III were performed during treatment. We trained classification algorithms to select features consisting of the MDS-UPDRS III item scores; the individual IMFT, IFT, and TIFT; and all three tapping tasks collectively to classify treatment effects. Furthermore, we trained regression algorithms to estimate the MDS-UPDRS III total score using the tapping task features individually and collectively. The IFT composite biomarker had the best classification performance (83.50% accuracy, 93.95% precision) and outperformed the MDS-UPDRS III composite biomarker (75.75% accuracy, 73.93% precision). It also achieved the best performance when the MDS-UPDRS III total score was estimated (mean absolute error: 7.87, Pearson's correlation: 0.69). We demonstrated that the IFT composite biomarker outperformed the combined tapping tasks and the MDS-UPDRS III composite biomarkers in detecting treatment effects. This provides evidence for adopting the IFT composite biomarker for detecting antiparkinsonian treatment effect in clinical trials.

Introduction

Parkinson's disease (PD) motor impairments can be characterized as slow and rigid and can lead to a gradual reduction in movement speed over time.[1] The recommended instrument for assessing the severity of PD motor symptoms is the Movement Disorder Society's revised version of the Unified Parkinson's Disease Rating Scale, Part III (MDS-UPDRS III).[2] The MDS-UPDRS III offers a reliable and valid metric for evaluating motor manifestations in each body area affected by PD.[3-5] There are two main limitations of the MDS-UPDRS III. First, the MDS-UPDRS III requires approximately 15 minutes to complete with a trained rater, therefore making it time consuming and labor intensive.[6] Thus, MDS-UPDRS III is not ideal for demonstrating the time of onset of fast-acting dopaminergic drugs, such as the inhaled and intranasal forms of levodopa (L-dopa)/carbidopa and apomorphine.[7,8] Second, the MDS-UPDRS III provides only a coarse rating of motor function and therefore cannot identify or differentiate between specific kinematics of finger movements.[3] As fine motor control abnormalities are typically the first manifestations of motor impairments in PD patients, it is important to develop composite biomarkers that are sensitive to these changes.[9] To address these limitations, there is a demand for biomarkers that detect fine-grained changes in motor function and are congruent with the MDS-UPDRS.

Finger tapping tasks provide insights into fine motor activity [10,11] and have been shown to be quick, effective, and simple assessments for estimating MDS-UPDRS motor disability[12,13] and assessing antiparkinsonian drug effects.[14-19] These tasks provide insights into fin- ger and forearm movement speed, accuracy, amplitude, frequency, rhythm, and fatigue.[10,14,20,21] PD patients often experience tremors, stiffness, and difficulty with movement, which can significantly impact their ability to perform daily activities, including buttoning a shirt, typing on a keyboard, or using utensils.[22,23] As patients want treatments that will improve their ability to carry out daily activities, measuring motor function through tapping biomarkers can provide a more direct and meaningful assessment of

the impact of treatments on patients' lives. Therefore, the tapping tasks could be considered of interest to both clinicians and patients.

The complexity of parkinsonism motor impairment manifestations cannot be captured by a single biomarker. By exploiting machine learning algorithms, we can combine multiple objective biomarkers into a single composite biomarker that would represent a multi-dimensional characterization of PD.[24] Previous studies have demonstrated that composite biomarkers could effectively differentiate between PD and healthy controls and estimate MDS-UPDRS III symptom severity.[25-27] This study investigates the accuracy and sensitivity of composite tapping biomarkers to detect drug effects and to estimate disease severity among PD patients.

Patients and Methods

This is an extension of a previous study that investigated the reliability of tapping tasks to detect the longitudinal effects of L-dopa/carbidopa and to determine the correlation of the tapping features with the MDS-UPDRS III.[14] The study was conducted at the Centre for Human Drug Research (CHDR) in Leiden, the Netherlands, between July and November 2020 and is registered in the Netherlands Trial Register (trial NL8617).

STUDY OVERVIEW

We conducted a double-blind, placebo-controlled, randomized, two-way crossover study with L-dopa/carbidopa in 20 PD patients that had recognizable off episodes (symptoms not adequately controlled by their medication).[28] Patients received a semi-individual dose of the investigational drug. To ensure an off-on transition, the patients were given a supramaximal dose that was at least 25% higher than their usually administered morning dose.[29]

PATIENT CRITERIA

Enrolled patients had a clinical diagnosis of PD, as confirmed by a neurologist, and a classification of a Hoehn–Yahr stages I to III during their on state by an investigator. Patients were included if they were between ages 20 and 85 years during screening, experienced self-described motor fluctuations, and were taking oral antiparkinsonian medication. Patients were excluded if they had known conditions that would affect L-dopa/carbidopa treatment or study compliance, such as previous intolerance, drug dependence, or psychiatric disease.

ASSESSMENTS

MDS-UPDRS III We selected the MDS-UPDRS III as the gold standard for the purposes of this study. The MDS-UPDRS III was conducted by trained raters at CHDR. The examination took on average 15 minutes to complete. It was performed pre-dose and at 10, 30, 60, and 90 minutes after dosing.

FINGER TAPPING TASKS All the tapping tasks were performed twice pre-dose and once at 10, 25, 45, 60, 75, 90, and 105 minutes after dosing. If the tapping tasks and MDS-UPDRS III were planned simultaneously, then tapping tasks were performed first.

ALTERNATE INDEX AND MIDDLE FINGER TAPPING AND ALTERNATE INDEX FINGER TAPPING Each patient was provided with a touchscreen laptop equipped with the alternate index and middle finger tapping (IMFT) and alternate index finger tapping (IFT) tasks.[10] The patients were instructed to use the hand that was most affected (if both hands were equally affected, to use their dominant hand) and to perform each task as fast and accurately as possible for 30 seconds. For the IMFT, patients were asked to tap between the two targets (2.5 cm apart) with their index and middle fingers. For the IFT, patients were asked to tap the targets (20 cm apart) with their index finger. The IMFT and IFT require two different movements; the IMFT and IFT are dependent on fine finger and forearm movements, respectively.[10] Each of the two tasks generated 43 features relating to speed (eg, total number of taps), accuracy (eg, spatial error), rhythm (eg, intertap interval), and fatigue (eg, change in velocity) (Table S1).[10,14]

THUMB–INDEX FINGER TAPPING A wireless goniometer (Biometrics Ltd, Newport, UK) was placed on the metacarpal and proximal phalanx of the index finger of the most affected hand (if both hands were equally affected, to use their dominant hand).[10,14,30] Each patient was instructed to sit comfortably, hold up the hand, and tap the index finger on the thumb as widely and quickly as possible continuously for 15 seconds. The thumb–index finger tapping (TIFT) assesses unilateral sequential fine finger movements. The 25 features of the TIFT include progressive changes in amplitude, hesitations, and tapping speed during the task (Table S1).[14]

STATISTICAL ANALYSIS

All data preprocessing and statistical analyses were conducted using Python (version 3.8.0) (31) and the Scikit-Learn library (version 1.0.1).[32]

DATA PREPROCESSING All features were visually and statistically inspected for normality using histograms and Shapiro–Wilk tests, respectively. Log or square root transformations were applied when the features were not normally distributed. Only features that were normally distributed were included in the analysis. Missing values were not imputed, and only complete cases were considered.

As the tapping composite biomarker is designed to be a proxy for overall motor function, we did not account for laterality of the tapping task in the biomarkers. The need for assessing the tapping tasks with both hands is therefore avoided, which could streamline the assessment process and reduce the burden on patients.

COMPOSITE BIOMARKERS We developed 10 composite biomarkers. The composite biomarkers represented the baseline-uncorrected or baseline-corrected MDS-UPDRS III 18-item scores; all three tapping tasks combined; and the IFT, IMFT, and TIFT tasks individually. From a statistical viewpoint, we corrected for baseline to remove any concomitant variability in the treatment response, which would therefore improve the precision of the treatment detection.[33] From a practical viewpoint, we considered using the baseline-uncorrected values to reduce the number of measurements needed for treatment classification. The baseline-uncorrected model would require only a single tapping assessment, whereas the baseline-corrected model would require two.

CROSS-VALIDATION We applied a nested k-fold cross-validation strategy to assess the performance and the generalizability of the composite biomarkers.[34] In nested cross-validation, the outer fold assesses the performance of the model, whereas the inner fold performs the model and hyper-parameter selection. In our study, the outer-fold step was repeated 100 times, with each iteration containing a different combination of training (80% of the data) and test sets (20%). Each outer training set was further split into an inner training (80% of the data) and validation sets (20%). The inner-fold step was repeated 50 times, and the best-performing inner model would be evaluated in the outer fold. The final results would be represented as the averaged and standard deviation of the models selected by each outer fold.[34] For the classification and regression models, we applied a group-shuffle split (same distribution of placebo and active treatments in each split) and a stratified-shuffle split (same distribution of MDS-UPDRS III scores in each split), respectively. To stratify the MDS-UPDRS III scores, we assigned each score to one of three binned ranges (eg, the baseline-corrected MDS-UPDRS III binned ranges were [-13, -8.76], [-8.76, -4.53], and [-4.53, 0.3]). Each outer fold had the same distribution of binned ranges. Stratification was not applied to the inner fold, as the small sample size would limit the number of samples available per bin. Within each inner fold, all features were standardized by subtracting the mean and scaling to the unit variance. To identify the features that were predictive of the outcomes, we identified features that were selected at least once by all outer-fold models.[34]

CLASSIFICATION OF ACTIVE OR PLACEBO TREATMENTS Classification models were trained to classify the active or placebo treatments. As we intended to predict the probability of treatment at all time points, we

chose the last measurements to train the models. The MDS-UPDRS III classification model was trained on the 90-minute MDS-UPDRS III item scores.[14] The tapping classification models were trained on measurements taken immediately after the MDS-UPDRS III starting at 105 minutes.
To identify the optimal classification model, we compared three classification models: support vector machines, logistic regression, and linear discriminant analysis (LDA). These classification models were selected as they are easy to implement and to interpret.[35-37] Previous studies have also used these algorithms to classify PD diagnosis or estimate MDS-UPDRS III.[38-41] Models were compared based on their mean accuracy, precision, and F1 scores.[40]

In addition, each model selected by the outer folds was used to predict the treatment at the other time points, with 20% of patients who were not used for training. This would allow researchers to identify at which time point treatment effects are detected. For each time point, the mean and standard deviation of the class probabilities were based on the predicted log-odd ratios from each fold. Additionally, these probabilities were used to estimate the repeatability and effect size. The repeatability was assessed by calculating the intraclass correlation coefficients (ICC) using the placebo results only. Using a random intercept model with the intercept and time point as fixed effects, the ICC was calculated by dividing the between-subject variance by the sum of the between-subject and within-subject variances. The effect size was calculated using all available data and a random intercept model with intercept, time point, treatment, and interaction between time point and treatment as fixed effects. In addition, the effect size was calculated as the contrast between the probabilities after treatment and the averaged baseline probabilities divided by the square root of the sum of the between-subject and within-subject variations.

ESTIMATION OF THE MDS-UPDRS III TOTAL SCORE To assess if the tapping composite biomarkers (baseline uncorrected and baseline corrected) could estimate the MDS-UPDRS III total score, linear regression with elastic-net regularization (optimized for α and the l1 ratio) was used to predict the MDS-UPDRS III total score at 90 minutes using the 105-minute tapping biomarkers. These two time points were compared, as it was previously shown that the IFT and TIFT showed significant and moderate-to-strong correlations with the MDS-UPDRS III.[14] Further, the 90- and 105-minute tapping tasks were equally as close to the 90-minute MDS-UPDRS III in timing and therefore we assumed would perform equally well.

To assess the performance of the models, we estimated the mean absolute error (MAE) of the outer-fold models. We evaluated the correlation between the predicted and true MDS-UPDRS III scores at all timepoints for each outer-fold model. Like the classification models, the MDS-UPDRS III scores were estimated at other time points with the 20% patients who were not used for training. Additionally, as for the classification models, those data were also used to estimate the repeatability and effect size.

Results

DATA COLLECTED

Twenty PD patients participated in this study. An overview of the demographic and disease characteristics of the patients was published previously ;[14] 14 patients were male, and their ages ranged from 48 to 70 years. Patients received one to four capsules of 100/25 mg L-dopa/carbidopa as they had a supramaximal morning levodopa equivalent dose (LED) ranging from 47 to 391 milligrams. The median MDS-UPDRS III score when using regular medication was 23 and 22 on their placebo and active treatment days, respectively.[14]

We analyzed 31 IMFT, 31 IFT, and 25 TIFT features. No features were excluded due to nonnormal distribution. Due to goniometer damage, we had missing data for 1 patient in the placebo condition and 2 patients in the active condition. As 6 patients had difficulties performing the IMFT, this led to missing data. However, the missing data were equally distributed across the treatment conditions and therefore deemed missing at random.

CLASSIFICATION OF PLACEBO AND ACTIVE TREATMENTS

We found that the LDA classifier consistently yielded the highest accuracy for all models (for both baseline uncorrected and baseline corrected); thus, we reported only the LDA results.

CLASSIFICATION OF TREATMENT EFFECTS The best-performing baseline-uncorrected composite biomarker, the IFT, yielded an accuracy, precision, F1 score, and large effect size of 68.50%, 70.23%, 68.93%, and 1.60 respectively (Table 1). The best-performing baseline-corrected composite biomarker, the IFT, achieved a higher average accuracy, precision, F1 score, and large effect size of 83.50%, 93.95%, 80.09%, and 2.58. Both models outperformed the MDS-UPDRS III classification models across all metrics. The IFT features that were mutually identified as important features for the baseline-uncorrected and baseline-corrected classification models were related to accuracy (e.g., spatial errors and the bivariate contour ellipse area), fatigue (e.g., velocity changes), and velocity (e.g., intertap intervals) (Figure 1).

CLASSIFICATION OF TREATMENT EFFECTS AT ALL TIME POINTS In Figure 2, the classification models were applied to all time points, showing the mean predicted probability of an active (>0.5) or placebo treatment (<0.5). In the baseline-corrected IFT, TIFT, and MDS-UPDRS III models, the mean predicted probability of a patient receiving a placebo treatment was consistently less than 0.5. In contrast, when active treatment was administered, the baseline-corrected IFT and MDS-UPDRS III model had a mean predicted probability above 0.5 from 60 minutes onward. The baseline-corrected IMFT and TIFT models crossed the 0.5 thresholds after 45 minutes. We found that the baseline-corrected IFT biomarker determined a large effect size (0.81) at 30 minutes, whereas the baseline- uncorrected IFT biomarker reached a large effect size of 0.84 at 60 minutes. The MDS-UPDRS III achieved a large effect size at 60 minutes (1.69 and 1.04 for baseline corrected and baseline uncorrected, respectively) (Figure S2). The MDS-UPDRS III demonstrated higher repeatability than the tapping tasks. Whereas the baseline-uncorrected MDS-UPDRS III biomarker obtained an excellent ICC, the IFT and TIFT both achieved good ICCs (0.78, 0.80) (42). However, the ICCs of the baseline-corrected MDS-UPDRS III and the IFT, IMFT, and TIFT biomarkers decreased to a moderate ICC range between 0.52 and 0.66.[42]

ESTIMATION OF MDS-UPDRS III

The mean MDS-UPDRS III total scores at 90 minutes for the placebo and active treatments were 33.5 and 22.0, respectively. When baseline-corrected, the mean MDS-UPDRS III scores for the placebo and active treatments were 0.3 and -13.0, respectively (Figure 3).

The best-performing baseline-uncorrected regression models were the TIFT and IFT composite biomarkers, which achieved the lowest average MAE of 10.31 and 10.36, respectively. In addition, the TIFT and IFT showed large effect sizes of 1.47 and 2.23, respectively, when estimating the MDS-UPDRS III. The best-performing baseline-corrected model was the IFT composite biomarker, which yielded the lowest average MAE of 7.87. For both the baseline-uncorrected and baseline-corrected models, the best-performing composite biomarkers outperformed that of the composite biomarkers of the three tasks. For the IFT features, the features that were mutually selected by both models were similar to that of the IFT classification features (Figure 2; Figure S1).

ESTIMATION OF MDS-UPDRS III AT ALL TIME POINTS The predicted and true MDS-UPDRS III scores were significantly correlated for the baseline-corrected and baseline-uncorrected models (Table 2). Once again, the best positive correlations were achieved by the TIFT baseline-uncorrected composite biomarker ($r = 0.58$, $P < 0.01$) and the IFT baseline-corrected composite biomarker ($r = 0.69$, $P < 0.01$). The greatest difference in the true MDS-UPDRS III scores between the placebo and active treatment interventions was at 90 minutes (Fig. 3). The tapping tasks achieved a moderate to good ICC (Table 2).

Discussion

DETECTION OF TREATMENT EFFECTS

The IFT biomarker (baseline corrected and baseline uncorrected) was, on average, more predictive of and more sensitive to treatment effects than the MDS- UPDRS III biomarker in terms of accuracy, precision, and clinical significance (as supported by the effect-size performances) (Table 1). This is significant as the ability to detect changes in aspects of motor function that may be missed by traditional assessments allows for a more sensitive measure of treatment efficacy. This can be valuable for detecting small and early changes in motor function that are indicative of a treatment response. The most important IFT features used to classify treatment effects are in concert with previous studies (Figure 1) that also identified that forearm movements relating to velocity, amplitude, and rhythm are sensitive to anti- parkinsonian drug effects.[10,15,43,44] We demonstrated that treatment effects were detected at 45 and 60 minutes for the TIFT and IFT composite biomarkers, respectively (Figure 2). This finding is notable as the mean onset of L-dopa/carbidopa action is about 50 minutes (45). This suggests that tapping tasks can detect the onset of oral L-dopa/carbidopa. The MDS-UPDRS III was not performed at 45 minutes, so it could not be determined whether the MDS-UPDRS III biomarker could detect treatment effects at 45 minutes. These findings further propound that the tapping tasks are practical and sensitive composite biomarkers for detecting motor response changes induced by anti- parkinsonian drugs (46). Further, the large effect sizes can potentially reduce sample size requirements and enhance power for future tapping task trials that assess treatment effects.

The performance of the classification models (except for the ICC) improved when the features were baseline corrected. Despite this, both models provide practical and clinical value. The baseline-uncorrected models required only a single measurement and represented the current motor function status. The baseline-corrected models require two measurements and represent the changes in motor function over time. The increased performance suggests that treatment response is dependent on the patient's tapping profile during their off state and adjusting for baseline removes variation in the L-dopa/carbidopa response.

ESTIMATION OF MDS-UPDRS III

We found that the baseline-corrected IFT biomarker, despite yielding the best performance among all the biomarkers, achieved a prediction error of approximately eight points and was significantly moderately correlated using the MDS-UPDRS III. The prediction error is comparable to existing sensor-based composite biomarkers used to estimate the MDS-UPDRS III. Studies using data sourced from an Axitvity AX3 (placed on the wrist and back or only the wrist) to estimate the gold standard achieved an MAE ranging from 4.29 to 6.29 points.[47,48] The tapping biomarkers predicted a smaller range of MDS-UPDRS III scores compared to that of the true MDS-UPDRS III scores (Figure 3). It is likely due to using only hand and forearm motor function assessments to predict the MDS-UPDRS III total scores, which includes motor assessments of other regions affected by PD, such as gait, facial expression, and speech.[4] As the correlations of the true and predicted MDS-UPDRS III scores were moderate (Table 2), the tapping biomarkers still showed concurrent validity with the gold standard. This suggests that the tapping biomarkers could provide clinicians with an understanding of the acute effects of drugs on motor fluctuations within a short monitoring period.

Despite the discrepancies between the true and predicted MDS-UPDRS III total scores, with the advancements in technology, it is not unusual for the performance of new clinical assessments to outperform the current gold standard. However, the discrepancy between the two assessments influences the accuracy estimates of the new clinical assessments, and as it would be interpreted as a prediction error.[49] Therefore, we argue that accurate estimation of the MDS-UPDRS III score is not essential for the adoption of the composite biomarker as a new complementary assessment for estimating symptom severity. Rather, the consequences resulting from the disagreement between the gold standard and the tapping composite biomarkers should be investigated.

FUTURE WORK

We demonstrated that the tapping composite biomarkers could detect the onset of oral L-dopa/carbidopa at 45 minutes. A follow-up study could investigate if the tapping composite biomarkers could detect an earlier onset of an even faster-acting antiparkinsonian drug, such as inhaled apomorphine that has an onset as early as 8 minutes.[8] This would further validate the sensitivity of the tapping composite biomarker to detect fast-acting dopaminergic drug effects.

Our sample size may limit the generalizability of this study's findings as a small sample size may not be representative of the broader population of patients with PD, making it difficult to generalize its results to a larger population.[50] This is particularly relevant for PD studies, where the disease can manifest in different ways and progress at different rates in different patients. To mitigate the effect of the small sample sizes, we employed cross-validation to bootstrap and validate the models against different groups of patients. We propose conducting a follow-up trial to implement the tapping tasks among more PD patients with more diverse MDS-UPDRS III profiles. The data collected from the trial can be used as an independent data set to assess the validity, reliability, and generalizability of our current methods. Although composite biomarkers have the advantage of capturing multiple aspects of motor function, the effects of individual components within the composite biomarker must be carefully examined to avoid misleading interpretations of the results. For example, a treatment that improves tapping speed but worsens tapping rhythm may result in an overall neutral effect, making it difficult to interpret the treatment's efficacy. Like other composite measures, such as the MDS-UPDRS III total score, it is crucial to examine the effects of each feature of the composite biomarker separately, as well as in conjunction with the overall composite score, to better understand the treatment's impact on finger motor function.

Conclusion

In conclusion, the IFT biomarker was more predictive of and sensitive to the detection of treatment effects than the MDS-UPDRS III biomarker; therefore, the tapping biomarkers appear to hold promise for evaluating the early and rapid effects of antiparkinsonian drugs. Moreover, the tapping task is easy to perform and can be done in clinical settings as well as at home by patients themselves, making it a practical and convenient method for monitoring disease progression and treatment response. Using tapping biomarkers, clinicians can obtain accurate and reliable data that can inform treatment decisions in real time.

TABLE 1 The mean and standard deviations of the accuracy, precision, F1 score, and effect size for each biomarker (at 90 minutes for MDS-UPDRS III and 105 minutes for the tapping task) are based on the 100 outer folds of the nested cross-validation

	Tasks	Accuracy	Precision	F1-score	ICC	Effect-size
BASELINE-UNCORRECTED	IMFT	56.90% (±15.09%)	61.67% (±22.53%)	56.56% (±18.07%)	0.60 (± 0.25)	0.64 (± 0.57)
	IFT	**68.50%** (±12.56%)	**70.23%** (±16.31%)	**68.93%** (±14.9%)	0.78 (± 0.21)	**1.60** (± 0.82)
	TIFT	67.72% (±15.84%)	65.55% (±21.03%)	67.51% (±18.22%)	0.78 (± 0.22)	1.14 (± 0.80)
	All 3 Tasks	63.0% (±16.91%)	64.35% (±27.32%)	59.82% (±23.16%)	0.68 (± 0.29)	0.91 (± 0.68)
	MDS-UPDRS III item scores	63.75% (±11.25%)	61.20% (±10.9%)	68.90% (±11.52%)	**0.92** (± 0.10)	1.03 (± 0.60)
BASELINE-CORRECTED	IMFT	66.86% (±15.23%)	70.83% (±17.25%)	69.01% (±15.04%)	0.57 (± 0.17)	1.44 (± 0.98)
	IFT	**83.50%** (±10.74%)	**93.95%** (±11.25%)	**80.09%** (±14.92%)	0.53 (± 0.16)	**2.58** (± 0.90)
	TIFT	77.86% (±14.97%)	82.32% (±21.43%)	74.72% (±18.44%)	0.52 (± 0.17)	1.14 (± 0.80)
	All 3 Tasks	77.98% (±13.26%)	81.85% (±21.15%)	74.66% (±19.17%)	0.48 (± 0.18)	0.91 (± 0.61)
	MDS-UPDRS III item scores	75.75% (±14.45%)	79.95% (±17.64%)	73.93% (±16.42%)	**0.66** (± 0.11)	2.12 (± 1.25)

TABLE 2 Average correlation and ICC (95% CI) between the true and predicted MDS-UPDRS scores across all time points for the repeated nested cross-validation 100 outer-fold predictions.

	Tasks	Correlation coefficient (r)	p-value	ICC	Effect-size
BASELINE-UNCORRECTED	IMFT	0.10 [0.03, 0.16]	p<.05 [<.05, 0.05]	0.69 [0.65, 0.73]	0.67 [0.53, 0.81]
	IFT	0.52 [0.45, 0.59]	p<.01 [<.01, <.01]	0.80 [0.76, 0.83]	1.02 [0.91, 1.14]
	TIFT	**0.58 [0.53, 0.63]**	p<.05 [<.01, <.05]	0.78 [0.74, 0.82]	**1.47** [1.27, 1.67]
	All 3 Tasks	0.11 [0.04, 0.18]	p<.05 [<.05, 0.05]	0.66 [0.61, 0.71]	0.75 [0.62, 0.88]
BASELINE-CORRECTED	IMFT	0.34 [0.27, 0.40]	p<.05 [<.01, 0.06]	0.48 [0.44, 0.52]	1.10 [0.92, 1.28]
	IFT	**0.69 [0.65, 0.73]**	p<.001[<.001,<.005]	0.45 [0.42, 0.48]	**2.23** [2.01, 2.45]
	TIFT	0.65 [0.60, 0.69]	p<.001 [<.001, <.001]	0.50 [0.46, 0.54]	1.37 [1.20, 1.54]
	All 3 Tasks	0.56 [0.52, 0.61]	p<.05 [<.001, <.05]	0.43 [0.39, 0.47]	1.06 [0.91, 1.21]

SUPPLEMENTARY TABLE 1 Overview of features for the Alternate Index and Middle Finger Tapping (IMFT), Alternate Index Finger Tapping (IFT), Thumb-Index Finger Tapping (TIFT)(8)

Task	Endpoint (UNIT)	Acronyms
TIFT	Amplitude: Slope from linear regression of each tap's amplitude against time. (degrees and degrees/seconds)	Mean (TAM) Change (TAC)
TIFT	Angle frequency change: Change in peak tapping frequency over time (Hz/min) Angle change (degrees^2/s)	Frequency Mean (AFM) Frequency Change (AFC) Angle Mean (AAM), Angle Change (AAC)
IMFT, IFT	Bivariate contour ellipse angle (degree) Bivariate contour ellipse area (mm^2) BCEA represents the area of an ellipse which encompasses the fixation points	BCEA angle (BCT) BCEA area (BCA)
IMFT, IFT	Distance travelled between consecutive taps (centimetres)	Total (DTT) Average (DTA) Standard Deviation (DTS) Covariance (DTV) Change between first/last (DTD) Change between intervals (DTC)
IMFT, IFT, TIFT	Inter-Tap Interval: Time between two consecutive taps (milliseconds)	Average (ITA) Standard Deviation (ITS) Covariance (ITV) Change between (ITC) Change between first/last (ITD)
IMFT, IFT	Missed Taps: Total number of double/missed taps (DBLTT) Ratio good taps: total taps (DBLTR) (count)	Total number of double/missed taps (DBLTT) Ratio good taps: total taps (DBLTR)
IMFT, IFT	Number of Halts: Number of taps where the inter-tap interval is larger than 2 * ITM (count)	NOH
TIFT	Peak frequency area under the curve: The total power around the peak frequency in the power spectrum around the peak frequency (degrees^2)	Amplitude (FPA) Frequency (FPF) Area under the curve (FPP)
IMFT, IFT	Ratio good taps:total taps: Taps on the correct side (left/right) of the screen	TNT
IMFT, IFT	Spatial error: Sum of the Euclidean distances between each tap and the center of the target (millimeters)	Total (SET) Average (SEA) Standard Deviation (SES) Covariance (SEV) Change between (SED) Change between first/last (SEC)
IMFT, IFT, TIFT	Total number of taps	TNT
IMFT, IFT	Total taps inside and outside target	Taps within the target circle (TIT) Taps outside the target circle (TOT)

[continuation of Supplementary Table 1]

Task	Endpoint (UNIT)	Acronyms
IMFT, IFT	Mean of each finger tap's velocity (centimetres/minute)	Average (VEA) Standard Deviation (VES) Covariance (VEV) Change between first/last (VED) Change between intervals (VEC)
TIFT	Mean of each finger tap's velocity (degrees/second)^2	Mean (TVM) Change (TVC)
TIFT	Velocity Amplitude (degrees/second)^2	Velocity Amplitude Mean (VAM) Change (VAC)
TIFT	Velocity Closing: Average of the amplitude (i.e. angle) travelled per second for each tap when moving the index finger towards the thumb (closing); velocity extracted from the derivative of the amplitude (degrees/second)	Mean (CVM) Change (CVC)
TIFT	Velocity Frequency (Hz)	Mean (VFM) Change (VFC)
TIFT	Velocity Opening: Average of the amplitude (i.e. angle) travelled per second for each tap when moving the index finger away from the thumb (opening); velocity extracted from the derivative of the amplitude (degrees/s)	Mean (OVM) Change (OVC)

FIGURE 1 The average feature coefficients of the respective features selected by the LDA (linear discriminant analysis) classifier for each finger tapping task feature and the MDS-UPDRS III (Movement Disorder Society-Unified Parkinson's Disease Rating Scale, Part III) item score features (baseline-uncorrected and baseline-corrected models). The error bars represent the 95% confidence interval.

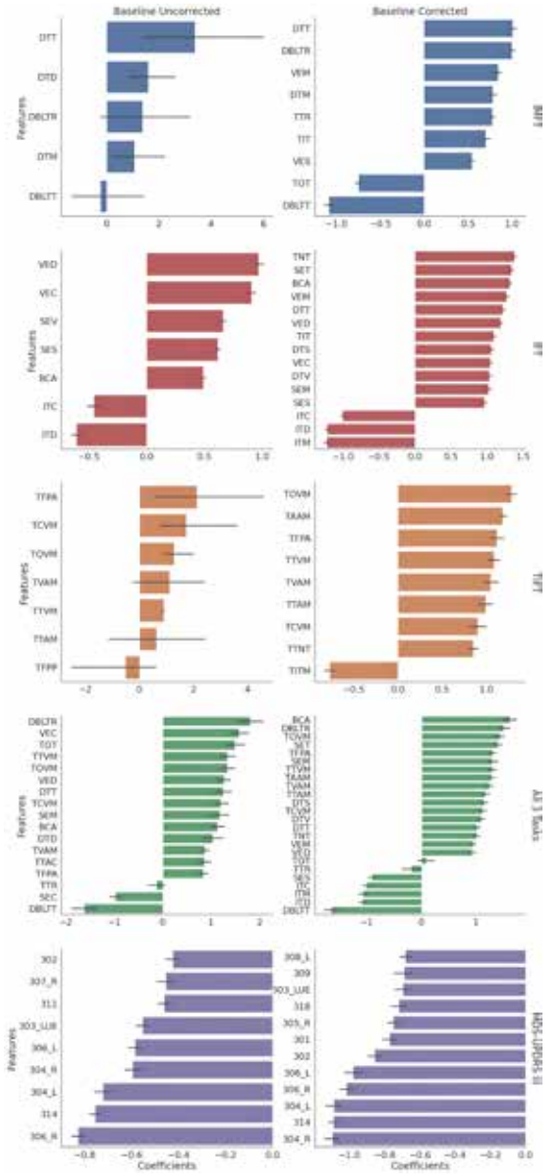


FIGURE 2 The mean predicted probability that active treatment was administered in the placebo (blue) and active (orange) treatment groups. The green dotted line represents the 0.5 decision boundary. The bands represent the 95% confidence interval.

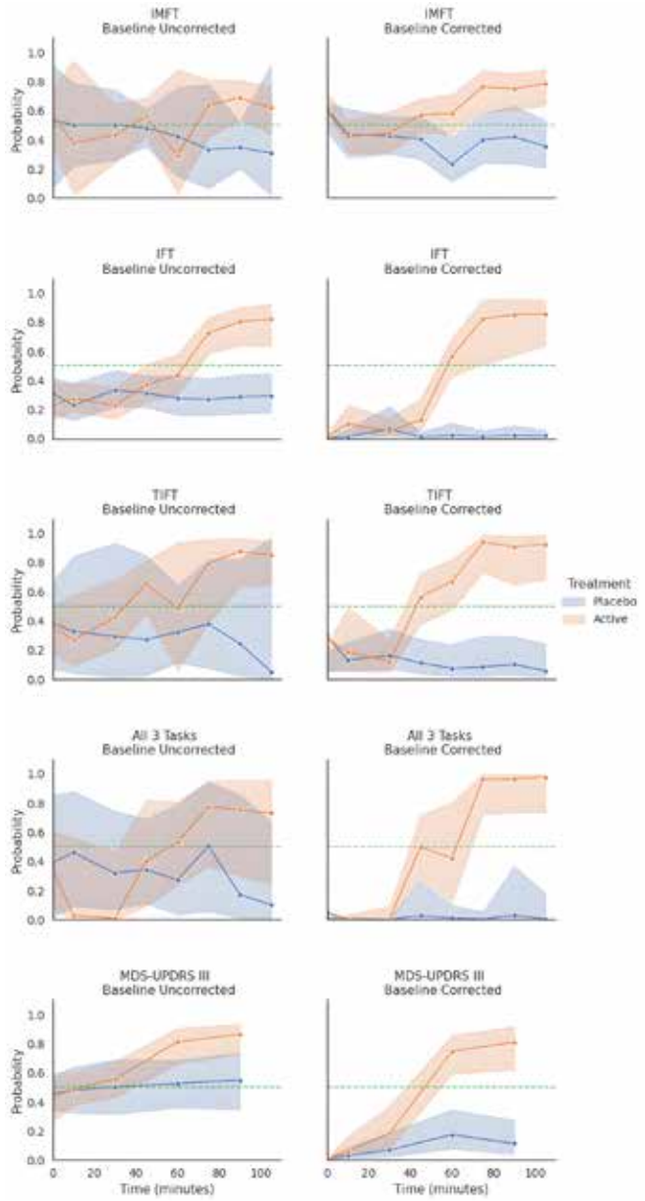
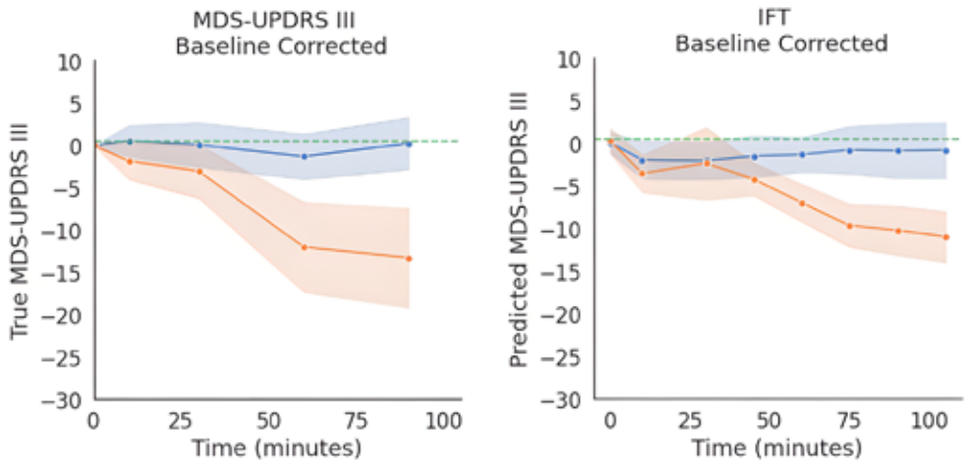
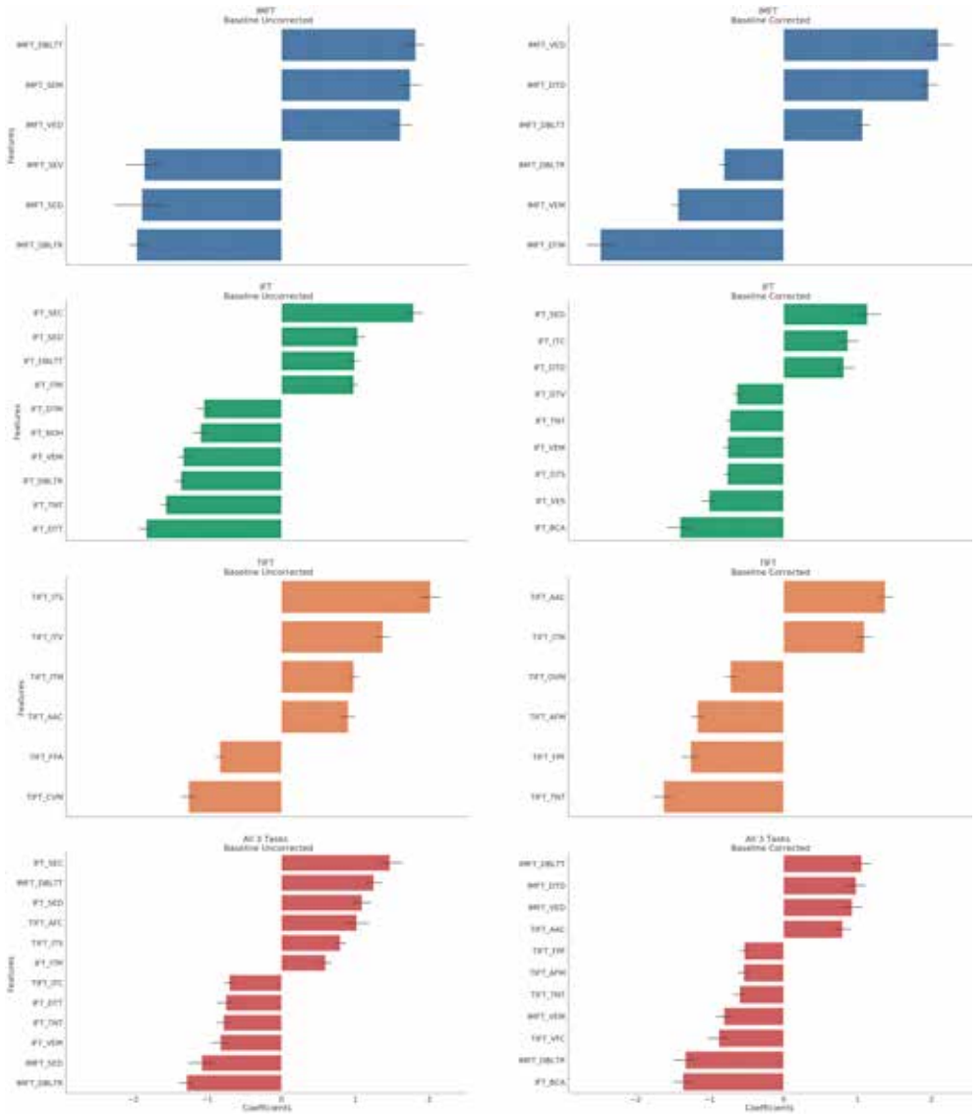


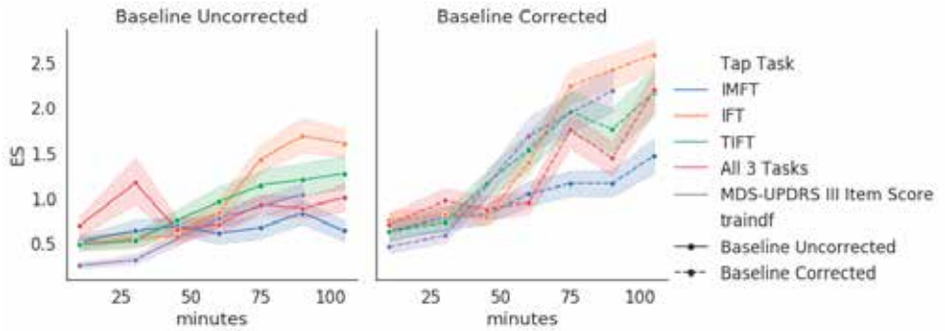
FIGURE 3 Average true and predicted MDS-UPDRS III (Movement Disorder Society-Unified Parkinson's Disease Rating Scale, Part III) scores with standard deviation from 0 to 105 minutes post dose for the placebo (blue) and active (orange) treatment interventions when baseline corrected.



SUPPLEMENTARY FIGURE 1 The average feature coefficients selected by the elastic-net linear regression models for each of the composite biomarkers under baseline-uncorrected and baseline-corrected conditions. The errors represent the 95% confidence intervals.



SUPPLEMENTARY FIGURE 2 Effect sizes of each of the tapping tasks and the Movement Disorder Society-Unified Parkinson's Disease Rating Scale, Part III, composite biomarkers at each time point.



REFERENCES

1 Davie CA. A review of Parkinson's disease. Br Med Bull 2008;86(1): 109–127. https://doi.org/10.1093/bmb/ldn013

2 Jankovic J. Parkinson's disease: clinical features and diagnosis. J Neurol Neurosurg Psychiatry 2008;79(4):368–376. https://doi.org/10.1136/jnnp.2007.131045

3 Regnault A, Boroojerdi B, Meunier J, Bani M, Morel T, Cano S. Does the MDS-UPDRS provide the precision to assess progression in early Parkinson's disease? Learnings from the Parkinson's progression marker initiative cohort. J Neurol 2019;266:1927–1936. https://doi.org/10.1007/s00415-019-09348-3

4 Goetz CG et al. Movement Disorder Society-sponsored revision of the unified Parkinson's disease rating scale (MDS-UPDRS): scale presentation and clinimetric testing results. Mov Disord 2008;23(15): 2129–2170. https://doi.org/10.1002/mds.22340

5 Martinez-Martin P, Rodriguez-Blazquez C, Alvarez-Sanchez M, et al. Expanded and independent validation of the Movement Disorder Society–unified Parkinson's disease rating scale (MDS-UPDRS). J Neurol 2013;260(1):228–236. https://doi.org/10.1007/s00415-012-6624-1

6 Ramsay N, Macleod AD, Alves G, et al. Validation of a UPDRS-/MDS-UPDRS-based definition of functional dependency for Parkinson's disease. Parkinsonism Relat Disord 2020;76:49–53. https://doi.org/10.1016/j.parkreldis.2020.05.034

7 Patel AB, Jimenez-Shahed J. Profile of inhaled levodopa and its potential in the treatment of Parkinson's disease: evidence to date. Neuropsychiatric Disease and Treatment 2018;14:2955–2964. https://doi.org/10.2147/NDT.S147633

8 Grosset KA, Malek N, Morgan F, Grosset DG. Inhaled apomorphine in patients with 'on-off' fluctuations: a randomized, double-blind, placebo-controlled, clinic and home based, parallel-group study. J Parkinsons Dis 2013;3(1):31–37. https://doi.org/10.3233/JPD-120142

9 Koop MM, Shivitz N, Brontë-Stewart H. Quantitative measures of fine motor, limb, and postural bradykinesia in very early stage, untreated Parkinson's disease. Mov Disord 2008;23(9):1262–1268. https://doi.org/10.1002/mds.22077

10 Makai-Bölöni S, Thijssen E, van Brummelen EMJJ, Groeneveld GJ, Doll RJ. Touchscreen-based finger tapping: repeatability and configuration effects on tapping performance. PLoS One 2021;16(12): e0260783. https://doi.org/10.1371/journal.pone.0260783

11 Nalçaci E, Kalaycioggl C, Çiçek M, Genç Y. The relationship between handedness and fine motor performance. Cortex 2001; 37(4):493–500. https://doi.org/10.1016/S0010-9452(08)70589-6

12 Taylor Tavares AL, Jefferis GSXE, Koop M, Hill BC, Hastie T, Heit G, Bronte-Stewart HM. Quantitative measurements of alternating finger tapping in Parkinson's disease correlate with UPDRS motor disability and reveal the improvement in fine motor control from medication and deep brain stimulation. Mov Disord 2005; 20(10):1286–1298. https://doi.org/10.1002/mds.20556

13 Fukawa K, Okuno R, Yokoe M, Sakoda S, Akazawa K. Estimation of UPDRS finger tapping score by using artificial neural network for quantitative diagnosis of Parkinson's disease. Proceedings of the IEEE/EMBS Region 8 International Conference on Information Technology Applications in Biomedicine, ITAB; IEEE, New York City; 2007:259–260. https://doi.org/10.1109/ITAB.2007.4407396.

14 Thijssen E, Makai-Bölöni S, van Brummelen E, den Heijer J, Yavuz Y, Doll RJ, Groeneveld GJ. A placebo-controlled study to assess the sensitivity of finger tapping to medication effects in PD. Mov Disord Clin Pract 2022;9:1074–1084. https://doi.org/10.1002/mdc3.13563

15 Espay AJ, Giuffrida JP, Chen R, et al. Differential response of speed, amplitude, and rhythm to dopaminergic medications in Parkinson's disease. Mov Disord 2011;26(14):2504–2508. https://doi.org/10.1002/mds.23893

16 Hasan H, Burrows M, Athauda DS, et al. The Bradykinesia Akinesia Incoordination (BRAIN) tap test: capturing the sequence effect. Mov Disord Clin Pract 2019;6(6):462–469. https://doi.org/10.1002/mdc3.12798

17 Wissel BD, Mitsi G, Dwivedi AK, et al. Tablet-based application for objective measurement of motor fluctuations in Parkinson disease. Digit Biomark 2018;1(2):126–135. https://doi.org/10.1159/000485468

18 Lipp MM, Batycky R, Moore J, Leinonen M, Freed MI. Preclinical and clinical assessment of inhaled levodopa for OFF episodes in Parkinson's disease. Sci Transl Med 2016;8(360):360ra136–360ra136. https://doi.org/10.1126/scitranslmed.aad8858

19 Arora S et al. Smartphone motor testing to distinguish idiopathic REM sleep behavior disorder, controls, and PD. Neurology 2018;91(16): E1528–E1538. https://doi.org/10.1212/WNL.0000000000006366

20 Kimber TE, Tsai CS, Semmler J, Brophy BP, Thompson PD. Voluntary movement after pallidotomy in severe Parkinson's disease. Brain 1999;122(5):895–906. https://doi.org/10.1093/brain/122.5.895

21 Yokoe M, Okuno R, Hamasaki T, Kurachi Y, Akazawa K, Sakoda S. Opening velocity, a novel parameter, for finger tapping test in patients with Parkinson's disease. Parkinsonism Relat Disord 2009;15(6):440–444. https://doi.org/10.1016/j.parkreldis.2008.11.003

22 Espay AJ, Hausdorff JM, Sanchez-Ferro ´A, et al. A roadmap for implementation of patient-centered digital outcome measures in Parkinson's disease obtained using mobile health technologies. Mov Disord 2019;34(5):657–663. https://doi.org/10.1002/mds.27671

23 Nisenzon AN, Robinson ME, Bowers D, Banou E, Malaty I, Okun MS. Measurement of patient-centered outcomes in Parkinson's disease: what do patients really want from their treatment? Parkinsonism Relat Disord 2011;17(2):89–94. https://doi.org/10.1016/j.parkreldis.2010.09.005

24 Sikap P et al. Perancangan Prototipe Sistem Pemesanan Makanan dan Minuman Menggunakan Mobile Device. Indonesia Journal on Networking and Security 2015;1(2):1–10. https://doi.org/10.1145/242224.242229

25 Zhan A, Mohan S, Tarolli C, et al. Using smartphones and machine learning to quantify Parkinson disease severity the mobile Parkinson disease score. JAMA Neurol 2018;75(7):876–880. https://doi.org/10.1001/jamaneurol.2018.0809

26 Mei J, Desrosiers C, Frasnelli J. Machine learning for the diagnosis of Parkinson's disease: a review of literature. Frontiers in Aging Neuroscience 2021;13:184. https://doi.org/10.3389/fnagi.2021.633752

27 Yang N, Liu DF, Liu T, et al. Automatic detection pipeline for accessing the motor severity of Parkinson's disease in finger tapping and postural stability. IEEE Access 2022;10:66961–66973. https://doi.org/10.1109/access.2022.3183232

28 Kalia LV, Lang AE. Parkinson's disease. The Lancet 2015;386(9996): 896–912. https://doi.org/10.1016/S0140-6736(14)61393-3

29 Tomlinson CL, Stowe R, Patel S, Rick C, Gray R, Clarke CE. Systematic review of levodopa dose equivalency reporting in Parkinson's disease. Mov Disord 2010;25(15):2649–2653. https://doi.org/10.1002/mds.23429

30 Biometrics Ltd. Twin-Axis goniometers for dynamic joint movement analysis; 2020.

31 Van Rossum G, Drake FL Jr. Python 3 Reference Manual, Version 3.7.3. Scotts Valley, CA: CreateSpace; 2009.

32 Pedregosa F. Scikit-learn: machine learning in {P}ython. Journal of Machine Learning Research 2011;12:2825–2830.

33 Kaiser L. Adjusting for baseline: change or percentage change? Stat Med 1989;8(10):1183–1190. https://doi.org/10.1002/sim.4780081002

34 Parvandeh S, Yeh HW, Paulus MP, McKinney BA. Consensus features nested cross-validation. Bioinformatics 2020;36(10): 3093–3098. https://doi.org/10.1093/bioinformatics/btaa046

35 Navia-Vazquez A, Parrado-Hernandez E. Support vector machine interpretation. Neurocomputing 2006;69(13–15):1754–1759. https://doi.org/10.1016/j.neucom.2005.12.118

36 Deng Y, Liu X, Xin C, Jia W. An interpretable classifier with linear discriminant analysis based on AFS theory. 2019 Chinese Control Conference (CCC). IEEE, New York City; 2019:7583–7588. https://doi.org/10.23919/ChiCC.2019.8866096.

37 Stiglic G, Kocbek P, Fijacko N, Zitnik M, Verbert K, Cilar L. Interpretability of machine learning-based prediction models in healthcare. WIREs Data Mining and Knowledge Discovery 2020;10(5):e1379. https://doi.org/10.1002/widm.1379

38 Moon S, Song HJ, Sharma VD, Lyons KE, Pahwa R, Akinwuntan AE, Devos H. Classification of Parkinson's disease and essential tremor based on balance and gait characteristics from wearable motion sensors via machine learning techniques: a data-driven approach. J Neuroeng Rehabil 2020;17(1):125. https://doi.org/10.1186/s12984-020-00756-5

39 Wu Y, Krishnan S. Statistical analysis of gait rhythm in patients with Parkinson's disease. IEEE Trans Neural Syst Rehabil Eng 2010;18(2):150–158. https://doi.org/10.1109/TNSRE.2009.2033062

40 Geetha R, Sivagami G. Parkinson Disease Classification using Data Mining Algorithms; 2011.
41 Yadav G, Kumar Y, Sahoo G. Predication of Parkinson's disease using data mining methods: a comparative analysis of tree, statistical and support vector machine classifiers. 2012 National Conference on Computing and Communication Systems. IEEE, New York City; 2012:1–8. https://doi.org/10.1109/NCCCS.2012.6413034
42 Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016; 15(2):155–163. https://doi.org/10.1016/j.jcm.2016.02.012
43 Trager MH, Velisar A, Koop MM, Shreve L, Quinn E, Bronte- Stewart H. Arrhythmokinesis is evident during unimanual not bimanual finger tapping in Parkinson's disease. J Clin Mov Disord 2015;2(1):8. https://doi.org/10.1186/s40734-015-0019-2
44 Giovannoni G, Van Schalkwyk J, Fritz VU, Lees AJ. Bradykinesia akinesia incoordination test (BRAIN TEST): an objective computerized assessment of upper limb motor function. J Neurol Neurosurg Psychiatry 1999;67(5):624–629. https://doi.org/10.1136/jnnp.67.5.624
45 Hauser RA, Ellenbogen A, Khanna S, Gupta S, Modi NB. Onset and duration of effect of extended-release carbidopa-levodopa in advanced Parkinson's disease. Neuropsychiatr Dis Treat 2018;14:839–845. https://doi.org/10.2147/NDT.S153321
46 Contin M, Riva R, Martinelli P, Albani F, Avoni P, Baruzzi A. Levodopa therapy monitoring in patients with Parkinson disease: akinetic-dynamic approach. Ther Drug Monit 2001;23(6):621–629. https://doi.org/10.1097/00007691-200112000-00005
47 Lobo V, Branco D, Guerreiro T, Bouça-Machado R, Ferreira J. Machine-learning models for MDS-UPDRS III prediction: a comparative study of features, models, and data sources. Information Society 2022.
48 Ur Rehman RZ, Rochester L, Yarnall AJ, Del Din S. Predicting the progression of Parkinson's disease MDS-UPDRS-III motor severity score from gait data using deep learning. Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society. EMBS, Institute of Electrical and Electronics Engineers Inc., New York City; 2021:249–252. https://doi.org/10.1109/EMBC46164.2021.9630769.
49 Walsh T. Fuzzy gold standards: approaches to handling an imperfect reference standard. J Dent 2018;74:S47–S49. https://doi.org/10.1016/j.jdent.2018.04.022
50 Berisha V, Krantsevich C, Hahn PR, Hahn S, Dasarathy G, Turaga P, Liss J. Digital medicine and the curse of dimensionality. npj Digital Medicine 2018;4(1):1–8. https://doi.org/10.1038/s41746-021-00521-5

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

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

Nederlandse samenvatting

Inleiding

De traditionele methoden voor het monitoren van aandoeningen van het centrale zenuwstelsel (CZS) zijn vaak afhankelijk van sporadische klinische beoordelingen in een klinische omgeving, die een onvolledige of vertekende weergave van de toestand van een patiënt kunnen bieden.[1,2] Deze episodische en persoonlijke aanpak kan schommelingen in de toestand van een patiënt missen en geeft geen volledig beeld van zijn of haar dagelijkse leven. Deze episodische en persoonlijke benadering kan schommelingen in de toestand van een patiënt missen en geeft geen volledig beeld van het dagelijks leven van de patiënt. De vooruitgang in mobiele gezondheid (mHealth) technologieën, waaronder smartphones, wearables en tablets, bieden echter een potentiële oplossing om deze beperkingen aan te pakken door continue, real-time gegevensverzameling over het dagelijks leven van een patiënt mogelijk te maken.[3] Deze mHealth-technologieën kunnen een verscheidenheid aan gezondheidsgegevens monitoren, zoals hartslag, slaappatronen en dagelijkse fysieke activiteit, dag en nacht, ongeacht de locatie van de patiënt. Door mHealth-technologieën te gebruiken om onopvallend gegevens op afstand te verzamelen, kan een arts een completer overzicht krijgen van de klinische status van een patiënt. De integratie van mHealth en ML in klinische studies moet worden gezien als een aanvulling op, en niet als een vervanging van, de traditionele klinische methodologie. De klinische expertise van mensen, waaronder klinische ervaring en menselijke rapportages, blijft onvervangbaar. Naarmate zowel mHealth-technologieën, ML en klinische praktijken zich blijven ontwikkelen, maakt deze geïntegreerde aanpak een meer dynamische en datagestuurde aanpak mogelijk, die ervoor kan zorgen dat het ontwerp van klinische proeven in de voorhoede blijft van zowel technologische als medische vooruitgang.

Alleen al het volume en de complexiteit van de gegevens die worden gegenereerd door mHealth-apparaten kunnen nieuwe uitdagingen met zich meebrengen. Niet alleen de omvang, maar ook de heterogeniteit van de gegevens maakt handmatige analyse niet alleen arbeidsintensief, maar ook moeilijk te modelleren.[4,5] Dit is waar Machine Learning (ML) voor kan zorgen. Dit is waar Machine Learning (ML) om de hoek komt kijken. **Hoofdstuk 2** onderstreept het potentieel van ML-algoritmen om gevalideerde, op mHealth gebaseerde biomarkers te ontwikkelen die kunnen worden ingezet in klinische onderzoeken.[6] ML-algoritmen kunnen op efficiënte wijze enorme en veelzijdige datasets doorzeven om patronen of correlaties te identificeren die kunnen helpen bij de klinische interpretatie van de gegevens. Door ML-algoritmen te combineren met mHealth-gegevens om op afstand gecontroleerde biomarkers te creëren, kunnen we mogelijk nieuwe mHealth-biomarkers creëren die kunnen worden gebruikt voor diagnoseclassificatie, inschatting van de ernst van symptomen en kwantificering van behandelingseffecten. Deze biomarkers kunnen mogelijk nieuwe inzichten genereren die mogelijk gemist worden door de klinische gouden standaardbeoordelingen, waardoor het mogelijk wordt om een dieper inzicht te krijgen in ziektetoestanden.[4] Dit relatief jonge veld vereist echter nog verder onderzoek en standaardisatie om de toepassing van deze technologieën in klinische studies te stimuleren.

In de volgende paragrafen zal ik een samenvatting geven van de bevindingen en discussies in mijn vorige hoofdstukken over de verschillende toepassingen en uitdagingen van mHealth biomarkers in klinisch onderzoek. Ik zal ingaan op hoe deze biomarkers kunnen worden ontwikkeld en toegepast voor diagnoseclassificatie, en als gevolg daarvan nieuwe inzichten bieden in ziektegerelateerde gedragsprofielen die moeilijk te vinden zijn in conventionele klinische settings. Daarnaast zal de rol van mHealth biomarkers bij het inschatten van de ernst van symptomen worden besproken, en ik zal het belang onderzoeken van het ontwikkelen van mHealth biomarkers die betrouwbaar zijn bij verschillende aandoeningen en populaties. Ik zal het ook hebben over hoe deze biomarkers kunnen worden ontworpen voor de detectie van behandelingen, waarmee de weg wordt vrijgemaakt voor longitudinale monitoring van de werkzaamheid van behandelingen. Tot slot zal ik ingaan op de beperkingen van mHealth biomarkers en gebieden identificeren die verder onderzoek en standaardisatie vereisen.

Classificatie van ziekten

In de context van klinische studies bieden biomarkers voor de classificatie van de ernst van de ziekte niet alleen een kwantificeerbare maatstaf om de uitgangswaarde van de ernst van een ziekte bij deelnemers aan de studie te bepalen, maar ze kunnen ook dienen als referentie om de evolutie van de ziekte in de tijd te volgen. Bij het evalueren van de effectiviteit van onderzoeksgeneesmiddelen zijn deze biomarkers van onschatbare waarde. Als het geneesmiddel tot doel heeft het ziekteverloop te beïnvloeden, kan een verandering in het verloop van de biomarker na verloop van tijd een indicatie zijn van het effect van het geneesmiddel. Als gevolg hiervan kan het gebruik van biomarkers voor de classificatie van de ernst van de ziekte de precisie en betrouwbaarheid van de resultaten van klinische onderzoeken verbeteren, door ervoor te zorgen dat potentiële behandelingen worden beoordeeld op zowel hun onmiddellijke effect als hun invloed op de progressie van de ziekte op de langere termijn.

Hoofdstuk 3 onderzocht de haalbaarheid van het classificeren van FSHD-patiënten (Facioscapulohumerale dystrofie) en gezonde controles met behulp van het Trial@Home-platform van het CHDR. Belangrijke kenmerken, zoals slaapactiviteit en locatiepatronen, werden geïdentificeerd die onderscheid maakten tussen FSHD-patiënten en controles 9. Dit suggereert dat significante variaties in slaap- en locatiepatronen kunnen dienen als potentiële nieuwe klinische biomarkers omdat deze momenteel niet worden vastgelegd door de gouden standaard beoordelingen van FSHD.[10] Deze biomarkers, op hun beurt, kunnen essentieel zijn in het begeleiden van het proces van geneesmiddelenontwikkeling, mogelijk bieden ze een gerichte aanpak voor geneesmiddelen interventies in de behandeling of het beheer van de bijbehorende aandoeningen. [11]

Het bereiken van een optimale classificatienauwkeurigheid vereist een delicaat evenwicht tussen de hoeveelheid kenmerken en de duur van de monitoring. Het introduceren van een breder scala aan kenmerken van verschillende sensoren, zoals die van smartwatches en smartphone GPS-systemen, kan de nauwkeurigheid van de voorspellingen verbeteren. Het vergroten van de hoeveelheid informatie in een model maakt het klinisch begrip van deze mHealth-biomarkers echter ook complexer en vergroot de last voor de patiënt als gevolg van de toegenomen gegevensverzameling.[12,13]

Inschatting van symptoomernst

mHealth biomarkers, indien gebruikt voor het schatten van de ernst van de symptomen, bieden een innovatieve aanpak voor het beoordelen van de effecten van medicijninterventies in klinische studies. Als onderzoekers nieuwe medicijnen beoordelen in fase 2 studies, is het begrijpen van de relatie tussen een medicijn, de dosering en de resulterende effecten in de tijd cruciaal.[14] mHealth biomarkers kunnen een duidelijk beeld geven van deze relatie, en helpen bij het vaststellen van een veilige en effectieve dosering. mHealth biomarkers hebben ook het potentieel om te dienen als directe indicatoren van de werkzaamheid van een medicijn. Ze kunnen symptoomschommelingen in de loop van de tijd kwantificeren, wat een uitgebreider beeld geeft dan arbeidsintensieve methoden zoals klinische interviews. Deze frequente monitoring kan vooral waardevol zijn bij het onderscheiden van zelfs de meest subtiele veranderingen in de ernst van de symptomen, wat fundamenteel is voor een vroegtijdige identificatie van de werkzaamheid van een behandeling. Door veranderingen in de biomarkers continu te monitoren, kunnen onderzoekers waardevolle feedback krijgen over de vraag of het medicijn het beoogde effect heeft, wat vooral cruciaal is tijdens fase 2-onderzoeken waar de therapeutische effecten onder de loep worden genomen. Om deze biomarkers als klinisch valide te beschouwen, is het noodzakelijk dat ze correleren met erkende klinische eindpunten. Of deze eindpunten nu ziekteprogressie, symptoomverlichting of andere klinisch relevante maatregelen betreffen, een sterke associatie verzekert dat de biomarker een betrouwbare maatstaf is voor het effect van het geneesmiddel.

Hoofdstuk 4 onderzocht de prestaties van multi-taak modellen om gelijktijdig de scores van twee klinische beoordelingen te schatten, de

FSHD klinische score en de Timed Up and Go (TUG) test.[15] Traditionele enkelvoudige taakmodellen zijn weliswaar betrouwbaar, maar niet altijd. Traditionele single-task modellen kunnen effectief zijn voor het voorspellen van één uitkomst, maar schieten tekort als ze worden toegepast op de multidimensionale symptoomprofielen die vaak voorkomen in klinische settings. Daarom is het belangrijkste voordeel van multi-taak modellen ten opzichte van hun single-taak tegenhangers hun vermogen om gebruik te maken van gedeelde representaties en inzichten over meerdere klinische beoordelingen.[16-18] Bovendien is het vermogen van multi-taak modellen om gedeelde representaties en inzichten over meerdere klinische beoordelingen [16-18] te gebruiken. Bovendien kan het vermogen van multi-taak modellen om te generaliseren van de ene klinische beoordeling naar de andere cruciaal zijn bij het evalueren van de ernst van de ziekte over een spectrum van beoordelingen. Als het model bijvoorbeeld een verslechtering in de FSHD klinische score vaststelt, kan het ook een parallelle afname in de TUG score voorspellen. Tot slot kunnen multi-taakmodellen een meer holistisch beeld geven van de gezondheid van de patient, door verschillende facetten van de ernst van de ziekte in één enkel kader te vatten. Door de parallelle beoordeling van meerdere beoordelingen mogelijk te maken, kunnen deze modellen een vollediger, genuanceerder beeld geven van de ziektestatus, waardoor gerichtere en effectievere interventies mogelijk worden.

In **hoofdstuk 5** kwam het belang van zelfgerapporteerde uitkomsten, met name de Depression Anxiety Stress Scale (DASS) en de Positive and Negative Affect Schedule (PANAS), naar voren als doorslaggevende kenmerken voor de depressiemodellen. Hun opname diende als een robuuste indicator voor subjectieve psychologische toestanden en benadrukte de onvervangbare waarde van patiënteninput bij het vastleggen van de nuances van psychische aandoeningen. Interessant is dat, hoewel passief verzamelde kenmerken zoals loopsnelheid en locatie niet zo voorspellend waren als DASS en PANAS, ze toch een waardevolle bijdrage leverden aan de algehele effectiviteit van de modellen. Deze bevinding onderstreept ook het belang van het integreren van passief verzamelde gegevens uit de echte wereld, omdat deze patronen en inzichten lijken te onthullen die mogelijk over het hoofd worden gezien in meer gecontroleerde klinische settings. Bovendien werd het vermogen van de modellen om het volledige spectrum van depressiezwaarte nauwkeurig weer te geven vergroot door gezonde controles op te nemen. Deze inclusie verbeterde niet alleen de robuustheid van de modellen, maar breidde ook de representatie van de potentiële remissietoestanden van depressie in de modellen uit. Deze multidimensionale aanpak, die zowel actieve als passieve gegevensverzameling combineert, zorgt dus voor een uitgebreider en genuanceerder begrip van psychische aandoeningen.

Het schatten van de ernst van de symptomen met behulp van mHealth biomarkers brengt specifieke uitdagingen met zich mee, vooral wanneer rekening wordt gehouden met de inherente variabiliteit van zowel de apparaten als de patiënten zelf. Een belangrijk punt van zorg is de inter-device variabiliteit.[2] Verschillen in mHealth-apparaten kunnen licht verschillende metingen produceren, wat leidt tot inconsistenties in de verzamelde gegevens. Deze variatie kan ruis introduceren in de analyses, wat de resultaten kan vertekenen of de precisie van de schatting van de ernst van de symptomen kan verminderen. Bovendien kunnen de ernst en de expressie van de symptomen zelf variëren binnen en tussen patiënten, wat nog een laag complexiteit toevoegt aan de modellering. Externe factoren die niet kunnen worden gecontroleerd of waar geen rekening mee kan worden gehouden, kunnen ook metingen in de war sturen. Bijvoorbeeld, terwijl een mHealth apparaat een verhoogde hartslag zou kunnen detecteren als een potentieel symptoom van een gezondheidstoestand, zou deze verhoging echter kunnen worden toegeschreven aan externe invloeden zoals angst, lichaamsbeweging, of andere niet-medische oorzaken. Het onderscheid maken tussen echte symptoomschommelingen en deze externe factoren blijft dus een uitdaging bij het gebruik van mHealth biomarkers voor een nauwkeurige inschatting van de ernst van de symptomen.

Behandelingseffecten

Om behandelingseffecten te detecteren, moeten mHealth biomarkers aantonen dat ze veranderingen in ziekteactiviteit kunnen detecteren na een medicamenteuze interventie. In essentie kan deze benadering van het ontwerpen en valideren van mHealth biomarkers hen waardevolle hulpmiddelen maken, niet alleen voor het begrijpen van een ziekte, maar ook voor het aanpassen en evalueren van behandelingsstrategieën. Hier ligt de focus niet alleen op de biomarker als voorspellend of diagnostisch hulpmiddel, maar ook op zijn gevoeligheid en doeltreffendheid bij het detecteren van behandelingseffecten ten opzichte van de gouden standaard. Door hun gevoeligheid voor door behandeling veroorzaakte veranderingen kunnen deze biomarkers dienen als meer dynamische eindpunten in onderzoeken, waardoor het effect van een behandeling directer en nauwkeuriger kan worden beoordeeld.

Hoofdstuk 8 bespreekt de ontwikkeling van mHealth biomarkers voor het monitoren van de effecten van antiparkinsonmedicijnen en het schatten van de ernst van Parkinson symptomen.[19] De alternatieve index vinger tapping (IFT) biomarker bleek voorspellender en gevoeliger voor behandelingseffecten in de motoriek dan de traditionele MDS-UPDRS III score, zowel wat betreft nauwkeurigheid als klinische significantie. Behandeleffecten werden gedetecteerd na 45 minuten voor de TIFT-biomarker (thumb-index finger tapping) en na 60 minuten voor de samengestelde IFT-biomarkers. Dit komt goed overeen met het gemiddelde begin van de werking van het geneesmiddel L-dopa/carbidopa, dat ongeveer 50 minuten duurt. De bevindingen suggereren dat IFT en TIFT gevoelige instrumenten zijn voor het beoordelen van de motorische functie in de context van symptomatische behandelingen voor aandoeningen zoals de ziekte van Parkinson. De grote effectgroottes die in deze studie werden gevonden, zouden de vereiste steekproefgrootte kunnen verkleinen en de statistische power voor toekomstige studies met taptaken kunnen vergroten. Deze pilotstudie kan bijdragen aan een beter begrip van hoe behandeleffecten op de fijne motoriek nauwkeurig kunnen worden gedetecteerd en gemeten, met name bij aandoeningen zoals de ziekte van Parkinson. Het valideert niet alleen de werkzaamheid van nieuwe biomarkers, maar biedt ook methodologische richtlijnen voor het valideren van nieuwe biomarkers in toekomstig onderzoek dat zich richt op het onderzoeken van medicijneffecten.

Herhaalbaarheid van voorspellingen over tijd en instellingen

In de context van klinisch onderzoek verwijst de term 'herhaalbaarheid' naar het vermogen van een test, meting of algoritme om consistente resultaten op te leveren wanneer deze meerdere keren onder dezelfde omstandigheden wordt uitgevoerd.[20,21] In zowel klinische als thuissituaties moet de herhaalbaarheid van voorspellingen consistent zijn. In zowel klinische als thuissituaties is consistente monitoring van vitaal belang voor het volgen van de progressie of verlichting van symptomen. Als bijvoorbeeld een algoritme voor hoestdetectie wordt gebruikt om de effectiviteit van een nieuw astmamedicijn bij kinderen te controleren, zouden inconsistente resultaten de integriteit van het onderzoek in gevaar brengen en tot onjuiste conclusies kunnen leiden. Voor algoritmen die zijn ontworpen om biologische signalen of gebeurtenissen te monitoren, zoals hoesten of schreeuwen, is herhaalbaarheid in verschillende instellingen voor gegevensverzameling en bij verschillende patiënten een belangrijk kenmerk dat de betrouwbaarheid van het algoritme onderstreept. Op het gebied van informatica en ML kan herhaalbaarheid worden verwisseld met 'robuustheid' en 'externe validiteit'. In wezen verwijzen deze termen - herhaalbaarheid, robuustheid en externe geldigheid - naar de consistente prestaties van een algoritme onder verschillende omstandigheden en datasets. **Hoofdstuk 6** en **7** richtten zich op de ontwikkeling van een smartphonegebaseerd algoritme voor geautomatiseerde hoest- en huildetectie bij baby's en kinderen.[22,23] Beide algoritmen vertonen een sterke herhaalbaarheid. Beide algoritmen vertonen een sterke herhaalbaarheid, wat cruciaal is voor consistente monitoring in de tijd. Het huilalgoritme

lijkt robuust tegen verschillende soorten fysieke barrières en kan op verschillende afstanden worden gebruikt, waardoor het flexibel is voor toepassingen in de echte wereld. Hoewel beide algoritmen een zekere mate van inter-device variabiliteit vertonen, ligt deze binnen een acceptabel bereik dat hun bruikbaarheid niet ernstig in gevaar brengt. Beide algoritmen worden beïnvloed door achtergrondruis, zij het in verschillende mate. Dit wijst op een gebied dat voor verbetering vatbaar is. Deze bevindingen suggereren dat beide algoritmen robuust genoeg zijn voor potentieel gebruik bij het monitoren van huilen en hoesten in een klinische setting of voor thuiszorg, hoewel aanpassingen nodig kunnen zijn afhankelijk van het gebruikte apparaat of de omgevingscondities.

Beperkingen

Veel aandoeningen, zoals psychische stoornissen of chronische ziekten, hebben vele facetten en kunnen mogelijk niet volledig worden vastgelegd door een enkele gouden standaard beoordeling of een enkel apparaat. In deze gevallen kan het zijn dat zowel de gouden standaard als de mHealth apparaten de complexiteit van de ziekte niet vastleggen, wat leidt tot discrepanties bij het vergelijken van de werkelijke en voorspelde klinische scores. Deze discrepanties kunnen het gevolg zijn van drie oorzaken. Ten eerste, beperkingen van de mHealth apparaten om al het klinisch relevante gedrag vast te leggen. Bijvoorbeeld, de mHealth apparaten slaagden er niet in om de bovenarm functionaliteit van FSHD patiënten vast te leggen en dus ook niet te voorspellen, zoals te zien is in **Hoofdstuk 3** en **4**.[9,15] Ten tweede, tekortkomingen van de gouden standaarden in het vastleggen van alle klinisch relevante gedragingen. Zoals te zien in Hoofdstuk 5, vonden we dat loop- en reisgedrag voorspellend zijn voor MDD, maar deze kenmerken worden niet behandeld door de SIGH-D IDSC. Verder kunnen de beperkingen van de gouden standaard, zoals interbeoordelaarsvariabiliteit of het niet vastleggen van de volledige complexiteit van een ziekte, vooroordelen introduceren die de betrouwbaarheid van de biomarker beïnvloeden. In sommige gevallen is de gouden standaard een menselijke beoordeling, die kan variëren afhankelijk van de deskundigheid van de beoordelaar of zelfs de dagelijkse omstandigheden. Bijvoorbeeld, in **Hoofdstuk 8**, zouden de vingertaptaken waarbij meerdere tikgerelateerde kenmerken worden gevolgd, inzichten kunnen bieden in motorische functionaliteit die uitgebreider zouden kunnen zijn dan traditionele onderzoeken naar de ziekte van Parkinson die uitsluitend gebaseerd zijn op klinische observatie.[19] Ten derde kunnen er verschillen zijn tussen de objectieve gedragsbiomarkers en subjectieve eindpunten. Een depressieve patiënt kan bijvoorbeeld melden dat hij zich rustelozer voelt als hij in bed ligt, maar de objectieve slaapgegevens die zijn vastgelegd door de smartwatch laten zien dat de patiënt 8 uur heeft geslapen. Het resultaat is dat de objectieve meting van de slaap mogelijk niet goed correleert met de subjectieve ervaring van de slaap, zoals we in **hoofdstuk 5** hebben gezien. Daarom is het cruciaal om zowel objectieve metingen als subjectieve ervaringen in overweging te nemen bij het evalueren van de effectiviteit van mHealth-apparaten voor het monitoren en beheren van aandoeningen zoals depressie. Objectieve metingen zijn niet altijd een representatief eindpunt voor subjectieve ervaringen.

De discrepanties tussen mHealth-sensoren en de gouden standaard kunnen van invloed zijn op hoe betrouwbaar clinici en onderzoekers deze sensoren vinden. Om een nieuwe technologie te integreren in klinische studies, moet deze ofwel dicht in de buurt komen van de gouden standaard of duidelijk zijn superioriteit aantonen. Het is de moeite waard om op te merken dat een lagere correlatie tussen mHealth biomarkers en de gouden standaard misschien niet duidt op een slechte klinische validiteit van de nieuwe biomarker; in plaats daarvan kan het mHealth systeem aspecten vastleggen die door traditionele methoden over het hoofd worden gezien. Daarom is het begrijpen van de beperkingen en vertekeningen die inherent zijn aan zowel de mHealth biomarker als de gouden standaard cruciaal voor het maken van nauwkeurige klinische beslissingen. Als clinici zich bewust zijn van deze factoren, kunnen ze de gegevens genuanceerder interpreteren.

Conclusie

Concluderend kan worden verwacht dat mHealth biomarkers en ML een paradigmaverschuiving zullen veroorzaken in het monitoren en beheren van CNS ziekten. Deze geavanceerde technologieën, gefaciliteerd door smartphones, wearables en tablets, kunnen zorgen voor een meer directe, continue en accurate beoordeling van ziekte. Daarom kunnen deze mHealth biomarkers traditionele episodische evaluaties veranderen in genuanceerde, longitudinale gegevensgestuurde analyses. De onderzoeksresultaten tonen de robuuste voorspellende capaciteiten, nauwkeurigheid, betrouwbaarheid en klinische relevantie van deze ontwikkelde biomarkers aan. Het is echter belangrijk om te erkennen dat verder onderzoek, ontwikkeling en standaardisatie nodig zijn om de voordelen van deze innovaties volledig te realiseren. Uiteindelijk bieden deze ontwikkelingen niet alleen een beter begrip van de ernst en progressie van de ziekte, maar ook betere hulpmiddelen om de potentiële werkzaamheid van farmacologische interventies te bepalen

.

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

中文摘要

导言

监测中枢神经系统（CNS）疾病的传统方法通常依赖于在临床环境下进行的零星现场临床评估，这可能无法全面或歪曲地反映患者的病情。[1,2]这种偶发的当面评估方法可能会错过患者病情的波动，也无法全面了解患者的日常生活。然而，移动医疗（MHEALTH）技术（包括智能手机、可穿戴设备和平板电脑）的发展为解决这些局限性提供了一个潜在的解决方案，它可以对患者的日常生活进行连续、实时的数据收集。[3]这些移动医疗技术可以监测各种健康指标，如心率、睡眠模式和全天候的身体活动，而不受患者所在位置的限制。利用移动医疗技术以不显眼的方式远程收集数据，可以让临床医生更全面地了解病人的临床状况。移动医疗和 ML 与临床试验的结合应被视为传统临床方法的补充，而不是替代。人类的临床专业知识，包括临床经验和人际关系，仍然是不可替代的。随着移动医疗技术、人工智能和临床实践的不断发展，这种综合方法允许采用更加动态和数据驱动的方法，从而确保临床试验的设计始终走在技术和医学进步的前沿。

移动医疗设备所产生的数据量之大、复杂程度之高可能会带来新的挑战。不仅是数据量大，数据的异质性也使得人工分析不仅耗费大量人力，而且难以建模。[4,5]这正是机器学习（ML）发挥作用的地方。第 2 章强调了 ML 算法在开发可用于临床试验的基于移动医疗的有效生物标记物方面的潜力。[6] ML 算法可以有效地筛选大量多方面的数据集，找出有助于临床解读数据的模式或相关性。通过将 ML 算法与移动医疗数据相结合来创建远程监测的生物标记，我们有可能创建新型移动医疗生物标记，用于诊断分类、症状严重程度估计和治疗效果量化。这些生物标记物有可能产生临床金标准评估可能遗漏的新见解，从而加深对疾病状态的了解。[4] 然而，这个相对年轻的领域仍需要进一步的研究和标准化，以鼓励将这些技术应用于临床试验。

在下面的章节中，我将总结前几篇论文中的发现和讨论，探讨移动医疗生物标记在临床试验中的各种应用和挑战。我将讨论如何开发这些生物标记物并将其应用于诊断分类，从而为传统临床环境中可能难以捉摸的疾病相关行为特征提供新的见解。此外，我还将讨论移动医疗生物标记在估计症状严重程度方面的作用，并探讨开发可靠的移动医疗生物标记在不同疾病和人群中的重要性。我还将讨论如何将这些生物标记设计用于治疗检测，为纵向监测治疗效果创造条件。最后，我将深入探讨移动医疗生物标记的局限性，确定需要进一步研究和标准化的领域。

疾病分类

在临床试验中，疾病严重程度分类生物标志物不仅能提供一种可量化的测量方法来评估试验参与者的基线疾病严重程度，还能作为跟踪疾病随时间进展的参考。在评估研究药物的有效性时，这些生物标记物就变得非常宝贵。如果药物旨在影响疾病的发展轨迹，那么生物标志物随时间推移而发生的变化就能说明药物的效果。因此，利用疾病严重程度分类生物标志物可以提高临床试验结果的准确性和可靠性，确保对潜在治疗方法进行评估时，既能考虑其直接影响，也能考虑其对疾病长期发展的影响。

第3章研究了利用CHDR的TRIAL@HOME平台对面阔肌营养不良症（FSHD）患者和健康对照组进行分类的可行性。研究发现，睡眠活动和位置模式等关键特征可以区分面岬-肱骨营养不良症患者和对照组。[9]这表明，在睡眠和位置模式中观察到的重大差异可作为潜在的新型临床生物标志物，因为目前FSHD的金标准评估并未捕捉到它们。[10]反过来，这些生物标志物对指导药物开发过程也至关重要，有可能为治疗或控制相关疾病的药物干预提供有针对性的方法。[11]

要达到最佳的分类准确性，需要在特征数量和监测持续时间之间取得微妙的平衡。从智能手表和智能手机 GPS 系统等各种传感器中引入更广泛的特征，可以提高预测的准确性。但是，增加模型的信息量也会增加临床理解这些移动医疗生物标记的复杂性，并增加患者因数据收集增加而产生的负担。[12,13]

症状严重程度估计

移动医疗生物标记用于症状严重程度评估时，为评估临床试验中药物干预的效果提供了一种创新方法。研究人员在第二阶段试验中对新药进行评估时，了解药物、药物剂量及其随时间产生的效果之间的关系至关重要。与临床访谈等劳动密集型方法相比，它们可以量化症状随时间的变化，提供更全面的视

图。这种频繁的监测对于辨别症状严重程度最细微的变化尤为重要，而这正是早期识别疗效的基础。通过持续监测生物标记物的变化，研究人员可以获得关于药物是否达到预期效果的宝贵反馈，这在治疗效果受到严格审查的第二阶段试验中尤为重要。要使这些生物标志物在临床上有效，它们必须与公认的临床终点相关联。无论这些终点是涉及疾病进展、症状缓解还是其他临床相关指标，强相关性都能确保生物标记物是衡量药物影响的可靠指标。

第 4 章研究了多任务模型在同时估算 FSHD 临床评分和定时起立行走 (TUG) 测试这两项临床评估得分时的性能。[15]传统的单任务模型虽然可以有效地预测单一结果，但在应用于临床环境中经常遇到的多维症状特征时，可能会出现不足。因此，与单任务模型相比，多任务模型的主要优势在于能够利用多个临床评估的共享表征和洞察力。[16-18] 此外，多任务模型能够从一种临床评估推广到另一种临床评估，这对于评估各种评估的疾病严重程度至关重要。例如，如果模型能识别 FSHD 临床评分的恶化，那么它也能预测 TUG 评分的平行下降。最后，多任务模型可以提供更全面的患者健康视图，在单一、统一的框架内涵盖疾病严重程度的各个方面。通过对多项评估进行并行评估，这些模型可以更全面、更细致地反映疾病状况，从而指导采取更有针对性和更有效的干预措施。

在第 5 章中，自我报告结果的重要性，特别是抑郁焦虑压力量表 (DASS) 和积极与消极情绪表 (PANAS)，成为抑郁模型的决定性特征。纳入这两项量表可作为主观心理状态的可靠指标，凸显了患者意见在捕捉心理健康状况细微差别方面不可替代的价值。有趣的是，尽管步行速度和位置等被动收集的特征不像 DASS 和 PANAS 那样具有预测性，但它们仍然对模型的整体有效性做出了宝贵的贡献。这一发现也强调了整合现实世界中被动收集的数据的重要性，因为这些数据似乎揭示了在更受控的临床环境中可能被忽视的模式和见解。此外，纳入健康对照组也增强了模型准确表现抑郁症严重程度的能力。健康对照组的加入不仅增强了模型的稳健性，还扩展了模型对抑郁症潜在缓解状态的表征。因此，这种结合主动和被动数据收集的多维方法可以更全面、更细致地了解心理健康状况。

使用移动医疗生物标记物估计症状严重程度面临着特殊的挑战，特别是考虑到设备和患者本身固有的变异性。一个重要的问题是设备间的可变性。[2] 移动健康设备之间的差异可能会产生略微不同的测量结果，从而导致所收集的数据不一致。这种差异会在分析中引入噪音，可能会使结果出现偏差或降低症状严重程度估计的精确度。此外，症状严重程度和表现本身在患者内部和患者之间也可能存在差异，这就给建模工作增加了另一层复杂性。无法控制或考虑的外部因素也会干扰读数。例如，移动医疗设备可能会检测到心率增快作为健康状况的潜在症状，但这种增快可能是由于焦虑、体育锻炼或其他非医疗原因等外部影响造成的。因此，将真正的症状波动与这些外部因素区分开来仍然是利用移动医疗生物标记准确估计症状严重程度的一个挑战。

治疗效果

为了检测治疗效果，移动健康生物标记需要证明其有能力检测药物干预后疾病活动的变化。从本质上讲，这种设计和验证移动健康生物标志物的方法不仅能使它们成为了解疾病的重要工具，还能使它们成为定制和评估治疗策略的重要工具。在这里，重点不仅在于生物标记物作为预测或诊断工具，还在于其相对于金标准检测治疗效果的灵敏度和有效性。通过展示对治疗引起的变化的敏感性，这些生物标志物可以作为试验中更动态的终点，从而有助于对治疗效果进行更即时、更准确的评估。

第 8 章讨论了用于监测抗帕金森病药物效果和估计帕金森病症状严重程度的移动医疗生物标记物的开发。[19] 研究发现，与传统的 MDS-UPDRS III 评分相比，替代性食指敲击 (IFT) 生物标记在准确性和临床意义方面对运动功能的治疗效果更有预测性和敏感性。拇指-食指敲击 (TIFT) 生物标志物在45分钟时检测到治疗效果，IFT复合生物标志物在60分钟时检测到治疗效果。这与左旋多巴/卡比多巴药物的平均起效时间 (约 50 分钟) 非常吻合。研究结果表明，IFT 和 TIFT 是评估帕金森病等疾病症状治疗过程中运动功能的灵敏工具，有可能识别出传统方法所遗漏的早期微小变化。本研究中还发现了较大的效应大小，这可以降低样本量要求，提高未来涉及敲击任务研究的统计能力。这项试验性研究可以加深

人们对如何准确检测和测量精细运动功能治疗效果的理解，尤其是在帕金森病等疾病中。它不仅验证了新生物标记物的有效性，还为今后重点调查药物效果的研究中验证新生物标记物提供了方法指导。

预测结果在不同时间和环境下的重复性

在临床研究中，‘可重复性’一词指的是测试、测量或算法在相同条件下多次执行时产生一致结果的能力[20,21]。在临床和家庭环境中，持续监测对于跟踪症状的发展或缓解至关重要。例如，如果使用咳嗽检测算法来监测儿童哮喘新药的疗效，不一致的结果会损害研究的完整性，并可能导致错误的结论。对于旨在监测生物信号或事件（如咳嗽或哭声）的算法来说，在不同的数据收集环境和患者中的可重复性是强调算法可靠性的关键属性[20]。在计算机科学和人工智能领域，可重复性可以与 ‘鲁棒性’ ‘和’ ‘外部有效性’ ‘互换’。从本质上讲，这些术语--可重复性、稳健性和外部有效性--指向算法在不同条件和数据集下的一致表现。第 6 章和第 7 章的重点是开发基于智能手机的婴幼儿咳嗽和哭声自动检测算法[22-23]。这两种算法都显示出很强的可重复性，这对长期持续监测至关重要。哭声算法对不同类型的物理障碍具有很强的抵御能力，并可在不同距离内使用，因此在实际应用中非常灵活。虽然两种算法都显示出一定程度的设备间变异性，但都在可接受的范围内，不会严重影响其实用性。两种算法都受到背景噪声的影响，只是程度不同而已。这就指出了一个潜在的改进领域。这些研究结果表明，这两种算法都足够强大，可用于监测临床环境或家庭护理中的哭声和咳嗽声，但可能需要根据所使用的设备或环境条件进行调整。

局限性

许多疾病，如精神障碍或慢性疾病，都是多方面的，单一的金标准评估或单一的设备可能无法完全捕捉。在这种情况下，金标准和移动医疗设备可能都无法捕捉到疾病的复杂性，从而导致在比较真实和预测的临床评分时出现差异。造成这些差异的原因有三个。首先，移动医疗设备在捕捉所有临床相关行为方面存在局限性。例如，如第 3 章和第 4 章所述，移动医疗设备未能捕捉 FSHD 患者的上臂功能，因此也就无法预测 FSHD 患者的上臂功能。[9,15] 其次，黄金标准在捕捉所有临床相关行为方面存在缺陷。如第 5 章所述，我们发现步行和旅行行为可预测 MDD，但 SIGH-D IDSC 并未涉及这些特征。此外，金标准的局限性，如评定者之间的差异或无法捕捉疾病的全部复杂性，可能会带来影响生物标志物可靠性的偏差。在某些情况下，金标准涉及人工评估，而人工评估可能因评估者的专业知识甚至日常条件而异。例如，在第 8 章中，手指敲击任务可追踪多种敲击相关特征，与仅依赖临床观察的传统帕金森病研究相比，可提供更全面的运动功能洞察。[19] 第三，客观行为生物标志物与主观终点之间可能存在差异。例如，抑郁症患者可能会报告说躺在床上时感觉更加烦躁不安，但智能手表捕获的客观睡眠数据却显示患者睡了 8 小时。因此，睡眠的客观测量结果可能与睡眠的主观体验并不十分相关，如第 5 章所述。因此，在评估移动医疗设备监测和管理抑郁症等疾病的有效性时，必须同时考虑客观测量结果和主观体验。客观测量结果并不总是主观体验的代表终点。

移动医疗传感器与黄金标准之间的差异会影响临床医生和研究人员对这些传感器可靠性的看法。新技术要想被纳入临床试验，就必须与黄金标准密切匹配，或者明确显示出其优越性。值得注意的是，移动医疗生物标志物与黄金标准之间的相关性较低可能并不表明新型生物标志物的临床有效性较差；相反，移动医疗系统可能捕捉到了传统方法所忽略的方面。因此，了解移动医疗生物标志物和金标准固有的局限性和偏差对于做出准确的临床决策至关重要。如果临床医生了解这些因素，他们就能对数据做出更细致的解释。

结论

总之，移动医疗生物标志物和 ML 可望引起中枢神经系统疾病监测和管理模式的转变。在智能手机、可穿戴设备和平板电脑的推动下，这些先进技术可以提供更加即时、连续和准确的疾病评估。因此，这些移动医疗生物标记可将传统的偶发性评估转变为细致入微的纵向数据驱动分析。研究结果表明，这些开发的生物标记物具有强大的预测能力、准确性、可靠性和临床相关性。然而，重要的是要认识到需要进一步研究、开发和标准化，以充分实现这些创新的益处。最终，这些进步不仅能让人们更全面地了解疾病的严重程度和进展，还能提供更好的工具来确定药物干预的潜在疗效。

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

PUBLICATIONS

Wevers, R., & **Zhuparris, A.** (2024). Embodying data, shifting perspective: a conversation with ahnjili zhuparris on future wake. In K. Leurs & S. Ponzanesi (Eds.), *Doing Digital Migration Studies: Theories and Practices of the Everyday* (pp. 89–106). Amsterdam University Press. https://doi.org/10.2307/jj.11895524.9

Hijma, H.J., **Zhuparris, A.**, van Hoogdalem, E.-J. and Cohen, A.F. (2024). Disproportional inflation of clinical trial costs: why we should care, and what we should do about it. *Nature Reviews Drug Discovery*. doi:https://doi.org/10.1038/d41573-024-00002-w.

Zhuparris A, Thijssen E, Elzinga WO, et al. Treatment detection and movement disorder society-unified parkinson's disease rating scale, part iii estimation using finger tapping tasks. Movement disorders. Published online July 4, 2023. doi:10.1002/mds.29520

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

Saghari, M., Gal, P., Grievink, H. W., Klaassen, E. S., **Zhuparris, A.**, Itano, A., Bodmer M., McHale D., Moerland, M. (2024). Evaluation of single-strain Prevotella histicola on KLH-driven immune responses in healthy volunteers: A randomized controlled trial with EDP1815. Elsevier BV. https://doi.org/10.1016/j.medmic.2023.100088

Rousel J., Nădăban A., Saghari M., Pagan L., **Zhuparris, A.**, Theelen B., Gambrah T., et al. 2023. 'Lesional skin of seborrheic dermatitis patients is characterized by skin barrier dysfunction and correlating alterations in the stratum corneum ceramide composition.' *Experimental Dermatology* 33 (1). https://doi.org/10.1111/exd.14952.

Zhuparris A, Maleki G, Koopmans I, et al. smartphone and wearable sensors for the estimation of facioscapulohumeral muscular dystrophy disease severity: cross-sectional study. JMIR Form Res. 2023;7:e41178. doi:10.2196/41178

ten Voorde, W., Saghari, M., Boltjes, J., de Kam, M. L., **Zhuparris, A.**, Feiss, G., Buters T.P., Prens, E.P., Damman J., Niemeyer-van der Kolk T., Moerland, M., Burggraaf J., van Doorn M.B.A., Rissmann, R. (2023). A multimodal, comprehensive characterization of a cutaneous wound model in healthy volunteers. Wiley. https://doi.org/10.1111/exd.14808

Zhuparris, A., Maleki, G., van Londen, L., Koopmans I., Aalten V., Yocarini, I. R., Exadaktylos, V. van Hemert, A, Cohen A., Gal, P., Doll, RJ, Groeneveld, G., Jacobs, G, Kraaij, W. A smartphone- and wearable-based biomarker for the estimation of unipolar depression severity. *Sci Rep* **13**, 18844 (2023). https://doi.org/10.1038/s41598-023-46075-2

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

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

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

Kruizinga MD, Moll A, **Zhuparris A**, et al. Postdischarge recovery after acute pediatric lung disease can be quantified with digital biomarkers. Respiration. 2021;100(10):979-988. doi:10.1159/000516328

Kruizinga, M. D., Heide, N. van der, Moll, A., **Zhuparris, A.**, Yavuz, Y., Kam, M. L. de, Stuurman, F. E., Cohen, A. F., Driessen, G. J. A. (2021). Towards remote monitoring in pediatric care and clinical trials—Tolerability, repeatability and reference values of candidate digital end-points derived from physical activity, heart rate and sleep in healthy children (J. Harezlak, Ed.). Public Library of Science (PLoS). https://doi.org/10.1371/journal.pone.0244877

Sverdlov, O., Curcic, J., Hannesdottir, K., Gou, L., de Luca, V., Ambrosetti, F., Zhang, B., Praestgaard, J., Vallejo, V., Dolman, A., Gomez-Mancilla, B., Biliouris, K., Deurinck, M., Cormack, F., Anderson, J. J., Bott, N. T., Peremen, Z., Issachar, G., Laufer, O., Joachim, D., Jagesar, R. R., Jons, N., Kas, M. J., ZhuParris, A., Zuiker, R., Recourt, K., Jacobs, G. E. (2021). A study of novel exploratory tools, digital technologies, and central nervous system biomarkers to characterize unipolar depression. Frontiers in Psychiatry, 12. https://doi.org/10.3389/fpsyt.2021.640741

Prins S, **Zhuparris A**, Hart EP, Doll RJ, Groeneveld GJ. A cross-sectional study in healthy elderly subjects aimed at development of an algorithm to increase identification of Alzheimer pathology for the purpose of clinical trial participation. Alzheimers Res Ther. 2021;13(1):132. doi:10.1186/s13195-021-00874-9

Prins S, **Zhuparris A**, Groeneveld GJ. Usefulness of plasma amyloid as a prescreener for the earliest alzheimer pathological changes depends on the study population. Ann Neurol. 2020;87(1):154-155. doi:10.1002/ana.25634

Kruizinga, M. D., Essers, E., Stuurman, F. E., **Zhuparris, A.**, van Eik, N., Janssens, H. M., Groothuis, I., Sprij, A. J., Nuijsink, M., Cohen, A. F., & Driessen, G. J. A. (2020). Technical validity and usability of a novel smartphone-connected spirometry device for pediatric patients with asthma and cystic fibrosis. Pediatric Pulmonology, 55(9), 2463–2470.

Hupli A, Berning M, **Zhuparris A**, Fadiman J. Descriptive assemblage of psychedelic microdosing: Netnographic study of YoutubeTM videos and on-going research projects. Perform Enhanc Health. 2019;6(3-4):129-138. doi:10.1016/j.peh.2019.01.001

Curriculum Vitae

Ahnjili ZhuParris (New York, 1990) graduated from the Chinese International School in Hong Kong in 2009 and subsequently pursued a bachelor's degree in biomedical sciences (with Honours in Neuroscience) from the University of Edinburgh, United Kingdom. Her interest in psychedelic research led her to pursue a master's degree in cognitive neuroscience at Radboud University, where her thesis focused on comparing the effects of LSD, Methylphenidate, and mindfulness on cognitive flexibility.

After completing her master's, Ahnjili worked as a data science consultant for e-commerce companies, which sparked her curiosity in the intersection of data science and clinical research. She joined the Centre for Human Drug Research (CHDR) in Leiden as a data scientist, where she contributed to the development of a data science library for the analysis of Trial@Home data. This experience inspired her to pursue a Ph.D., with a focus on developing novel biomarkers using smartphone and wearable data, under the guidance of prof.dr. G.J. Groeneveld (CHDR, Leiden University Medical Center), dr. R.J. Doll (CHDR), and prof. dr. W. Kraaij (Leiden Institute of Advanced Computer Science).

As of 2024, Ahnjili is primarily based in the Netherlands, where she works as a machine learning engineer in the aesthetic treatment industry. She is also an AI artist and facilitator of AI workshops centered on the exploration of AI applications. Her work predominantly revolves around the critical examination of AI's role, questioning its application and integration in various contexts. Ahnjili's work has garnered support from the Mozilla Foundation (US), IMPAKT(NL), and Constant (BE) and has showcased her work at Ars Electronica (AU), Articulating Data (UK), and the CICA Museum (KR).

Acknowledgements

I sincerely want to extend my deepest gratitude and appreciation to the incredible individuals who have been the pillars of my journey. To Michel, Grandma, LaRose, LaVerne, and Nicky, this journey, filled with challenges and triumphs, would not have been possible without the collective support you all have shown me.

我真诚地想要向那些一直是我旅程中支柱的了不起的人们表达我最深的感激和欣赏。对于朱志坚、我亲爱的爷爷奶奶,和刘新，你们坚定不移的支持和爱是我成就的基石。你们的鼓励和对我的信任让我前进，使得曾经看似不可能的事情变为可能。这段充满挑战和胜利的旅程，如果没有你们所有人的集体支持、爱和信念，是不可能完成的。

To my supportive supervisors, Geert Jan, Robert Jan, and Wessel, I extend my sincere gratitude for your invaluable guidance and mentorship. Your expertise and encouragement have been pivotal in shaping my intellectual growth. A heartfelt gratitude to my daily supervisor, Robert-Jan, whose supervision has been instrumental in cultivating my critical thinking as an academic and data scientist.

A special shoutout to my boyfriend, Peter, for being my rock during this academic pursuit and for designing the cover of my thesis. Your unwavering emotional support and infectious laughter have been a source of strength, especially during challenging times.

I am also indebted to my exceptional friends and colleagues, whose camaraderie and insights have enriched this academic adventure.

Lastly, I would like to thank the Centre for Human Drug Research, for making this PhD trajectory possible.