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

Zhuparris, A.

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