Ecological validity of biomarkers in drug research

Koopmans, I.W.

ECOLOGICAL VALIDITY OF BIOMARKERS IN DRUG RESEARCH INGRID KOOPMANS

ECOLOGICAL VALIDITY OF BIOMARKERS IN DRUG RESEARCH

Ingrid Koopmans

ECOLOGICAL VALIDITY OF BIOMARKERS IN DRUG RESEARCH

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 6 november 2025
klokke 14:30 uur

door
Ingrid Wilhelmina Koopmans
geboren te Wageningen
in 1991

DESIGN
Caroline de Lint, Den Haag (caro@delint.nl)

The publication of this thesis was financially supported by the foundation
Centre for Human Drug Research (CHDR), Leiden, the Netherlands

Promotor

Prof. Dr. G.J. Groeneveld

Copromotor

Dr. Ir. R.J. Doll

Doctorate committee

Prof. Dr. J.M.A. van Gerven
Prof. Dr. M. Fiocco (Leiden University)
Prof. Dr. Ir. J.R. Buitenweg (UTwente)
Prof. Dr. F.C.T van der Helm (TU Delft)

CHAPTER 1

INTRODUCTION

Drug development typically consists of conducting multiple studies to evaluate a compound's efficacy, side-effect profile, and risk management strategies. The initial phase involves preclinical testing in vitro -primarily to confirm mechanistic target engagement, such as receptor binding and affinity- followed by animal studies to assess desired pharmacological effects, pharmacokinetic properties, and toxicology. Once preclinical findings sufficiently demonstrate both efficacy and safety, the compound can proceed to evaluation in humans.

Clinical trials are conducted in human subjects and are typically categorized into three types, as defined by the European Medicines Agency (EMA): human pharmacology, safety and efficacy, and special populations.[1] The earliest human study, known as a First-in-Human (FIH) trial, is designed to closely monitor safety and pharmacokinetics (i.e., absorption, distribution, metabolism, and excretion). Desired pharmacodynamic outcomes can also be evaluated during this stage.[2] To minimize confounding factors, these FIH trials are often conducted in healthy participants, thereby offering a clearer assessment of side effects unclouded by disease symptoms. Additionally, healthy subjects are typically more readily recruited, facilitating a faster start to clinical development. These studies in healthy volunteers are commonly referred to as Phase 1 studies.

Following favourable Phase 1 results—demonstrating safety and suitable pharmacokinetics—trials move on to the patient population. Early patient studies aim to identify an effective dose, evaluate safety in the target group, and verify the drug's underlying concept. Such trials are generally called Phase 2 studies. Larger, adequately powered studies that definitively test a drug's efficacy are referred to as Phase 3. By the time a drug reaches advanced clinical development, the combined outcomes of Phases 1, 2, and 3 must demonstrate both safety and efficacy, as well as a clear advantage over existing treatments.

To achieve this proof, the final stage before registration typically involves a trial in the intended patient population, using endpoints that drug regulators recognize as clinically meaningful. Known as clinical outcome assessments (COAs), these endpoints are often tied directly to the patient's quality of life. COAs may be patient-reported, observer-reported, clinician-rated, or part of a standardized performance measure.[3] For instance, in patients with epilepsy, a reduction in seizure frequency is widely accepted as a relevant outcome. Other examples include improved scores on daily-life questionnaires completed by asthma patients or increased walking distance in the 6-minute walk test for individuals with Duchenne muscular dystrophy.[3]

Progressing from a FIH trial to these large-scale registration studies is both lengthy and expensive. Moreover, the likelihood of success is relatively low—only around 10% of drugs entering clinical development ultimately achieve registration.[4] Although precise figures are infrequently disclosed, costs are estimated at around 500 million euros per approved drug, excluding expenses related to failed compounds.[5]

The clinical 'funnel' can be illustrated by the transition probabilities between phases—Phase 1 to Phase 2, Phase 2 to Phase 3, and Phase 3 to registration—with the majority of terminations occurring in Phase 2.[4] The relatively higher success rate in progressing from Phase 1 to Phase 2 is often attributed to the narrower focus of Phase 1 studies on safety and pharmacokinetics, as well as the solid pharmacological data collected from animal models.

In contrast, moving from Phase 2 to the large Phase 3 registration trials is less frequently successful. This juncture is considered critical because trial sizes and associated costs escalate significantly, making failures more financially consequential.[5] To mitigate risks, drug developers strive to gather robust evidence early in development and conduct smaller, proof-of-concept studies with sensitive assessments. These approaches help reduce the likelihood of unfavourable outcomes during registration trials.

One strategy aimed at curbing development costs is a phase-agnostic, question-based framework called Question-Based Drug Development (QBDD). This method emphasizes systematically asking the most critical—and riskiest—questions early. It encompasses six core questions spanning the drug's path through the human body, from 'Does the drug reach its site of action?' to 'What are the on-target and off-target pharmacological effects?' By directly translating these questions into study objectives and prioritizing them according to each drug's financial risks, early failures can be identified, thus reducing unnecessary later-stage expenditures.[6]

Even with QBDD principles, a considerable gap remains between initial patient studies and large, costly registration trials. Early patient studies (often termed Phase 2a) typically focus on target engagement—demonstrating that the drug modifies the disease's underlying pathophysiology.[7] Subsequent Phase 2b studies frequently resemble small-scale registration trials, employing the same COAs as in Phase 3. Unfortunately, these trials may be underpowered due to smaller sample sizes. The potential for false-positive (Type I) errors could lead to failure in Phase 3 and major financial repercussions, while false-negative (Type II) errors might cause premature abandonment of promising therapies.

Ideally, Phase 2b studies should incorporate endpoints that balance the mechanistic accuracy of Phase 2a measures with the real-world relevance of Phase 3 COAs. These 'bridging' endpoints are more precise and less variable than standard COAs -thus requiring fewer participants- yet are still closer to real-life outcomes than basic mechanistic readouts. In essence, they serve as an intermediary step between early proof-of-concept and full-scale clinical outcome assessments, aiding in a more reliable transition to successful Phase 3 trials.

BIOMARKERS

Biomarkers are objective measures of biological processes, states, or conditions that play a central role in evaluating safety and efficacy throughout drug development.[8] Typically classified into two main groups – 'safety biomarkers' and 'response biomarkers'- they support clinical decision-making by providing early and reliable indicators of a drug's performance.[8,9] An ideal biomarker should be safe, easy to measure, and cost-effective, while also meeting key technical requirements such as sensitivity, specificity, reproducibility, repeatability, and cross-species translatability.

Pharmacodynamic biomarkers quantify the biological response to a therapeutic intervention, providing direct evidence of target engagement and effect on the disease pathway. By reflecting the mechanism of action of a drug, these markers facilitate dose selection and enable monitoring of treatment efficacy. Pharmacodynamic biomarkers can help researchers bridge the gap between preclinical findings and human studies, as they deliver measurable endpoints that validate the compound's intended action early in clinical development. This is called proof-of-concept.

Surrogate endpoints, derived from validated biomarkers, serve as proxy measures that can predict or correlate with clinically meaningful outcomes, potentially accelerating drug development and reducing costs by providing an earlier readout of treatment efficacy or disease progression.[9] However, these endpoints are only referred to as surrogate endpoints when drug registration authorities formally accept their use in place of COAs that reflect patient symptoms and quality of life. Such acceptance requires extensive evidence demonstrating a strong correlation between the surrogate endpoint and the traditional clinical endpoints, ensuring that it reliably reflects the ultimate clinical benefit or risk. As a consequence, surrogate endpoints are seldom approved and rarely used. Nevertheless, in instances where a surrogate endpoint is thoroughly validated and mechanistically linked to the disease pathway, it can facilitate smaller, more efficient proof-of-concept studies without compromising scientific rigor.

COAs in drug registration studies evaluate clinically meaningful benefits, such as improvements in symptoms, function, or quality of life. Because COAs often rely on clinician-based evaluations, subjective patient-reported outcomes, or a combination of the two, they inherently introduce variability in the data. As a result, larger sample sizes are typically required in (phase 3) registration trials to ensure sufficient statistical power to detect true treatment effects.

Biomarkers used in the above-mentioned bridging trials should present a more real life setting and/or have a high relevance to the real-world context. The term often used to describe this characteristic is ecological validity.

ECOLOGICAL VALIDITY

The term ecological validity is frequently used in the field of (neuro)psychology and often confused with external validity and mundane realism.[10,11] Whereas external validity concerns the generalizability of findings to various populations, settings, and points in time, mundane realism relates specifically to how closely an experimental setting mirrors everyday life. Ecological validity may incorporate elements of mundane realism to determine whether the study's variables and conclusions are truly relevant and applicable to real-world contexts. By contrast, external validity extends beyond ecological validity by examining how well a study's results can be applied to a different target population. These differences are often misunderstood and can complicate the interpretation of the literature on ecological validity.[11]

Within this thesis, ecological validity refers to the degree to which biomarkers used in earlier phase trials can be generalized to the clinical and demographic conditions of the COAs used in registration trials, encompassing both trial settings and study populations. Although the use of the term ecological validity for this phenomenon is not common, the term is used across studies – spanning aviation research,[12] mild cognitive impairment,[13] Parkinson's disease,[14,15] and treatments involving benzodiazepines[16] or medication for opioid use disorder[17]- and reflects the pursuit of more realistic approaches to testing pharmacological interventions. Proposed strategies for enhancing ecological validity include integrating mobile phone data,[18] employing more detailed gait analyses,[19] and utilizing virtual reality.[20,21] However, the ecological validity of biomarkers in the early stages of

drug development is rarely examined, limiting insights into biomarker potential and potentially affecting how results are interpreted.

RESEARCH OBJECTIVE AND STRUCTURE OF THIS THESIS

The aim of this thesis is to identify highly ecologically valid biomarkers for early phase clinical drug development and evaluate the ecological validity.

An example of translating clinical findings into measurements more relevant to real life is the study examining driving behaviour, described in **Chapter 2** of this thesis. This research focuses on biomarkers that assess the potential effects of pharmacological compounds on driving performance, as impaired driving carries significant safety risks. On-the-road driving tests, widely regarded as the gold standard with strong ecological validity, have traditionally been used to measure these effects.[22] However, these assessments are both time-consuming and expensive, as well as logistically complex to implement.

This study proposes using a driving simulator as an intermediate tool that bridges the gap between fundamental, laboratory-based evaluations of aspects of driving behaviour and the real-world on-the-road tests mandated by regulatory bodies such as the FDA.[23] While simulators inherently exhibit lower ecological validity than on-road driving tests, they offer higher ecological validity than psychomotor tests focusing on hand–eye coordination. Through the use of sleep deprivation to induce impaired driving performance, the study investigates how well results from these methods translate to one another. This tiered approach provides a structured pathway for assessing medication-related effects on driving before advancing to on-the-road trials.

Another bridging study between the clinical research unit and a real-life setting, described in **Chapter 3** of this thesis, uses a biomarker that evaluates a potential increase in fall risk. A commonly employed biomarker of postural stability in early-phase clinical trials is body sway, measured while individuals stand still with eyes closed. The total horizontal sway over a defined period is compiled into a single endpoint, and an increase in this metric generally signifies diminished stability, assumed to be associated with a greater likelihood of falling. However, actual fall risk is not directly measured, introducing uncertainty about potential safety concerns in later-phase drug development. An outcome measure that can be assessed in healthy volunteers but has higher ecological validity can be expected to have stronger predictive value for real-world fall risk in future trials.

In this chapter, we aimed to validate the previously described Interactive Walkway (IWW) for evaluating drug-induced effects. In the study described, we induced a mild fall risk by administering a benzodiazepine to healthy older adults and comparing its effects to a more selective hypnotic agent and placebo. The IWW offers higher ecological validity and greater specificity to fall risk compared to the body sway test. Additionally, the Timed Up and Go (TUG) test was included – a brief walking assessment in which participants stand from a chair, walk three meters, return, and sit down again. Finally, we assessed the ecological validity of each task within the context of the study findings.

Chapter 4 describes our efforts to develop a more ecologically valid method for quantifying muscle handgrip strength. Although grip strength is routinely used as an objective measure of muscular function in clinical trials and is often abnormal in patient populations, it does not adequately capture the functional outcomes relevant to patients' daily lives. An increase in grip strength alone is insufficient for regulatory approval; regulatory agencies require COAs, such as the 6-Minute Walk Test (6MWT), survival metrics, or timed chair-stand tests, which typically involve larger patient populations to detect clinically meaningful changes.

Emerging tools like PowerJar offer a better balance between quantifiability and ecological validity. Opening a jar, for instance, demands more than just hand grip and can also reflect the challenges posed by a unilateral impairment. PowerJar provides a more functional, contextually relevant measure of muscle strength than traditional grip strength, while offering higher resolution – and thus greater precision- than tests like the 6MWT. By more closely mirroring real-world tasks, PowerJar may serve as a more meaningful tool for assessing and demonstrating clinically relevant improvements in strength.

Chapters 5 and **6** of this thesis describe our initial efforts to establish a pain model that incorporates an affective component into a purely nociceptive task. Traditional pain assessments, which focus primarily on nociceptive processes,[24] may overlook drug effects that alter pain perception via emotional modulation. Anxiolytics, for instance, could be beneficial for pain management but are not typically evaluated with biomarkers that capture the emotional response component in healthy volunteers. By integrating an affective dimension into an evoked pain task, we enhance its ecological validity.

In many chronic pain conditions, the emotional response is believed to significantly influence the perception of pain.[25,26] To replicate this aspect in

healthy volunteers, we increased the emotional response to pain through a virtual reality simulation that displays skin damage at the site of the painful stimulus. This novel addition to the nociceptive testing battery was initially examined for repeatability and validated using patient-reported outcome measures. Subsequently, a first clinical trial was conducted with a drug known to suppress emotional responses, advancing our understanding of how the affective component interacts with pain modulation.

Finally, **Chapter 7** presents the findings from these studies and discusses the evaluation of the biomarkers. Current literature on frameworks for assessing ecological validity is reviewed, and a new structured framework is proposed. This framework enables a standardized approach to quantifying the ecological validity of both existing and novel biomarkers, thereby promoting more effective use of biomarkers in early-phase clinical drug development and enhancing overall efficiency in the drug development process.

REFERENCES

1 European Medicines Agency. ICH guideline E8 (R1) on general considerations for clinical studies [Internet]. 2021. Available from: www.ema.europa.eu/contact

2 Cohen AF, Burggraaf J, Van Gerven JMA, Moerland M, Groeneveld GJ. The use of biomarkers in human pharmacology (Phase I) studies. Vol. 55, Annual Review of Pharmacology and Toxicology. Annual Reviews Inc.; 2015. p. 55–74.

3 US Food and Drug Administration. Clinical outcome assessment (COA) compendium [Internet]. 2021. Available from: https: //www.fda.gov/regulatory-

4 Thomas D, Chancellor D, Micklus A, LaFever S, Hay M, Chaudhuri S, et al. Clinical Development Success Rates and Contributing Factors 2011–2020. 2021.

5 Sertkaya A, Beleche T, Jessup A, Sommers BD. Costs of Drug Development and Research and Development Intensity in the US, 2000-2018. JAMA Netw Open. 2024 Jun 28;7(6).

6 de Visser SJ, Cohen AF, Kenter MJH. Integrating scientific considerations into R&D project valuation. Nat Biotechnol. 2020 Jan 1;38(1): 14–8.

7 Torres-Saavedra PA, Winter KA. An Overview of Phase 2 Clinical Trial Designs. Int J Radiat Oncol Biol Phys. 2022 Jan 1;112(1): 22–9.

8 Califf RM. Biomarker definitions and their applications. Exp Biol Med. 2018 Feb 1;243(3): 213–21.

9 US Food and Drug Administration. BEST (Biomarkers, EndpointS, and other Tools) Resource. 2016.

10 Beechey T. Ecological Validity, External Validity, and Mundane Realism in Hearing Science. Ear Hear. 2022 Sep 1;43(5): 1395–401.

11 Hammond KR. Ecological Validity: Then and Now. 1998.

12 Ehlert AM, Wilson PB. Stimulant Use as a Fatigue Countermeasure in Aviation. Aerosp Med Hum Perform. 2021;92(3): 190–200.

13 Piau A, Wild K, Mattek N, Kaye J. Current state of digital biomarker technologies for real-life, home-based monitoring of cognitive function for mild cognitive impairment to mild Alzheimer disease and implications for clinical care: Systematic review. J Med Internet Res. 2019 Aug 1;21(8).

14 Foreman KB, Addison O, Kim HS, Dibble LE. Testing balance and fall risk in persons with Parkinson disease, an argument for ecologically valid testing. Parkinsonism Relat Disord. 2011 Mar;17(3): 166–71.

15 Ramsperger R, Meckler S, Heger T, van Uem J, Hucker S, Braatz U, et al. Continuous leg dyskinesia assessment in Parkinson's disease – clinical validity and ecological effect. Parkinsonism Relat Disord. 2016 May 1;26: 41–6.

16 Curran HV. Psychopharmacology Benzodiazepines, memory and mood: a review. Psychopharmacology (Berl). 1991;105–6.

17 Marsch LA, Chen CH, Adams SR, Asyyed A, Does MB, Hassanpour S, et al. The Feasibility and Utility of Harnessing Digital Health to Understand Clinical Trajectories in Medication Treatment for Opioid Use Disorder: D-TECT Study Design and Methodological Considerations. Front Psychiatry. 2022 Apr 29;13.

18 Asselbergs J, Ruwaard J, Ejdys M, Schrader N, Sijbrandij M, Riper H. Mobile phone-based unobtrusive ecological momentary assessment of day-to-day mood: An explorative study. J Med Internet Res. 2016 Mar 1;18(3).

19 Ilg W, Seemann J, Giese M, Traschütz A, Schöls L, Timmann D, et al. Real-life gait assessment in degenerative cerebellar ataxia: Toward ecologically valid biomarkers. Neurology. 2020 Sep 1;95(9): E1199–210.

20 Segawa T, Baudry T, Bourla A, Blanc JV, Peretti CS, Mouchabac S, et al. Virtual Reality (VR) in Assessment and Treatment of Addictive Disorders: A Systematic Review. Front Neurosci. 2020 Jan 10;13.

21 Höller Y, Höhn C, Schwimmbeck F, Plancher G, Trinka E. Effects of Antiepileptic Drug Tapering on Episodic Memory as Measured by Virtual Reality Tests. Front Neurol. 2020 Feb 20;11.

22 Verster J, Roth T. Standard operation procedures for conducting the on-the-road driving test, and measurement of the standard deviation of lateral position (SDLP). Int J Gen Med. 2011;359.

23 FDA-CDER. Evaluating drug effects on the ability to operate a motor vehicle – guidance for industry [internet]. 2017. Available from: https: //www.Fda.Gov/drugs/guidancecomplianceregulatoryinformation/guidances/default.Htm

24 Siebenga PS, van Amerongen G, Okkerse P, Denney WS, Dua P, Butt RP, et al. Reproducibility of a battery of human evoked pain models to detect pharmacological effects of analgesic drugs. Eur J Pain [Internet]. 2019/02/23. 2019; Available from: https: //onlinelibrary.wiley.com/doi/pdf/10.1002/ejp.1379

25 Lumley MA, Cohen JL, Borszcz GS, Cano A, Radcliffe AM, Porter LS, et al. Pain and emotion: A biopsychosocial review of recent research. J Clin Psychol. 2011;67(9): 942–68.

26 Ruscheweyh R, Nees F, Marziniak M, Evers S, Flor H, Knecht S. Pain Catastrophizing and Pain-related Emotions. Clin J Pain. 2011;27(7): 578–86.

CHAPTER 2

FIT FOR PURPOSE OF ON-THE-ROAD DRIVING AND SIMULATED DRIVING: A RANDOMISED CROSSOVER STUDY USING THE EFFECT OF SLEEP DEPRIVATION

PLoS One. 2023; 18(2): e0278300. doi: 10.1371/journal.pone.0278300

Ingrid Koopmans,[1,2*] Robert-Jan Doll,[1] Hein van der Wall,[1] Marieke de Kam[1], Geert Jan Groeneveld,[1,2] Adam Cohen,[1,2] and Rob Zuiker[1,2]

1 Centre for Human Drug Research, Leiden, The Netherlands.
2 Leiden University Medical Center, Leiden, The Netherlands.

ABSTRACT

Drivers should be aware of possible impairing effects of alcohol, medicinal substance, or fatigue on driving performance. Such effects are assessed in clinical trials, including a driving task or related psychomotor tasks. However, a choice between predicting tasks must be made. Here, we compare driving performance with on-the-road driving, simulator driving, and psychomotor tasks using the effect of sleep deprivation.

This two-way cross over study included 24 healthy men with a minimum driving experience of 3000km per year. Psychomotor tasks, simulated driving, and on-the-road driving were assessed in the morning and the afternoon after a well-rested night and in the morning after a sleep-deprived night. Driving behaviour was examined by calculating the Standard Deviation of Lateral Position (SDLP).

SDLP increased after sleep deprivation for simulated (10cm, 95%CI: 6.7–13.3) and on-the-road driving (2.8cm, 95%CI: 1.9–3.7). The psychomotor test battery detected effects of sleep deprivation in almost all tasks. Correlation between on-the-road tests and simulator SDLP after a well-rested night (0.63, $p < .001$) was not present after a night of sleep deprivation (0.31, $p = .18$). Regarding the effect of sleep deprivation on the psychomotor test battery, only adaptive tracking correlated with the SDLP of the driving simulator (-0.50, $p = .02$). Other significant correlations were related to subjective VAS scores.

The lack of apparent correlations and difference in sensitivity of performance of the psychomotor tasks, simulated driving and, on-the-road driving indicates that the tasks may not be interchangeable and may assess different aspects of driving behaviour.

INTRODUCTION

In the last decades, public and private organisations tried to improve automobile safety and decrease unsafe driving practices by addressing impaired driving.[1] Despite efforts, road trauma is still a significant public health issue.[2] A major cause of driving crashes and deaths is drowsy driving and/or driving under influence.[3,4] This has recently been confirmed in a systematic review and meta-analysis in which a strong association between sleepiness and car accidents was established.[5] Additionally, several research groups demonstrated that sleep deprivation impairs driving performance in simulated driving and on-the-road driving tests.

Driving performance must be captured in a reliable, repeatable, and sensitive manner to study the effect of interventions (e.g., sleep deprivation, medication, or distractions in the car). Limiting the environmental variables (e.g., interaction with other road users) and standardising road conditions (e.g., length, number of lanes, or speed limit) supports the creation of a standardised test. Many on-the-road driving studies are performed on a highway where subjects stay in a single lane and thus limit their interactions with other drivers. While this standardised procedure results in reliable and repeatable results, one could argue that this costly and time-consuming measurement can be easily replaced with a driving simulator.

Simulated driving is a widely used alternative for on-the-road driving. Besides cost-effectiveness, high levels of drugs and alcohol can be tested in a laboratory setting with medical assistance nearby. This makes the driving simulator an attractive and safer method to study healthcare interventions on driving behaviour. Moreover, in 2017, the United States Food and Drug Administration (FDA) started accepting driving simulator studies for the registration of (new) drugs in some conditions.[6] However, as there is a wide variety in the validity of simulators ranging from simple single-monitor desk setups to hydraulic based widescreen setups, it is known that the driving experience in a simulator can be considered unrealistic. The lack of consequences (e.g., after a car crash) might preclude a sense of fear and vigilance, resulting in a distorted representation of the drug effect. Even though simulators are not always found to have similar sensitivity to drug effects as the on-the-road task,[7,8] driving simulators can detect impaired driving performance induced by sleep deprivation and well-known drugs.[9-11]

Driving performance is often quantified by measuring a single measure describing the swaying/waving of the car and expressed with the standard

deviation of the lateral position (SDLP). This measure is sensitive to CNS modulators (e.g., sleep deprivation, alcohol, and drugs),[12-14] and is regarded as reflecting overall driving.[15] However, when testing skills in isolation, such as hand-eye coordination, concentration, and decision making, correlations for these individual tests with the SDLP are modest at the most.[9,13,16] Therefore, it remains intriguing to better understand the contribution of cognitive domains to driving performance. Integrating isolated skills with those derived from the driving task could provide detailed information on intervention effects. In fact, combining cognitive/motor performance, on-the-road driving, and simulated driving in a single study, is rarely explored. Here, we aim to compare the impact of sleep deprivation on driving performance using on-the-road driving, simulator driving, and psychomotor tasks in healthy subjects.

METHODS

This was a single-centre randomised, two-way cross-over study. The study was conducted at the clinical research unit of the Centre for Human Drug Research (CHDR) in Leiden, The Netherlands. The study was approved by the Medical Ethics Committee Stichting Beoordeling Ethiek Biomedisch Onderzoek (Assen, the Netherlands) and registered under NL68626.056.19. The study was conducted according to the Dutch Act on Medical Research Involving Human Subjects (WMO) and in compliance with all International Conference on Harmonisation Good Clinical Practice (ICH-GCP) guidelines.

Participants

All participants provided written informed consent before screening and study-related activities. This study was a part of a more extensive research on the effects of sleep deprivation> Next to the impact of sleep deprivation on driving, effects on pain thresholds were assessed. Because of the influence of the ovarian cycle on pain thresholds, initially, only men were invited to participate. For the latter assessment, only men between 23 and 35 years of age were invited to participate in the study. Only experienced drivers, defined as participants having a valid driving licence for at least five years and having, on average, a self-reported annual mileage of at least 3000 km, were included in the study. No history or presence of sleep disorders was allowed. Participants had to remain in the same time zone as the Netherlands at least seven days before the first visit and during the trial period. Additionally, participants showing signs of Simulator Sickness Syndrome before or during a screening of the simulated driving sessions were not eligible to participate. See Figure 1 for the CONSORT flow diagram.

Experiment design

During the screening period, all participants were trained on the study assessments by completing the full or a shortened version of each test. After inclusion, participants attended the clinic on two visits: a sleep-deprivation visit, and a well-rested visit. Participants were randomized on the order of visits, with at least 5 days to recover from the sleep deprivation if performed first (see Figure 2). Participants were not allowed to consume caffeine, alcohol or drugs during the clinical trial starting 4, 24 hours and 3 days prior to the first visit, respectively. To make sure the participants were well rested prior to the study, they were asked to maintain a normal sleep rhythm (at least 8 hours between 22:00 and 8:00) for two nights prior to the visits. For the sleep deprived visit, participants arrived in the late afternoon and could leave the day after. Participants were kept awake throughout the night after arrival. All assessments were performed in the morning after sleep deprivation. During the visit, participants were allowed only light physical activities (e.g., foosball). During the well-rested visit, participants were instructed to sleep at home for at least 8 hours before they were admitted to the clinical unit. All assessments were subsequently completed twice: once in the morning and once in the afternoon.

Assessments

The assessments contain simulated driving, on-the-road driving, the performance of a cognitive test battery (NeuroCart), and general questionnaires. The cognitive test battery consists of six tests: eye movement test (both smooth pursuit and saccadic), adaptive tracker, VAS Bond & Lader, Karolinska Sleepiness Scale and the body sway test. All measurements were performed in a quiet room with dimmed lightning. There was only one subject per session in the same room. The order of the assessments for each round of assessments is visualized in Figure 3.

On-the-road driving

For on-the-road driving, a car (Volkswagen caddy) was specifically modified with safety and measurement equipment (GRIDT). The location of the car was recorded using a GPS sensor mounted on the roof of the car. A Mobileye system (MobilEye Vision Technologies LTD., Israel) was used to

determine the relative position of the car on the road and log the speedometer (both sampled with a frequency of 13 Hz). This data was used to determine the Standard Deviation of Lateral Position (SDLP) of the driving session.[17] See Figure 4 for a visual impression of the SDLP. For safety reasons, a certified driving instructor sat on the passenger seat during all on-the-road assessments and had access to dual controls. Subjects were instructed to drive on a predefined section of a public road (N11, the Netherlands) and to maintain a steady speed of 95 km/h. Subjects were instructed to only overtake other vehicles when this was required for maintaining a steady speed. The chosen trajectory was a 40 km long two-lane highway starting at 15-minute drive from the clinical unit with a speed limit of 100 km/h. The road contained two sections including traffic lights with a speed limit of 70 km/h for 0.5 km, each. Subjects were to drive on this road in both directions turning the vehicle after 26 km. The subjects received instructions identical to the published SOP by Verster et al.[15] prior to the start of the drive and these instructions were repeated on request. A research assistant was present in the car to operate the data logging system.

Cleaning of the on-the-road data. Data outside the trajectory of interest was removed using the GPS prior to data analysis. Additionally, measurements outside the speed range of 85–110 km/h and during successful lane switches were excluded from the analysis data set. A successful lane switch is defined as the crossing of the white stripes in the road with the middle of the car. The start of a lane switch is a deviation of the middle lane of at least 100 cm. The end of a lane switch is defined as the moment when the car is within 100 cm of the middle of the new lane. To make sure pre-lane and post-lane switch behaviour (e.g., the intention to switch lanes) is also excluded from the data, three seconds before the start and after the end of a lane switch is included in the removal of data. Therefore, the data collected between two-lane switches (i.e., to the left lane and back) can be kept for analysis. The data removal is illustrated in Figure 5.

Simulated driving

The simulated driving test was performed on a fix-based driving simulator (Drivemaster, Green Dino B.V., the Netherlands).[9,18] Each experiment session lasted for 20 minutes of driving on a two-lane highway with traffic. Subjects were instructed to maintain a steady speed of 100 km/h on the outer lane. Overtaking was only allowed for maintaining a steady speed. The first five minutes were removed prior to data analysis.[19] A test drive of 15 minutes was performed on the same simulated highway trajectory during screening. Lane switches were removed following the same procedure as described above.

Driving questionnaires

After each (simulated) driving task a combined perceived driving effort and quality scale was used to record a self-assessment of the subjects driving performance.[15] The performance and motivation score is a VAS scale running from 1 (worst driving performance possible) to 15 (best driving performance possible). The perceived effort scale is a labelled VAS scale with 1 (no effort at all) and 15 (most effort possible). The driving instructor provided an opinion on subjects driving behaviour on the road with a scale of 1 to 10, with 10 representing perfect driving behaviour, on 11 aspects of driving: scanning, change of gear, steering, breaking, use of clutch, speed, rounding corners, anticipation on surroundings, applying traffic regulations, attention and reaction time. The total score was used for further analysis.

Eye movement measurements

Recording and analysis of saccadic eye movements is conducted with a microcomputer-based system that samples and analyses eye movements. The program for signal collection and the AD-converter is from Cambridge Electronic Design (CED LTD., Cambridge, UK), the signal amplification using Grass (Grass-Telefactor, An Astro-Med, INC. Product Group, Braintree, USA) and the sampling and analysis scripts are developed at the CHDR (Leiden, the Netherlands). Disposable silver-silver chloride electrodes (Ambu Blue Sensor N) will be applied on the forehead and beside the lateral canthi of both eyes of the subject for registration of the electro-oculographic signals. Skin resistance is reduced to less than 5 kOhm before measurements by scrubbing the skin and using electrolyte gel. Head movements are restrained using a fixed head support. The target consists of a moving dot that is displayed on a computer screen. This screen is fixed at 58 cm in front of the head support.

Saccadic eye movements are recorded for approximately 15 degrees to either side for stimulus amplitudes. Fifteen saccades are recorded with interstimulus intervals varying randomly between 3 and 6 seconds. Average values of latency (reaction time), saccadic peak velocity of all correct saccades and inaccuracy of all saccades will be used as parameters. Saccadic inaccuracy is calculated as the absolute value of the difference between

the stimulus angle and the corresponding saccade, expressed as a percentage of the stimulus angle. Saccadic peak velocity is one of the most sensitive parameters for sedation.[20,21] The use of a computer for measurement of saccadic eye movements was originally described by Baloh et al.,[22] and has been validated at CHDR by Van Steveninck et al..[20]

For smooth pursuit eye movements, the target moves at a frequency ranging from 0.3 to 1.1 Hz, by steps of 0.1 Hz. The amplitude of target displacement corresponds to 22.5 degrees eyeball rotation to both sides. Four cycles are recorded for each stimulus frequency. The time in which the eyes are in smooth pursuit of the target will be calculated for each frequency and expressed as a percentage of stimulus duration. The average percentage of smooth pursuit for all stimulus frequencies will be used as parameter. The method has been validated at CHDR by Van Steveninck et al.[23,24] based on the work of Bittencourt et al.[25] and the original description of Baloh et al..[26]

Adaptive tracker

The adaptive tracking test was performed using customised equipment and software (based on TrackerUSB hard-/software (Hobbs, 2004, Hertfordshire, UK)). This 3.5-minute period is including a run-in time of 0.5 minute, in this run-in time the data is not recorded. Adaptive tracking is a pursuit-tracking task. A circle moves randomly about a screen. The subject must try to keep a dot inside the moving circle by operating a joystick. If this effort is successful, the speed of the moving circle increases. Conversely, the velocity is reduced if the test subject cannot maintain the dot inside the circle. The average performance and the standard deviation of scores over 3.5 minutes will be used for analysis. The adaptive tracking test has proved sensitive for measurement of CNS effects of alcohol,[27] various pharmacological compounds,[28] and sleep deprivation.[29]

Body sway

The body sway meter allows measurement of body movements in a single plane, providing a measure of postural stability. Body sway is measured with a pot string meter (Celesco) based on the Wright ataxiameter.[30] With a string attached to the waist, all body movements are integrated and expressed as mm sway. Before starting a measurement, subjects were asked to stand still and comfortable, with their feet approximately 10 cm part and their hands in a relaxed position alongside the body and eyes closed. The total sway during two minutes is used as a parameter for body sway. The method has been used to demonstrate effects of sleep deprivation,[31] alcohol,[32] and several pharmacological compounds.[28]

VAS Bond & Lader

Visual analogue scales as originally described by Norris have often been used previously to quantify subjective effects of a variety of sedative agents.[33,34] Subjects indicate (with a mouse click on the computer screen) on sixteen horizontal visual analogue scales how he feels. From these measurements, three main factors are the calculated as described by Bond and Lader:[35] alertness (from nine scores), contentedness (often called mood; from five scores), and calmness (from two scores).

Karolinska Sleepiness Scale (KSS)

The Karolinska Sleepiness Scale (KSS)[36] measures the participant's state of sleepiness at a given moment in time. Participants were asked: 'Use the following scale to indicate how sleepy you are feeling at this moment. Write the number in the box.' Nine numerical response alternatives are listed vertically with verbal labels assigned to alternate numbers: 1. Extremely Alert; 2; 3 Alert; 4; 5 Neither Alert nor Sleepy; 6; 7 Sleepy But Not Fighting Sleep; 8; 9 Extremely Sleepy, Fighting Sleep, Effort to Stay Awake.

Statistical analysis

All statistical analyses were performed using SAS version 9.4 (SAS Institute, INC., Cary, NC, USA). A sample size calculation was performed using previous results of the simulator with a two-sided paired t-test.[9] A total sample size of n = 20 would be sufficient to determine significant differences in SDLP measured with the driving simulator of 2.5 cm with a significance level of 0.05 and a power of 0.80. Accounting for technical malfunctions and subjects dropping out, we aimed to include 24 subjects.

Each variable was analysed with a mixed model analysis of variance with fixed factor condition (separate for each set of assessments) and random factor subject. Simulator mean speed, simulator standard deviation speed, GRIDT mean speed, and body sway were log-transformed to correct for a log-normal distribution before statistical analysis.

The repeatability was quantified by the coefficient of variation (COV) within and between subjects, as estimated from the subject (between-subject) variability and residual (within-subject) variability of the mixed model analysis and the mean over the three conditions of the estimated least square

means. The common variance is the sum of the inter and intrasubject variability. For log-transformed variables the COV is calculated from the same estimated variabilities of the mixed model analyses which are back transformed by 100*evariability−1−−−−−−−−−−√ to a COV.

Pearson correlations were calculated for the SDLP and each variable in each condition (well rested and sleep deprived). In case of log normal distribution of a variable the log values of the variable are used.

RESULTS

Participants

A total of 25 participants were enrolled in the study from March to June 2019. Two subjects stopped participation during the night of sleep deprivation due to illness unrelated to the sleep deprivation. Because one of these subjects had sleep deprivation as his first visit, the subsequent well-rested visit was also not performed. Twenty-four subjects are included for statistical analysis (age mean (SD) is 25.7 (1.6) years, BMI is 24.3 (3.4) kg/m^2). Data could not be collected during 3 GRIDT assessments (two during the sleep-deprivation visit and one afternoon session during the well-rested visit) due to technical difficulties.

Repeatability of driving parameters

GRIDT

The repeatability of the GRIDT driving parameters (i.e., SDLP, mean speed, SD-speed) during the well-rested visit (morning and afternoon) is presented in Table 1. Results of driving parameters are presented in Table 2. The mean (SD) SDLP for the GRIDT was 21.33 cm (2.3) and 22.26 cm (2.4) during the morning and afternoon, respectively. The coefficient of variation (COV) was 8.9% and 6.5% for the inter-and intra-subject variability, respectively. The common variance was 6.13 (COV: 11.0%).

DRIVING SIMULATOR

The simulator driving parameters are presented in Table 2. The SDLP(SD) is 30.14 cm (5.2) and 32.16 cm (5.5) during the well-rested visit for morning and afternoon driving assessments, respectively. The COV for the simulator was 11.8% and 16.5% for inter and intra subject variability, respectively (Table 1). The common variance of the SDLP measured in the simulator was 47.99 (COV: 20.3%). The Bland-Altman plot (Figure 6) shows the bias and limits of agreement for the simulator (bias: -2.01 and 95% limits of agreement: -11.10 and 7.09) as well as for the GRIDT (bias: -0.860 and 95% limits of agreement: -4.33 and 2.61).

Sensitivity of driving parameters

GRIDT

The GRIDT SDLP was significantly increased (2.76 cm, $p < .001$) after a night of sleep deprivation compared to the morning assessments after a well-rested night. While the mean speed was reduced after sleep deprivation, the SD speed was not (Table 2). Subjects report lower driving performance and motivation and increased driving effort after sleep deprivation compared to the well-rested morning. In line with the results of these self-reported questionnaires, the instructor assessment also indicated lower driving performance scores after sleep deprivation.

DRIVING SIMULATOR

The SDLP measured by the simulator was significantly increased (9.97 cm, $p < .001$) after a night of sleep deprivation compared to the morning assessment after a well-rested night. The SD-speed for the simulator increased significantly after a night of sleep deprivation compared to the assessments on the well-rested morning, while the mean speed was not. Like the GRIDT, subjects rate a decrement in their driving performance and motivation, and an increase in effort scores after a night of sleep deprivation compared to the well-rested morning.

Correlation between simulator and psychomotor test battery

The mean and 95% CI results of the cognitive tests are presented in Table 3. Except for the smooth eye pursuit ($p = .34$), all tests performed using the NeuroCart® were significantly affected by sleep deprivation. The Pearson correlations are calculated between the parameters in each condition (i.e., well-rested and sleep-deprived) and the SDLP of both the simulator and the GRIDT in that same condition (Table 4). The correlations are included in the overview when at least one of the driving assessments has a significant correlation ($p < .05$). For these cases the correlation with the other driving assessment (i.e., GRIDT or diving simulator) is shown, except for the test related subjective scores. The highest correlation (-0.73, $p < .001$) was between the driving simulator SDLP and corresponding subjective driving performance and motivation score. The smooth eye pursuit significantly correlated with the simulator SDLP (0.49 with $p = .01$), but not with the GRIDT SDLP ($p = .41$).

The correlation between the driving assessment in the simulator and with the GRIDT was significant for the well-rested morning (0.63 with $p < .001$) and the well-rested afternoon (0.58 with $p = .003$). Although not presented in Table 4, the correlation was not significant for the SDLP after a night of sleep deprivation (0.31 with $p = .18$).

The scatter plot in Figure 7 shows a linear correlation between the SDLP measured by the GRIDT and the simulator for each set of assessments. The correlation between both well-rested assessments was rather similar, the trendline flattens for the sleep deprived measurements.

Additional Pearson correlations were calculated for the difference in the well-rested morning measurement, and the sleep deprived measurement (see Table 5). Like Table 4, only significant correlations were shown. Four correlations were ±0.50 or stronger, but most correlations were not significant ($p > .05$). The adaptive tracker, which is the only NeuroCart® test included in the table, had a significant correlation with the simulator of -0.50 with $p = .02$, but not with the GRIDT (correlation of -0.13 with $p = .58$).

DISCUSSION

To obtain a more detailed overview of the effect of healthcare interventions on driving performance, it is important to assess driving performance and assess both cognitive and motor performance. Here, we present the results of a study where the driving performance (both on the road and in a simulator) and the performance on a variety of psychomotor tasks were affected by sleep deprivation. Additionally, we compared the correlations between the different tasks.

Repeatability of on-the-road and simulated driving

Instead of a camera mounted on the left backside on the car's roof as often used in on-the-road driving test,[15] we installed a camera system (Mobileye) behind the front window to capture the SDLP during the on-the-road task. Using this method, we observe slightly higher SDLP values under well-rested conditions than reported in other studies.[16] This difference might be explained by the different position of the camera systems and differences in the processing of raw video to the estimated lateral position. However, we observe similar SDLP values and variability in the SDLP values compared to another clinical trial using the Mobileye.[37] Additionally, the Mobileye was successfully used in other driving studies.[38,39] Therefore, we conclude that the on-the-road values presented here are reliable.

The SDLP, mean speed, and the SD speed were repeatable during on-the-road and simulated driving (see Table 1). While the repeatability of the mean speed was similar for both simulated and on-the-road driving, the repeatability of the SDLP and SD speed were better during on-the-road driving than during simulated driving. An explanation could be that environmental conditions were constrained in the simulator (e.g., road conditions, other cars) and identical for all sessions. These constraints were not present during the on-the-road task, in which participants' constant vigilance is required to anticipate unpredictable events. This could indicate that the driving simulator has a faster habituation effect. As suggested by Helland et al., subjects have a lower sense of danger and gravitational cues when driving in a driving simulator which are normally used to adjust steering.[40] This might explain the higher SDLP values (Table 2) and the higher variability of the SDLP of the driving simulator (Table 1). Nonetheless, we conclude that on-the-road driving performance can be assessed with better repeatability than simulated driving.

Effect of sleep deprivation

Sleep deprivation affected all CNS task outcomes, except for the smooth eye movements, mean speed in the driving simulator, and the SD of the speed in the on-the-road car (Tables 2 and 3). This is in line with what was reported previously by others on simulated driving,[41] on-the-road driving,[16] and on the psychomotor test battery.[29] The psychomotor test results reported here are similar to those reported for medication-induced drowsiness.[42,43] The on-the-road task showed a smaller effect of sleep deprivation on the SDLP than the driving simulator (13% vs 33%). However, the minimal detectable effect size (MDES) of the SDLP in the on-the-road task is lower compared to the driving simulator (1.6 vs 6.0), which can be explained by the lower variability of the SDLP in the on-the-road driving task than in the driving simulator task (Table 2). The difference in sensitivity to the effects of sleep deprivation on simulator compared to on-the-road driving indicates that the tasks are not fully interchangeable and may assess different aspects of driving behaviour.

The simulator and on-the-road SDLP values were significantly correlated during both the well-rested morning and afternoon sessions (Table 4 and Figure 7). The slope value of the linear regression lines for the well-rested morning and afternoon sessions was gradual for both sessions. Interestingly, the simulator and on-the-road SDLP values were no longer

correlated during the sleep deprivation condition, while the correlation between the effect of sleep deprivation on the simulator and on-the-road SDLP values was significant (Table 5). Combined with the earlier discussed MDES, this could indicate that the simulator task is more sensitive to the effect of sleep deprivation or, that the driving tasks do not measure the same change in driving behaviour. Another explanation could be a test order effect because of the fixed order of tests in the study design. Even though both the simulator and the on-the-road task can be used to detect sleep deprivation induced changes, the lack of (strong) correlations between both tasks during sleep deprivation could indicate that simulated driving and on-the-road driving are affected by sleep deprivation differently.

Correlations between driving and psychomotor tasks

This study aimed to compare the effect of sleep deprivation on the driving tasks and the psychomotor test battery. High correlations between the tasks indicate a higher level of validity for the psychomotor test battery. Additionally, these correlations provide information on how sleep deprivation impairs driving performance. When assessing each set of measurements separately, none of the tasks were significantly correlated with the SDLP of the driving simulator, except for the smooth eye pursuit task at well-rested morning (Table 4). Interestingly, the smooth eye pursuit is the only task that was unable to detect the effect of sleep deprivation. The absence of a significant correlation between postural balance and the simulated driving SDLP confirms the findings of Jongen et al. in 2015.[16] In a study by Huizinga et al.[9] assessing the effect of alcohol and alprazolam using the same driving simulator, tracker task, and eye movement tasks, there was a significant correlation using linear regression between SDLP and the psychomotor tasks. The lack of (strong) correlations between the psychomotor and driving tasks in this study indicates that care must be taken when relating psychomotor performance to driving ability.

For the effect of sleep deprivation, only a few of the correlations between the driving simulator, on-the-road driving, and psychomotor tasks were significant. Of the psychomotor tasks, the adaptive tracker showed a significant correlation with the driving simulator, but not for the on-the-road driving, for the effect of sleep deprivation (Table 5). Park et al.[44] suggested that the driving simulator task might measure different effects because of the long monotonous task compared to the short psychomotor tasks. Another reason for the insignificant correlation between the psychomotor and on-the-road tasks is that the on-the-road steering wheel and other technical properties of the car induced more heavy steering and a default position to steer straight ahead, which might help maintain a straight course, thereby effectively reducing the SDLP. These technical properties are not present in the simulator and tracker. Another explanation could be a possibly heightened sense of danger during the on-the-road assessment due to the inability of a serious crash in the simulator. This cannot be confirmed since no assessments of fear, stress or stress hormone levels were performed during the day. Any of those measurements should be included when assessing possible test-effects. The difference between the assessments is the possible anxiety for a crash and the external visual/audiological stimuli during the on-the-road driving task (such as weather and special vehicles). Even though no radio or conversations were allowed during on-the-road driving, the surroundings were less repetitive and stable than the psychomotor test battery and driving simulator. All other significant correlations for the sleep deprivation effect and the driving simulator are found with the subjective assessments of driving (performance and effort) and the VAS Bond-Lader (Table 5). However, it should be noted that the participants in this study were not blinded, which might have influenced this study's subjective measures.

The test duration for both driving tasks was around 30 minutes. This made it possible to compare both tasks, but it does deviate from the standard length of the on-the-road tasks, which is 100 km.[15] Increasing the length of the trajectory, and therefore the duration of the task, might show a more prominent effect of sleep deprivation on the SDLP. The data for this study has not been analysed in the same way as done by Verster et al..[15] A different cut-off for speed is used and removing of lane switches. The absolute increase of the SDLP found in this study must therefore be compared to other on-the-road driving studies with care.

Limitations

Several limitations should be noted when interpreting the results of this study. First, the study design did not include a habituation night for both the well-rested and sleep deprivation visits. Lifestyle restrictions demanded that each participant have a stable sleep pattern with a bedtime between 22:00 and 23:00 hours and awakening between 7:00 and 8:00 hours. Although all participants confirmed adherence at the start of each visit, it cannot be ruled out that a few participants had not followed up on these restraints. While the effect of sleep deprivation was detected for most measures, the result of the study could have been optimised by adding a habituation night.[45]

Secondly, differences in the interval between the two study visits between subjects may have influenced study outcomes. A 5-day interval was applied between the sleep deprivation visit and the subsequent well-rested visit for one group of participants. In contrast, the other group performed the well-rested and sleep-deprived visits contiguously. This may have led to a difference in familiarisation or learning effect between the two groups. If we had introduced a 5-day interval between the well-rested and the sleep deprivation periods, this potential bias could have been prevented. However, all subjects were trained during the screening period on all study procedures. Therefore, the learning and familiarisation effect during the actual study periods is expected to be small.

This study was performed in young, healthy men who were considered experienced drivers. This population was chosen for pragmatic reasons as a second part of the study included determining the effects of sleep deprivation on evoked pain tests, which had to be initially restricted to a male population.[46,47] Results may have been different if elderly or female drivers had also been included.[14,48] Additionally, less experienced drivers can overestimate their driving performance[49,50] which might influence the results of the driving tasks and the correlations between driving variables and the psychomotor test battery. In conclusion, the selection of the study populations may limit the generalisability of the study results.

The current selection of psychomotor tasks does not cover all cognitive domains which might be influenced by sleep deprivation. Other tests often used in sleep deprivation studies, such as the PVT-192,[51] might have better correlations between the isolated testing of psychomotor functioning and the SDLP measured with either the simulator or the on-the-road car.

CONCLUSION

In general, this study demonstrates that the psychomotor test battery, driving simulator and on-the-road driving tasks, are sensitive to sleep deprivation. However, the lack of apparent correlations between test variables under well-rested and sleep-deprived conditions indicates that each task in this study measures driving impairment differently. This study supports the need for studies in early drug development, including driving tasks as close to real-life driving as possible and indicates the complexity of (impaired) driving behaviour.

ACKNOWLEDGEMENTS The authors would like to thank Floortje van Dixhoorn (intern) and all research assistants for their help with this study.

REFERENCES

1 Higgins J. S. et al. , 'Asleep at the Wheel-The Road to Addressing Drowsy Driving,' Sleep, vol. 40, no. 2, 2017, doi: 10.1093/sleep/zsx001

2 CBS, 'Centraal Bureau voor de Statistiek—overledenen: doden door verkeersongeval in Nederland.,' 2017.

3 SWOV, 'SWOV factsheet rijden onder invloed.'

4 A. Foundation, 'Prevalence drowsy driving crashes estimates large scale naturalistic driving study.'

5 Bioulac S. et al. , 'Risk of motor vehicle accidents related to sleepiness at the wheel: A systematic review and meta-analysis,' Sleep, vol. 40, no. 10, 2017, doi: 10.1093/sleep/zsx134

6 DA—CDER, 'Evaluating Drug Effects on the Ability to Operate a Motor Vehicle—Guidance for Industry,' 2017. [Online]. https: //www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/default.htm

7 Helland A. et al. , 'Comparison of driving simulator performance with real driving after alcohol intake: A randomised, single blind, placebo-controlled, cross-over trial,' Accid Anal Prev, vol. 53, pp. 9–16, 2013, doi: 10.1016/j.aap.2012.12.042

8 Veldstra J. L., Bosker W. M., de Waard D., Ramaekers J. G., and Brookhuis K. A., 'Comparing treatment effects of oral THC on simulated and on-the-road driving performance: Testing the validity of driving simulator drug research,' Psychopharmacology (Berl), vol. 232, no. 16, pp. 2911–2919, Aug. 2015, doi: 10.1007/s00213-015-3927-9

9 Huizinga C. R. et al. , 'Evaluation of simulated driving in comparison to laboratory-based tests to assess the pharmacodynamics of alprazolam and alcohol,' J Psychopharmacol, p. 269881119836198, 2019, doi: 10.1177/0269881119836198

10 Simen A. A. et al. , 'A randomized, crossover, placebo-controlled clinical trial to assess the sensitivity of the CRCDS Mini-Sim to the next-day residual effects of zopiclone,' Ther Adv Drug Saf, vol. 6, no. 3, pp. 86–97, 2015, doi: 10.1177/2042098615579314

11 Mets M. A. J., Kuipers E., de Senerpont Domis L. M., Leenders M., Olivier B., and Verster J. C., 'Effects of alcohol on highway driving in the STISIM driving simulator,' Hum Psychopharmacol, vol. 26, no. 6, pp. 434–439, Aug. 2011, doi: 10.1002/hup.1226

12 Verster J. C., Taillard J., Sagaspe P., Olivier B., and Philip P., 'Prolonged nocturnal driving can be as dangerous as severe alcohol-impaired driving,' J Sleep Res, vol. 20, no. 4, pp. 585–588, Dec. 2011, doi: 10.1111/j.1365-2869.2010.00901.x

13 J. C. Verster, D. W. Spence, A. Shahid, S. R. Pandi-Perumal, and T. Roth, 'Zopiclone as Positive Control in Studies Examining the Residual Effects of Hypnotic Drugs on Driving Ability,' 2011.

14 Jongen S. et al. , 'A pooled analysis of on-the-road highway driving studies in actual traffic measuring standard deviation of lateral position (i.e., 'weaving') while driving at a blood alcohol concentration of 0.5 g/L,' Psychopharmacology (Berl), vol. 234, no. 5, pp. 837–844, Mar. 2017, doi: 10.1007/s00213-016-4519-z

15 Verster J. and Roth T., 'Standard operation procedures for conducting the on-the-road driving test, and measurement of the standard deviation of lateral position (SDLP),' Int J Gen Med, p. 359, 2011, doi: 10.2147/IJGM.S19639

16 Jongen S., Perrier J., Vuurman E. F., Ramaekers J. G., and Vermeeren A., 'Sensitivity and validity of psychometric tests for assessing driving impairment: Effects of sleep deprivation,' PLoS One, vol. 10, no. 2, Feb. 2015, doi: 10.1371/journal.pone.0117045

17 Verster J. and Roth T., 'Standard operation procedures for conducting the on-the-road driving test, and measurement of the standard deviation of lateral position (SDLP),' Int J Gen Med, p. 359, 2011, doi: 10.2147/IJGM.S19639

18 Jacobs M., Hart E. P., Miranda Y. M., Groeneveld G. J., van Gerven J. M. A., and Roos R. A. C., 'Altered driving performance of symptomatic Huntington's disease gene carriers in simulated road conditions,' Traffic Inj Prev, vol. 19, no. 7, pp. 708–714, 2018. doi: 10.1080/15389588.2018.1497796

19 Huizinga C. R. et al. , 'Evaluation of simulated driving in comparison to laboratory-based tests to assess the pharmacodynamics of alprazolam and alcohol,' J Psychopharmacol, p. 269881119836198, 2019, doi: 10.1177/0269881119836198

20 AL. van Steveninck, 'Methods of assessment of central nervous system effects of drugs in man.,' 1993.

21 Van Steveninck A. L., Schoemaker H. C., Pieters M. S. M., Kroon R., Breimer D. D., and Cohen A. F., 'A comparison of the sensitivities of adaptive tracking, eye movement analysis, and visual analog lines to the effects of incremental doses of temazepam in healthy volunteers,' Clin Pharmacol Ther, vol. 50, no. 2, pp. 172–180, 1991, doi: 10.1038/clpt.1991.122

22 Baloh R., Sills A., Kumley W., and Honrubia V., 'Quantitative measure¬ment of saccade amplitude, duration and velocity.,' Neurology, vol. 25, pp. 1065–1070, 1975.

23 Van Steveninck A., Cohen A., and Ward T., 'A microcomputer based system for recording and

analysis of smooth pursuit and saccadic eye movements.,' Brit.J.Clin.Pharmaco., vol. 27, no. 5, pp. 712–713, 1989.
24 AL. Van Steveninck, 'Methods of assessment of central nervous system effects of drugs in man.,' State University Leiden, 1993.
25 Bittencourt P., Wade P., Smith A., and Richens A., 'Benzodiazepines impair smooth pursuit eye movements.,' Br J Clin Pharmacol, 1983, doi: 10.1111/j.1365-2125.1983.tb01495.x
26 Baloh R., Sills A., Kumley W., and Honrubia V., 'Quantitative measurement of saccade amplitude, duration and velocity.,' Neurology, vol. 25, pp. 1065–1070, 1975.
27 van Steveninck A. L. et al. , 'Pharmacodynamic interactions of diazepam and intravenous alcohol at pseudo steady state,' Psychopharmacology (Berl), 1993, doi: 10.1007/BF02244655
28 Groeneveld G. J., Hay J. L., and Van Gerven J. M., 'Measuring blood–brain barrier penetration using the NeuroCart, a CNS test battery,' Drug Discovery Today: Technologies, vol. 20. Elsevier Ltd, pp. 27–34, Jun. 01, 2016. doi: 10.1016/j.ddtec.2016.07.004
29 van Steveninck A. L., van Berckel B. N. M., Schoemaker R. C., Breimer D. D., van Gerven J. M. A., and Cohen A. F., 'The sensitivity of pharmacodynamic tests for the central nervous system effects of drugs on the effects of sleep deprivation,' Journal of Psychopharmacology, vol. 13, no. 1, pp. 10–17, 1999, doi: 10.1177/026988119901300102
30 Wright BM., 'A simple mechanical ataxia-meter.,' J Physiol, vol. 218, pp. 27P–28P, 1971.
31 Van Steveninck A. L., Van Berckel B. N. M., Schoemaker R. C., Breimer D. D., Van Gerven J. M. A., and Cohen A. F., 'The sensitivity of pharmacodynamic tests for the central nervous system effects of drugs on the effects of sleep deprivation,' Journal of Psychopharmacology, 1999, doi: 10.1177/026988119901300102
32 van Steveninck A. L. et al. , 'Pharmacodynamic interactions of diazepam and intravenous alcohol at pseudo steady state,' Psychopharmacology (Berl), vol. 110, no. 4, pp. 471–478, 1993, doi: 10.1007/BF02244655
33 Norris H., 'The action of sedatives on brain stem oculomotor systems in man,' Neuropharmacology, 1971, doi: 10.1016/0028-3908(71)90039-6
34 de Visser S. J., van der Post J., Pieters M. S. M., Cohen A. F., and van Gerven J. M. A., 'Biomarkers for the effects of antipsychotic drugs in healthy volunteers,' Br J Clin Pharmacol, 2001, doi: 10.1111/j.1365-2125.2001.01308.x
35 Bond M., Lader A., 'The use of analogue scales in rating subjective feelings,' Br J Med Psychol, vol. 47, pp. 211–218, 1974.
36 Åkerstedt T. and Gillberg M., 'Subjective and objective sleepiness in the active individual,' International Journal of Neuroscience, vol. 52, no. 1–2, pp. 29–37, 1990, doi: 10.3109/00207459008994241
37 Anund A., Fors C., Hallvig D., Åkerstedt T., and Kecklund G., 'Observer Rated Sleepiness and Real Road Driving: An Explorative Study,' PLoS One, vol. 8, no. 5, May 2013, doi: 10.1371/journal.pone.0064782
38 Sun Y., Wu C., Zhang H., Zhang Y., Li S., and Feng H., 'Extraction of Optimal Measurements for Drowsy Driving Detection considering Driver Fingerprinting Differences,' J Adv Transp, vol. 2021, 2021, doi: 10.1155/2021/5546127
39 Zhang H., Wu C., Huang Z., Yan X., and Qiu T. Z., 'Sensitivity of lane position and steering angle measurements to driver fatigue,' Transp Res Rec, vol. 2585, pp. 67–76, 2016, doi: 10.314½585-08
40 Helland A. et al. , 'Comparison of driving simulator performance with real driving after alcohol intake: A randomised, single blind, placebo-controlled, cross-over trial,' Accid Anal Prev, vol. 53, pp. 9–16, 2013, doi: 10.1016/j.aap.2012.12.042
41 Garner A. A., Miller M. M., Field J., Noe O., Smith Z., and Beebe D. W., 'Impact of experimentally manipulated sleep on adolescent simulated driving,' Sleep Med, vol. 16, no. 6, pp. 796–799, Jun. 2015, doi: 10.1016/j.sleep.2015.03.003
42 Schrier L. et al. , 'Pharmacokinetics and pharmacodynamics of a new highly concentrated intranasal midazolam formulation for conscious sedation,' Br J Clin Pharmacol, vol. 83, no. 4, pp. 721–731, 2017, doi: 10.1111/bcp.13163
43 Groeneveld G. J., Hay J. L., and van Gerven J. M., 'Measuring blood–brain barrier penetration using the NeuroCart, a CNS test battery,' Drug Discovery Today: Technologies, vol. 20. Elsevier Ltd, pp. 27–34, Jun. 01, 2016. doi: 10.1016/j.ddtec.2016.07.004
44 D. Park, J. C. Ware, J. F. May, T. J. Rosenthal, M. R. Guibert, and R. W. Allen, 'The Effects of Sleep Deprivation on Simulator Driving as Compared with Other Psychomotor Tests George,' 2007, pp. 257–264.
45 P. Alhola and P. Polo-Kantola, 'Sleep deprivation: Impact on cognitive performance,' 2007.
46 H. J. Hijma, I. W. Koopmans, E. Klaassen, R. J. Doll, R. G. J. A. Zuiker, and G. J. Groeneveld, 'A randomized, crossover study to evaluate the gender and sensitizing effects of sleep deprivation using a nociceptive test battery in healthy volunteers,' submitted.
47 van den Berg B. et al. , 'Simultaneous measurement of intra-epidermal electric detection thresholds and evoked potentials for observation of nociceptive processing following sleep deprivation,' Exp Brain Res, Jan. 2022, doi: 10.1007/s00221-021-06284-5
48 Michaels J. et al. , 'Driving simulator scenarios and measures to faithfully evaluate risky driving behavior: A comparative study of different driver age groups,' PLoS One, vol. 12, no. 10, Oct. 2017, doi: 10.1371/journal.pone.0185909
49 de Craen S., Twisk D. A. M., Hagenzieker M. P., Elffers H., and Brookhuis K. A., 'Do young novice drivers overestimate their driving skills more than experienced drivers? Different methods lead to different conclusions,' Accid Anal Prev, vol. 43, no. 5, pp. 1660–1665, Sep. 2011, doi: 10.1016/j.aap.2011.03.024
50 Rosenbloom T., Beigel A., Perlman A., and Eldror E., 'Parental and offspring assessment of driving capability under the influence of drugs or alcohol: Gender and inter-generational differences,' Accid Anal Prev, vol. 42, no. 6, pp. 2125–2131, 2010, doi: 10.1016/j.aap.2010.07.002
51 Lim J. and Dinges D. F., 'Sleep deprivation and vigilant attention,' in Annals of the New York Academy of Sciences, 2008, vol. 1129, pp. 305–322. doi: 10.1196/annals.1417.002

FIGURE 1 CONSORT flow chart of screening, participation and analysis.

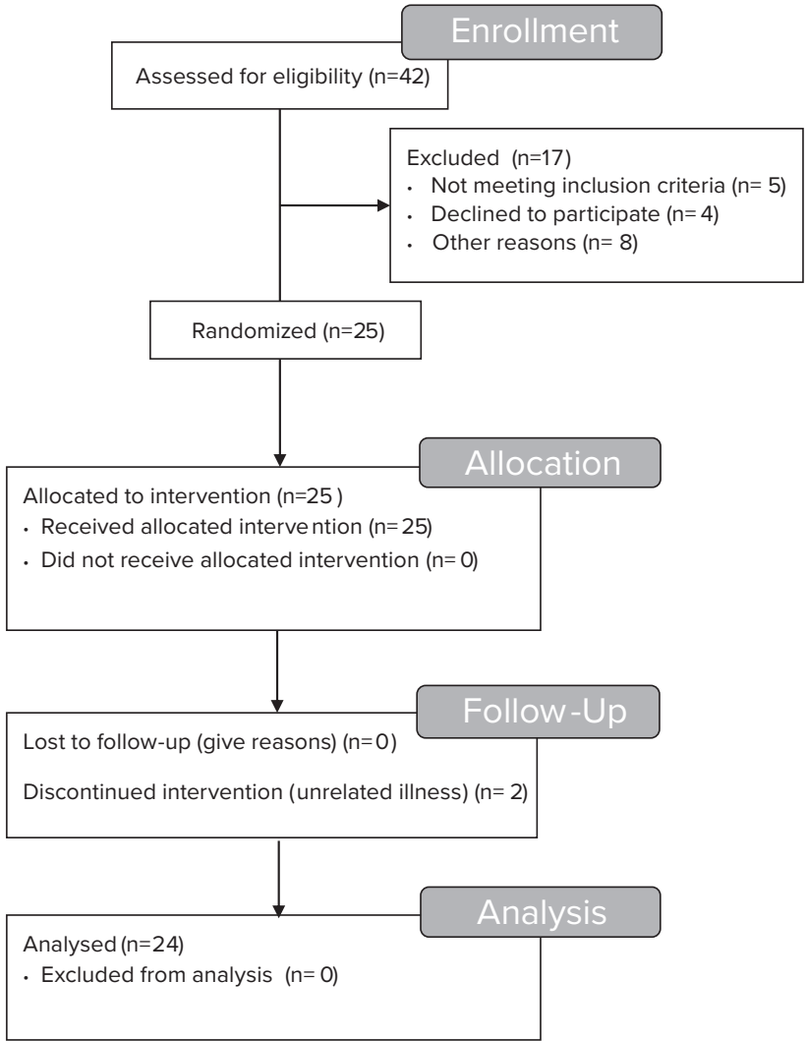


FIGURE 2 Schematic overview of the study design.

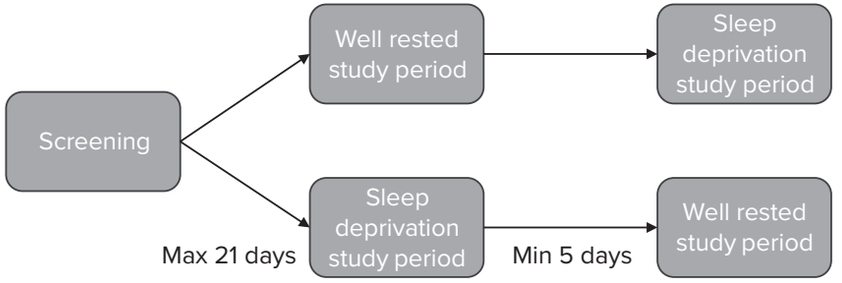


FIGURE 3 Schematic overview of the order of tests.

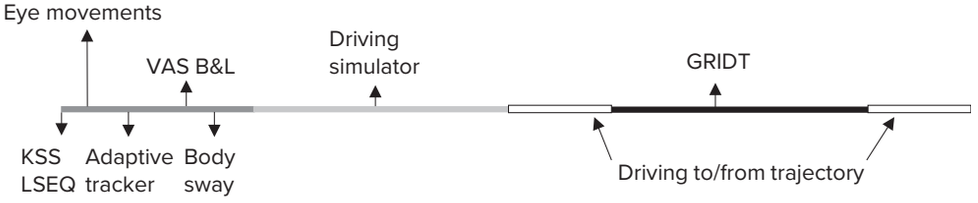


FIGURE 4 Visual impression of the car weaving within a lane. The Standard Deviation of Lateral Position (SDLP) is calculated as the standard deviation of the sway around the average position within the lane.

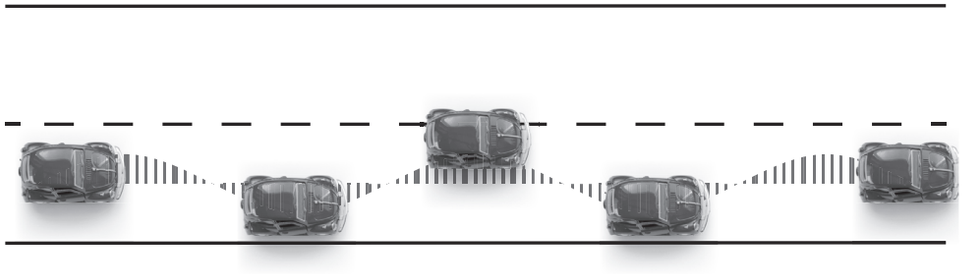


FIGURE 5 Schematic overview of data removal during a lane switch. All data between three seconds before until three seconds after a lane switch is removed prior to analysis.

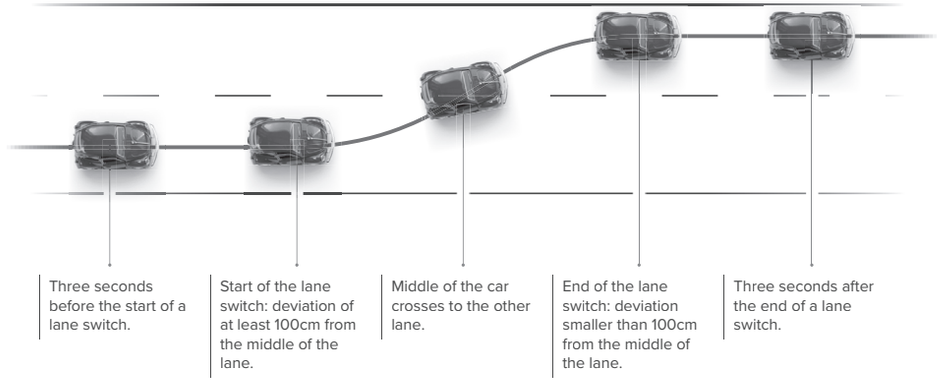


FIGURE 6 Combined Bland-Altman plot for simulator (bias: -2.01 and 95% limits of agreement: -11.10 and 7.09) and GRIDT (bias: -0.860 and 95% limits of agreement: -4.33 and 2.61).

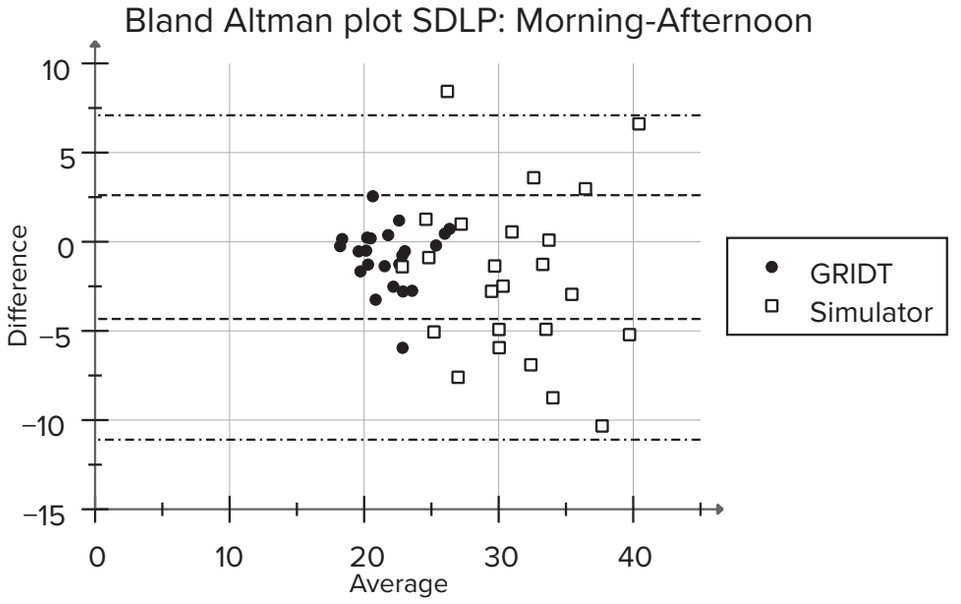


FIGURE 7 Scatter plot of the SDLP measured by the simulator and the GRIDT for the well-rested morning (WRM), well-rested afternoon (WRA) and the sleep deprived (SD) assessments.

TABLE 1 Inter-subject, intra-subject, common variance, and minimal detectable effect size (MDES), calculated at the well-rested visit.

Variable		Inter-subject Variance (COV)	Intra-subject Variance (COV)	Common Variance (COV)	MDES N=16 cross-over
SDLP (cm)	Simulator	16.32 (11.8%)	31.67 (16.5%)	47.99 (20.3%)	6.0
	GRIDT	3.99 (8.9%)	2.14 (6.5%)	6.13 (11.0%)	1.6
Mean speed[1] (km/h)	Simulator	1.2%	1.4%	1.9%	1.5%
	GRIDT	1.2%	1.0%	1.6%	1.1%
SD-speed[1] (km/h)	Simulator	25.6%	21.3%	33.7%	25.0%
	GRIDT	14.7%	9.9%	17.8%	11.0%

COV: Coefficient of Variation.
1. Geometric mean based on logarithmic transformed data

TABLE 2 Results of driving-related parameters for both the simulator and GRIDT.

Parameter	SDLP [cm]		Mean Speed (km/h)		SD Speed (km/h)		Performance and motivation (cm)		Driving effort (cm)		Instructor assessment
	SIMU-LATOR	GRIDT	SIMU-LATOR[1]	GRIDT[1]	SIMU-LATOR[1]	GRIDT[1]	SIMU-LATOR	GRIDT	SIMU-LATOR	GRIDT	GRIDT
Well rested morning (SD)	30.14 (5.2)	21.33 (2.3)	96.96 (1.6)	95.72 (1.4)	2.57 (0.9)	3.55 (0.7)	8.4 (1.8)	9.1 (2.1)	3.4 (2.1)	3.3 (1.6)	67.0 (5.5)
Well rested afternoon (SD)	32.16 (5.5)	22.26 (2.4)	96.92 (1.9)	95.64 (1.4)	2.65 (1.0)	3.59 (0.7)	7.9 (2.2)	8.4 (2.4)	4.6 (2.9)	3.4 (2.2)	66.2 (6.3)
Sleep deprived (SD)	40.26 (9.4)	24.08 (2.7)	97.33 (1.9)	94.99 (1.7)	3.36 (1.1)	3.73 (0.5)	4.6 (2.6)	5.6 (2.4)	10.3 (3.4)	9.2 (4.4)	61.1 (6.6)
Contrasts[2] [95% CI]	9.97 (6.65 – 13.29)	2.76 (1.87, 3.66)	0.4% (-0.4%, 1.3%)	-0.7% (-1.3%, -0.1%)	28.7% (13.7%, 45.7%)	5.0% (-1.1%, 11.6%)	-3.8 (-5.0, -2.6)	-3.5 (-4.8, -2.2)	6.9 (5.7, 8.1)	5.9 (4.4, 7.4)	-5.9 (-8.1, -3.7)
p-value[2]	< .001	< .001	0.34	< 0.05	< .001	0.11	< .001	< .001	< .001	< .001	< .001

1 Geometric mean based on logarithmic transformed data
2 Contrasts between the well-rested morning and sleep deprived

TABLE 3 Mean values (SD) of NeuroCart parameters.

Parameter	Karolinska Sleepiness Scale	Saccadic peak velocity (deg/s)	Saccadic reaction time (sec)	Smooth pursuit (%)	Body sway (mm)[1]	Adaptive tracking (%)	VAS Alertness (mm)	VAS Calmness (mm)	VAS Mood (mm)
Well rested morning (SD) n=24	3.5 (1.2)	528.8 (62.0)	0.214 (0.03)	45.23 (9.3)	211.0 (80.8)	32.7 (4.2)	53.8 (5.3)	54.3 (5.5)	54.9 (6.7)
Well rested afternoon (SD) n=24	4.0 (1.2)	514.4 (55.6)	0.216 (0.03)	44.63 (9.9)	204.3 (97.6)	33.1 (5.2)	52.1 (6.2)	56.5 (6.1)	55.3 (5.8)
Sleep deprived (SD) n=23	6.8 (1.2)	489.1 (61.7)	0.222 (0.03)	44.56 (12.2)	249.0 (130.1)	26.8 (5.9)	34.7 (9.2)	61.5 (11.0)	49.7 (7.2)
Contrasts (95% CI) of Well rested morning vs Sleep deprived	3.3 (2.7, 4.0)	-40.71 (-53.25, -28.16)	0.0084 (0.0007, 0.0162)	-1.04 (-3.20, 1.13)	17.6% (2.3%, 35.1%)	-5.847 (-7.581,-4.113)	-19.18 (-23.07, -15.29)	7.23 (3.57, 10.89)	-5.27 (-7.96, -2.58)
p-value of Well rested morning vs Sleep deprived	p< .001	p< .001	p< 0.05	p=0.34	p< 0.05	p< .001	p< .001	p< .001	p< .001

1 Geometric mean based on logarithmic transformed data

TABLE 4 Pearson correlations for the variables in different conditions. Only correlations with at least one significant correlation (p<.05) are included in this overview.

Variable	SDLP	Condition	(Pearson) correlation	p-value	Intercept	Slope
SDLP GRIDT	Simulator	Well rested morning	0.63	<.001	12.79	0.28
		Well rested afternoon	0.58	.003	13.68	0.26
Mean Speed simulator	Simulator	Well rested afternoon	0.21	.33	94.60	0.07
Mean Speed GRIDT	GRIDT	Sleep deprived morning	-0.54	.01	103.18	-0.34
Subj driving performance and motivation simulator	Simulator	Sleep deprived morning	-0.73	<.001	12.57	-0.20
		Well rested afternoon	-0.40	.05	13.00	-0.16
Subj driving effort simulator	Simulator	Sleep deprived morning	0.56	.01	2.05	0.20
		Well rested afternoon	0.49	.02	-3.53	0.25
Subj driving effort GRIDT	GRIDT	Sleep deprived morning	0.44	.05	-7.54	0.70
Smooth eye pursuit	Simulator	Well rested morning	0.49	.01	18.41	0.89
	GRIDT	Well rested morning	0.18	.41	30.10	0.71

TABLE 5 Pearson correlations for the delta between sleep deprivation and well-rested morning for any parameter and the SDLP of the simulator and GRIDT. Only correlations with at least one significant value (i.e., $p \leq .05$) are presented.

Parameter	SDLP	(Pearson) correlation	P-value	Intercept	Slope
SDLP GRIDT	Simulator	0.51	.02	1.60	0.15
Subj driving performance and motivation simulator	Simulator	-0.50	.01	-2.03	-0.18
Subj driving effort GRIDT	GRIDT	0.45	.04	3.31	0.97
Adaptive Tracker	Simulator	-0.50	.02	-3.72	-0.22
	GRIDT	-0.13	.58	-4.81	-0.24
VAS Alertness	Simulator	-0.44	.04	-14.25	-0.51
	GRIDT	0.48	.03	-24.45	21.60
VAS Calmness	Simulator	-0.10	.64	8.28	-0.11
	GRIDT	-0.44	.05	13.66	-22.25
VAS Mood	Simulator	-0.53	.01	-1.32	-0.41
	GRIDT	-0.02	.95	-4.81	-0.06

THE FOLLOWING SUPPLEMENTS ARE AVAILABLE ONLINE

S1 (study protocol), **S2** (CONSORT Checklist) and **S3** (Data per subject).
https: //doi.org/10.1371/journal.pone.0278300

CHAPTER 3

THE INTERACTIVE WALKWAY PROVIDES FIT-FOR-PURPOSE FALL RISK BIOMARKERS IN THE ELDERLY: COMPARISON OF ZOLPIDEM AND SUVOREXANT

Clin Transl Sci. 2024 Jul;17(7): e13875. doi: 10.1111/cts.13875.

Ingrid Koopmans,[1,2] Daphne Geerse,[2,3] Lara de Ridder,[1] Melvyn Roerdink,[3] Maria Joanna Juachon,[1] Clemens Muehlan,[4] Jasper Dingemanse,[4] Joop van Gerven,[1,2] Geert Jan Groeneveld,[1,2] Rob Zuiker[1,2]

1 **Centre for Human Drug Research, Leiden, The Netherlands.**
2 **Leiden University Medical Center, Leiden, The Netherlands.**
3 **Department of Human Movement Sciences, Faculty of Behavioural and Movement Sciences, Vrije Universiteit Amsterdam, Amsterdam Movement Sciences, The Netherlands**
4 **Department of Clinical Pharmacology, Idorsia Pharmaceuticals Ltd, Allschwil, Switzerland**

ABSTRACT

Dynamic balance assessments such as walking adaptability may yield a more realistic prediction of drug-induced falls compared to postural stability measurements, as falls often result from limited gait adjustments when walking. The Interactive Walkway (IWW) measures walking adaptability but sensitivity to medication effects is unknown. If proven sensitive and specific, IWW could serve as a biomarker for targeted fall-risk assessments in early clinical drug development.

In this 3-way crossover study, 18 healthy elderly (age: 65-80 years) subjects received 5 mg zolpidem, 10 mg suvorexant or placebo in the morning. Assessments were performed pre-dose and approximately hourly until 9 h post-dose. IWW assessments included an 8-meter walking test, goal-directed stepping, obstacle-avoidance, and tandem-walking. Other pharmacodynamic measurements were the Timed-Up-and-Go (TUG) test at a comfortable and fast pace, adaptive tracking, and body sway.

A decline in performance was observed for zolpidem compared to placebo for 3 h post-dose in IWW walking adaptability outcome measures, TUG, adaptive tracking, and body sway. For the IWW tasks, a decrease in walking speed (among others) was observed. IWW parameters were not affected by suvorexant compared to placebo at any time point. However, an increase of 9.8% (95%CI: 1.8%,18.5%) in body sway was observed for suvorexant compared to placebo up to 3 h post-dose.

The IWW successfully quantified drug effects of two hypnotic drugs and distinguished between zolpidem and suvorexant regarding their effects on walking. As a biomarker, the IWW demonstrated sensitivity in assessing dynamic balance and potential fall risk in early phase clinical drug development.

INTRODUCTION

Sleep disturbances can have serious adverse consequences in older adults. Increased risk of falls is among the most prevailing.[1] Approximately 28-35% of people aged 65 years and over fall annually, increasing to 32-42% for those over 70 years of age.[2] This risk of falling is increased with medications commonly prescribed for insomnia that are known to affect psychomotor functioning. Sleep medications such as benzodiazepines have been demonstrated to affect standing balance, increasing the risk of falls in the elderly.[3-6] (benzodiazepines OR, 1.57 (95%CI: 1.43; 1.72).[7] Benzodiazepines are positive allosteric modulators of the gamma-aminobutyric acid-A (GABA-A) receptors involved in the basal ganglia-thalamocortical systems and affect fine-tuning of motor commands.[8] In an elderly population, the consequences of a fall due to a loss of balance caused by drugs can be severe.

During early clinical drug development, postural stability is typically measured through anterior-posterior body sway during a quiet, upright stance. We have previously used the body sway test to demonstrate the effects of sleep deprivation,[9] alcohol,[10] benzodiazepines,[11,12] and other psychoactive agents[13] on body stability. Even though this body sway is a biomarker sensitive to drug effects, its relationship with common causes of falling is unclear.

Falls during walking and transfers predominantly result from inadequate interactions with the environmental context, leading to balance loss due to a trip, slip, or a misplaced step.[14-15] Walking adaptability thus seems to be an essential determinant of fall risk. Assessing fall-risk biomarkers incorporating such walking-adaptability interactions may, therefore, result in a stronger predictor of falls compared to body sway during quiet stance or other clinical tests such as the Timed-Up-and-Go (TUG) test. The Interactive Walkway (IWW) is an instrument developed to assess walking adaptability and walking-related fall risk by augmenting a multi-Kinect-v2 walkway with projected visual context (stepping targets, suddenly appearing obstacles) and parameterizing various fall-risk biomarkers, such as obstacle-avoidance margins and success rates based on markerless full-body 3D motion tracking.[16,17] IWW fall-risk assessment protocols comprise complementary environmental-context tasks such as avoiding suddenly appearing obstacles, precision stepping and tandem-walking tasks.

IWW outcome measures are sensitive for discriminating between freezing and non-freezing people with Parkinson's disease and healthy controls, as well as between people with stroke and healthy controls,

with differences in expected directions.[18,19] More relevant to the current study is the observation that IWW assessments improved the identification of prospective fallers compared to generic fall-risk factors and standard clinical test scores such as TUG, with poor obstacle-avoidance success rates and insufficient slowing down in tasks that demand precise foot placement as key predictors.[20] However, it is unknown whether such IWW fall-risk biomarkers are sensitive to sleep-medication effects. If so, the IWW could qualify as a valuable instrument for a targeted fall-risk assessment in early clinical drug development.

Orexin receptor antagonists are a new class of hypnotics, of which suvorexant and lemborexant are the first representatives registered for the treatment of insomnia in the United States, while daridorexant is the first available to patients in both US and Europe.[21] Suvorexant, lemborexant, and daridorexant are dual orexin receptor antagonists (DORAs) and inhibit orexin receptors alleviating the potential hyperarousal effect of orexin -A and orexin-B neuropeptides.[22–24] Benzodiazepines have a broader effect on the central nervous system (CNS) compared to DORAs with a specific target at the orexin system, reducing wakefulness. Based on differences in the mechanism of action, differences in fall-risk and balance can be expected.

This placebo-controlled study aimed at evaluating the use of the IWW as a biomarker for fall-risk studying the effect of 10 mg suvorexant and 5 mg zolpidem in 18 healthy elderly subjects. We expect that both drugs increase fall-risk biomarkers, especially in the first hours after intake, but less so for suvorexant, given that it is expected to have a smaller influence on psychomotor functioning than zolpidem.

METHODS

The study was registered at ToetsingOnline under NL76600.056.21 and approved by Foundation Beoordeling Ethiek Biomedisch Onderzoek, Assen, The Netherlands. All subjects gave written informed consent before the study started. The study was performed according to ICH-GCP guidelines, and the Declaration of Helsinki and its latest amendments. The study was conducted from 26 April 2021 to 25 June 2021 at CHDR and the Leiden University Medical Center (LUMC) in Leiden, The Netherlands.

Design

This study was a single-center, randomized, double-blind, placebo-controlled, three-way crossover exploratory study in 18 healthy elderly subjects. The study consisted of a medical screening visit and three one-day treatment periods (Figure S1). Subjects received a single dose of suvorexant 10 mg, zolpidem 5 mg or placebo during the study periods. An in-between period of at least six days was chosen to ensure sufficient washout from the drug with the longest half-life (i.e., 12 h for suvorexant).

A complete medical screening was performed at CHDR to assess a subject's eligibility for this study. All subjects underwent screening 21 to 1 day before the first dosing, consisting of medical history, physical examination, Mini-Mental State Examination (MMSE), clinical laboratory tests (blood chemistry, hematology, serology, and urinalysis), supine vital signs, urine drug screen, breath alcohol test, and electrocardiogram. In addition, the subjects were familiarized with all study activities, including the IWW, body sway, and adaptive tracking, to minimize learning effects during study execution.

At treatment periods, subjects arrived in the morning and stayed at the LUMC until they were discharged in the evening, approximately 11 h after admission. At check-in, eligibility was re-checked based on a urine drug screen, breath alcohol test, concomitant medication, and adverse event (AE) review. Vital signs and AEs were repeatedly recorded throughout the study to assess safety. Blood samples measuring suvorexant and zolpidem plasma concentrations were collected before and at 1, 2, 3, 5, 7, 8, and 9 h after dosing. Assessments for the IWW, adaptive tracking, and body sway were performed at 2 and 1 h pre-dose and at 1, 2, 3, 5, 7, 8, and 9 h after dosing.

Subjects

Healthy elderly female and male subjects between 65 and 80 years were recruited via media advertisement or from the CHDR subject database. Subjects were only included if they had a regular sleeping pattern, scored 25 or higher on the MMSE, did not have a fall more than three times during the past year or had neurological diseases and/or orthopaedic problems that could interfere with normal gait function. Potential subjects performed all assessments during screening and were excluded if there was any doubt on their ability to complete the task during the study. Furthermore, it was not allowed to use concomitant medications with a pronounced effect on the CNS. Subjects were also asked not to consume alcohol, caffeine or xanthine-containing beverages or use any nicotine-containing products within 24 h before each study visit. Treatment administration was done around eleven o'clock in the morning.

Treatments

The recommended dose of suvorexant is 10 mg, according to the prescribing information.[22] It is approved by the FDA as a treatment for insomnia. Its median peak concentration (C_{max}) occurs at around 2 h post-dose in the fasted state which was also implemented in this study. The elimination half-life ($T_{½}$) is approximately 12 h.[22]

Zolpidem is often used as an active comparator in studies with sleep-inducing agents that measure coordination, residual effects, and/or postural stability.[25,26] Zolpidem is a hypnotic to treat insomnia, with 5 mg being the recommended starting dose for the elderly. Zolpidem is a ligand of high-affinity positive modulator sites of GABA-A receptors. It selectively binds to α1-subunits of this ion channel. Following oral administration, zolpidem is rapidly absorbed, with the time to attain C_{max} reached within 0.5-3 h. The $T_{½}$ is approximately 2.4 h.[27]

Suvorexant was provided as 10 mg tablets, and zolpidem was supplied as 5 mg tablets. Both study medications were over encapsulated in Swedish orange capsules to maintain blinding. Placebo consisted of identical, lactose-filled capsules. Subjects began fasting minimally 2 h before until 2 h after each study drug administration. Water was allowed ad libitum.

Randomisation and blinding

Study staff and subjects remained blinded until the database was locked. A statistician not involved in the clinical study conduct performed block-randomization using SAS (Cary, NC, USA) version 9.4. Subjects were randomly assigned to one of six treatment sequences in a balanced study design. Subject numbers were sequentially assigned to participants after medical screening by blinded study staff.

Pharmacodynamic Assessments

Interactive Walkway

Fall-risk biomarkers were derived from various walking (adapt)ability assessments with the IWW. The IWW comprises four spatially and temporally integrated Kinect-v2 sensors with optimized inter-sensor distances,[17,28] providing markerless 3D full-body kinematics of various body points (e.g., ankles, spine base, and spine shoulder). The IWW was equipped with a projector (EPSON EB-585W, ultra-short-throw 3LCD projector, Epson Europe B.V., Amsterdam, The Netherlands) to augment the entire 8-m walkway with gait-dependent visual context, such as obstacles or a narrow beam, for the walking-adaptability tasks. Using a spatial calibration grid, the sensors and projector coordinate systems were spatially aligned to a standard coordinate system.[16] IWW data was sampled at 30 Hz using custom-written software utilizing the Kinect-for-Windows Software Development Kit (SDK 2.0). IWW fall-risk biomarkers were validated for unconstrained walking and walking-adaptability assessments[16,17] and were better able to identify fallers prospectively than standard clinical test scores such as TUG.[20]

Subjects performed the following IWW tasks (see Figure 1), outcome measures for each task were averaged over the repetitions:

8-METER WALKING TASK (8MWT) This included walking at a self-selected walking speed. Outcome measures were walking speed (cm/s), step length (cm), step width (cm), cadence (steps/min), and step time (s). Two repetitions were performed.

OBSTACLE-AVOIDANCE TASK This included avoiding suddenly appearing obstacles. Outcome measures were obstacle-avoidance margins (cm), success rate (%), and (normalized) walking speed (%). Five repetitions were performed, more than the other tasks, to provide more data on the success rate for the obstacles.

GOAL-DIRECTED STEPPING (GDS) TASK This included precision stepping onto a sequence of shoe-size-matched stepping stones in an irregular pattern. Outcome measures were stepping accuracy (cm) and (normalized) walking speed (%). Two repetitions were performed.

TANDEM-WALKING TASK This included walking on a line. Outcome measures were success rate (defined as the percentage of steps on the line, %), (normalized) walking speed (%), and mediolateral sway (cm). Two repetitions were performed.

TUG This included rising from a standard armchair, walking to a line on the floor 3 m away, turning, returning, and sitting down again. The outcome measure was completion time (sec). Two repetitions were performed at a comfortable and two at a fast-walking speed.

Subjects always started with the 8MWT, which enabled the researcher(s) to adjust the settings of the walking-adaptability tasks to one's gait characteristics to obtain a similar level of difficulty for each subject and

measurement time. All IWW tasks were performed at a self-selected comfortable walking speed, except for the TUG, which was also performed at a fast-walking speed.

Adaptive tracking

The adaptive tracking test was performed as described initially by Borland and Nicholson[29] using customized equipment and software (based on TrackerUSB hard-/software (Hobbs, 2004, Hertfordshire, UK)). Adaptive tracking is a pursuit-tracking task susceptible to many psychoactive drugs.[10,11,30–32] During the test, a circle moves randomly on a screen, and the subject is instructed to keep a dot inside the moving circle by operating a joystick. When successful, the speed of the moving circle increases. Conversely, the velocity is reduced if the subject cannot maintain the dot inside the circle. The average speed of the moving circle as a percentage of the maximum speed of the circle over 3.5 min was used for analysis.

Body Sway

Body sway during quiet standing was used to assess postural stability as previously described.[10,12] Anterior-posterior body sway was measured with closed eyes using a body sway meter (Celesco) based on the Wright taximeter.[33] All body movements over 2 min were integrated and expressed as millimetres of sway and recorded. This relatively simple test shows deteriorations of postural stability with CNS-depressants[10] and some improvements with stimulants.[13]

Pharmacokinetic Assessments

Plasma samples were analyzed by an independent bioanalytical laboratory (Analytisch Biochemisch Laboratorium BV, Assen, The Netherlands). Concentrations of suvorexant and zolpidem were quantified using validated liquid chromatography with tandem mass spectrometry methods with a lower limit of quantification of 1.00 ng/mL and 0.50 ng/mL, respectively, and coefficient of variation between 2.1 and 8.3%, and 0.2% and 3.3% respectively. More detailed description of the analysis is available in the Supplementary information.

Analysis

Pharmacodynamics

Statistical analyses were performed using the SAS Version 9.4 (SAS Institute INC., Cary, NC, USA)

Each parameter was analyzed with a mixed-model analysis of covariance with treatment, time, period, sex, and treatment by time as fixed factors and subject, subject by treatment, and subject by time as random factors and the (average) baseline measurement as a covariate. Assessments of the first three h post-dose (recorded at 1, 2, and 3 h) were combined to increase the power of the analysis for the contrasts covering the T_{max} of both drugs. The sample size for this study is based on a precision estimate using previous collected data of effects on body sway by DORA compounds and benzodiazepines.

The Kenward-Roger approximation was used to estimate denominator degrees of freedom, and model parameters were estimated using the restricted maximum likelihood method.

The general treatment effect and specific contrasts were reported with the estimated difference and the 95% CI, the least square mean (LSM) estimates, and the p-value. Graphs of the LSM estimates over time by treatment were presented with 95% CI as error bars and change from baseline LSM estimates.

The following contrasts were calculated within the model: Suvorexant up to 3 h – Placebo up to 3 h; Suvorexant at 5 h – Placebo at 5 h; Suvorexant 7 to 9 h – Placebo 7 to 9 h; Zolpidem up to 3 h – Placebo up to 3 h; Zolpidem at 5 h – Placebo at 5 h; Zolpidem 7 to 9 h – Placebo 7 to 9 h; Suvorexant up to 3 h – Zolpidem up to 3 h; Suvorexant at 5 h – Zolpidem at 5 h, and Suvorexant 7 to 9 h – Zolpidem 7 to 9 h. The results were not corrected for multiple testing.

Body sway (anterior-posterior sway in mm/2 min) data was natural log-transformed before entering the Mixed Model Repeated Measures (MMRM). LSM, LSM difference, and 95% CI were transformed back to their original scale (i.e., to geometric mean and geometric mean ratio expressed in percentage change).

Pharmacokinetics

PK variable programming was conducted with R 3.6.1 for Windows (R Foundation for Statistical Computing/R Development Core Team, Vienna, Austria, 2019). PK parameters were calculated from concentration data in mass/volume units. Parameters were calculated using noncompartmental analysis, using actual elapsed time from dosing to estimate individual plasma PK parameters. These parameters were: C_{max}, T_{max}, $T_{½}$, and the area under the concentration-time curve from time zero to the last quantifiable concentration time point (AUC_{last}). All PK data were summarised by treatment group

using descriptive statistics. Values were expressed as the mean ± SD for all parameters except T_{max}, which was presented as the median (range).

RESULTS

Participants

18 (nine male and nine female) healthy subjects were enrolled in the study. The mean age (range) of all subjects was 71.9 (66-88) years, and their mean BMI (range) was 25.44 (21.5-29.6) kg/m^2. Seventeen subjects were white, and one subject was African-American. Eleven participants were excluded during the medical screening. Reasons for exclusion were: high blood pressure (n=6), abnormal ECG (n=3) or other reasons (n=2). No subjects were excluded based on the training of the assessments. All subjects completed the study; therefore, the safety, PK-, and PD-analysis set consisted of 18 subjects (Table S1).

Pharmacodynamics

All analyses were performed on the change from baseline, with baseline defined as an average of the first two assessments on that day. The results of the body sway, adaptive tracking, TUG, and IWW tasks up to 3 h compared to placebo post-dose are presented in Table 1. The table includes the variables with at least one significant contrast for IWW. The table, including all results of the body sway, adaptive tracking, TUG, and IWW tasks compared to baseline, is presented in the Supplement (Table S2). None of the tasks included in this study showed any statistically significant results for the contrasts at 5 h post-dose and 7-9 h post-dose; these time points are therefore not presented here.

In general, the average walking speed by treatment was fastest in the 8MWT (115.8 – 121.4 cm/s) and slowest in the tandem-walking task (94.2 – 103.6 cm/s). For all IWW tasks, walking speed decreased significantly for zolpidem compared to both placebo and suvorexant. Between these two contrasts, the effect was stronger for zolpidem compared to placebo. The greatest and smallest differences in walking speed for zolpidem compared to placebo were found for the tandem-walking task (estimate of difference (ED): -9.41 cm/s (95% CI: -13.74; -5.07), Figure 2) and the obstacle-avoidance task (ED: -5.25 cm/s (-7.55; -2.96)), respectively. Speed differences for zolpidem compared to suvorexant were again greatest for the tandem-walking task (ED: -7.87 cm/s (-12.12; -3.61)) and smallest for the 8MWT (ED: -3.45 cm/s (-5.68; -1.23)). None of the IWW outcome measures differed significantly between suvorexant and placebo conditions. Shorter step lengths during the 8MWT were observed for zolpidem compared to both placebo (ED: -2.26 cm (-3.21; -1.30)) and suvorexant (ED: -1.85 cm (-2.80; -0.89)) conditions. Likewise, smaller leading-limb margins during the obstacle-avoidance task were observed for zolpidem compared to placebo (ED: -0.03 cm (-0.04; -0.01)) and suvorexant (ED: -0.02 cm (-0.03; -0.00)) conditions (see Figure 2). Finally, participants swayed more mediolaterally during the tandem-walking test with zolpidem than with suvorexant (ED: 0.54 cm (0.23; 0.84)), see Figure 2.

Compared to placebo (see Table S2), body sway during quiet standing was significantly increased for both zolpidem (ED: 122.3 mm, 35.2% (25.3%; 45.8%)) and suvorexant (ED: 34.1 mm, 9.8% (1.8%; 18.5%)) in the 3 h post-dose (Figure 2). Adaptive-tracking performance decreased significantly in the 3 h post-dose for zolpidem compared to placebo (ED: -3.60, (-4.52; -2.68)), but not significantly for suvorexant compared to placebo (ED: -0.77, (-1.70; 0.15). The TUG (both at comfortable (see Figure 2) and fast speed) increased significantly for zolpidem compared to placebo (ED: 0.68 sec, (0.38; 0.99) and ED: 0.43 sec (0.26; 0.60), respectively). The TUG did not increase for suvorexant compared to placebo but was significantly longer for zolpidem compared to placebo and suvorexant for both comfortable and fast speed (ED: 0.68 sec, (0.38; 0.99) and ED: 0.43 sec (0.26; 0.60), ED: 0.59 sec, (0.28; 0.90) and ED: 0.41 sec (0.24; 0.59), respectively).

Pharmacokinetics

For suvorexant and zolpidem, the mean concentration-time curve is depicted in supplementary figures. A summary of the PK parameters is provided in Table 2. The median T_{max} of zolpidem was 1h (Figure S2) and of suvorexant around 2 h (Figure S3). Individual concentrations showed accurate assessment of C_{max} in the first three hours for both drugs. The geometric mean C_{max} of zolpidem was 80 ng/mL (range 48 – 171 ng/mL) and of suvorexant 228 ng/mL (range 117 – 366 ng/mL). The AUC_{last} of zolpidem was 296 h*ng/mL (range 180 – 624 h*ng/mL) and the AUC_{last} of suvorexant was 1075 h*ng/mL (range 609 – 1896 h*ng/mL). The $T_{½}$ of zolpidem based on nine subjects was 2.3 h. The $T_{½}$ of suvorexant could not be calculated accurately as the estimation is only based on two subjects.

The concentration-effect curve of walking speed during the 8MWT (average per session) (Figure S4) and body sway (Figure S5) for zolpidem and suvorexant with a linear trendline shows a steeper relationship for zolpidem on both assessments. Other parameters were not analysed for this relation between concentration and effect.

Safety

Generally, both suvorexant and zolpidem were well tolerated. 19 AEs were reported for suvorexant, of which somnolence (n=12) was the most prevalent. 29 AEs were reported for zolpidem, of which somnolence (n=12), balance disorder (n=5), and dizziness (n=4) were the most prevalent. All AEs judged as related to the study medication were mild in intensity and self-limiting.

DISCUSSION

To evaluate the effect on walking adaptability of (newly) registered drugs, a sensitive biomarker with higher validity to real-life circumstances is preferred. Here, we present the results of a study in which two types of sleep-promoting drugs were evaluated versus placebo on three different levels of ecological validity using four biomarkers for postural stability and walking adaptability: the body sway, adaptive tracking, the TUG test, and the IWW. The PK, safety, body sway, and adaptive tracker data are all in accordance with the literature and the prescribing information of zolpidem and suvorexant. This underlines the reliability of the study. The walking speed of all IWW tasks and the performance parameters of the TUG, body sway, and adaptive tracking were all affected by zolpidem, indicating that zolpidem affected balance. In contrast, only body sway was affected by suvorexant, while IWW outcomes were not, indicating that this drug had a smaller effect on walking balance. After 3 h, no effects on stability and walking adaptability were detected for both sleep-inducing agents.

The IWW is a standardized test battery to measure fall-risk biomarkers, which was shown to differentiate between zolpidem and placebo with robust results. This biomarker, which not only considers postural stability during standing but also during walking, can therefore be used in early clinical drug development to detect effects on walking, postural stability, and possibly fall risk. The IWW might be preferred to static stability measurements, such as body sway, because of the high validity to daily activities. Additionally, the IWW can include different tasks, including stepping over obstacles or challenging participants to increase step length and width with stepping stones. This creates the opportunity to target specific drug effects relevant to drug development. Such options are limited or absent in the Step Quick Turn Test (SQTT, described below), TUG, or body sway task. On average, the walking speed under zolpidem was significantly decreased for all involved IWW tasks. The relation between walking speed and the GDS task was studied in a previous paper.[34] It can be argued that subjects prevented mistakes by reducing walking speed and that a translation to baseline walking speed is needed to accurately assess walking adaptability during this task. In addition to a significant change in walking speed, the significant change in the 8MWT step length indicates that participants took smaller steps, possibly to prevent the consequences of disbalance. The decrease in the margin of the leading limb during the obstacle-avoidance task indicates that participants stepped more closely to the obstacle. This could increase the risk of falling when stepping over real objects (and not 2D projections). The increased sway during the tandem-walking task indicates that more prominent movements of the upper body were made, which could be interpreted as more imbalance.

Although suvorexant numerically decreased the walking speed during the IWW tasks, the endpoints were not significantly altered compared to placebo. Similarly, the TUG task was significantly affected by zolpidem for up to 3 h post-dose but not by suvorexant (Table 1). Based on the data collected in this study, a sample size of 17 would suffice to differentiate between zolpidem and suvorexant (see statement in supplement). Because the IWW has not yet been fully validated, the current benchmark for clinical relevance is the well-known effect of zolpidem on increased fall-risk.[35] With the current study, a statistical difference between zolpidem and suvorexant has been shown with sufficient power, which indicates that suvorexant has a considerably lower impact on walking adaptability than zolpidem, which is clinically known to affect walking adaptability.

The psychomotor test battery for this study contained the adaptive tracking task and the body sway task. Both psychomotor tasks have been used before in clinical trials involving DORAs, and the results confirm previous findings.

In this study, for zolpidem and suvorexant compared to placebo, a significant increase in body sway was found for the first 3 h post-dose compared to placebo, while there was also a significant difference between zolpidem and suvorexant. Adaptive tracking showed a significant difference for zolpidem up to 3 h post-dose, the task did not show a significant difference for suvorexant.

The increase in body sway and decrease in adaptive tracking performance conform to previous findings in a study assessing morning dosing of daridorexant in healthy elderly subjects.[36] In that study, a single dose of 25 mg daridorexant showed an increased body sway in the first two hours

post-dose compared to the lower dose of 5 mg daridorexant or placebo. A similar result was found for almorexant 400 mg at 2 h post-dose; a return to baseline was seen after 6-8 h.[37]

In another previous study in healthy elderly, 10 mg zolpidem was compared with 8 mg ramelteon (melatonin receptor agonist) in a middle-of-the-night (MOTN) study using (among others) the SQTT.[38] This test is a shorter version of the TUG; it consists of two steps forward, a quick 180 degrees turn, and two steps back to the starting point. Only zolpidem caused a significant prolongation of the time to complete the SQTT and increased the sway during the task, which corresponds to the results found in this study.

In a MOTN-study, 10 mg zolpidem was compared to 5 mg and 10 mg Lemborexant.[39] A significant effect of zolpidem and lemborexant on body sway was detected in the MOTN and just after morning awakening. This supports the increase in body sway for zolpidem and suvorexant found in this study.

Comparing the results of the IWW in this trial with TUG and body sway, it may be concluded that the IWW does not have a greater sensitivity to detect the effects of zolpidem. However, the IWW is closer to fall-risk associated with activities of daily living, i.e., walking and transfers, and IWW therefore has more and direct clinical relevance for medication effects than body sway and other comparable lab-based assessments with more quantitative endpoints compared to the TUG. Because of the significant difference between suvorexant and placebo in the first 3 h post-dose, one may hypothesize that body sway is more sensitive than the IWW to the effects of suvorexant. The TUG showed similar results to the walking speed of the IWW tasks, and because of the lack of difference for other variables of the IWW tasks, the additional value of the IWW seems to be minor.

The difference in the effect of both drugs on the IWW and other assessments can be mainly explained by differences in the mechanism of action. Where suvorexant affects specifically the orexin system reducing wakefulness, zolpidem activates GABA-A receptors and neurotransmitters, which results in general sedation and muscle relaxation. This is also clearly visible in the concentration-effect graphs using the linear trendlines. Zolpidem shows a stronger impairing effect on both body sway and walking speed (see Figure S4 and S5).

Zolpidem and suvorexant were administered in the morning, which is different from the indication of sleep medication prescribed in the evening. Nevertheless, the PK parameters (Table 2) and safety reports were in line with the manufacturer's United States Prescribing Information (USPI) for zolpidem and suvorexant.[22,27]

Study limitations

Several limitations are to be considered when interpreting the results of this study. First, and perhaps most importantly, this study included freely and independently moving elderly participants, i.e., not young subjects or subjects with a real risk of falling. During the screening, participants answered questions regarding fall history, the use of walking aids, and any medical history events related to impaired walking or balance. Indicators of imbalance or a history of falling led to exclusion. The study population might not show characteristics of typical fallers (such as not adapting gait, not slowing down, or taking risks) when under the influence of medication. Even though this might not be the population naturally at risk, fall-risk biomarkers were sufficiently sensitive to show a structural effect of sleep-inducing agents in this relatively small sample size. In a population with a higher risk propensity, the effects of the drugs tested on walking ability could be expected to be larger rather than smaller.

Secondly, a previously proven relationship between walking speed and parameters indicated a higher risk of falling with higher walking speed.[34] Therefore, it might be argued that subjects prevented mistakes by reducing walking speed and that a translation to baseline walking speed is needed to accurately assess walking adaptability. Again, this nevertheless did not lead to absence of drug effects on walking ability so at higher walking speeds the drug effects would have been larger rather than smaller.

The effect of suvorexant might be different due to daytime dosing when the orexin system is more active, and additionally, hormone systems inducing sleep are naturally less active. However, fall-risk assessments in the previously mentioned MOTN studies show a similar relationship between benzodiazepine and DORA compounds.[39]

In conclusion, the effect of drugs affecting body stability were well detected using the IWW, both in terms of placebo-controlled differential effects of two types of sleep medication, as well as in terms of PD over time. This is the first study using the IWW to demonstrate the fit-for-purpose of the instrument to study the influence a single dose zolpidem and suvorexant on walking-adaptability related fall-risk parameters, showing it may be used as a new biomarker in clinical drug development to provide an early indication of drug-induced increased risk of fall.

REFERENCES

1 Hill EL, Cumming RG, Lewis R, Carrington S, Le Couteur DG. Sleep disturbances and falls in older people. Journals of Gerontology – Series A Biological Sciences and Medical Sciences. 2007;62(1): 62–6.

2 World Health Organization. WHO global report on falls prevention in older age [Internet]. 2007 [cited 2023 Jun 30]. Available from: https: //www.who.int/publications/i/item/9789241563536

3 Castleden CM, Allen JG, Altman J, St John-Smith P. European Journal of Clinical Pharmacology A Comparison of Oral Midazolam, Nitrazepam and Placebo in Young and Elderly Subjects. Eur J Clin Pharmacot. 1987;32: 253–7.

4 Nikaido AM, Ellinwood EH, Heatherly DG, Gupta SK. Age-related increase in CNS sensitivity to benzodiazepines as assessed by task difficulty. Psychopharmacology (Berl). 1990;100: 90–7.

5 Patat A, Brohier S, Zieleniuk I, Rosenzweig P, Perault MC, Vandel B, et al. Assessment of the interaction between a partial agonist and a full agonist of benzodiazepine receptors, based on psychomotor performance and memory, in healthy volunteers. Journal of Psychopharmacology. 1995;9(2): 91–101.

6 Swift C, Swift M, Ankier S, Pidgen A, Robinson J. Single dose pharmacokinetics and pharmacodynamics of oral loprazolam in the elderly. J clin Pharmac. 1985;20: 119–28.

7 Woolcott JC, Richardson KJ, Wiens MO, Patel B, Marin J, Khan KM, et al. Meta-analysis of the Impact of Nine Medication Classes on Falls in Elderly Persons. Essential Reviews in Geriatric Psychiatry. 2009;169(21): 1952–60.

8 Griffin CE, Kaye AM, Rivera Bueno F, Kaye AD. Benzodiazepine pharmacology and central nervous system-mediated effects. Ochsner Journal. 2013;13: 214–23.

9 Van Steveninck AL, Van Berckel BNM, Schoemaker RC, Breimer DD, Van Gerven JMA, Cohen AF. The sensitivity of pharmacodynamic tests for the central nervous system effects of drugs on the effects of sleep deprivation. Journal of Psychopharmacology. 1999;13(1): 10–7.

10 van Steveninck AL, Gieschke R, Schoemaker HC, Pieters MSM, Kroon JM, Breimer DD, et al. Pharmacodynamic interactions of diazepam and intravenous alcohol at pseudo steady state. Psychopharmacology (Berl). 1993;110(4): 471–8.

11 Van Steveninck AL, Schoemaker HC, Pieters MSM, Kroon R, Breimer DD, Cohen AF. A comparison of the sensitivities of adaptive tracking, eye movement analysis, and visual analog lines to the effects of incremental doses of temazepam in healthy volunteers. Clin Pharmacol Ther. 1991;50(2): 172–80.

12 Van Steveninck AL, Wallnöfer AE, Schoemaker RC, Pieters MSM, Danhof M, Van Gerven JMA, et al. A study of the effects of long-term use on individual sensitivity to temazepam and lorazepam in a clinical population. Br J Clin Pharmacol. 1997;44(3): 267–75.

13 Cohen AF, Burggraaf J, Van Gerven JMA, Moerland M, Groeneveld GJ. The use of biomarkers in human pharmacology (Phase I) studies. Vol. 55, Annual Review of Pharmacology and Toxicology. Annual Reviews Inc.; 2015. p. 55–74.

14 Niino N, Tsuzuku S, Ando F, Shimokata H. Frequencies and circumstances of falls in the national institute for longevity sciences, longitudinal study of aging (NILS-LSA). J Epidemiol. 2000;10: S90–4.

15 Talbot LA, Musiol RJ, Witham EK, Metter EJ. Falls in young, middle-aged and older community dwelling adults: Perceived cause, environmental factors and injury. BMC Public Health. 2005 Aug 18;5(86).

16 Geerse DJ, Coolen BH, Roerdink M. Kinematic validation of a multi-Kinect v2 instrumented 10-meter walkway for quantitative gait assessments. PLoS One. 2015;10(10): 1–15.

17 Geerse DJ, Coolen BH, Roerdink M. Walking-adaptability assessments with the Interactive Walkway: Between-systems agreement and sensitivity to task and subject variations. Gait Posture. 2017;54: 194–201.

18 Geerse DJ, Roerdink M, Marinus J, van Hilten JJ. Assessing walking adaptability in stroke patients. Disabil Rehabil. 2021;43(22): 3242–50.

19 Geerse DJ, Roerdink M, Marinus J, van Hilten JJ. Assessing Walking Adaptability in Parkinson's Disease: 'The Interactive Walkway.' Front Neurol. 2018 Dec 12;9.

20 Geerse DJ, Roerdink M, Marinus J, van Hilten JJ. Walking adaptability for targeted fall-risk assessments. Gait Posture. 2019 May 1;70: 203–10.

21 Hoyer D, Allen A, Jacobson LH. Hypnotics with novel modes of action. Br J Clin Pharmacol. 2020 Feb 1;86(2): 244–9.

22 FDA. USPI Suvorexant [Internet]. 2014. Available from: www.fda.gov/medwatch.

23 FDA. USPI Lemborexant [Internet]. 2019. Available from: www.fda.gov/medwatch.

24 FDA. USPI Daridorexant [Internet]. 2022. Available from: www.fda.gov/medwatch.

25 Drake CL, Durrence H, Cheng P, Roth T, Pillai V, Peterson EL, et al. Arousability and fall risk during forced awakenings from nocturnal sleep among healthy males following administration of zolpidem 10 mg and doxepin 6 mg: A randomized, placebo-controlled, four-way crossover trial. Sleep. 2017;40(7).

26 Greenblatt DJ, Roth T. Zolpidem for insomnia. Expert Opin Pharmacother. 2012 Apr;13(6): 879–93.

27 FDA. USPI Zolpidem [Internet]. 2007. Available from: http: //www.fda.gov/medwatch

28 Geerse D, Coolen B, Kolijn D, Roerdink M. Validation of foot placement locations from ankle data of a Kinect v2 sensor. Sensors (Switzerland). 2017;17(10): 6–9.

29 Borland R, Nicholson A. Comparison of the residual effects of two benzodiazepines (nitrazepam and flurazepam hydrochloride) and pentobarbitone sodium on human performance. Br J Clin Pharmacol. 1975;2(1): 9–17.

30 De Haas SL, De Visser SJ, Van Der Post JP, Schoemaker RC, Van Dyck K, Murphy MG, et al. Pharmacodynamic and pharmacokinetic effects of MK-0343, a GABAA α2,3 subtype selective agonist, compared to lorazepam and placebo in healthy male volunteers. Journal of Psychopharmacology. 2008 Jan;22(1): 24–32.

31 Gijsman HJ, Van Gerven JMA, Pieters MSM, Schoemaker HC, Kroon R, Tieleman MC, Ferrari MD, Spinhoven Ph, Van Kempen GMJ CAF. Pharmacokinetic and pharmacodynamic profile of oral and intravenous -meta-chlorophenylpiperazine in healthy volunteers. J Clin Psychopharmacol. 1998;18: 289–94.

32 Cohen A, Ashby L, Crowley D, Land G, Peck A, Miller A. Lamotrigine (BW430C), a potential anticonvulsant. Effects on the central nervous system in comparison with phenytoin and diazepam. Br J Clin Pharmacol. 1985;20(6): 619–29.

33 Wright BM. A simple mechanical ataxia-meter. J Physiol. 1971;218: 27P-28P.

34 Roerdink M, Geerse DJ, Peper C (Lieke) E. 'Haste makes waste': The tradeoff between walking speed and target-stepping accuracy. Gait Posture. 2021;85(January): 110–6.

35 Kolla BP, Lovely JK, Mansukhani MP, Morgenthaler TI. Zolpidem is independently associated with increased risk of inpatient falls. J Hosp Med. 2013;8(1): 1–6.

36 Muehlan C, Boehler M, Brooks S, Zuiker R, van Gerven J, Dingemanse J. Clinical pharmacology of the dual orexin receptor antagonist ACT-541468 in elderly subjects: Exploration of pharmacokinetics, pharmacodynamics and tolerability following single-dose morning and repeated-dose evening administration. Journal of Psychopharmacology. 2020 Mar 1;34(3): 326–35.

37 Hoever P, Hay J, Rad M, Cavallaro M, Van Gerven JM, Dingemanse J. Tolerability, pharmacokinetics, and pharmacodynamics of single-dose almorexant, an orexin receptor antagonist, in healthy elderly subjects. J Clin Psychopharmacol. 2013 Jun;33(3): 363–70.

38 Zammit G, Wang-Weigand S, Rosenthal M, Peng X. Effect of Ramelteon on middle-of-the-night balance in older adults with chronic insomnia. Journal of Clinical Sleep Medicine. 2009;5(1): 34–40.

39 Murphy P, Kumar D, Zammit G, Rosenberg R, Moline M. Safety of lemborexant versus placebo and zolpidem: Effects on auditory awakening threshold, postural stability, and cognitive performance in healthy older participants in the middle of the night and upon morning awakening. Journal of Clinical Sleep Medicine. 2020;16(5): 765–73.

FIGURE 1 Schematic overview of the Interactive Walkway tasks.

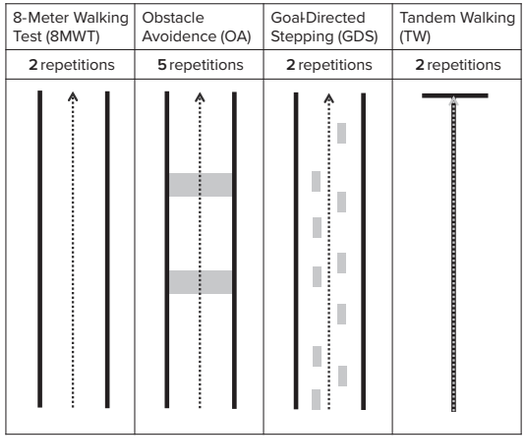


FIGURE 2 Graphical presentation of estimated means and 95% Confidence Intervals of results for placebo, zolpidem, and suvorexant. Top left: walking speed during tandem walking task. Top right: Margins leading limb during Obstacle Avoidance task. Bottom left: Time to complete the Timed Up and Go test. Bottom right: Total sway during the body sway task.

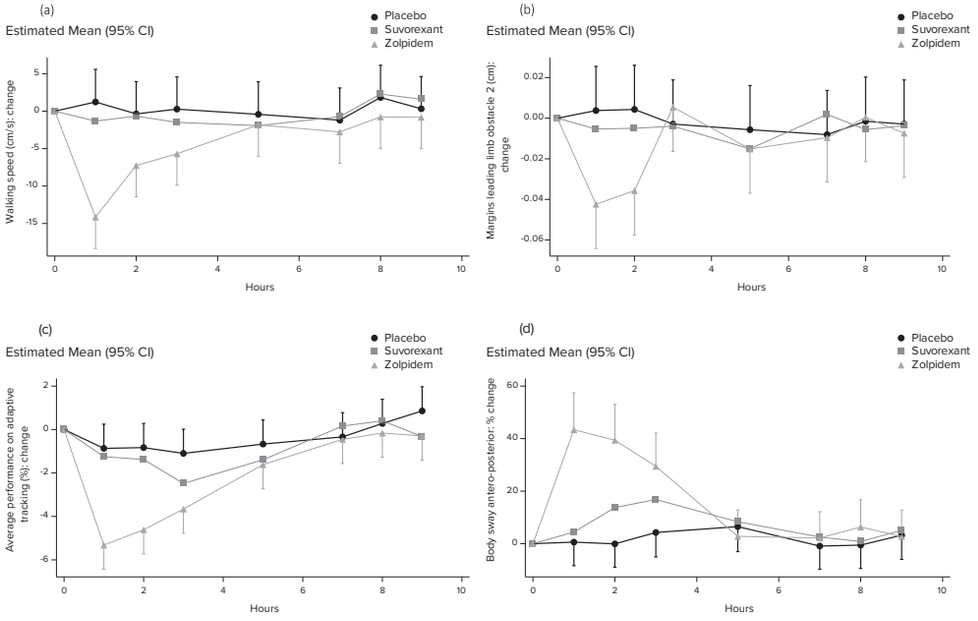


TABLE 1 Contrasts (95% CI) and p-value up to 3 h post-dose.

	zolpidem vs placebo	suvorexant vs placebo	zolpidem vs suvorexant	LSM placebo	LSM zolpidem	LSM suvorexant
Body sway (mm)	35.2% (25.3%,45.8%) p<.001	9.8% (1.8%,18.5%) p=.017	23.1% (14.1%,32.8%) p<.001	347.8	470.1	381.9
Adaptive tracking (%)	-3.60 (-4.52,-2.68) p<.001	-0.77 (-1.70,0.15) p=.101	-2.82 (-3.75,-1.91) p<.001	16.9	13.3	16.1
Interactive Walkway walking speed (cm/s)						
8-meter walking test	-5.56 (-7.77,-3.36) p<.001	-2.11 (-4.32,0.10) p=.061	-3.45 (-5.68,-1.23) p=.003	121.4	115.8	119.3
Goal-directed stepping	-8.12 (-11.14,-5.10) p<.001	-2.07 (-5.09,0.95) p=.173	-6.05 (-9.08,-3.02) p<.001	110.8	102.7	108.7
Obstacle-avoidance	-5.25 (-7.55,-2.96) p<.001	-1.43 (-3.71,0.86) p=.215	-3.83 (-6.12,-1.53) p=.002	115.1	109.8	113.7
Tandem-walking	-9.41 (-13.74, -5.07) p<.001	-1.54 (-5.87, 2.79) p=.475	-7.87 (-12.12,-3.61) p<.001	103.6	94.2	102.1
IWW 8MWT – Step Length (cm)	-2.26 (-3.21,-1.30) p<.001	-0.41 (-1.37,0.55) p=.392	-1.85 (-2.80,-0.89) p<.001	69.8	67.5	69.4
IWW OA – Margins Leading Limb (cm)	-0.03 (-0.04,-0.01) p<.001	-0.01 (-0.02,0.01) p=.381	-0.02 (-0.03,-0.00) p=.011	0.12	0.09	0.11
IWW TW – Sway (cm)	0.29 (-0.01,0.60) p=.062	-0.24 (-0.55,0.06) p=.116	0.54 (0.23,0.84) p<.001	3.13	3.42	2.89
Timed-Up and Go (sec)	0.68 (0.38,0.99) p<.001	0.09 (-0.22,0.40) p=.559	0.59 (0.28,0.90) p<.001	9.86	10.54	9.95
Timed-Up and Go fast (sec)	0.43 (0.26,0.60) p<.001	0.01 (-0.16,0.18) p=.895	0.41 (0.24,0.59) p<.001	7.25	7.68	7.26

IWW: Interactive Walkway; 8MWT: 8-meter walking test; OA: obstacle-avoidance; TW: Tandem-Walking

TABLE 2 Pharmacokinetic parameters of zolpidem and suvorexant.

	C_{max} (ng/mL) Mean(+/-SD)	T_{max} (h) Median(min-max)	$T_{½}$ (h) Mean(+/-SD)	AUC_{last} (h*ng/mL) Mean(+/-SD)
Zolpidem	82.8 (+/- 26.8) N=18	1 (1 – 3) N=18	2.9 (+/- 1.0) N=9	296 (+/- 104.5) N=18
Suvorexant	235.3 (+/- 61.1) N=18	2 (1 – 3) N=18	-	1074 (+/- 335.2) N=18

THE FOLLOWING SUPPLEMENTS ARE AVAILABLE ONLINE

Table S1 (Demographic characteristics), **Table S2** (All constrasts), **Figure S1** (Study design), **Figure S2** (Plasma concentration suvorexant), **Figure S3** (plasma concentration zolpidem), **Figure S4** (concentration-effect for walking speed), **Figure S5** (concentration-effect for body sway), and description of concentration analysis.
https://doi.org/10.1111/cts.13875

CHAPTER 4

THE IMPACT OF A VIRTUAL WOUND ON PAIN SENSITIVITY: INSIGHTS INTO THE AFFECTIVE DIMENSION OF PAIN

Front. Pain Res., 26 February 2025. Sec. Pain Research Methods.
Volume 6 – 2025 | https: //doi.org/10.3389/fpain.2025.1502616

Ingrid Koopmans*[1,2], Robert-Jan Doll[1,2], Maurice Hagemeijer[3], Robert van Barneveld[3], Marieke de Kam[1], Geert Jan Groeneveld[1,2]

1 Centre for Human Drug Research, Leiden, The Netherlands.
2 Leiden University Medical Center, Leiden, The Netherlands.
3 Righteous Games, Eindhoven, The Netherlands.

ABSTRACT

The perception of pain is difficult to assess due to the complex combination of various components related to nociception, experience, and cognition. There are currently no biomarkers to assess the affective component of pain in healthy volunteers. Using Virtual Reality (VR), it may be possible to assess changes in pain perception when adding an affective component to painful stimulation.

In this two-visit feasibility study, we assess the effect of a simulated wound in VR on the electrical pain detection (PDT) and tolerance (PTT) threshold in 24 healthy male study participants. The VR simulation presented a copy of the research room from first person view. Prior to each VR assessment, study participants were primed by interacting with the VR environment. Two conditions were assessed: (1) VR-Wound: a burn-wound, smoke, and electrical sparks become visible and audible with increasing stimulus intensity, and (2) VR-neutral: no additional aspects. The PDT and PTT to electrical stimuli were recorded during both VR conditions and outside of VR. VAS-Questionnaires were used to assess unpleasantness and fear.

The PDT decreased when a virtual wound is presented compared to a neutral condition. Study participants experienced the electrical stimulation as more painful and more intense during the wound simulation than during the neutral condition. The effect was more pronounced during the second visit.

VR enhanced the perception of pain, thereby providing new insights into the affective component of pain. Further testing of this methodology is warranted by performing a clinical study that evaluates drug effects on the affective component of pain.

INTRODUCTION

The affective component of pain plays an important role in pain and is linked as an important factor to cases of chronic pain.[1,2] Emotions can modulate the experience of pain. However, singling out the affective component of pain in a clinical research setting remains difficult.[3,4] As a result, demonstrating that a (drug) treatment is effective at alleviating pain by addressing the affective component is challenging.[1,2]

The analgesic effects of new drugs are commonly assessed in early phase clinical drug studies using various well-established tests, and are ideally conducted in healthy study participants. These tests are particularly valuable if they provide early indications of a drug's efficacy, which is strongly dependent on the availability of pharmacodynamic biomarkers to be used for proof of pharmacology, proof-of- mechanism, or proof-of-concept. Analgesic effects in healthy study participants can be assessed by changes in pain detection thresholds (PDT) and pain tolerance threshold (PTT) to stimuli (e.g., electrical, heat, or pressure).[5] To attribute the effect of the change in threshold to the studied intervention, these tests are performed in a controlled environment, minimizing external interferences and distraction. However, healthy study participants in this setting will unlikely show sufficient variation in the affective component of pain – without being challenged- , making it difficult to study the effects of analgesic compounds that target the affective component in an isolated fashion. By adding a challenge that adds an affective component to a pure nociceptive task, the task becomes more susceptible to the effects of new analgesic compounds that target pain syndromes in an affective pain component plays an important role. Such a task may produce a suitable pharmacodynamic biomarker, which can be used in early phase clinical drug studies of analgesics influencing the affective component of pain.

It is well known that the perception of pain can be altered due to distraction or anxiety. When distracted, both children and adults report less pain.[6,7] In contrast, inducing anxiety can increase pain intensity and unpleasantness.[8] Interpreting a (painful) stimulus as potentially harmful influences the reported levels of pain.[9] Additionally, creating an illusion for the study participants within reasonable limits, such as the rubber arm paradigm is found effective suggesting threat without a nociceptive stimulus.[10] It therefore seems clear that it is possible to modulate pain experience in a controlled pain experiment. A promising possibility to modulate a person's

pain experience by a combination of focus and anxiety might be by using Virtual Reality (VR).

Current research on VR in relation to pain is primarily focussed on alleviating the perception of pain by deep immersion in a distracting setting.[11,12] Others have studied the fundamental aspects of the effect of VR on pain. For example, it was demonstrated that the level of virtual ownership of an avatar (simulated person) affects the pain experience.[13-15] The simulated size of affected body parts and transparency of these body parts also influence pain experience.[16,17] Using VR to introduce a coloured area on the location of a painful stimulus was demonstrated to be effective to modulate the pain experience.[16] Due to the used heat paradigm, PTT recordings were not feasible due to the risk of skin damage. Another study including a burning hand simulation in augmented reality also showed a reduction in PDT.[18] Pain experience questionnaires, which is the current standard for emotional responses on pain, are not yet included in a study with VR.

A VR simulation with a realistic visual enhancement of consequences of the stimuli combined with audio related to the pain experiment has never been performed. In this study, we combine an electrical pain test with VR. In VR, the electrical stimulation is accompanied with sounds and visuals of electrical sparks, and an increasingly damaging skin underneath the stimulating electrodes. With this, we aim to add an affective component to a nociceptive stimulus with the purpose to try to exacerbate the pain in a setting closer to real life. In addition to capturing the pain detection and pain tolerance thresholds, qualitative aspects (e.g., subjective scales for anxiety and fear, and personality questionnaires) were also recorded. This setup could potentially provide biomarkers (pain thresholds) to study effects of analgesic drugs that target the affective component of pain.

METHODS

This was an exploratory single-centre two-visit cross-over study. The study was conducted between March and July 2021 at the clinical research unit of the Centre for Human Drug Research (CHDR) in Leiden, The Netherlands. The study was approved by the Medical Ethics Committee Stichting Beoordeling Ethiek Biomedisch Onderzoek (Assen, the Netherlands). Study conductance was according to the Dutch Act on Medical Research Involving Human Subjects (WMO) and in compliance with all International Conference on Harmonisation Good Clinical Practice (ICH-GCP) guidelines and the Declaration of Helsinki. The study was prospectively registered in the International Clinical Trials Registry Platform as NL-OMON28178.

Study design

Potential study participants underwent medical screening and training on the pain task without VR. In this same visit, study participants filled in the personality questionnaires as described below. After inclusion, the study consisted for each study participant of two visits, each one day starting at around 9:00 and finalized around 16:00. Study participants were admitted to the clinical research unit for the duration of the visit and discharged after completion of all study assessments. A rest period of 7 days was included in between both visits. All study participants underwent the same procedures during each visit. Each visit started with a urine drug test and alcohol breath analysis after which the anxiety inventory was performed. During the study visit, four sets of pain tasks were completed. One set of pain tasks included one VR assessment (either with wound or neutral) and a before and after assessment without VR. This to function as a baseline and to potentially capture the long-term effect of the VR simulation. A total of 12 pain tasks were completed for each study visit. The first visit started with the neutral VR simulation followed by three sets of assessments with the wound simulation. The second visit also contained three sets of pain tasks with the wound but had the neutral simulation included in as second set. The assessments are repeated during each visit to increase the power of the study. After each assessment, study participants reported the pain experience using VAS questionnaires. A rest period of one hour separated each set of assessments. See Figure 1 for a schematic overview of the study activities.

Study participants

All study participants provided written informed consent prior to undertaking any study-related activities. To match the avatar in VR, only healthy male study participants between 18 and 40 years of age were invited to participate. Only light to medium skin tones (Fitzpatrick ≤ IV) were allowed and no deformations or (dis)colouring of the skin was allowed in upper and lower limbs. Study participants with a pain tolerance threshold >80% of the maximum stimulation of the test (without VR) were excluded in the study. No history of psychiatric illness or visionary disorders were allowed. Study participants who smoked more than 5 cigarettes per day on average or consumed more than 8 units of (methyl)-xanthine a day were excluded because of possible withdrawal symptoms during study participation. Study participants who had previously experienced Simulator Sickness Syndrome with either VR or another simulator were not eligible to participate. During screening, study participants were neither trained on the VR simulation nor

given information about the content of the VR simulation. The sample size of this study was not based on a formal sample size calculation due to the exploratory nature. As it is our aim to use this method in early phase clinical drug studies we chose a sample size that is typically used in in early phase clinical drug studies of analgesics.

Assessments

All measurements were performed in a quiet room. During all assessments only the study participant and a research assistant was present in the room. To prevent infection with COVID-19, all study participants wore face masks throughout the study and the equipment was cleaned with disinfectant in between study participants.

Temperament and Character Inventory

The Temperament and character inventory (TCI) was developed by Cloninger et al. and widely accepted for personality assessments.[19,20] The TCI contains 240 items which needs to be answered with 'correct' or 'incorrect'. The Dutch translation was provided in digital form by Datec and used during the screening visit. Endpoints include seven dimensions of temperament and character: Novelty seeking (NS), harm avoidance (HA), reward dependence (RD), persistence (PS), self-directedness (SD), cooperativeness (CO) and self-transcendence (ST). Each of these dimensions are divided in multiple sub-factors resulting in a total of 24 subscales.

Pain Catastrophizing Scale

Dutch language version of the Pain Catastrophizing scale (PCS) evaluates the pain-related thoughts and emotional distress related to pain.[21,22] The questionnaire consists of 12 self-report questions with a 4-point scale measuring three components of catastrophic thinking: rumination, magnification, and helplessness. The PCS is only performed at screening.

Pain-Anxiety Symptoms Scale

The Pain-Anxiety Symptoms scale short form (PASS-20) was presented to the study participant during the screening visit.[23] The questionnaire consists of 20 self-report questions with a 5 Likert-scale measuring four dimensions of pain-related fear and anxiety: Cognitive anxiety responses, escape and avoidance, fearful thinking, and physiological anxiety responses. The PASS-20 is only assessed during the screening visit.

Spielberger State-Trait Anxiety Inventory

State and trait Anxiety was measured with the Spielberger State-Trait Anxiety Inventory, trait scale (STAI-DY).[24] The STAI-DY consists of 40 questions with a 1 to 4 scale: 20 items are related to the State-Anxiety (STAI-DY-1) and 20 items are related to the Trait-Anxiety (STAI-DY-2) subscales. The STAI-DY-1 is assessed both during the screening visit and at the start of each study visit. The STAI-DY-2 is assessed only during the screening visit.

Electrical stair test

The electrical stair test uses two electrodes (AG-AGCL) on the tibial bone to assess cutaneous electrical pain. Single electrical stimuli are provided with a duration of 0.2 ms, increasing from 0 mA to a maximum of 50 mA in steps of 0.5 mA. Study participants were provided with an electronic version of the Visual Analogue Scale (EVAS) and instructed to start moving the slider when the stimulus became painful. The intensity of this pain detection threshold (PDT) is the first endpoint of this pain task. The second endpoint recorded is the pain tolerance threshold (PTT), the intensity at which the study participant indicates the maximum value on the EVAS which corresponds to the maximum pain tolerated. If the study participant does not indicate the PTT before 50mA, the maximum duration of the test is 120 seconds after which the stimulation is stopped automatically. After each electrical stair test, four electronic VAS assessments were used to evaluate the level of pain, unpleasantness and intensity of pain, and fear.

Virtual Reality

EQUIPMENT Study participants wore a VR headset with headphones (Vive Pro, HTC) during the VR-Pain measurement. The VR environment emulates the room in which study participants performed all assessments. The VR includes an avatar of the study participant from a first-person view. The chair and equipment of the electrical stair pain test (including the electrodes on the leg and a VAS slider) were included as well. To ensure embodiment, the position of the legs, hands, and VAS slider were tracked using HTC Vive trackers and a leap motion sensor. Additionally, the skin colour of the avatar matched the most frequently occurring skin colour in the Netherlands (i.e., Fitzpatrick II-III).

PRIMING AND PERCEPTION OF EMBODIMENT Prior to each VR assessment, study participants were primed by performing a set of instructions encouraging interaction with the VR environment. The instructions included

asking the study participant to grab the VAS slider from the sky (handed by the assistant) and describe objects located in the room. The study participant controlled both the start and stop of the test, including the simulation, using the VAS slider.

After each VR assessment, the study participants' perception of embodiment was evaluated. Six statements related to embodiment were presented to which could be answered using a 7-point Likert scale (1: completely disagree, 7: completely agree). The statements were: (1) the virtual body parts felt like my own body parts, (2) it felt like the virtual body was my own, (3) when I saw the wound appearing on my leg it felt like the wound was a part of me, (4) the movements of the virtual body appeared like my own movements, (5) I felt I had control over the movements of the virtual body, and (6) I had the illusion owning a different body than my own.

VR CONDITIONS There were two different VR conditions: (1) VR-Wound and (2) VR-Neutral (see Figure 2). The VR-Wound condition shows a burn wound around the electrodes on the leg. The intensity of the wound increases simultaneously with the intensity of the pain test. This simulation is accompanied by sounds of electrical sparks through the VR headset. The simulation started directly at the beginning of the test and reached maximum intensity at 40 seconds. After 40 seconds, the intensity of the audio-visual simulation no longer increases but continues until the test is stopped. This to make sure most of the study participants experience the full simulation. During the VR-Neutral condition, no additional visual or auditory stimulations were applied.

Subjects were instructed to look at the electrodes which was monitored by the research assistant via a mirror image of the VR view on the computer. It was not possible to confirm if the subjects had their eyes open during the assessment. After each assessment including the VR-Wound simulation, study participants scored the simulation on realism, unpleasantness, and their focus on the wound during the pain task.

Statistical Analysis

The statistical analysis was preceded by a data review which consisted of visual inspection of individual graphs per visit of all efficacy measurements by time. To establish whether significant effects can be detected on the repeatedly measured pain parameters, the change from baseline of each parameter was analysed with a mixed effects model with condition (pre-VR, VR-Neutral, VR-Wound or Post-VR), visit (day 1 or day 2), session (1, 2, 3, or 4 within day), condition by visit, condition by session, visit by session and condition by visit by session as fixed factors and study participant, study participant by visit and study participant by session as random factor. The Kenward-Roger approximation was used to estimate denominator degrees of freedom and model parameters were estimated using the restricted maximum likelihood method.

The TCI was compared with external data using a two-sided t-test, as was the difference between baseline assessments with the STAI on each visit.

RESULTS

Study participants

A total of 25 healthy male study participants were enrolled in the study. One study participant stopped participation after the first visit due to COVID-19 quarantine requirements in the Netherlands. He was replaced and excluded for statistical analysis. The other 24 study participants were included for statistical analysis (age mean (SD) is 23.3 (5.0), range 18 – 34). No relevant datapoints were excluded based on the blinded data review. Data could not be collected for two VR measurements of separate study participants due to technical difficulties and two VR measurements of two other study participants were lost because of an emergency evacuation practice drill of the clinical research unit. Additionally, answers to related questionnaires for these measurements were not collected.

Effects of Virtual Reality

Pain Thresholds electrical stair task

The least square means of all PDT and PTT values are presented in Figure 3. The mean PDT during the VR-Wound condition (4.85 mA) was significantly lower (-18.4%, 95%CI: (-26.9%, -9.0%) p<.001) than the PDT during the VR-Neutral simulation (5.95mA). This was more pronounced during the second visit, see Figure 3. The pre-VR neutral measurement was significantly lower (17.2%, 95%CI: (3.0%, 33.4%) p<.016) compared to the VR-Neutral simulation (5.08mA and 5.95mA, respectively).

For the PTT, no statistical difference was found between the VR-Neutral and VR-Wound simulation (18.11 mA vs 17.52mA, p=.21). Additionally, no significant difference was observed for the PTT between de pre-VR neutral measurement and the VR-Neutral test (17.42mA vs 18.11mA, p=.21). However, there was a significant difference for the PTT between the VR-Neutral

(18.68mA) and VR-Wound (16.04mA) simulation for the second visit (-14.1%, 95%CI: (-21.5%, -6.1%) p<.001).

Pain perception during electrical stair task

Study participants rated the pain during the VR-Wound simulation (70 mm) significantly more intense (4.5, 95%CI: (1.8, 7.2) p=.0013) compared to the VR-Neutral simulation (65 mm). Additionally, the pain was rated significantly more unpleasant (5.9, 95%CI: (2.1, 9.8) p=.0028) for the VR-Wound simulation (71 mm) compared to the VR-Neutral simulation (65 mm).

The pain intensity and unpleasantness were not scored significantly different between the pre-VR pain test and the VR-Neutral pain test (intensity: 95%CI: -0.8 (-4.0; 2.4) p=0.6, unpleasantness: 95%CI: -2.7(-6.8; 1.4) p=0.2). See Figure 2 for an overview.

Embodiment and subjective experience of the wound simulation

The level of embodiment during the VR simulations was for both the neutral and the wound simulation on average 21.73 points with a standard deviation of 5.15 and 5.34 points, respectively (see Figure S1). The mean of the VAS Wound questions ranged between 92.3 and 94.5 for the focus, between 51.6 and 56.7 for realism, and between 50.4 and 62.6 for unpleasantness (see Figure 3 and Table S1).

Personality characteristics

Temperament and Character Inventory

Table 1 shows the TCI characteristics for the included study participants and a norm dataset provided by Datec (Leiden, the Netherlands). A student's t-test demonstrated that the study participants in our study showed different characteristics when compared to the norm group on three TCI characteristics: study participants showed lower scores for HA (-2.9 (-5.7; -0.1) p=.04) and ST (-3.0 (-5.7; -0.3) p=.03), and higher scores for NS (3.5 (1.0; 6.0) p=.006) and PS (1.1 (0.3; 1.9) p=.0087). Identified differences for healthy volunteers compared to the chronic pain group are in general for the same personality characteristics: HA (-9.1 (-12.4; -5.8) p<.001), NS (4.3 (1.7; 6.9) p=.0015), PS (1.4 (0.6; 2.2) p=.0005), SD (4.3 (0.7; 7.9) p=.0185).

Pain Catastrophizing scale

Supplementary Table S2 shows an overview of the PCS results. Study participants scored on average 14.1 points (SD = 7.2), the lowest score was 0 and the highest score was 28.

Pain-Anxiety Symptoms Scale

Supplementary Table S3 contains the overview of the PASS-20 results. The average total score of the PASS-20 questionnaire for all study participants was 27.1 (SD = 14.0), the lowest total score was 3 and the highest total score was 52.

Spielberger State-Trait Anxiety Inventory

Summary data of the Spielberger Trait/Stage Anxiety Inventory is added to the supplement in Table S4. On the trait questionnaire (STAI-DY2), study participants had a mean score of 50.2 (SD: 4.3).

The mean STAI-DY1 total score was slightly lower in the second visit (27.2) compared to the first (29.9). However, this difference was not statistically significant (95%CI: -2.7 (-0.1, 5.5) p=.055).

DISCUSSION

Here, we present the results of a study where VR was used to modulate the pain experience during a pain task. We demonstrate that VR can be used to enhance pain in the context of an evoked pain test. By introducing a virtual wound on the location of a painful stimulus, the PDT was lowered when compared to a neutral VR condition. Additionally, we demonstrated that study participants experienced the electrical stimulation as more unpleasant and more intense during the wound simulation, while the electrical simulation paradigm remained identical.

Effects of Virtual Reality

A difference in PDT between the pre-VR neutral and VR-Neutral conditions was observed (see Figure 2). The PDT during the VR-Neutral condition was significantly higher than the PDT during the pre-VR neutral condition, suggesting a higher pain tolerability in the VR environment. Such effects were reported in previous studies as well, where, for example, wound treatment was perceived as less painful in VR than outside VR.[25] Interestingly, in our study no differences were observed for perceived pain intensity and unpleasantness for the VR-Neutral condition compared to the pre-VR neutral measurement. This might be caused by the relatively low number of assessments and the relatively high intra-subject variability. In conclusion, the immersion into (non-wound simulating) VR can be considered to increase pain detection thresholds.

The VR-Wound condition resulted in lower pain detection thresholds compared to the neutral VR simulation. Additionally, the perceived pain inten-

sity and unpleasantness were increased during the VR-Wound condition. Combined, these observations indicate enhanced pain perception when immersed into a VR condition simulating a wound and thereby intensifying the stimulation by adding an affective component to the painful stimulus. Interestingly, the effect was more robust during the second visit during which the VR-Wound condition was the first VR condition tested on that day. This suggests that there might be an effect of the order of VR conditions or stronger responses for the first assessment of a visit. We could not confirm this hypothesis due to the limited number of visits and the chosen order of the measurements. Nonetheless, the effect was overall large enough to allow us to demonstrate a statistically significant enhancement of the pain experience with the VR-Wound condition.

We found no effect of VR simulation (i.e., VR-Neutral vs VR-Wound) on the embodiment score (Figure S1). Others suggested that the level of virtual body ownership could be considered a confounder when differences in outcomes exist.[13,14] However, study participants had a similar perception of body ownership in both the neutral and wound condition. The most likely explanation for this finding is that we used an extensive priming procedure in all VR simulations. The embodiment score was in general quite low and might be improved when the avatar can be more customized to the study participants or with a longer priming session before the measurements.

As mentioned earlier, few studies have adopted a similar approach to studying the effect of a simulation on the location of evoked pain in healthy volunteers.[16,18] These studies were both not executed in the settings typically used in drug development, but did show similar direction of results in lowering the PDT. In early drug development, repeated measures and assessments of concentration effects over the course of the day, as implemented in this study, are the standard. By testing this paradigm in these conditions with similar results, it becomes more feasible to use such a task in early drug development. Additionally, both studies did not include any questionnaires on pain experience, limiting the possibility to relate the findings to the affective component and pain experience.

The aim of this study was to add an affective(-motivational) component to a nociceptive stimulus to create a task more prone to respond to (dug induced) changes in the affective component of pain. Other studies often conclude that untangling the different domains of pain is not possible.[3] If this conclusion holds true, experimental settings may lack sensations, emotions and cognitive processes due to their controlled laboratory nature. With this setup, we aimed to capture more dimensions of pain, including the affective dimension. Talbot recommends that future studies should ensure blinding of all involved and clear instructions for the study participants to prevent unintentional biases in questionnaire responses – a practice we also advocate based on our findings.

Generalisation to chronic pain patients

This study shows how the pain experience during a pain task can be enhanced in healthy study participants, however how this relates to people with chronic pain is unclear. Because personality traits are often related to pain responses,[26] we determined these in the study study participants in this study. This comparison provides information on the ecological validity of this study. The study participants in this study show significantly different personality traits on the TCI questionnaire compared to the norm group (Table 1). Significantly lower scores are found for HA and ST, and higher scores for NS. Other studies have already demonstrated that people with chronic pain are different from a normal population, with higher scores for HA and lower scores for NS, SD, and Cooperativeness (CO).[27] It therefore is possible that in a population with personality characteristics that are more similar to people with chronic pain, the effects of the pain test enhanced with VR may be different.

This study aimed to modulate the response on a painful stimulus possibly resulting in a challenge model for pain which includes an affective component. When properly validated, such a model could yield a biomarker that can be used in healthy study participants for early proof of concept of analgesic drugs aiming to reduce the affective component of pain. An early proof of concept in drug development can provide more insight in possible applications and patient stratification for future studies. Drugs that influence pain processing on a more central level may be beneficial for pain syndromes that currently remain untreatable.[28,29] A future study with the VR-Pain setup will include an intervention to reduce the affective component of pain to provide the next step in validation. An example of such intervention could be an emotional altering drug, e.g. an anxiolytic.

Study limitations

Due to the nature of the conditions, blinding of study participants was not possible in this study. As a result, we could not control for potential confounders including socially desirable responses. However, study participants were not told in advance which VR condition they would be presented with, and they were not informed on the hypothesis of the wound sim-

ulation. Future studies are recommended to include procedures that allow some way of blinding. For example, assessing the effect of medication on VR can be performed in a (double-) blind fashion in a crossover study allowing for a balanced number of VR-Neutral and VR-Wound assessments. Additionally, changing some of the stimulation procedures may aid in limiting anticipation effects (e.g., varying the rate of electrical stimulation).

All visual enhancements and progression of the wound were identical for each session with the VR-Wound simulation. However, unpredicted painful stimuli were experienced as more painful in an earlier study.[8] It is therefore possible that repeated confrontation with the virtual wound would reduce the impact of the simulation. However, study participants reported consistent focus, realism, and unpleasantness throughout the session (see Figure 4). None of the three parameters of the VAS Wound questionnaire (i.e., focus, realism, and unpleasantness) showed significant variation over the visits or the different measurements. Also, the focus parameter indicates that study participants followed the instructions to look at the wound most of the time. Validation of this parameter can be done in future studies by including eye tracking in the VR setup and, for example, creating a heat map of visual focus.

To avoid the uncanny valley (relation between human likeness and a viewer's affinity toward it), a photo realistic wound was avoided. This resulted in the rather low (but stable) realism score. It can be imagined that different (possible improved) results may be obtained with a more realistic version of a wound or a simulation that has a better fit with the specific feeling of this test. A specific study aimed at determination of the optimal simulation may be considered for future research.

CONCLUSION

This study is the first demonstrating the potential of VR in combination with a pain task to provide a challenge model highlighting the affective component of pain in a setting used in early phase drug development. The perceived level of immersion in the VR simulation was stable throughout the study making this setup feasible to use in drug studies with multiple visits and multiple measurements per day. Future studies should aim at validation for the use of proof of concept in early drug development.

ACKNOWLEDGEMENTS The authors would like to thank Femke de Graaf (intern) and all research assistants for their help with this study.

REFERENCES

1 L. N. Mory, D. de O. Fernandes, C. Mancini, M. Mouthon, and J. N. Chabwine, 'The Affective Dimension of Pain Appears to Be Determinant within a Pain–Insomnia–Anxiety Pathological Loop in Fibromyalgia: A Case-Control Study,' J Clin Med, vol. 11, no. 12, 2022, doi: 10.3390/jcm11123296.

2 D. D. Price and S. W. Harkins, 'The affective-motivational dimension of pain A two-stage model,' APS Journal, vol. 1, no. 4, pp. 229–239, 1992, doi: 10.1016/1058-9139(92)90054-G.

3 K. Talbot, V. J. Madden, S. L. Jones, and G. L. Moseley, 'The sensory and affective components of pain: are they differentially modifiable dimensions or inseparable aspects of a unitary experience? A systematic review,' Br J Anaesth, vol. 123, no. 2, pp. e263–e272, 2019, doi: 10.1016/j.bja.2019.03.033.

4 M. Auvray, E. Myin, and C. Spence, 'The sensory-discriminative and affective-motivational aspects of pain,' Neurosci Biobehav Rev, vol. 34, no. 2, pp. 214–223, 2010, doi: 10.1016/j.neubiorev.2008.07.008.

5 P. S. Siebenga et al., 'Reproducibility of a battery of human evoked pain models to detect pharmacological effects of analgesic drugs,' European Journal of Pain, vol. 23, no. 6, pp. 1129–1140, Jul. 2019, doi: 10.1002/ejp.1379.

6 K. A. Birnie, C. T. Chambers, and C. M. Spellman, 'Mechanisms of distraction in acute pain perception and modulation,' Pain, vol. 158, no. 6, pp. 1012–1013, 2017, doi: 10.1097/j.pain.0000000000000913.

7 R. Ruscheweyh, F. Nees, M. Marziniak, S. Evers, H. Flor, and S. Knecht, 'Pain Catastrophizing and Pain-related Emotions,' Clin J Pain, vol. 27, no. 7, pp. 578–586, 2011, doi: 10.1097/ajp.0b013e31820fde1b.

8 K. Carlsson, J. Andersson, P. Petrovic, K. M. Petersson, A. Öhman, and M. Ingvar, 'Predictability modulates the affective and sensory-discriminative neural processing of pain,' Neuroimage, vol. 32, no. 4, pp. 1804–1814, 2006, doi: 10.1016/j.neuroimage.2006.05.027.

9 A. Arntz and L. Claassens, 'The meaning of pain influences its experienced intensity,' Pain, vol. 109, no. 1–2, pp. 20–25, 2004, doi: 10.1016/j.pain.2003.12.030.

10 H. H. Ehrsson, K. Wiech, N. Weiskopf, R. J. Dolan, and R. E. Passingham, 'Threatening a rubber hand that you feel is yours elicits a cortical anxiety response,' Proc Natl Acad Sci U S A, vol. 104, no. 23, pp. 9828–9833, 2007, doi: 10.1073/pnas.0610011104.

11 H. G. Hoffman et al., 'Virtual reality as an adjunctive non-pharmacologic analgesic for acute burn pain during medical procedures,' Annals of Behavioral Medicine, vol. 41, no. 2, pp. 183–191, 2011, doi: 10.1007/s12160-010-9248-7.

12 D. Viderman, K. Tapinova, M. Dossov, S. Seitenov, and Y. G. Abdildin, 'Virtual reality for pain management: an umbrella review,' Front Med (Lausanne), vol. 10, Jul. 2023, doi: 10.3389/fmed.2023.1203670.

13 M. Martini, D. Perez-Marcos, and M. V. Sanchez-Vives, 'Modulation of pain threshold by virtual body ownership,' European Journal of Pain (United Kingdom), vol. 18, no. 7, pp. 1040–1048, 2014, doi: 10.1002/j.1532-2149.2014.00451.x.

14 F. Mancini, M. R. Longo, M. P. M. Kammers, and P. Haggard, 'Visual distortion of body size modulates pain perception,' Psychol Sci, vol. 22, no. 3, pp. 325–330, 2011, doi: 10.1177/0956797611398496.

15 Y. Yim, Z. Xia, Y. Kubota, and F. Tanaka, 'The proteus effect on human pain perception through avatar muscularity and gender factors,' Sci Rep, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-61409-4.

16 M. Martini, D. Perez-Marcos, and M. V. Sanchez-Vives, 'What colour is my arm? Changes in skin colour of an embodied virtual arm modulates pain threshold,' Front Hum Neurosci, vol. 7, no. JUL, pp. 1–5, 2013, doi: 10.3389/fnhum.2013.00438.

17 K. Kilteni, J. M. Normand, M. V. Sanchez-Vives, and M. Slater, 'Extending body space in immersive virtual reality: A very long arm illusion,' PLoS One, vol. 7, no. 7, 2012, doi: 10.1371/journal.pone.0040867.

18 D. Eckhoff, C. Sandor, G. L. Y. Cheing, J. Schnupp, and A. Cassinelli, 'Thermal pain and detection threshold modulation in augmented reality,' Front Virtual Real, vol. 3, Sep. 2022, doi: 10.3389/frvir.2022.952637.

19 R. Cloninger, D. Svrakic, and T. Przybeck, 'Cloninger CR, Svrakic DM, Przybeck TR. A psychobiological model of temperament and character. Arch Gen Psychiatry 50: 975-990,' Arch Gen Psychiatry, vol. 50, no. January 1994, pp. 975–990, 1994.

20 I. J. Duijsens, P. Spinhoven, J. G. Goekoop, T. Spermon, and E. H. M. Eurelings-Bontekoe, 'The Dutch temperament and character inventory (TCI): Dimensional structure, reliability and validity in a normal and psychiatric outpatient sample,' Pers Individ Dif, vol. 28, no. 3, pp. 487–499, 2000, doi: 10.1016/S0191-8869(99)00114-2.

21 S. Van Damme, G. Crombez, P. Bijttebier, L. Goubert, and B. Van Houdenhove, 'A confirmatory factor analysis of the Pain Catastrophizing Scale: Invariant factor structure across clinical and non-clinical populations,' Pain, vol.

96, no. 3, pp. 319–324, 2002, doi: 10.1016/S0304-3959(01)00463-8.

22 M. J. L. Sullivan, S. R. Bishop, and J. Pivik, 'The Pain Catastrophizing Scale: Development and Validation,' Psychol Assess, vol. 7, no. 4, pp. 524–532, 1995, doi: 10.1037/1040-3590.7.4.524.

23 L. M. McCracken, C. Zayfert, and R. T. Gross, 'The pain anxiety symptoms scale: development and validation of a scale to measure fear of pain,' Pain, vol. 50, no. 1, pp. 67–73, 1992, doi: 10.1016/0304-3959(92)90113-P.

24 H. M. Van der Ploeg, P. B. Defares, and C. D. Spielberger, 'A Dutch-Language Adaptation of the Spielberger State-Trait Anxiety Inventory,' De Psycholoog, vol. 15, no. 8, pp. 460–467, 1980.

25 H. G. Hoffman, D. R. Patterson, E. Seibel, M. Soltani, L. Jewett-Leahy, and S. R. Sharar, 'Virtual reality pain control during burn wound debridement in the hydrotank,' Clinical Journal of Pain, vol. 24, no. 4, pp. 299–304, 2008, doi: 10.1097/AJP.0b013e318164d2cc.

26 B. Naylor, S. Boag, and S. M. Gustin, 'New evidence for a pain personality? A critical review of the last 120 years of pain and personality,' Scand J Pain, vol. 17, pp. 58–67, 2017, doi: 10.1016/j.sjpain.2017.07.011.

27 R. Conrad et al., 'Temperament and character personality profiles and personality disorders in chronic pain patients,' Pain, vol. 133, no. 1–3, pp. 197–209, 2007, doi: 10.1016/j.pain.2007.07.024.

28 D. Kaufmann, B. C. Yarns, and P. Gazerani, 'Editorial: The affective aspects of chronic pain and potential treatments,' Front Behav Neurosci, vol. 17, 2023, doi: 10.3389/fnbeh.2023.1209561.

29 V. Neugebauer, Mariacristina Mazzitelli, B. Cragg, G. Ji, E. Navratilova, and F. Porreca, 'Amygdala, neuropeptides, and chronic pain-related affective behaviors,' Neuropharmacology, vol. 170, no. June, pp. 1–33, 2020, doi: 10.1016/j.neuropharm.2020.108052.Amygdala.

FIGURE 1 **Schematic overview of study design and the assessments. The questionnaires are left out of the figure for clarity. Each set contained the following assessments: pain task without VR – VAS Pain – STAI-6 – VAS Fear – pain task with VR (either neutral or wound) – VAS Pain – STAI-6 – VAS Fear – Embodiment – pain task without VR – VAS Pain.**

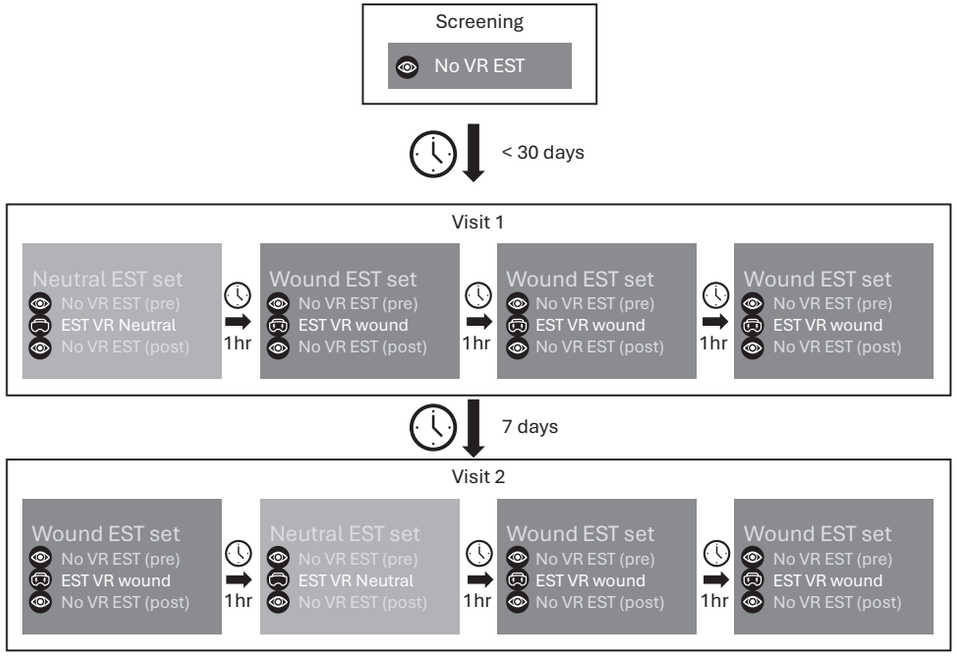


FIGURE 2 Virtual Reality simulation of the burn wound on the leg increasing in severity from left to right.

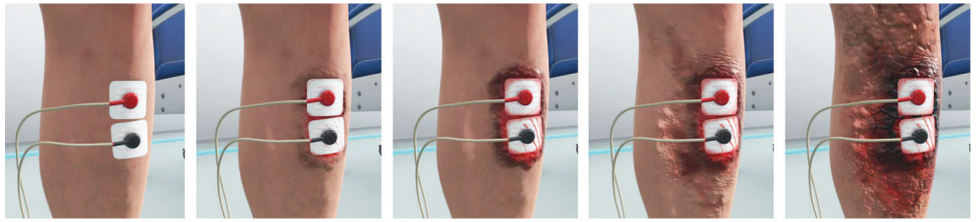


FIGURE 3 Graphical overview of Pain Detection Threshold (a), Pain Tolerance Threshold (b), Visual Analogue Scale (VAS) score of the pain intensity (c) and experienced unpleasantness of the painful stimuli (d).

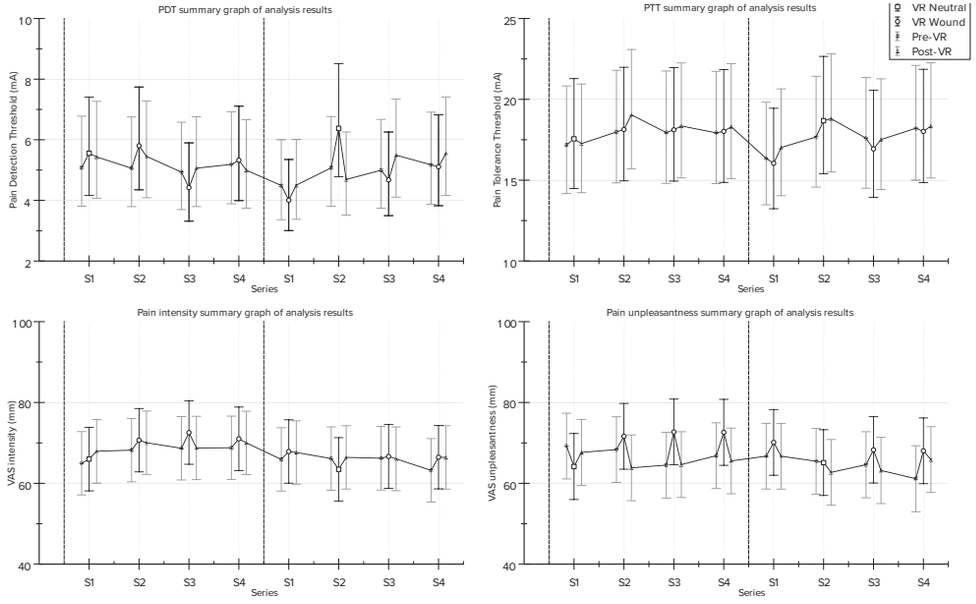


FIGURE 4 Graphical overview of Visual Analogue Scale (VAS) scores regarding the virtual wound: Focus, Realism and Unpleasantness.

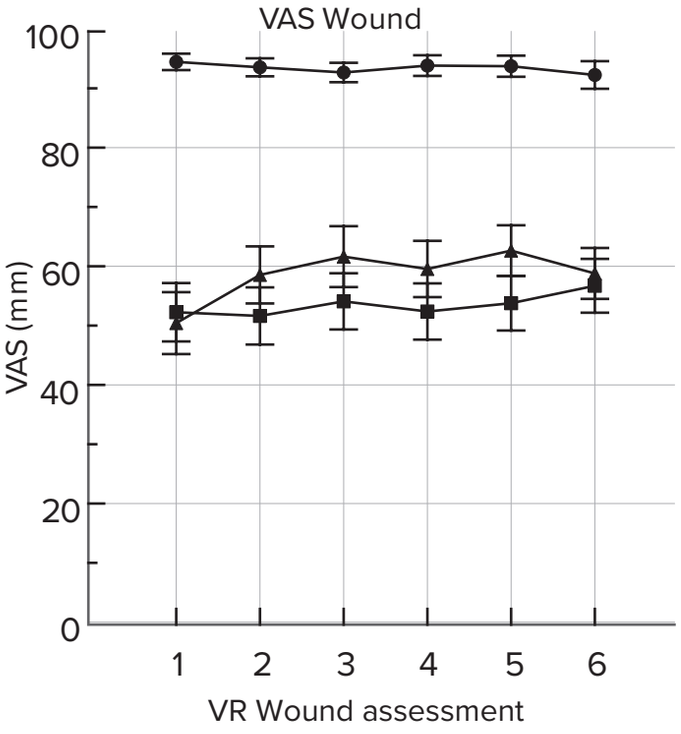


TABLE 1 Temperament and Character Inventory (TCI) for the study participants, norm group (data provided by Datec) and chronic pain patients (data from Conrad et al).[6] The asterisk (*) indicates a significant difference ($p<.05$) between groups.

	Subjects (N=24) Mean (SD)	Norm (N=167) Mean (SD)	Chronic pain (N=207) Mean (SD)
Harm Avoidance	9.6 (6.5)	12.5* (6.5)	18.7* (7.8)
Novelty Seeking	22.5 (6.3)	19.0* (5.7)	18.2* (6.1)
Reward dependence	16.0 (4.1)	15.0 (3.8)	14.6 (4.2)
Persistence	5.6 (1.9)	4.5* (1.9)	4.2* (1.8)
Cooperativeness	32.2 (4.2)	32.4 (6.2)	30.6 (6.7)
Self-directedness	33.8 (5.0)	32.5 (7.0)	29.5* (8.7)
Self-Transcendence	9.2 (4.5)	12.2* (6.5)	10.6 (5.6)

THE FOLLOWING SUPPLEMENTS ARE AVAILABLE ONLINE

Figure S1 (Embodiment scores), **Table S1** (VAS Wound), **Table S2** (Pain Catastrophizing Scale), and **Table S3** (Pain Anxiety Symptoms Scale-20).

https://doi.org/10.3389/fpain.2025.1502616

CHAPTER 5

VIRTUAL REALITY IN A NOCICEPTIVE PAIN TEST BATTERY: A RANDOMIZED, PLACEBO CONTROLLED TWO-WAY CROSSOVER STUDY WITH DIAZEPAM

I.W. Koopmans*[1,2], K.P.W. Rietdijk[1,2], R. Bohoslavsky[1], R.J. Doll[1,2], G.J. Groeneveld[1,2]

1 Centre for Human Drug Research, Leiden, The Netherlands.
2 Leiden University Medical Center, Leiden, The Netherlands.

ABSTRACT

Pain is a complex experience influenced by many psychological factors such as emotion, mood, time of day, and stress. We developed a virtual reality pain task that can modulate pain experience, providing possible biomarkers for the affective component in healthy volunteers. Diazepam, a benzodiazepine used for anxiety, may affect this component. We hypothesize that VR-PainCart can assess drug effects on the affective component of pain.

In a randomized crossover study with 24 healthy male participants, we evaluated the effect of a simulated wound in VR on electrical pain detection (PDT) and tolerance (PTT) thresholds during an electrical pain task. Participants underwent pre-dose tests, followed by 5 mg diazepam or placebo, and six rounds of post-dose tests. Each round included an electrical pain test and two VR conditions: (1) VR-wound that increases with stimulus intensity, and (2) VR-neutral: no additional aspects. PDT and PTT were recorded during both VR conditions and without VR. VAS-Questionnaires assessed pain intensity and unpleasantness, and the McGill Pain Questionnaire (MPQ) investigated pain characteristics.

Diazepam increased PDT in the VR-wound environment (ED: 6.0%, CI 2.4–53.2, $p<0.05$). A trend in PTT increase with diazepam in VR-wound was observed but not statistically significant (ED: 6.5%, CI -3.1–17.0%, $p=0.179$). VAS pain intensity and unpleasantness differences between diazepam and placebo were not significantly different.

VR simulated wound enhanced pain perception in an electrical nociceptive task. Diazepam increased PDT in VR-wound, indicating pharmacological modulation of the affective pain component. Future research will include diverse populations and drugs targeting the affective component, such as antidepressants, to evaluate new analgesic compounds.

INTRODUCTION

Pain is complex and cannot be exclusively defined by its intensity. The affective-motivational model suggests that pain includes not only the well-known nociceptive component but also emotional and cognitive dimensions that shape the pain experience.[1] This is consistent with the existence of drugs that, while not directly affecting nociception, still offer analgesic effects due to their anxiolytic or antidepressant properties. Precise pharmacodynamic biomarkers are crucial for determining proof-of-pharmacology, target engagement, and possible efficacy.[2] However, effective biomarkers that quantify the contribution of emotional aspects to pain remain unavailable. Current patient-reported outcome measures (PROMS) that assess the emotional dimension of pain fall short in terms of content validity and psychometric accuracy.[3] Therefore, developing biomarkers that can accurately evaluate this emotional component—commonly referred to as the affective dimension of pain—is of significant interest.

Human pain models are an important tool for evaluating the analgesic effects of drugs and gaining insights into the mechanisms of pain. Nonetheless, no single experimental model can fully capture the complexity of clinical pain.[4] The 'PainCart®' contains several sensitive and specific tests for measuring different modalities of nociception and is developed to test analgesics in healthy participants. During testing, the emotional processing of pain is minimized as much as possible by using a standardized silent room with no distractions or interactions.[5] This approach results in a nociceptive test battery with high repeatability and sensitivity to drug concentration effects.[6] However, due to limited emotional processing included in the pain tasks, results may not reflect effects on the inherently subjective affective component of pain.[7] As a result, drugs influencing this component may show no or underestimated effects on this nociceptive test battery.

In a previous study, the painful experience of an electrical stimulation task was successfully modulated using Virtual Reality (VR).[8] During an electrical stimulation task, a simulated wound was presented at the location of the electrodes via a VR-headset. The virtual wound increased in severity with the increase of the stimulus. When comparing the response to this 'enhanced stimulation' task to a neutral VR pain task (i.e., without wound), healthy participants had 1) decreased pain detection thresholds, 2) increased perception of VAS pain intensity, and 3) increased perception of VAS pain unpleasantness.

In this study, we administered diazepam, an anxiolytic drug that binds to the GABA-A receptor, increasing the affinity of the receptor and enhancing GABA's inhibitory effects. The anterior cingulate cortex (ACC) is sensitive to changes in the GABA system and plays a crucial role in pain experience.[9,10] Effects on cerebral blood flow into the temporal regions can already be detected after a single dose of diazepam, and higher pain threshold to a cold pressor test. Additionally, studies have found that a low dose of diazepam influences emotional processing, with limited side effects that might influence study execution (e.g., dizziness or headache)[11] other than a decrease in anxiety.[12]

To assess the sensitivity of the VR pain model to quantify the effects of a pharmacological intervention, a single dose of diazepam (5 mg) was used to reduce emotional processing and the affective pain component.

METHODS

This was a randomized, double-blind, placebo-controlled, two-way crossover study in healthy participants. The study was conducted in accordance with the Declaration of Helsinki and approved by the local ethical committee (Stichting BEBO, Assen, The Netherlands), and all participants provided written informed consent prior to any study related activities. Before commencing the study visits, participants were medically screened during a separate visit and when found eligible they were included in the study. Participants received a single dose of diazepam 5 mg and matching placebo in randomized order on two separate study visits. Drug administrations were separated by a washout period of 7 days. VR-PainCart assessments were performed twice pre-dose and repeated hourly up to 6 h after drug administration. Participants were required to fast (only allowed to drink water) 2 hours prior to drug administration up to 1.5 hours post-dose. Drug administration occurred in the morning between 10:00 and 11:30 for all participants after which relative mealtimes were standardized. No blood samples were collected to analyse diazepam serum concentrations due to its well-known pharmacokinetic parameters. A follow-up phone-call was performed 7 to 10 days after the last drug administration to record any adverse events and medication usage. See Figure S1 for a schematic overview of the study design.

Participants

Healthy male participants aged 23 to 35 were enrolled in the study. To ensure avatar realism in the VR simulation, only participants with light to medium skin tones (i.e., Fitzpatrick scale ≤ IV) were included. Additionally, participants were required to have no skin deformations or discolorations on their upper and lower limbs. Eligibility was further restricted to those with a pain tolerance threshold below 80% of the maximum stimulation in the test conducted without VR, any history of psychiatric illness or vision disorders, and a history of simulator sickness based on previous experience in VR or other simulators. Additionally, participants who smoked more than 5 cigarettes per day or consumed more than 8 units of (methyl)-xanthines daily were excluded due to potential withdrawal symptoms during the study periods and to reduce possible effects on pain thresholds.[13] During the screening process, participants were not trained on or informed about the contents of the VR simulation.

Assessments

All assessments were performed in a quiet room with controlled lightning and temperature. During all assessments only the participant and the research assistants were present in the room. Materials and procedures of the electrical stair test and VR enhancement were identical to the version used in the previous study.[8]

Electrical stair test

The electrical stair test[14] used two AG-AGCL electrodes placed on the tibial bone to evaluate cutaneous electrical pain. Single electrical stimuli, each lasting 0.2 ms, were administered, starting at 0 mA and incrementing by 0.5 mA up to a maximum of 50 mA. Participants were provided with an electronic Visual Analog Scale (EVAS) and instructed to move the slider as soon as the stimulus became painful. The intensity at which pain is first detected, is defined as the pain detection threshold (PDT), and the first endpoint of this assessment. The second endpoint, the pain tolerance threshold (PTT), is recorded when the participant indicates the maximum value on the EVAS, representing the highest level of pain they can tolerate. If the PTT is not reached before 50 mA, the test automatically stops after a maximum total duration of 120 seconds.

Virtual Reality

MATERIALS During the pain assessments that included VR, participants wore a VR headset with headphones (Vive Pro, HTC). The VR environment simulated the room in which participants performed all assessments including an avatar in the same sitting position which was viewed from first-person perspective (see Figure S2). Avatar size could be adjusted according to the

height of the participant. All equipment needed for the electrical stair pain test is included in the simulation, including the stimulator, electrodes on the leg, and an EVAS slider. The position of the legs, hands, and VAS slider were tracked using HTC VIVE trackers and a leap motion sensor.

Prior to each VR assessment, participants were primed by performing a set of instructions encouraging interaction with the VR environment. The instructions included asking the participant to grab the VAS slider from the sky (handed to them by the assistant) and describe objects located in the room.

VR CONDITIONS There were two different VR conditions: (1) VR-Wound and (2) VR-Neutral. The VR-Wound condition showed the progressive development of a burn wound with blood, burned skin, and smoke, around the leg electrodes (see Figure 1). The intensity of the wound increased simultaneously with the intensity of the pain test. This simulation was accompanied by sounds of electrical sparks and sizzling noises through the VR headset. The simulation started directly at the beginning of the test and reached maximum intensity at 40 seconds. After 40 seconds, the intensity of the audio-visual simulation no longer increased but continued until the test is stopped. This duration was chosen to make sure most of the participants experience the full simulation. During the VR-Neutral condition, no additional visual or auditory stimulations were applied.

SUBJECTIVE EXPERIENCE OF PAIN AND VR Subjective pain experience was assessed after each pain test including VR by the McGill Short Form[15] and two visual analogue scales (VAS) for the unpleasantness and intensity of pain. Additionally, after the VR-Wound condition, three VAS questions evaluating the (1) focus on the wound, (2) realism, and (3) unpleasantness of the wound were assessed.

EMBODIMENT After each assessment including VR, the level of embodiment was recorded with the embodiment questionnaire including 6 items on a 7-point Likert scale (1: completely disagree, 7: completely agree). The questions each focussed on a different aspect of the embodiment of the virtual body: ownership of body parts, ownership of the body, wound as part of the body (only after VR-Wound simulation), ownership of movement, control of the virtual body, illusion of another body.

Analysis

This is an exploratory study; therefore, the sample size is not based on a sample size calculation. The sample size is the same as the previous study which showed significant effects of the VR-Wound simulation with 24 participants. Statistical analyses were performed using the SAS Version 9.4 (SAS Institute INC., Cary, NC, USA)

Each parameter was analysed with a mixed-model analysis of covariance with treatment, period, condition (if applicable), time and treatment by time, condition by time (if applicable) and treatment by condition by time (if applicable), random factors participant, participant by treatment and participant by time and the average prevalue as covariate.

For wound specific parameters (VAS focus, VAS realism, and VAS unpleasantness of the wound), the VR setting effect and its interactions are not calculated since there is only one VR-Wound setting per timepoint and there are no degrees of freedom left.

The Kenward-Roger approximation was used to estimate denominator degrees of freedom, and model parameters were estimated using the restricted maximum likelihood method.

The general treatment effect and specific contrasts were reported with the estimated difference and the 95% CI, the least square mean (LSM) estimates, and the p-value. Graphs of the LSM estimates over time by treatment were presented with 95% CI as error bars and change from baseline LSM estimates.

The following contrasts are calculated within the models: Diazepam – Placebo. And where applicable: Diazepam – Placebo within no VR; Diazepam – Placebo within VR-Neutral; Diazepam – Placebo within VR-Wound; Diazepam – Placebo within VR-Neutral as first; Diazepam – Placebo within VR-Neutral as second; Diazepam – Placebo within VR-Wound as first; Diazepam – Placebo within VR-Wound as second. For the electric stair PDT and PTT, also: VR-Neutral – no VR within Placebo; VR-Wound – VR-Neutral within Placebo; VR-Neutral – no VR within Diazepam; VR-Wound – VR-Neutral within Diazepam. The results were not corrected for multiple testing.

RESULTS

Participants

A total of 24 healthy male participants were enrolled in the study. None of the participants discontinued participation or were excluded from the analysis (age mean (SD) is 22.0 (2.4), range 18 – 28, and BMI of 23.6 kg/m^2 (2.8),

range 19.8 – 29.3). Due to a technical error, VR-Neutral simulation data of the first visit for two participants was lost. Few adverse events (AE) were recorded. All AEs were mild, transient and confirm the known effects of diazepam at this dose level.

Pain thresholds

Table 1 presents the least square means of pain thresholds derived from the statistical model, along with the contrasts between placebo and diazepam. Additional contrasts within treatments are provided in Table 2.

Diazepam vs placebo

Compared to placebo, diazepam significantly increased the PDT for the VR-Wound condition (ED = 25.2%, 95% CI: 2.4 to 53.2, p = .030). However, diazepam did not significantly affect the PDT for the pain task outside VR (ED = -3.6%, 95% CI: -21.2 to 18.0, p = .715) or the neutral VR simulation (ED = -1.3%, 95% CI: -19.3 to 20.7, p = .897) (See Figure 2). Diazepam also did not significantly affect the PTT in any of the conditions.

Effects of VR on pain thresholds

During the placebo study period, the neutral VR simulation had no significant effect on the PDT or PTT compared to no VR (PDT: ED = 6.9%, 95% CI: -1.8 to 16.3, p = .123; PTT: ED = 1.7%, 95% CI: -0.3 to 3.7, p = .089) (see Table 2). However, in the diazepam study period, the neutral VR simulation significantly increased both PDT and PTT compared to the pain task outside of VR (PDT: ED = 9.4%, 95% CI: 0.6 to 19.1, p = .037; PTT: ED = 2.1%, 95% CI: 0.2 to 4.2, p = .033) (see Table 1).

When comparing the VR-Wound condition to the VR-Neutral condition, we observed a significant decrease in PDT during the placebo period (ED = -13.8%, 95% CI: -20.7 to -6.3, p < .001) (see Table 2). In contrast, during the diazepam study period, the VR-wound condition significantly increased PDT (ED = 9.4%, 95% CI: 0.5 to 19.0, p = .037). No effect on PTT was observed during the placebo period, but following diazepam administration, PTT was significantly increased in the VR-Wound condition compared to the VR-Neutral condition (ED = 2.9%, 95% CI: 1.0 to 4.9, p = .003).

Questionnaires

There were no significant differences in VAS ratings for unpleasantness or intensity across any of the VR conditions or treatment effects. The VAS ratings for focus, realism, and unpleasantness of the wound were similar between the two treatments. The McGill questionnaire showed no treatment effects for either the total score or the subdomains (sensory, affective, present pain intensity, and pain score).

All questions related to the level of embodiment remained relatively stable across assessments and treatments, except for one item. The question assessing the feeling of movement control of the virtual body showed a significant increase for diazepam compared to placebo.

DISCUSSION

This study is the first to demonstrate that augmentation of the pain experience induced by an enhanced virtual reality simulation that was integrated into a nociceptive pain test battery, can be attenuated using an anxiolytic drug. Here we demonstrated that this reduction in PDT was significantly attenuated when the participant received a single oral dose of diazepam. In fact, the PDT in the VR-Wound condition was significantly higher than the PDT in the neutral VR condition. The VR-PainCart successfully isolated the affective pain component from changes in nociception, as virtual reality raised the pain detection threshold but did not affect the pain tolerance threshold after administration of diazepam. With this study, we replicated the previous findings that the addition of a wound in VR on the location of the painful stimuli significantly decreased the PDT compared to not presenting this wound. Additionally, we reproduce previous findings where a neutral VR simulation increased the PDT, but not the PTT.[8]

The administration of diazepam significantly mitigated the reduction in pain detection threshold to electrical pain caused by the VR-Wound simulation (See Figure 2). This finding not only confirms that diazepam alters the pain experience and can be demonstrated to have analgesic properties in this model. Additionally, the lack of effect on the other two VR conditions (no-VR and VR-Neutral) confirms that this effect is isolated from nociception, which was not influenced by diazepam. This isolated effect builds upon the hypothesis of a precious study assessing the effect of diazepam on pain. There, using a pressure cuff on the upper arm concluded that the effect of diazepam should be assigned to the emotional aspect of pain and not a change in nociception.[16]

There was no effect of diazepam or VR on the PTT. Pain-related emotions, cognitive interpretation, and subjective catastrophizing of future consequences can be triggered by the immediate sensory unpleasantness of a

pain stimulus.[17] Whilst diazepam could create an emotional disconnect between the two during onset of pain, pain-related emotions during the experience may already be active once the PTT is reached.

Participants did not experience the pain any differently following administration of diazepam as recorded with the different questionnaires, even though the PDT was elevated. This shows that the pain task using the VR-PainCart cannot be replaced with commonly used methods such as the McGill questionnaire, which includes an affective subscale. The lack of change in the McGill short form, also shows that the sensory characteristics of the stimuli are unaffected, maintaining the realism of the stimulus.

The level of embodiment experienced by the participants remained stable and was mostly unaffected by diazepam, except for one parameter: control of movement seemed to be improved by diazepam compared to placebo. One explanation could be related to reduced general motion activity caused by benzodiazepines.[18] Diazepam may also have led to a slight reduction of CNS processing resulting in less observations of the lag time and therefore a sense of better control.

While several significant differences in PDT and PTT were found between different VR condition, no effects were observed on the subjective rating scales for pain intensity and pain unpleasantness. This is contrary to the previous findings in the first VR-PainCart study were participants indicated a higher pain intensity and unpleasantness under the VR-Wound condition.[8] The previously found effects might have been different due to the lack of blinding resulting in socially desirable answers. On the other hand, the number of questionnaires in this study was quite large with the addition of the McGill questionnaire, and as such the increased time between pain task and questionnaire may have influenced the responses. Therefore, selection of the questionnaires might improve reliability.

Two limitations of the study were (1) the relatively small sample size and (2) the inclusion criteria focussed on healthy young men. A backwards power calculation demonstrates that the study was sufficiently powered to detect a difference of 1.076 with a standard deviation 1.27 (results of VR-Wound simulation) in PDT between diazepam and placebo treatment with power of 80% and alpha of 0.05. Because the study was only performed in men, we cannot generalize the results of this study to other populations (e.g., different age, gender, or personality characteristics). It may be possible that healthy women or elderly respond differently to the VR-PainCart and show different modulating effects. The personality characteristics and perhaps different emotional state of (chronic) pain patients[19] make it difficult to predict the modulating effect of the VR-PainCart on their pain experience. Additionally, the response to the VR challenge on the emotional processing might be absent or over-active in different clinical populations, e.g. in patients with pain and central sensitization.[20-22] Future research should include patients to identify the clinical relevance of the biomarker and to provide a predictive validity in early drug development.

The findings of this validation study provide valuable insights into the potential of the PainCart, a pain test battery already known for its high repeatability and sensitivity to nociceptive tests. Now, with the addition of VR simulations targeting the affective aspect of pain, the VR-PainCart addresses the need for precise pharmacodynamic biomarkers that are critical for establishing proof-of-pharmacology or target engagement. This is particularly significant as there is a growing demand for effective new analgesics, though recent efforts in drug discovery have unfortunately not resulted in effective treatment of chronic pain.

REFERENCES

1 D. D. Price and S. W. Harkins, 'The affective-motivational dimension of pain A two-stage model,' APS Journal, vol. 1, no. 4, pp. 229–239, 1992, doi: 10.1016/1058-9139(92)90054-G.

2 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, pp. 55–74, 2015, doi: 10.1146/annurev-pharmtox-011613-135918.

3 A. Heiberg Agerbeck et al., 'Validity of Current Assessment Tools Aiming to Measure the Affective Component of Pain: A Systematic Review,' Patient Relat Outcome Meas, vol. Volume 12, pp. 213–226, 2021, doi: 10.2147/prom.s304950.

4 J. Hay, P. Okkerse, G. Van Amerongen, and G. J. Groeneveld, 'Determining pain detection and tolerance thresholds using an integrated, multi-modal pain task battery,' Journal of Visualized Experiments, vol. 2016, no. 110, 2016, doi: 10.3791/53800.

5 P. Okkerse et al., 'The use of a battery of pain models to detect analgesic properties of compounds: a two-part four-way crossover study,' Br J Clin Pharmacol, vol. 83, no. 5, pp. 976–990, 2017, doi: 10.1111/bcp.13183.

6 P. S. Siebenga et al., 'Reproducibility of a battery of human evoked pain models to detect pharmacological effects of analgesic drugs,' Eur J Pain, 2019, doi: 10.1002/ejp.1379.

7 K. Talbot, V. J. Madden, S. L. Jones, and G. L. Moseley, 'The sensory and affective components of pain: are they differentially modifiable dimensions or inseparable aspects of a unitary experience? A systematic review,' Br J Anaesth, vol. 123, no. 2, pp. e263–e272, 2019, doi: 10.1016/j.bja.2019.03.033.

8 I. Koopmans, R.-J. Doll, M. Hagemeijer, R. van Barneveld, M. de Kam, and G. J. Groeneveld, 'The impact of a virtual wound on pain sensitivity: insights into the affective dimension of pain,' Frontier in Pain Research, vol. 6, no. February, pp. 1–6, 2025.

9 W. D. Hutchison, K. D. Davis, A. M. Lozano, R. R. Tasker, and J. O. Dostrovsky, 'Pain-related neurons in the human cingulate cortex,' Nat Neurosci, vol. 2, no. 5, pp. 403–405, 1999, doi: 10.1038/8065.

10 A. Galambos et al., 'A systematic review of structural and functional MRI studies on pain catastrophizing,' J Pain Res, vol. 12, pp. 1155–1178, 2019, doi: 10.2147/JPR.S192246.

11 S. E. Murphy, C. Downham, P. J. Cowen, and C. J. Harmer, 'Direct effects of diazepam on emotional processing in healthy volunteers,' Psychopharmacology (Berl), vol. 199, no. 4, pp. 503–513, 2008, doi: 10.1007/s00213-008-1082-2.

12 H. Friedman et al., 'Pharmacokinetics and pharmacodynamics of oral diazepam: Effect of dose, plasma concentration, and time,' Clin Pharmacol Ther, vol. 52, no. 2, pp. 139–150, 1992, doi: 10.1038/clpt.1992.123.

13 J. W. Ditre, B. W. Heckman, E. L. Zale, J. D. Kosiba, and S. A. Maisto, 'Acute analgesic effects of nicotine and tobacco in humans: A meta-analysis,' Pain, vol. 157, no. 7, pp. 1373–1381, 2016, doi: 10.1097/j.pain.0000000000000572.

14 H. J. Hijma, P. S. Siebenga, M. L. De Kam, and G. J. Groeneveld, 'A Phase 1, Randomized, Double-Blind, Placebo-Controlled, Crossover Study to Evaluate the Pharmacodynamic Effects of VX-150, a Highly Selective NaV1.8 Inhibitor, in Healthy Male Adults,' Pain Medicine (United States), vol. 22, no. 8, pp. 1814–1826, 2021, doi: 10.1093/pm/pnab032.

15 R. Melzack, 'The short-form McGill pain questionnaire,' Pain, vol. 30, no. 2, pp. 191–197, 1987, doi: 10.1016/0304-3959(87)91074-8.

16 C. R. Chapman and B. W. Feather, 'Effects of diazepam on human pain tolerance and pain sensitivity,' Psychosom Med, vol. 35, no. 4, pp. 330–340, 1973, doi: 10.1097/00006842-197307000-00007.

17 P. Rainville, Q. V. H. Bao, and P. Chrétien, 'Pain-related emotions modulate experimental pain perception and autonomic responses,' Pain, vol. 118, no. 3, pp. 306–318, 2005, doi: 10.1016/j.pain.2005.08.022.

18 S. J. De Visser, J. P. Van Der Post, P. P. De Waal, F. Cornet, A. F. Cohen, and J. M. A. Van Gerven, 'Biomarkers for the effects of benzodiazepines in healthy volunteers,' Br J Clin Pharmacol, vol. 55, no. 1, pp. 39–50, 2003, doi: 10.1046/j.1365-2125.2002.t01-10-01714.x.

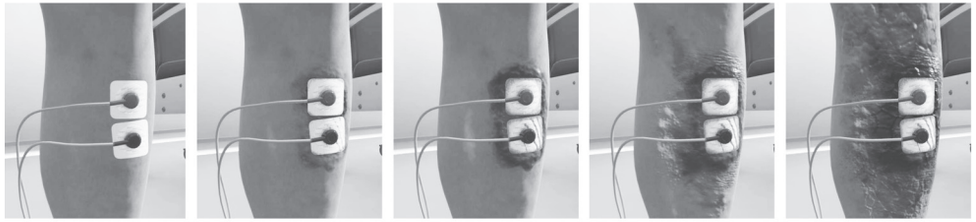
19 B. Naylor, S. Boag, and S. M. Gustin, 'New evidence for a pain personality? A critical review of the last 120 years of pain and personality,' Scand J Pain, vol. 17, pp. 58–67, 2017, doi: 10.1016/j.sjpain.2017.07.011.

20 H. B. Vaegter and T. Graven-Nielsen, 'Pain modulatory phenotypes differentiate subgroups with different clinical and experimental pain sensitivity,' Pain, vol. 157, no. 7, pp. 1480–1488, 2016, doi: 10.1097/j.pain.0000000000000543.

21 V. Oliva, R. Gregory, J. C. W. Brooks, and A. E. Pickering, 'Central pain modulatory mechanisms of attentional analgesia are preserved in fibromyalgia,' Pain, vol. 163, no. 1, pp. 125–136, 2022, doi: 10.1097/j.pain.0000000000002319.

22 A. Gil-Ugidos, A. Vázquez-Millán, N. Samartin-Veiga, and M. T. Carrillo-de-la-Peña, 'Conditioned pain modulation (CPM) paradigm type affects its sensitivity as a biomarker of fibromyalgia,' Sci Rep, vol. 14, no. 1, pp. 1–11, 2024, doi: 10.1038/s41598-024-58079-7.

FIGURE 1 The neutral simulation did not include the burn wound (left). During the wound simulation, the wound around the electrodes increased in size and severity (from left to right).



(for color image see page 85)

FIGURE 2 Estimated difference between diazepam and placebo treatment for the Electrical Stair PDT overall, without VR, within VR-Neutral simulation and VR-Wound simulation.

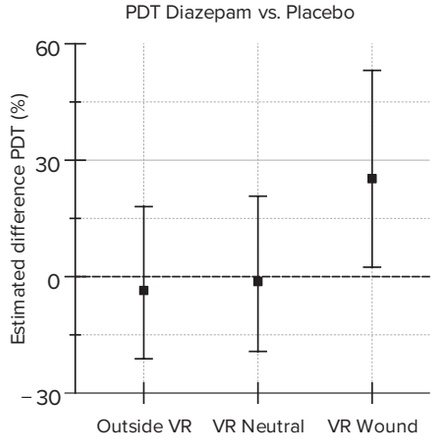


TABLE 1 Pain thresholds recorded in mA for each VR-condition and both treatments including the contrast between placebo and diazepam with the percentage of change in thresholds.

	No VR		VR-Neutral		VR-Wound	
	PDT	PTT	PDT	PTT	PDT	PTT
Placebo (mA)	4.63	19.95	4.95	20.29	4.26	20.08
Diazepam (mA)	4.46	20.34	4.88	20.78	5.34	21.38
Placebo vs diazepam	-3.6% (-21.2; 18.0) p=.715	1.9% (-7.3; 12.0) p=.676	-1.3% (-19.3; 20.7) p=.897	2.4% (-6.8; 12.5) p=.608	25.2% (2.4; 53.2) **p=.030**	6.5% (-3.1; 17.0) **p=.179**

VR: Virtual Reality. PDT: Pain Detection threshold. PTT: Pain Tolerance Threshold. Bold indicates statistical difference (p<.05).

TABLE 2 Statistical contrasts (% of change) within treatment.
VR: Virtual Reality. PDT: Pain Detection Threshold.

Contrast	Placebo		Diazepam	
	PDT	PTT	PDT	PTT
No VR – VR-neutral	6.9% (-1.8; 16.3) p=.123	1.7% (-0.3; 3.7) p=.089	9.4% (0.6; 19.1) **p=.037**	2.1% (0.2; 4.2) **p=.033**
VR-neutral – VR-wound	-13.8% (-20.7; -6.3) **p<.001**	-1.1% (-3.0; 0.9) p=.277	9.4% (0.5; 19.0) **p=.037**	2.9% (1.0; 4.9) **p=.003**

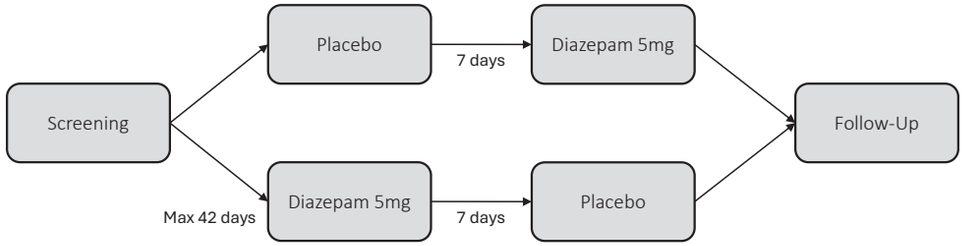
PTT: Pain Tolerance Threshold. Bold indicates statistical difference (p<.05).

SUPPLEMENTS

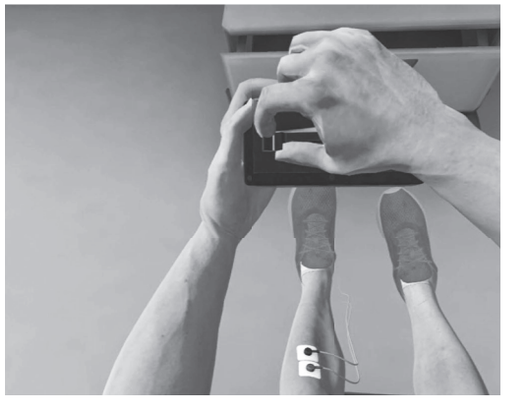
SUPPLEMENTARY TABEL S1 Overview of adverse events

Summary of Number of Subjects with TEAEs by Treatment, SOC, PT and Severity, Safety population						
	Diazepam (N=24)			Placebo (N=24)		
System Organ Class/ Preferred Term	Mild N	Moderate N	Severe N	Mild N	Moderate N	Severe N
ANY EVENTS	12	-	-	7	-	-
GASTROINTESTINAL DISORDERS	-	-	-	1	-	-
Gastrooesophageal reflux disease	-	-	-	1	-	-
GENERAL DISORDERS AND ADMINISTRATION SITE CONDITIONS	4	-	-	1	-	-
Fatigue	4	-	-	1	-	-
INFECTIONS AND INFESTATIONS	-	-	-	1	-	-
Nasopharyngitis	-	-	-	1	-	-
NERVOUS SYSTEM DISORDERS	7	-	-	4	-	-
Dizziness	1	-	-	-	-	-
Sedation	-	-	-	1	-	-
Somnolence	6	-	-	3	-	-
PSYCHIATRIC DISORDERS	-	-	-	2	-	-
Flat affect	-	-	-	1	-	-
Insomnia	-	-	-	1	-	-

SUPPLEMENT FIGURE S1 Schematic overview of study design



SUPPLEMENT FIGURE S2 View of the participant during the electrical stair assessments with Virtual Reality.



CHAPTER 6

POWERJAR, A NOVEL DEVICE FOR QUANTITATE ASSESSMENT OF JAR OPENING: AN EXPLORATORY TECHNICAL VALIDATION STUDY.

Ingrid Koopmans*[1,2], Ingrid van den Heuvel*[1], Anil Tarachandani[3], Marcus van Diemen[1,2], Gary Kath[4], Ajay Verma[3], Geert Jan Groeneveld[1,2], and Robert-Jan Doll[1,2]

1 Centre for Human Drug Research, Leiden, The Netherlands.
2 Leiden University Medical Center, Leiden, The Netherlands.
3 Biogen, Cambridge, USA
4 Design-To-Prototype, LLC, Scotch Plains, USA

ABSTRACT

Grip strength, assessed with a handheld dynamometer, is commonly used to monitor disease progression and evaluate healthcare interventions. However, grip strength alone does not fully reflect the complexity of daily tasks, which require a combination of strength, coordination, and fine motor control. This study introduces the PowerJar, a novel device designed to quantify grip and rotational forces during simulated jar-opening tasks, providing a more complete assessment of hand function.

This observational study included healthy volunteers and patients with neuromuscular diseases. Healthy participants performed PowerJar tasks four times at 60-minute intervals, while patients performed tasks once during each of two visits. Usability was assessed through a questionnaire after each visit. Grip strength was measured using both the PowerJar and a handheld dynamometer. Repeatability was evaluated by assessing the consistency of PowerJar measurements across multiple sessions.

The study included 62 healthy participants and 18 patients. Usability assessments indicated that the PowerJar tasks were reasonably challenging but manageable. A strong positive correlation was found between handheld dynamometer and PowerJar measurements, although the dynamometer recorded on average higher grip strength values. Repeatability analysis showed moderate to good repeatability for both grip and angle parameters.

The PowerJar demonstrated usability across different populations and provided additional information beyond standard grip strength assessments. The device's moderate repeatability suggests its potential for early-phase drug development and clinical trials. However, further research is needed to explore the PowerJar's sensitivity to changes in neuromuscular diseases and responses to interventions.

INTRODUCTION

Neuromuscular diseases often lead to progressive impairments in muscle strength and functional independence, reducing the patients' quality of life. Clinical research frequently uses grip strength assessed with a handheld dynamometer to monitor disease progression. Additionally, it is used to evaluate the efficacy of healthcare interventions such as drugs and physical therapy.[1-5] Grip strength is an appealing procedure to be used in both clinical and research settings due to its simplicity, reliability, and validity.[5]

While grip strength is often used to assess muscle function, it provides an isolated measure of force which does not fully reflect the complexity of daily life tasks. These tasks often require a combination of strength, coordination, and fine motor control.[6] As a result, grip strength does not fully reflect the challenges patients face in their daily activities or the true impact of neuromuscular disease. This discrepancy can lead to situations where grip strength improves, yet patients might experience no significant improvement in daily functioning. In drug development, enhancing daily life quality is essential for regulatory approval. Consequently, quality of life is frequently used as a clinical outcome measure in registration trials to assess the effect of treatments on patients' everyday activities.

To increase efficiency and minimize costs and risks, predictive biomarkers are often utilized in early-phase trials to identify potential outcomes before progressing to registration trials. Such a biomarker should quantify a daily task that impacts daily life of patients. Functional tasks, such as opening jars or bottles, represent a critical aspect of daily living and require both grip strength and rotational force. These tasks are significant challenges for individuals with neuromuscular diseases, and easy relatable to daily life function. Approximately 14% of the elderly were found unable to open a screw cap bottle containing their medication.[7]

Unlike isolated strength measurements, assessing the ability to perform such tasks can provide deeper insights into the daily life burden of the disease and the efficacy of (drug) interventions. Earlier attempts to assess daily life activities considering hand function resulted in the Jebsen hand function test and the Duroz hand index. These assessments include tying shoelaces, cutting putty with knife and fork, manipulating coins into a slot or pouring a glass of water and showed differences between healthy volunteers and patients suffering from conditions like stroke and arthritis.[8-10] The outcome parameters of these tests are not ideal as these often

rely on observer ratings, making them dependent on individual observers. Additionally, these tests are labour-intensive and have relatively low resolution compared to the possibilities of computerized tests.

This contrasts with biomarkers used in clinical trials as part of early phase drug development, which should be quick, easy to assess, sensitive, and repeatable to ensure its reliability in placebo-controlled clinical trials with possible small dose-dependent effects. To bridge this gap in biomarkers, the PowerJar was developed, a novel device designed to quantify grip and rotational forces during simulated jar-opening tasks. By replicating this daily activity, the PowerJar provides an opportunity to measure outcomes that are more reflective of everyday challenges and closer related to clinical outcome assessments collected in drug registration trials.[11]

This study addressed two objectives. First, we evaluated the usability of the PowerJar device in both healthy volunteers and patients with neuromuscular diseases. The usability was evaluated by determining whether the device can be effectively utilized and integrated into clinical studies. Furthermore, the PowerJar's capability to capture grip and rotational forces in a standardized manner was assessed. Second, we investigated the repeatability and reliability of the PowerJar measurements in healthy volunteers. Repeatability was assessed by assessing the consistency of measurements across four sessions in one visit. Reliability was assessed by comparing grip strength assessments of the PowerJar to those recorded with a handheld dynamometer. Combined with usability, the repeatability and reliability were used for an evaluation of the endpoints generated by the PowerJar as possible biomarkers in early phase drug development clinical trials.

METHODS

This was a two-part, observational study in healthy volunteers and patients with a neurological disease and self-reported hand weakness. The study was approved by the BEBO Ethics Committee located in Assen (The Netherlands). All participants provided written informed consent prior to any study related activities. The first part of the study, part A, included healthy male and female participants who performed all tasks of the PowerJar (see methods section) four times at 60-minute intervals. The second part, part B, included male and female patients with a neurological disease and self-reported hand weakness who performed all tasks of the PowerJar once during each of two visits. There were at least 6 days between the visits.

Participants

Part A: Healthy male and female participants were invited to join the study and were assigned to one of three age groups. Each of the three age groups (between 21-40, 41-60, and 61-80 years old) consisted of 10 male and 10 female participants. After signing of the informed consent form, participants underwent medical screening consisting of a physical examination and recording of their medical history, height, and weight. When included in the study, participants were not allowed to consume alcohol within 24 hours of the study visit, use any form of medication or supplements during the study, and had to refrain from heavy physical exercise at least 48 hours prior to the study visit.

Part B: Male or female patients with a neurological disease, including Myasthenia Gravis, Myotonic Dystrophy, Parkinson's Disease, and Inclusion Body Myositis were invited for a brief medical screening. After signing of the informed consent form, participants underwent a brief medical screening consisting of recording medical history and current medication. Eligibility was limited to patients who self-reported hand weakness, assessed through the question: 'Do you have difficulties opening a jar, or has this task become more challenging than it was in the past?'. Patients with medical conditions that might interfere with the study results and/or inability to use the PowerJar device were excluded.

Assessments

Usability assessment of PowerJar

The usability of the PowerJar device was assessed by observing participants as they interacted with the device. Furthermore, after completing all PowerJar assessments, participants completed a questionnaire addressing fatigue and discomfort. Fatigue was assessed using a 15-point Likert scale, where 6 indicated very light fatigue and 20 indicated very heavy fatigue. Participants were asked to specify the task that induced the fatigue and the location in their body where they experienced it. Additionally, the participants reported any discomfort experienced with the PowerJar and if applicable the location of the discomfort.

Handheld Dynamometer

A handheld dynamometer (Smart Hand Dynamometer, Jamar) was used as the gold standard to measure maximum voluntary grip strength (i.e., the maximum amount of grip force exerted by the participant in kg). Assessments

were performed as recommended by Roberts et al.[3] The participant was seated in a chair with their forearms on the chair's fixed armrests, ensuring that the wrists were able to be moved freely just beyond the ends of the armrests. The participant's wrists were to be maintained in a neutral position with the thumb facing upwards. The participant held the dynamometer in the tested hand. See Figure 3 for a side view of the assessment. The participant was then encouraged to squeeze the dynamometer as tightly as possible. This was repeated three times for each hand.

PowerJar

The PowerJar (UsinLife LLC, Edison, NJ, USA) combines resistive torque producing electronics with an isometric grip measurement device that is shaped in a jar-and-lid configuration (Figure 1). The lid of the PowerJar has a diameter of 6.5cm and was designed to have similar dimensions as the lid of a typical jar. The PowerJar is connected to a computer that enables measurement of the grip force with a resolution of 0.06 kg and the rotation of the lid with a resolution of 0.01 degrees with a sample rate of up to 13Hz. Additionally, the computer is used to configure the resistive torque of the lid Nm. The resistive torque is generated by an electric motor combined with adjustable gearings rated to generate 0 – 5.7 Nm with a resolution of 0.01 Nm. The PowerJar kept the resistive torque at a constant level throughout the measurement.

As part of the PowerJar assessments, electromyography (EMG) was recorded with surface electrodes. As we found these data not to be useful in the assessment of muscle fatigue these data will not be further discussed in this report.

POWERJAR TASKS Participants were to complete a total of 6 different tasks using the PowerJar. For all these tasks, participants were to use both hands. In part A, the participant's dominant hand was put on the lid and the non-dominant hand was placed around the body of the device, with the thumb wrapped around one side and the remaining fingers wrapped around the opposite side (See Figure 2). In part B, the participant was free to choose which hand is placed on the lid and which hand was placed on the body. The observer used the connected laptop to run six different tasks. The laptop was also connected to an external monitor where the participants could receive live feedback on their performance. The exerted grip force in kg and rotation of the lid in degrees were sampled with a non-uniform sampling frequency ranging between 7.5 and 13 Hz.

Prior to the assessments, the participant was shown how to use the PowerJar. Additionally, the highest resistive torque a participant could withstand was determined by stepwise increasing the resistive torque. In case the highest possible resistive torque that the participant could withstand was lower than the expected highest level of resistive torque based on the age group (see Table 1), this resistive torque was used for the next tasks.

In each round of assessments, participants performed six distinct tasks. Tasks 2, 3, and 4 were designed to induce fatigue which should have been primarily assessed with the EMG analysis. However, upon data review and exclusion of the EMG results, these tasks were found less informative as addition to the standard grip strength assessment for the aim of this research. Therefore, the data of tasks 2, 3, and 4 were excluded from the primary analysis but added as supplementary materials.

TASK 1 Angle Control (20 degrees). Participants were instructed to rotate the lid to an angle of 20 degrees and maintain this position for as long as possible. Real-time feedback was provided on a monitor, allowing participants to track the rotational position of the lid. The non-rotating hand was placed on the body of the jar to provide stability, applying grip force when necessary. Participants were instructed to release the lid if the angle dropped below 15 degrees. This task was performed at three resistive torque levels (see Table 1).

TASK 2 Rhythmic Rotation (15–25 degrees). Participants rotated the lid between 15 and 25 degrees in sync with a 2 Hz metronome. This task was performed at three resistive torque levels (see Table 1) for a duration of 30 seconds.

TASK 3 Rhythmic Grip. Participants were instructed to grip the body of the jar as hard as possible and then release it in sync with a 2 Hz metronome for a duration of 30 seconds.

TASK 4 Rapid Grip-and-Release. Participants gripped and released the body of the jar as hard and as quickly as possible for a duration of 30 seconds.

TASK 5 Rapid Opening and Closing (40 degrees). Participants rotated the lid between 0 (open) and 40 (closed) degrees as frequently as possible within 30 seconds. Real-time feedback was displayed on an external monitor to guide rotational positioning. Participants used their non-rotating hand to stabilize the bottle, applying grip force when necessary. This task was performed at three resistive torque levels (see Table 1).

TASK 6 Maximal Grip. Participants were instructed to grip the body of the jar as hard as possible and maintain this grip for as long as possible. Participants were allowed to release the jar when their grip strength dropped below 75% of the initial maximum or after reaching a 60-second time limit.

Data analysis

Handheld Dynamometer

Participants performed three maximum grip strength measurements per hand using the handheld dynamometer. Only the highest value obtained from the non-dominant hand, which was the hand used during the maximal PowerJar grip assessment, was used for further analysis.

PowerJar

DATA PROCESSING To ensure a constant sampling rate, data is resampled to the average sampling rate of 11.11 Hz. After resampling, missing data points were filled using a linear interpolation. Additionally, the grip data was smoothed using a moving average with a window of 6 samples (i.e., approximately 0.54s).

PARAMETERS The parameters listed in Table 2 were computed for the assessment during the analysis. For Task 1, the computation window was defined as follows: from the point when the angle exceeded 15 degrees to when the angle dropped below 15 degrees. The following grip features were calculated: area under the curve, maximum value, linear term and root mean square error of a first order polynomial fit. The angle of the lid was characterized by the linear term and root mean square error of a first order polynomial fit. Furthermore, the total hold time was determined.

For Task 5, all features were computed between the first and last closing cycle of the lid. The following grip features were calculated: dominant frequency, frequency dominance, and power of the dominant frequency. Angle features included the quadratic and linear terms of a second order polynomial fit, dominant frequency, frequency dominance, and power at the dominant frequency. The combination of grip and angle was characterized by the correlation between their values and the number of successful openings. Peak-to-peak features consisted of the linear term of a first order polynomial fit, intercept, and root mean square error.

For Task 6, the computation window was defined from the point of maximum grip strength to when the grip strength decreased to 75% of the maximum value. The following grip features were calculated: area under the curve, maximum value, and linear term of a polynomial fit. Furthermore, the total hold time was determined.

Statistical analysis

Statistical analysis was performed in R statistical software (v4.4.1; R Core Team 2025) using the DPLYR (v1.1.4), TIDYR (v1.3.1), LME4 (v1.1-36), CAR (v3.1-3), EMMEANS (v1.10.6) and LMERTEST (v3.1-3) package.

Comparison of grip strength between the handheld dynamometer and the PowerJar

Grip strength performance measured with the handheld dynamometer was compared to the maximum grip strength obtained during Task 6 of the PowerJar assessments. Participants in Part B were free to choose whether to place their weakened or not-weakened hand on the lid and body for each assessment. However, hand placement was not registered for part B, therefore this analysis was limited to healthy participants (Part A). A Pearson correlation coefficient was calculated to assess the relationship between the two measurements, and a paired two-sided t-test was conducted to evaluate differences between the devices.

Repeatability and potential sensitivity task 1, 5 and 6

To assess the repeatability and potential sensitivity, we fitted a random intercept model with measurement as fixed effect for each parameter. Type-III F-statistics were used to assess statistical significance of measurement as fixed effect ($\alpha = .05$). To estimate the repeatability, we derived the intraclass correlation coefficient (ICC) based on the variance components by dividing the between-subject variance by the sum of the between-subject and within-subject variance. The repeatability is considered poor for ICC values below 0.50, moderate for values between 0.50 and 0.75, good for values between 0.75 and 0.90, and excellent for values above 0.90.[12] Furthermore, to assess the potential sensitivity, minimum detectable effect (MDE) values were calculated. The MDE was then calculated by multiplying the effect size by the pooled standard deviation (i.e., the square root of the sum of the within- and between-subject variance). The effect size used to calculate the MDE

was based on a paired sample t-test with a power of .80, a significance level of 5% (α = .05), and a sample size of 15 (a typical sample size for a clinical).

Data from Part B was excluded from this analysis. Unlike Part A, patients in Part B were allowed to switch hands between visits, and no quantification of symptom severity was available for each assessment day. As disease progression or variability could influence repeatability measurements, this part of the dataset was deemed unsuitable for inclusion in this analysis.

Resistive torque correction per age group

The resistive torque levels (Table 1) participants had to withstand were to correct for age. To determine whether this age-based resistive torque levels gave the desired results, a random intercept model was fitted with age group and measurements as fixed effects. To ensure normality, the values of the parameters Slope Angle (Task 1) and Quadratic Term Angle (Task 5) were exponentially transformed prior to model fitting. The parameter values were compared between each age group. Due to the explorative nature of this study, no p-value correction was applied.

RESULTS

Participants

In part A, 73 participants were screened, and 62 participants were included in the three pre-defined groups. The main reason for exclusion during screening was a disease or condition that affected upper limb function (n=11). All participants completed the full study, but two healthy female participants did not complete all tasks due to discomfort performing the tasks (one in the youngest age group and one in the middle age group) and were therefore excluded from analysis.

In part B, 21 participants were screened, and 18 patients were included (9 female). Three patients were excluded; two because of multiple interfering diseases in upper limbs and one because the disease did not affect their hands. There were 3 diseases present in the patient group: Parkinson's disease (n=8), Myasthenia Gravis (n=5) and body myositis (n=5). Two participants discontinued participation after the first visit, therefore only the data of their first visit were included in the analysis.

The demographics of the analysis population is summarized in Table 3.

Usability of the PowerJar

The questionnaire on usability was added later to the study, resulting in some missing data. A total of 46 healthy participants completed the questionnaire, and they reported an average (+/- standard deviation) fatigue level of 13.0 (+/- 1.5), which corresponds to reasonably heavy tasks. Task 5 was identified as the most fatiguing activity (n = 32), with fatigue primarily located in the arm (n = 14), hand (n = 10), and thumb (n = 9). Among healthy participants, 16 participants reported no discomfort from using the PowerJar. However, 14 experienced minor thumb pain, 8 developed the beginnings of a blister, and 4 reported hand discomfort during or immediately after the assessments.

Patients reported in their first visit an average fatigue level of 12.9 (+/- 1.5), with task 5 also inducing the most fatigue (n = 13). Fatigue was most commonly experienced in the thumb (n = 9), followed by the hand (n = 4), wrist (n = 3), and arm (n = 2). Regarding discomfort, 6 patients reported none, 8 experienced thumb discomfort, and 3 reported discomfort in other locations.

During their second visit, patients reported an average fatigue level of 12.6 (+/- 1.6). As in the first visit, task 5 was identified as the most fatiguing activity (n = 7). Fatigue was experienced in the hand (n = 6), followed by the thumb (n = 4), arm (n = 3) and other locations (n = 5). In terms of discomfort, 9 patients reported none, 3 experienced thumb discomfort, and 7 reported discomfort in other locations.

Missing and excluded PowerJar data

The number of missing measurements for the highest successfully completed resistive torque level per task were 1, 1, and 0 for task 1, task 5, and task 6, respectively. These measurements were missing due to technical reasons. Several additional measurements were excluded from data analysis to maintain validity of the dataset:

Task 1 (239 initially available measurements): Exclusions were made if participants failed to achieve a 15-degree rotation throughout the measurement (n = 22) or if the measurement was abruptly terminated without the angle dropping below 15 degrees (n = 6). Additionally, measurements with smoothed grip force values ≤ 1 kg were excluded, as these were considered measurement noise (n = 17). After these exclusions, a total of 194 measurements remained for analysis. See Figure S1 for an example of the data from an assessment.

Task 5 (239 initially available measurements): One measurement was excluded due to fewer than two complete opening-closing cycles (n = 1). Grip force data with smoothed values ≤ 1 kg were also excluded as noise (n = 12). After these exclusions, a total of 226 measurements remained for analysis. See Figure S2 for an example of the data from an assessment.

Task 6 (240 initially available measurements): Measurements shorter than 1.5 seconds (n = 8) or those ending abruptly without a sufficient drop in grip strength below 75% of the participant's maximum (n = 2) were excluded. No exclusions were necessary due to low grip force, as smoothed values ≤ 1 kg were not observed (n = 0). After these exclusions, a total of 230 measurements remained for analysis. See Figure S3 for an example of the data from an assessment.

Six healthy participants had difficulties completing measurements at the highest resistive torque levels, resulting in missing data. Consequently, the most challenging resistive torque level varied between participants, with some reaching the second level as their highest and others reaching the third level. To ensure consistent analysis, only data from the highest successfully completed resistive torque level for each participant was included in the analysis.

Comparison handheld dynamometer and PowerJar

Data from 57 healthy participants were analysed. Due to human error, data from 5 participants were not recorded for the handheld dynamometer. A strong positive correlation was found between the handheld dynamometer and PowerJar measurements (Pearson correlation coefficient: r = .79). The paired t-test indicated a significant difference between the two methods ($p < .001$), with a mean difference of 16.79 kg (handheld dynamometer larger than PowerJar). See Figure 4 for the scatterplot and boxplot of the results.

Repeatability Task 1, 5 and 6

All calculated ICCs are provided in the supplementary materials. A summary of the ICCs greater than 0.5 is presented in Table 4, the full tables are added in the supplements (Table S1, S2 and S3). For Task 1, the area under the grip curve, maximum grip, and hold time demonstrated moderate repeatability, with ICCs of 0.70, 0.66, and 0.59, respectively. In Task 5, two parameters, number of openings and dominant frequency, showed good repeatability, with ICCs of 0.79. The ratio of successful openings and the correlation between angle and grip demonstrated moderate repeatability, with ICCs of 0.67 and 0.64, respectively. Additionally, the dominant frequency and associated power of the grip, as well as the dominant frequency of the angle, exhibited moderate repeatability, with ICCs ranging from 0.51 to 0.59. For Task 6, maximum grip showed good repeatability and the area under the grip curve showed moderate repeatability, with ICCs of 0.87 and 0.71 respectively.

Resistive torque correction per age group

The correction performed on the resistive torque levels per age group aimed to result in comparable results between each group. In Table 5, the contrasts, 95% confidence intervals (CI) and their corresponding p-value are shown with at least one significant difference between groups (with the cut-off set at $p < .05$). The full tables for both Tasks 1 and 5 are added in as supplementary materials (Table S4). Three out of the seven parameters for task 1 showed a significant difference between age groups, including the Total Grip between 20-40 and 41-60 years (ED: -109 kg*s (-214.24, -5.46)), and the Maximum Grip between 20-40 and 61-80 years (ED: -4.01 kg (-6.38, -1.64)). Nine parameters of task 5 (total of 14) differ significantly between age groups. These nine include both grip and angle related parameters. For example, the dominant frequency of both grip (ED: -0.29 Hz (-0.57, -0.02)) and angle (ED: 0.45 Hz (0.11, 0.80)) showed a difference between the age groups 20-40 and 41-61 years. Whereas the Number of Openings showed a difference between all age groups (ED: 7.50 (0.56, 14.44), ED: 17.46 (10.52, 24.40), ED: 9.96 (3.02, 16.90).

DISCUSSION

In this study, we assessed the usability and repeatability of the PowerJar in a clinical setting. Usability was assessed in both healthy volunteers and patients with a neuromuscular disease, and repeatability was assessed in healthy volunteers only. The tasks selected to assess repeatability were chosen to resemble daily life activities that often require prolonged grip or the combined use of grip and torque. Both grip and torque during jar-opening tasks are often overlooked using handheld dynamometers to assess the grip strength. By including these more relevant tasks, the PowerJar provides valuable additional information beyond standard grip strength assessments.

Usability

Ninety-seven percent of the healthy participants and all patients completed participation in the study, which suggests that the device is easy to use and well-suited for various grip and rotation tasks. The average fatigue levels, 13.0 (i.e., reasonably heavy) for healthy participants and 12.8 for patients, indicated that the tasks were perceived as reasonably challenging. Combined with the fact that nearly all participants successfully completed participation in the study, it suggests that the tasks were challenging, but

not excessively demanding. These results are expected, as we aimed to design the tasks to induce fatigue, but not to induce failure to perform the test. Of all tasks, both groups reported the rapid opening and closing task (i.e., task 5) as the most fatiguing task, likely due to the repetitive dynamic movement it requires. Interestingly, while both groups experienced hand fatigue, arm fatigue was reported most frequently among healthy participants, while patients rarely reported arm fatigue. This may hint towards differences in techniques used by both groups. It might be expected that patients, especially those with hand weaknesses, would compensate by using larger muscle groups (e.g., the bigger arm muscles) to assist with the task. However, the data on perceived fatigue indicates that patients primarily relied on the thumb and the hand. This could imply that the rapid opening and closing task assessment can only provide insight in hand function and not accurately reflect lower and upper arm impairment.

Patients reported reduced discomfort during the second visit compared to the first. This may be explained by an increased device familiarity or by patients intentionally performing the task less intensively during the second visit. Regarding the device familiarity, the hand could have been positioned slightly different during the second visit to reduce discomfort. Furthermore, changes in disease symptoms (e.g., off-period for Parkinson) may have influenced the reported discomfort. However, neither the exact hand position nor disease symptom severity were recorded. Future studies would benefit from registering in more detail the hand positioning and symptom severity to improve data consistency and reliability in data collection. Despite this limitation, the PowerJar demonstrated its usability as task in the setting of early phase clinical trials across different populations.

Comparison with handheld dynamometer

The handheld dynamometer is frequently used tool in clinical studies to assess grip strength. The strong correlation between the PowerJar and the handheld dynamometer measurements confirms the PowerJar's potential as a reliable grip strength assessment tool. However, the dynamometer consistently recorded approximately 17 kg higher grip strength values than the PowerJar. This discrepancy is likely due to differences in hand and finger positioning between the two devices, which engage different muscles. This difference underscores the limitations of the handheld dynamometer in reflecting grip strength during everyday tasks, such as gripping a jar. On the other hand, typical handheld dynamometers allow for between-subject variabilities in hand size, a feature not present in the PowerJar. Regardless, given the strong correlation, using the PowerJar-obtained grip strength measurements could be used to determine the effects of interventions.[1-5] Combined with other PowerJar tasks, the PowerJar could provide a more detailed assessment of grip strength. Further research is required to confirm these findings and to explore the potential of the PowerJar in clinical settings.

Repeatability

The study assessed the repeatability of PowerJar tasks. Ensuring a good repeatability is essential for early-phase drug development, particularly in placebo-controlled trials where the pharmacodynamic response to the treatment is quantified. Therefore, the tasks must provide consistent parameters for the same participant within a single visit without any intervention.

The parameters derived from the various PowerJar tasks demonstrated moderate to good repeatability, with ICC values greater than 0.5 for multiple grip-related and angle-related features. In the angle control task, the total grip and maximal grip where the best repeatable parameters. Furthermore, the hold time showed moderate repeatability. For the rapid opening and closing task, parameters associated with the opening frequency and closing frequency showed moderate to good repeatability. In the maximal grip task, the maximum grip parameter showed good repeatability, while total grip showed moderate repeatability. In conclusion, these findings indicate that the PowerJar tasks can repeatably measure various grip and angle parameters, making them of interest for assessing hand function in clinical and research settings.

The repeatability analysis was limited to the highest achieved resistance level, because not all participants were able to perform all three pre-defined resistive torque levels. The pre-defined resistive torque levels aimed to account for possible age-related differences in strength. However, as demonstrated in the age group comparison, this adjustment did not eliminate intergroup differences, suggesting that other factors such as individual physical condition may also play a role in performance. Additionally, it suggests that resistive torque levels may be more effectively determined by participant performance or even standardized for better comparability. Even though intergroup differences were present and only the highest resistive torque was analysed, the PowerJar still demonstrated moderate repeatability. These findings indicate the PowerJar's potential to assess the effects of clinical interventions.

Limitations

This was the first study to use the PowerJar in a clinical setting, and we identified several limitations in both the study design and the interpretation of results. First, due to the exploratory nature of this study, the number of included patients, was moderate. Additionally, no other, clinical, assessment of hand function of the patients was included in the study. This made it difficult to assess the impact of hand function impairment on the performance within and between study visits. The next step would be to perform more studies to relate PowerJar tasks to daily life activities. Future studies can address this by implementing stricter screening criteria and more elaborate clinical assessments of disease severity to create both a well-defined and larger study population. A second limitation is related to the technical issues that resulted in the EMG data not being useful in the assessment of muscle fatigue. Nonetheless, the fatigue questionnaire results provided insight into the effort and experience of the study participants, but future studies should aim to include the EMG for both an objective assessment of fatigue and the identification of possible compensation mechanisms in patients.

CONCLUSION

The current study provides evidence that the PowerJar is moderately repeatable in healthy volunteers under controlled conditions. Furthermore, the study demonstrates the PowerJar's usability across different populations. This makes the PowerJar a promising tool for assessing hand function in a clinical study setting. Further research is required to explore its sensitivity to changes in neuromuscular diseases and responses to drug interventions, and its potential relation to clinical outcomes for patients.

REFERENCES

1 F. Louter et al., 'Instruments for measuring the neuromuscular function domain of vitality capacity in older persons: an umbrella review,' Eur Geriatr Med, vol. 15, no. 5, pp. 1191–1213, 2024, doi: 10.1007/s41999-024-01017-7.

2 D. P. Leong et al., 'Reference ranges of handgrip strength from 125,462 healthy adults in 21 countries: a prospective urban rural epidemiologic (PURE) study,' J Cachexia Sarcopenia Muscle, vol. 7, no. 5, pp. 535–546, 2016, doi: 10.1002/jcsm.12112.

3 H. C. Roberts et al., 'A review of the measurement of grip strength in clinical and epidemiological studies: Towards a standardised approach,' Age Ageing, vol. 40, no. 4, pp. 423–429, 2011, doi: 10.1093/ageing/afr051.

4 J. A. Allen et al., 'Safety, tolerability, and efficacy of subcutaneous efgartigimod in patients with chronic inflammatory demyelinating polyradiculoneuropathy (ADHERE): a multicentre, randomised-withdrawal, double-blind, placebo-controlled, phase 2 trial,' Lancet Neurol, vol. 23, no. 10, pp. 1013–1024, 2024, doi: 10.1016/S1474-4422(24)00309-0.

5 S. C. Higgins, J. Adams, and R. Hughes, 'Measuring hand grip strength in rheumatoid arthritis,' Rheumatol Int, vol. 38, no. 5, pp. 707–714, 2018, doi: 10.1007/s00296-018-4024-2.

6 U. S. L. Nayak and J. M. Queiroga, 'Pinch grip, power grip and wrist twisting strengths of healthy older adults,' Gerontechnology, vol. 3, no. 2, 2004, doi: 10.4017/gt.2004.03.02.003.00.

7 A. Beckman, C. Bernsten, M. G. Parker, M. Thorslund, and J. Fastbom, 'The difficulty of opening medicine containers in old age: A population-based study,' Pharmacy World and Science, vol. 27, no. 5, pp. 393–398, 2005, doi: 10.1007/s11096-005-7903-z.

8 L. Santisteban, M. Térémetz, J. P. Bleton, J. C. Baron, M. A. Maier, and P. G. Lindberg, 'Upper limb outcome measures used in stroke rehabilitation studies: A systematic literature review,' PLoS One, vol. 11, no. 5, pp. 1–16, 2016, doi: 10.1371/journal.pone.0154792.

9 M. T. Duruöz, K. Nas, S. A. Kasman, N. Öz, E. Uzun, and H. H. Gezer, 'Validity and reliability of the Duruöz Hand Index in patients with psoriatic arthritis,' Rheumatol Int, vol. 44, no. 3, pp. 535–542, 2024, doi: 10.1007/s00296-023-05517-w.

10 M. T. Duruöz et al., 'Development and VAlidation of a Rheumatoid Hand Functional Disability Scale That Assesses Functional Handicap,' J Rheumatol, vol. 23, no. 7, 1996.

11 US Food and Drug Administration, 'clinical outcome assessment (COA) compendium,' 2021. [Online]. Available: https: //www.fda.gov/regulatory-

12 T. K. Koo and M. Y. Li, 'A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research,' J Chiropr Med, vol. 15, no. 2, pp. 155–163, 2016, doi: 10.1016/j.jcm.2016.02.012.

FIGURE 1 PowerJar setup as used in the study. The computer that is attached and used for controlling the PowerJar is not shown.

FIGURE 2 A) The position of the participant while gripping the PowerJar. B) The grip position of the hands and fingers of a typical participant.

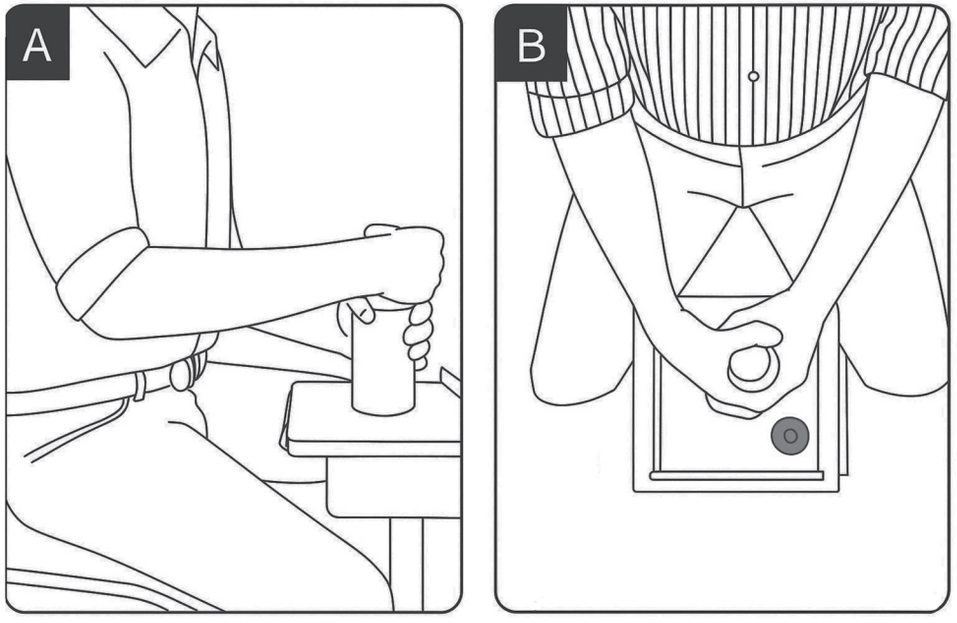


FIGURE 3 Grip position of the handheld dynamometer of a typical participant.

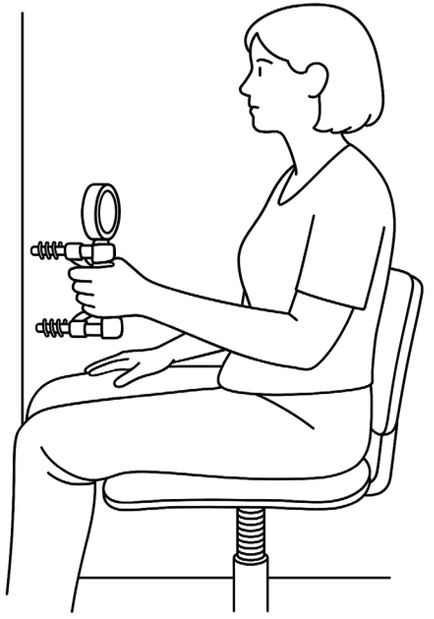


FIGURE 4 A) Boxplots of the Maximum Grip measured with the handheld dynamometer and the PowerJar. B) Correlation plot (r=.79) between the maximum grip values obtained by the handheld

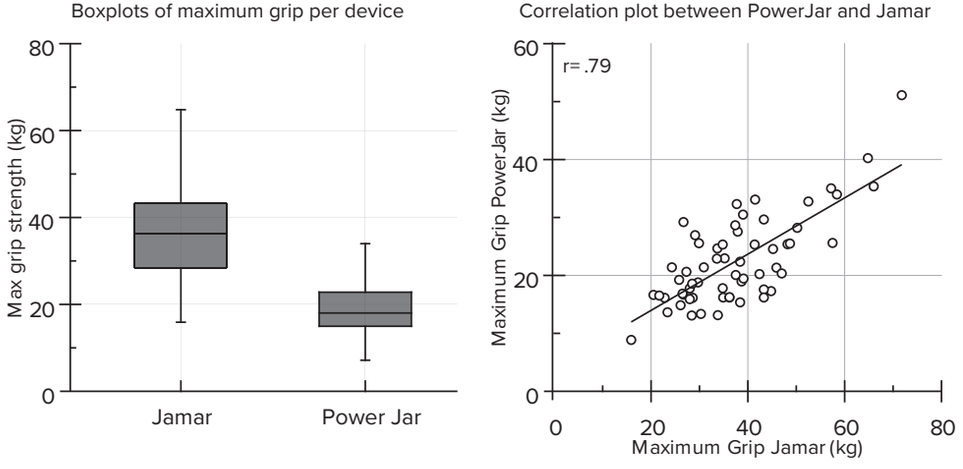


TABLE 1 Pre-determined resistive torques per age group.

Age group	First level	Second level	Third level
20 – 40 years old	2000 mNm	3500mNm	4991 mNm
41 – 60 years old	1500 mNm	3000mNm	4991mNm
61 – 80 years old	1000 mNm	2500 mNm	4000 mNm

TABLE 2 Calculated parameters per task.

Category	Parameter	1	5	6	Definition
Grip	Total [kgs]	x		x	Area under the curve (AUC) in the defined period.
	Slope [kg/s]	x		x	Linear term of a polynomial fit (1st order) of the grip values in the defined period.
	Max [kg]	x		x	Maximum grip value in the defined period.
	RMSE [kg]	x			Root mean square error of the polynomial fit (1st order) of grip values in the defined period.
	Dominant freq. [Hz]		x		Dominant frequency of the grip values.
	Freq. dominance [a.u.]		x		Peak AUC (around dominant frequency +- 0.25Hz) / total AUC (0-10Hz).
	Power frequency [kg^2]		x		Power at the dominant frequency.
Grip & Angle	Hold time [s]	x		x	Time of the defined period.
	Correlation grip and angle [a.u.]		x		Correlation coefficient between angle and grip values.
	Number of openings [#]		x		Opening is defined as angle >40.
	Number of Openings : total peaks [a.u.]		x		Number of successful openings divided by the total peaks.
Angle	Slope [deg/s]	x	x		Linear term of a polynomial fit (1st order task 1, 2nd order task 5) of the angles in the defined period.
	RMSE [deg]	x			Root mean square error of the polynomial fit (1st order) of the angles in the defined period.
	Quadratic term polyfit [deg/s2]		x		Quadratic term of a polynomial fit (2nd order task 5) of the angles in the defined period.
	Dominant freq. [Hz]		x		Dominant frequency of the angle values.
	Freq. dominance [a.u.]		x		Peak AUC (around dominant frequency +- 0.25Hz) / total AUC (0-10Hz).
	Power frequency [deg2]		x		Power at the dominant frequency.
Peak-to-Peak (P2P)	Slope P2P [s/#]		x		Linear term of a polynomial fit (1st order) of the peak-to-peak intervals in the defined period.
	Intercept P2P [s]		x		Intercept term of a polynomial fit (1st order) of the peak-to-peak intervals in the defined period.
	RMSE P2P [s]		x		Root means square error of a polynomial fit (1st order) of the peak to peak intervals in the defined period.

TABLE 3 Average age and standard deviation (SD) of age and BMI per age group of healthy volunteers and the patient group.

	Age in years (+/- SD)	BMI in kg/m^2 (+/- SD)
Healthy volunteers group 1	22.6 (+/-2.0)	23.6 (+/- 2.8)
Healthy volunteers group 2	48.5 (+/-5.3)	25.1 (+/- 3.7)
Healthy volunteers group 3	71.9 (+/- 4.0)	25.6 (+/- 2.7)
Patients	64.7 (+/- 9.4)	Not recorded

TABLE 4 Calculated mean values, intraclass correlation coefficients (ICC), minimal detectable effect (MDE), and p-value for the effect of trial, for parameters from tasks 1, 5, and 6. Only parameters which had an ICC value of 0.5 or larger were included in this table. with a value of 0.5 or larger for task 1, 5 and 6.

Parameter	Mean	ICC	MDE	p-value	F (NUMDF, DENDF)
TASK 1					
Total Grip	176.60 [kg*s]	0.70	146.20 [kg*s]	=.074	2.36 (3, 140.2)
Maximum Grip	7.53 [kg]	0.66	3.41 [kg]	=.060	1.21 (3, 140.6)
Hold Time	38.35 [s]	0.59	20.59 [s]	=.011	3.82 (3, 153.9)
TASK 5					
Frequency Dominance Grip	0.32 [a.u.]	0.55	0.09 [a.u.]	= .006	4.33 (3, 165.5)
Frequency Power Grip	3.55 [kg^2]	0.59	2.26 [kg^2]	=.083	2.27 (3, 165.4)
Correlation Grip and Angle	0.51 [a.u.]	0.64	0.23 [a.u.]	< .001	6.81 (3, 165.2)
Number of Openings	24.25 [#]	0.79	11.04 [#]	< .001	18.57 (3, 175.1)
Ratio Successful Openings	71.45 [a.u.]	0.67	21.41 [a.u.]	= .016	3.56 (3, 175.1)
Dominant Frequency Angle	1.29 [Hz]	0.79	0.53 [Hz]	< .001	19.34 (3, 175.1)
Frequency Dominance Angle	0.38 [a.u.]	0.51	0.08 [a.u.]	< .001	9.50 (3, 175.2)
TASK 6					
Maximum Grip	19.75 [kg]	0.87	5.57 [kg]	< .001	6.30 (3, 168.3)
Total Grip	4218.71 [kg*s]	0.71	3181.58 [kg*s]	< .001	6.01 (3, 168.6)

TABLE 5 Effect of age-group on parameter value. Contrasts (estimate of difference (95% CI)) and p-value between groups with at least one significant difference between groups (p<.05).

Parameter	Age-group		
	(20-40) vs (41-60)	(20-40) vs (61-80)	(41-60) vs (61-80)
TASK 1			
Total Grip [kg*s]	-109.45 (-214.03, -4.86) p=.041	-76.58 (-186.21, 33.06) p=.168	32.87 (-75.76, 141.50) p=.548
Slope Grip [kg/s]	0.05 (-0.05, 0.15) p=.316	-0.08 (-0.18, 0.03) p=.144	-0.13 (-0.23, -0.02) p=.017
Maximum Grip [kg]	-1.69 (-3.94, 0.57) p=.139	-4.00 (-6.37, -1.63) p=.001	-2.31 (-4.66, 0.04) p=.054
TASK 5			
Dominant Frequency Grip [Hz]	-0.29(-0.57, -0.02) p=.039	-0.20 (-0.47, 0.07) p=.142	0.09 (-0.18, 0.36) p=.517
Frequency Dominance Grip [a.u.]	-0.05 (-0.10, 0.01) p=.084	-0.10 (-0.15, -0.05) p<.001	-0.05 (-0.10, 0.00) p=.054
Frequency Power Grip [kg^2]	-0.83 (-2.22, 0.56) p=.239	-2.77 (-4.13, -1.40) p<.001	-1.94 (-3.32, -0.56) p=.007
Correlation Grip and Angle [a.u.]	-0.17 (-0.33, -0.01) p=.039	-0.13 (-0.29, 0.03) p=.110	0.04 (-0.12, 0.20) p=.605
NOO [#]	7.55 (0.58, 14.51) p=.034	17.53 (10.56, 24.49) p<.001	9.98 (3.01, 16.94) p=.006
Ratio Successful Openings [a.u.]	-4.85 (-19.43, 9.73) p=.508	12.70 (-1.88, 27.29) p=.087	17.55 (2.97, 32.13) p=.019
Dominant Frequency Angle [Hz]	0.46 (0.11, 0.80) p=.011	0.74 (0.39, 1.09) p<.001	0.28 (-0.06, 0.63) p=.108
Frequency Dominance Angle [a.u.]	-0.03 (-0.08, 0.02) p=.172	-0.05 (-0.10, 0.00) p=.037	-0.02 (-0.07, 0.03) p=.452
Intercept Peak to Peak [s]	-0.41 (-0.77, -0.04) p=.031	-0.55 (-0.92, -0.19) p=.004	-0.15 (-0.52, 0.22) p=.424

SUPPLEMENTS

SUPPLEMENTARY TABLE S1 ICC values for task 1 for the highest resistive torque for healthy participants.

Parameter	Mean	ICC	MDE	p-value	F (NUMDF, DENDF)
Total Grip	177.93 [kg*s]	0.71	146.33 [kg*s]	=.074	2.36 (3, 140.2)
Slope Grip	0.12 [kg/s]	0.34	0.17 [kg/s]	=.060	2.53 (3, 143.6)
Maximum Grip	7.53 [kg]	0.66	3.41 [kg]	=.069	1.21 (3, 140.6)
RMSE Grip	0.56 [kg]	0.27	0.33 [kg]	=.309	2.41 (3, 144.4)
Hold Time	38.60 [s]	0.60	20.57 [s]	=.011	3.82 (3, 153.9)
Slope Angle	-0.33 [deg/s]	0.37	0.62 [deg/s]	=.031	3.05 (3, 155.1)
RMSE Angle	1.77 [deg]	0.21	0.92 [deg]	=.534	0.73 (3, 156.2)

SUPPLEMENTARY TABLE S2 ICC values for task 5 for the highest resistive torque for healthy participants.

Parameter	Mean	ICC	MDE	p-value	F (NUMDF, DENDF)
Dominant Frequency Grip	0.61 [Hz]	0.33	0.48 [Hz]	=.694	0.48 (3, 166.2)
Frequency Dominance Grip	0.32 [a.u.]	0.55	0.09 [a.u.]	=.006	4.33 (3, 165.5)
Frequency Power Grip	3.55 [kg^2]	0.59	2.26 [kg^2]	=.083	2.27 (3, 165.4)
Correlation Grip and Angle	0.51 [a.u.]	0.64	0.23 [a.u.]	<.001	6.81 (3, 165.2)
NOO	24.25 [#]	0.79	11.04 [#]	<.001	18.57 (3, 175.1)
Ratio Successful Openings	71.45 [a.u.]	0.67	21.41 [a.u.]	=.016	3.56 (3, 175.1)
Quadratic Term Angle	0.00 [deg/s^2]	0.07	0.05 [deg/s^2]	=.516	0.76 (3, 175.7)
Slope Angle	-0.22 [deg/s]	0.19	1.21 [deg/s]	=.640	0.56 (3, 175.5)
Dominant Frequency Angle	1.29 [Hz]	0.79	0.53 [Hz]	<.001	19.34 (3, 175.1)
Frequency Dominance Angle	0.38 [a.u.]	0.51	0.08 [a.u.]	<.001	9.50 (3, 175.2)
Frequency Power Angle	44.76 [deg^2]	0.37	14.03 [deg^2]	=.109	2.05 (3, 175.3)
Slope Peak to Peak	-0.01 [s/#]	0.00	0.05 [s/#]	=.016	3.53 (3, 175.3)
Intercept Peak to Peak	1.17 [s]	0.46	0.62 [s]	<.001	13.67 (3, 174.4)
RMSE Peak to Peak	0.26 [s]	0.38	0.21 [s]	=.488	0.81 (3, 174.5)

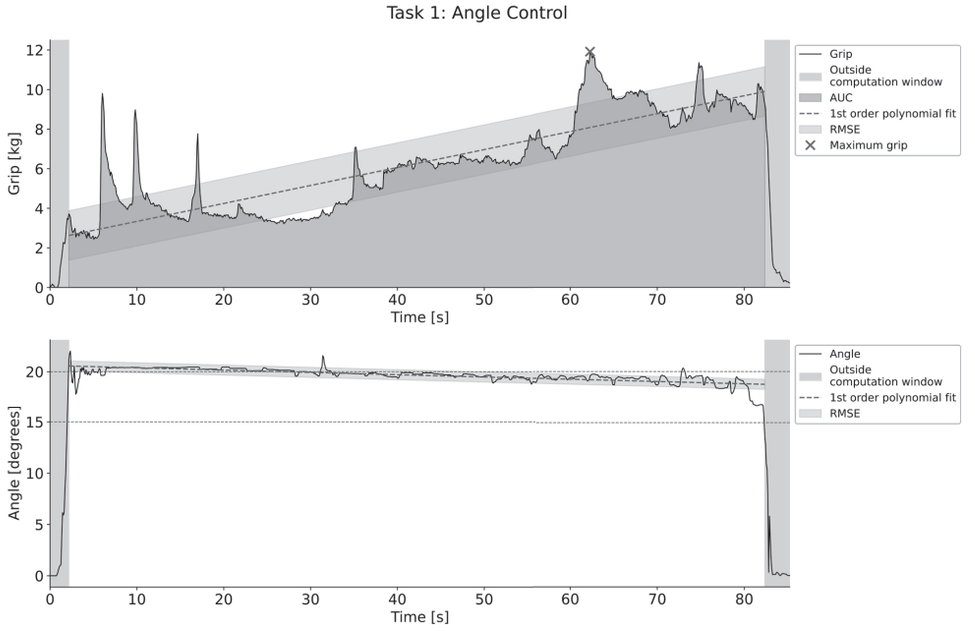
SUPPLEMENTARY TABLE S3 ICC values for task 6.

Parameter	Mean	ICC	MDE	p-value	F (NUMDF, DENDF)
Hold Time	31.03 [s]	0.48	16.68 [s]	=.335	1.14 (3, 169.3)
Maximum Grip	19.75 [kg]	0.87	5.57 [kg]	<.001	6.30 (3, 168.3)
Total Grip	4218.71 [kg*s]	0.71	3181.58 [kg*s]	=.001	6.01 (3, 168.6)
Slope Grip	-0.27 [kg/s]	0.25	0.30 [kg/s]	=.229	1.45 (3, 170.2)

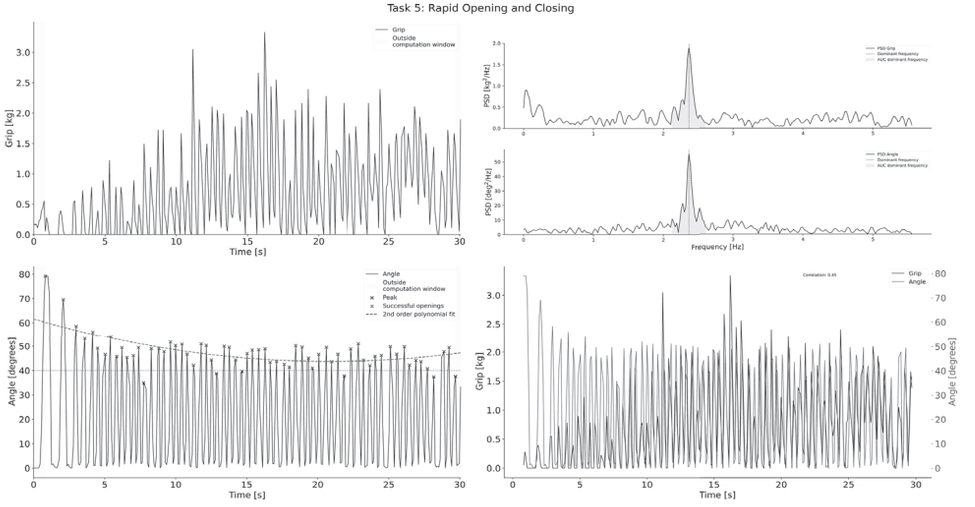
SUPPLEMENTARY TABLE S4 Effect of age-group on parameter value. Contrasts (95% CI) and p-value between groups. Parameters indicated with an asterisk are exp-transformed.

Parameter	Age Group		
	(20-40) vs (41-60)	(20-40) vs (61-80)	(41-60) vs (61-80)
Total Grip [kg*s]	-109.45 (-214.03, -4.86) p=.041	-76.58 (-186.21, 33.06) p=.168	32.87 (-75.76, 141.50) p=.548
Slope Grip [kg/s]	0.05 (-0.05, 0.15) p=.316	-0.08 (-0.18, 0.03) p=.144	-0.13 (-0.23, -0.02) p=.017
RMSE Grip [kg]	-0.02 (-0.21, 0.16) p=.803	-0.05 (-0.25, 0.15) p=.647	-0.02 (-0.22, 0.18) p=.824
Maximum Grip [kg]	-1.69 (-3.94, 0.57) p=.141	-4.00 (-6.37, -1.63) p=.001	-2.31 (-4.66, 0.04) p=.054
Hold Time [s]	-5.27 (-19.26, 8.72) p=.455	6.62 (-8.12, 21.36) p=.373	11.89 (-2.86, 26.64) p=.113
Slope Angle e^[deg/s] *	0.05 (-0.07, 0.17) p=.385	0.06 (-0.08, 0.17) p=.489	-0.01 (-0.14, 0.12) p=.895
RMSE Angle [deg/s]	-0.23 (-0.72, 0.26) p=.347	-0.15 (-0.67, 0.37) p=.569	0.08 (-0.44, 0.61) p=.751
TASK 5			
Dominant Frequency Grip [Hz]	-0.29(-0.57, -0.02) p=.039	-0.20 (-0.47, 0.07) p=.142	0.09 (-0.18, 0.36) p=.517
Frequency Dominance Grip [a.u.]	-0.05 (-0.10, 0.01) p=.084	-0.10 (-0.15, -0.05) p<.001	-0.05 (-0.10, 0.00) p=.054
Frequency Power Grip [kg^2]	-0.83 (-2.22, 0.56) p=.239	-2.77 (-4.13, -1.40) p<.001	-1.94 (-3.32, -0.56) p=.007
Correlation Grip and Angle [a.u.]	-0.17 (-0.33, -0.01) p=.039	-0.13 (-0.29, 0.03) p=.110	0.04 (-0.12, 0.20) p=.605
Number of Openings [#]	7.55 (0.58, 14.51) p=.034	17.53 (10.56, 24.49) p<.001	9.98 (3.01, 16.94) p=.006
Ratio Successful Openings [a.u.]	-4.85 (-19.43, 9.73) p=.508	12.70 (-1.88, 27.29) p=.087	17.55 (2.97, 32.13) p=.019
Quadratic Term Angle e^ [deg/s^2] *	0.00 (-0.02, 0.02) p=.974	0.01 (-0.01, 0.03) p=.217	0.01 (-0.01, 0.03) p=.228
Slope Angle [deg/s]	0.11 (-0.51, 0.73) p=.722	-0.29 (-0.91, 0.32) p=.343	-0.41 (-1.02, 0.21) p=.194
Dominant Frequency Angle [Hz]	0.46 (0.11, 0.80) p=.011	0.74 (0.39, 1.09) p<.001	0.28 (-0.06, 0.63) p=.108
Frequency Dominance Angle [a.u.]	-0.03 (-0.08, 0.02) p=.172	-0.05 (-0.10, 0.00) p=.037	-0.02 (-0.07, 0.03) p=.452
Frequency Power Angle [deg^2]	-6.60 (-14.84, 1.65) p=.115	-2.31 (-10.56, 5.95) p=.579	4.29 (-3.96, 12.54) p=.303
Slope Peak to Peak [s/#]	(-0.01, 0.03) p=.208	0.02 (0.00, 0.04) p=.135	0.00 (-0.02, 0.02) p=.801
Intercept Peak to Peak [s]	-0.41 (-0.77, -0.04) p=.031	-0.55 (-0.92, -0.19) p=.004	-0.15 (-0.52, 0.22) p=.424
RMSE Peak to Peak [s]	-0.03 (-0.16, 0.09) p=.579	-0.11 (-0.23, 0.01) p=.071	-0.08 (-0.20, 0.04) p=.203

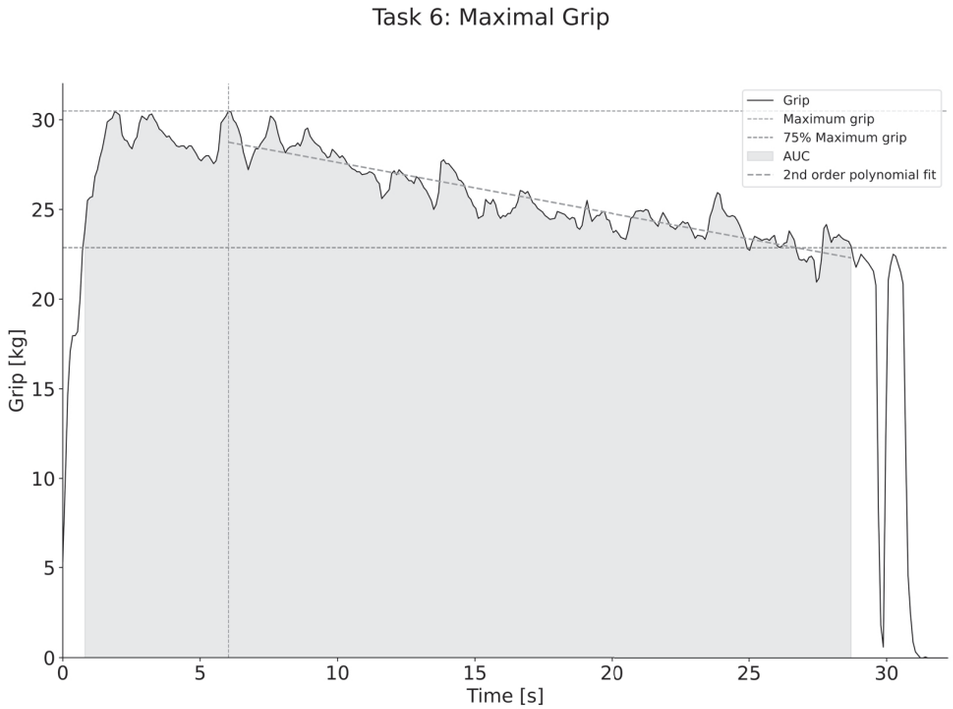
SUPPLEMENTARY FIGURE S1 Example of the data collected during Task 1.



SUPPLEMENTARY FIGURE S2 Example of the data collected during Task 5.



SUPPLEMENTARY FIGURE S3 Example of the data collected during Task 6.



CHAPTER 7

GENERAL DISCUSSION AND CONCLUSIONS

Ecological validity refers to the extent to which a biomarker, a biological measure used to assess health or disease states, reflects real-world outcomes or functional activities relevant to patient experiences and decision-making by registration authorities. In this context, ecological validity of a biomarker is defined as its ability to generalize to daily life activities and its relevance to clinical outcome assessments accepted by regulatory authorities as valid for the registration of drugs. As clinical drug trials aim to bridge controlled trial settings with real-world clinical contexts, it is increasingly important that outcome measures not only demonstrate statistically significant treatment effects but also to provide meaningful insights into how diseases impacts patients' lives.

This thesis aimed to identify biomarkers used in early-phase clinical trials for drug development and assess them on their level of ecological validity. This evaluation of ecological validity of biomarkers in early-phase drug development is a relatively new concept. Therefore, an objective and structured approach is essential. In this general discussion, we introduce a novel framework designed to assess the ecological validity of biomarkers used in clinical trials. This framework allows for evaluation of biomarkers from the stage proof of pharmacology through to later phases, where real-world evidence is generated. By applying this framework on the biomarkers used in this thesis, we can evaluate how well early-phase biomarkers predict real-world outcomes, bridging the gap between clinical trials and everyday life.

RATIONALE FOR NEW FRAMEWORK

Interest in ecological validity is growing, as is reflected by the increasing number of related scientific publications. A search on ecological validity in the PubMed database, shows an almost exponential increase since 1990 (Figure 1). With this growing interest, various frameworks have emerged across different research fields. There are frameworks for Human-Computer Interaction (HCI)[1], deployment of new systems in a clinical setting[2], specific biomarkers for neurocognitive research (e.g., brain activity markers in neuroimaging)[3], and pragmatic clinical trials.[4,5]

These frameworks provide insight into assessing the real-world relevance of interventions, tasks, and outcomes, and into the evaluation of biomarkers in clinical trials. However, they present clear limitations when used for the evaluation of biomarkers in drug trials. Most of these frameworks are more focussed on the task or system, which makes them less applicable for evaluating the biomarkers themselves in the context of clinical outcomes in drug trials. For instance, frameworks focused on task performance evaluate participant-task interactions but do not consider the clinical symptom that biomarkers must reflect, such as an immune response or disease progression.[1,3] Other frameworks are focused more on the design of clinical trials or the implementation of systems, and therefore do not directly assess the biological validity or clinical utility of biomarkers.[3-5] Additionally, some are domain-specific (e.g., neuroscience) and not generalisable across therapeutic areas.

In summary, although several frameworks address ecological validity, there is a lack of one tailored specifically to biomarkers in drug studies, particularly in predicting or reflecting real-world patient outcomes (e.g., symptom relief, disease progression, or quality of life). Therefore, a new framework is needed, one that considers biomarkers used in drug studies and their ability to predict outcomes that are relevant to the target population. This framework should address the complexity of biomarkers in clinical trials, their level of relevance to the real-world, and predictive value in relation to patient experiences and daily functioning. Importantly, it should also consider the dynamic and evolving relationship between ongoing scientific research and clinical symptoms, which supports the understanding of biomarker relevance.

DYNAMICS IN ECOLOGICAL VALIDITY

As described above, scientific evidence is an important support in evaluating ecological validity of biomarkers. The relationship between a biomarker and clinical symptoms that affect daily life can be suggested and logically assumed, but it requires scientific evidence to confirm and strengthen its ecological validity. Over time, this assumed relationship may evolve based on ongoing clinical studies, shifting positively (indicating a stronger correlation) or negatively (demonstrating a disconnection between the biomarker and the disease).

For example, Amyloid load in the brain was initially considered primarily a diagnostic marker of Alzheimer's disease (AD). In patients with mild cognitive impairment (MCI), an increased amyloid burden on PET scans was associated with a higher likelihood of progression to AD.[6] This led to the development of anti-amyloid therapies aimed at removing amyloid plaques. However, while early drugs successfully reduced amyloid accumulation, they failed to show meaningful cognitive improvement, raising

concerns about amyloid as a treatment target.[7,8] Despite these initial setbacks, amyloid plaque reduction was later accepted as a surrogate endpoint for accelerated approval of monoclonal antibody therapies by the FDA.[9] Aducanumab was the first Alzheimer's treatment to receive accelerated approval based on amyloid plaque reduction, but failed to convince with clinical improvement in later trials.[10] Lecanemab was the first reduced amyloid load and later demonstrated a significant (but modest) slowing of cognitive decline, leading to full regulatory approval.[11]

This example shows how the ecological validity of a biomarker can shift over time. Initially, amyloid PET imaging scored low on ecological validity due to the lack of observed clinical benefit despite biological effects. However, following the positive outcomes of the lecanemab trial, which demonstrated both amyloid reduction and a measurable slowing of cognitive decline, the ecological validity of amyloid as a biomarker increased accordingly.

In contrast, biomarkers such as DSDNA for systemic lupus erythematosus (SLE) and serum prolactin for epilepsy were once thought to be relevant but are now understood to have limited or no relevance to the diseases they were associated with. Anti-DSDNA antibodies are often used for the diagnosis and monitoring of systemic lupus erythematosus (SLE), as their levels often correlate with disease activity. However, some patients experience severe disease flares despite having low anti-DSDNA levels.[12] Additionally, some drugs, such as belimumab, lower anti-DSDNA levels, but this does not always translate into clinical improvement.[13] Anti-DSDNA is therefore not a suitable biomarker for treatment response in clinical trials.[12] Similarly, serum prolactin has been proposed as a biomarker for epileptic seizures, because it increases after generalized tonic-clonic and some focal seizures.[14,15] However, its clinical utility is limited by several factors, including variability in response and confounders affecting serum prolactin levels.[16] Furthermore, drugs that lower serum prolactin (e.g., dopamine agonists) do not reduce seizure frequency, indicating that prolactin elevation is a post-seizure effect and not suitable as a marker of disease activity or treatment response in clinical trials.[17]

These examples illustrate that the ecological validity of a biomarker is not fixed, but changes with the available scientific evidence. This dynamic interplay will be essential in evaluating the biomarkers discussed in this thesis.

TIERED APPROACH TO ECOLOGICAL VALIDITY OF BIOMARKERS

EVIB: Ecological Validity of Biomarkers

Here we introduce a framework designed to assess the ecological validity of biomarkers used in clinical trials with a clinical target population. It consists of six tiers, where a higher tier indicates greater ecological validity suggesting that the biomarker is more likely to predict the outcomes of registration trials and of real-world benefit. The six tiers of the framework are:

1. Proof of pharmacology: biomarkers in this tier confirm drug-target engagement with unknown clinical and/or functional relevance (e.g., drug-induced changes in cytokine levels without known implications for symptoms).
2. Mechanistic insight: biomarkers in this tier show mechanistic changes in disease-related pathways but has not been proven to predict clinical outcomes (e.g., amyloid-beta levels in CSF for Alzheimer's disease).
3. Clinical correlation: biomarkers in this tier shows a demonstrated link to clinical outcome. (e.g., bone mineral density for osteoporosis or spike wave discharges on EEG for epilepsy)
4. Functional impact: biomarkers in this tier provide insights into the effect on daily life but lack a direct assessment of daily life activities (e.g., FEV1 in patients with Pompe disease, finger tapping for Parkinson)
5. Daily-life quantification: biomarkers in this tier directly measure or predict quantifiable activities of daily life (e.g., six-minute walk test for aerobic capacity, MDS-UPDRS for PD)
6. Real-world evidence: biomarkers in this tier provide longitudinal observational data for broader insights (e.g., daily number of steps of patients with Parkinson Disease or phone use for patients with FSHD).

Biomarkers can be categorized into tiers based on available scientific evidence from clinical trials to patient practice. It is important to describe the biomarker and its relation to clinical symptoms or disease severity in the context of a specific patient population. For example, the UPDRS assessment can be considered a tier 5 biomarker for Parkinsons Disease. However, when applied to patients who broke a leg and is in recovery wearing a cast, the UPDRS has limited ecological validity.

The framework can be applied to biomarker studies through four steps: First, the relevant real-world activity or COA to which the ecological validity should be linked must be identified. Second, the evidence generated for the biomarker within the study should be summarised. Third, the available evidence supporting the biomarker's ecological validity should be evaluated. Finally, the biomarker is assigned a tier within the framework based on this evidence. Based on these steps, the biomarkers discussed in this thesis will be analysed and categorized based on their relevance to daily life activities. The chapter concludes with practical considerations for applying the framework, followed by an overall evaluation.

BIOMARKERS FOR DRIVING BEHAVIOUR

The extent to which a drug impairs driving ability (i.e., operating a motorized vehicle) can significantly influence its safety profile and, as a result, its regulatory approval. If a drug negatively affects driving performance, it may compromise patient safety, restrict clinical application, and reduce market potential. Discovering such effects late in the development process, typically during phase III trials or post-market surveillance, can limit the drug's usability and success. Therefore, if driving performance is a relevant factor in regulatory evaluation, it is essential to assess its impact as early as possible in the development. No single clinical symptom or disease state is directly predictive of driving ability. Instead, research in this area focuses on identifying impairments that result in deviations from normative driving behaviour that may pose a safety risk.

Summary of generated evidence

In **Chapter 2**, the impact of sleep deprivation on driving ability was evaluated. Sleep deprivation is a well-known intervention that can induce impaired driving. In this study, multiple methods were used to assess driving performance, including on-the-road driving, simulator driving, and various psychomotor tasks (including eye movements, body sway, and a pursuit tracking task). The primary outcome variables were the standard deviation of lateral position (SDLP) for both the on-the-road and the simulated driving tasks, and the performance on the pursuit tracking task. Sleep deprivation increased SDLP for both simulated (10cm, 95%CI: 6.7–13.3) and on-the-road driving (2.8cm, 95%CI: 1.9–3.7). Additionally, sleep deprivation affected almost all psychomotor test battery biomarkers. A moderate correlation was observed between on-the-road and simulator SDLP following a well-rested night (0.63, $p < .001$), but this correlation diminished after sleep deprivation. Among the psychomotor tasks, only pursuit tracking performance significantly correlated with simulator SDLP (-0.50, $p = .02$), but not with the SDLP of the on-the-road task. The lack of correlation between other objective biomarkers suggests that these tasks may not be interchangeable. Instead, each biomarker appears to assess different aspects that can be related to driving behaviour.

Summary of evidence and tier assignment

Pursuit tracking performance has been demonstrated to serve as a biomarker for sustained attention and hand-eye coordination. This performance can be influenced by various factors, including drug interventions (e.g., drowsiness-inducing drugs). Since driving also requires attention and hand-eye coordination, tracker performance could be considered a proxy for assessing certain aspects of driving. However, it is important to note that while the tracker performance captures some relevant factors, it does not fully reflect all elements of driving. For instance, driving involves elements such as sustained attention, decision making, and various risk factors (e.g. accidents) that are not captured by tracker performance alone.

Although there is limited evidence that improvements in tracker performance translate directly to improved driving behaviour, drugs known to impair driving are associated with decreased tracker performance. However, the strength and variability of this relationship across different mechanisms of action (MOA) remain unclear. Furthermore, due to the lack of strong correlations between tracker performance and biomarkers such as the SDLP, tracker performance should not be directly equated with actual driving ability. While mechanistically linked to the intervention, the ability of the tracker performance to predict driving impairment remains unproven. Therefore, within the ecological validity framework presented here, tracker performance is classified at tier 3; it is correlated with clinical symptoms but has not yet proven to predict driving behaviour directly.

In contrast to tracker performance, the SDLP derived from driving simulation reflects several realistic aspects of driving, such as driving on highways, long trip durations, and simulated traffic, all of which resemble real-world conditions more closely than tracker performance. However, driving simulators lack certain elements present in on-the-road driving. For example, simulated driving lacks the danger and dynamic factors of real-life driving (e.g., interactions with other vehicles, road conditions, and the physical

sensations of speed and motion). Despite these limitations, driving simulators are accepted by regulatory authorities as valid tools for reflecting daily driving behaviour in a representative manner.[18] Depending on the simulator's level of realism, the recorded SDLP may be considered in either tier 4 or tier 5.

The SDLP assessed in an on-the-road driving test, often referred to as the gold standard,[19,20] provides the most accepted and ecologically valid biomarker of driving behaviour. The SDLP recorded on the road is strongly correlated to real-life driving performance[21] and widely used to evaluate the effects of interventions on driving behaviour,[22] although it does not capture all aspects of everyday driving (e.g., urban traffic, intersections, and varying road conditions). For its alignment with actual driving, the on-the-road SDLP assessment is considered to have the highest ecological validity, placing it at tier 5 on the ecological validity scale. However, including a more realistic driving scenario, possibly using an actual commute of the participant, could increase the ecological validity.

Evaluation

This clinical trial, which assessed biomarkers across three levels of ecological validity, demonstrates the utility of the proposed framework. In some cases, stepping back to a lower tier on the scale can make research more feasible, safer, and more cost-effective. For example, using a driving simulator (tier 4) in early-phase trials offer a more accessible and manageable alternative to the on-the-road test (tier 5). Additionally, tracker performance (a lower-tier biomarker requiring only a few minutes to complete) could be used in early-phase trials, while reserving the SDLP assessment (a higher tier measure) for later-phase trials to more accurately study the effects of a drug on driving behaviour. This approach balances scientific evidence with the practical constraints of clinical trial design.

BIOMARKERS FOR FALL RISK

Falls predominantly occur in ambulatory settings (32-57%), during routine daily activities. Imbalance is an important factor in most of the cases (39–62%), but environmental elements such as uneven surfaces or steps also play a significant role (21–27%).[23] Therefore, if a drug negatively affects balance or motor control, it can increase the risk of falls. Historically, sleep-inducing drugs have been shown to increase the fall risk, particularly in elderly. Benzodiazepines, one of the most prescribed sleep aids, are well-known for increasing the risk of falls. Depending on the specific type, the risk of falling may double or even triple.[24,25] Therefore, when registering a new sleep medication, it is essential to assess any potential increase in fall risk, both for patient safety and general market acceptance.

Currently, no COA has been established for the assessment of increased fall risk. The most ecologically valid biomarker would be the actual number of falls occurring within a certain timeframe. However, this endpoint requires long-term monitoring of large patient cohorts, making it impractical for most clinical trial settings. To prevent the need for extended follow-up, proxy assessments are commonly used to estimate fall risk. These assessments may include isolated evaluations of muscle strength or postural stability, as well as dynamic assessments that integrate multiple aspects of gait and balance. Such proxy measures are easier to administer, involve short tasks, and yield immediate results, making them a more practical alternative for assessing fall risk for clinical development and regulatory review of drugs.

Summary of generated evidence

Dynamic balance assessments may provide a more realistic prediction of drug-induced falls compared to postural stability measurements, as falls often result from limited gait adjustments during walking. In **Chapter 3**, we explored the sensitivity of the Interactive Walkway (IWW) to drug effects, using it to assess walking adaptability. Healthy elderly participants were included in a 3-way crossover study, with treatments consisting of 5 mg zolpidem, 10 mg suvorexant, or placebo. It has been previously shown that benzodiazepines, such as zolpidem, increase the risk of falls, and zolpidem was therefore considered the positive control in this study. The more specific hypnotic drug suvorexant is believed to induce fewer side effects than zolpidem (e.g., sleepiness, reduction of fine motor control, reduction in muscle strength). The IWW assessments included an 8-meter walking test, a goal-directed stepping test, an obstacle-avoidance test, and a tandem-walking test. Other pharmacodynamic measurements included the Timed-Up-and-Go (TUG) test at both a comfortable and a fast pace, pursuit tracking, and body sway. Compared to placebo, zolpidem significantly reduced the performance on all biomarkers within 3 hours of the IWW walking adaptability test, TUG test, pursuit tracking test, and body sway test. Additionally, the effect of zolpidem on the IWW included a decrease in walking speed for all tasks. In contrast, suvorexant did not affect any parameters of the TUG test,

pursuit tracking, or IWW at any time point. It did, however, result in a significant 9.8% increase in postural stability, as assessed by the body sway test. This suggests that suvorexant may have an impact on balance without directly affecting dynamic walking performance. In summary, the IWW was successfully used to quantify the drug effects of zolpidem and suvorexant on walking performance.

Summary of evidence and tier assignment

The total amount of sway in the antero-posterior direction during quiet standing is a useful biomarker for assessing postural stability, capturing the summation of all unilateral sway over time. This biomarker is often referred to as the total body sway and is recorded over a few minutes. While an increase in sway during standing can logically suggest a potential decline in postural stability, it is not always correlated with real-world fall risk.[26-28] Falls often occur during movement and not while standing still.[29] This means that total body sway may miss key dynamic factors that influence fall risk.

The total body sway has been extensively validated in patient populations, particularly with drugs known to increase fall risk, including this trial.[30] We also observed that suvorexant induced an increase in body sway, with no other significant effects on the other biomarkers. This supports the idea that body sway may reflect certain aspects of postural stability. However, it should not be directly equated with fall risk. In conclusion, total body sway is a reliable measure of postural stability but does not capture the full range of dynamic movements associated with real-world falls. Therefore, it is classified as tier 3 on the ecological validity scale. This tier reflects a proven link to clinical symptoms of balance impairment but limited ability to predict real-world activities.

The time to complete the TUG test (i.e., TUG time) is a commonly used clinical assessment that evaluates mobility, balance, and functional ability. The TUG time is often used to assess mobility in populations at risk for falls, including the elderly or individuals with neurological disorders like Parkinson's disease.[31,32] The test simulates real-life scenarios: participants must stand up from a chair, walk a set distance, turn around, and sit back down on the chair. Any increase in the time to complete the test is believed to reflect imbalance and therefore potential fall risk. Therefore, we can state that the TUG time is correlated with daily activities such as walking and standing, but it remains a simulation of real-life mobility. While the TUG time is useful in assessing mobility and balance, it does not encompass the full range of dynamic factors that contribute to real-world falls, such as navigating uneven surfaces (e.g., steps causing tripping) or managing sudden environmental changes.[23] Thus, while the TUG time correlates with activities of daily living, it remains a simplified simulation of real-world mobility. Given these limitations, the TUG test is classified as tier 4 on the ecological validity scale, reflecting its relationship with real-world mobility while lacking the complexity of real-world environments.

The IWW test provides various biomarkers. In the contact of this thesis, the focus is on the relative margin to the obstacle during the obstacle avoidance task, as this particular biomarker cannot be assessed by any of the other available tasks. During the obstacle avoidance task, participants are to complete an 8-meter trajectory. While walking, 'obstacles' (i.e., visually projected flat blocks on the floor) suddenly appear in front of the participants. The participants were instructed to step over these obstacles without touching them and to continue completing the trajectory. This task simulates a real-world situation where individuals must adjust their gait to navigate hazards. It induces a stepping pattern that is different from the comfortable one, requiring adjustments and introducing an environmental hazard known to increase the risk of falls.[23] The margin of the leading limb to the obstacle reflects the likelihood that the participant might have tripped over the obstacle had it been 3D. In our trial, this margin to the object demonstrated the ability to differentiate between placebo and active treatment, as well as between two different mechanisms of action of hypnotic agents. Given the focus on a real-world scenario, despite being conducted in a controlled setting, the margin of the leading limb in the obstacle avoidance task is classified as tier 4 on the ecological validity scale. With additional evidence linking the margin to the object to real-life falls caused by stumbling over objects, it could potentially be elevated to tier 5.

Evaluation

This trial strengthened the ecological validity of the IWW because of the correlations found with zolpidem, a known fall-risk factor. The time to complete the TUG and total body sway assess only the factor of imbalance as cause for falls. In addition to assessing possible imbalance during walking (reflected in the walking speed), the IWW also includes the factor of uneven steps (involved in 25% of falls) and objects on surface (involved in 10% of the falls). Because of this more complete assessment, the margin of the leading limb reaches a higher level of ecological validity. This chapter also provides

insights into opportunities to enhance the Timed-Up-and-Go (TUG) test. For example, incorporating obstacles could elevate the task's ecological validity, making it more reflective of real-life scenarios.

BIOMARKERS FOR GRIP STRENGTH

Grip strength assessment is commonly used to evaluate general health and muscle strength in clinical studies. It can be quantified as a single biomarker, or as an integrative part of a combined performance assessment (e.g., the Quantitative Myasthenia Gravis (QMG)). A handheld dynamometer provides a relatively inexpensive, simple, and quick assessment of grip strength. Since daily life activities often involve gripping objects or items with our hands, it seems logical that a decrease in grip strength would impact daily activities, which is likely to be reflected in quality of life assessments in registration trials. However, while grip strength is a commonly used biomarker, it is limited in capturing the complexity of daily life tasks, which often require a combination of strength, coordination, and fine motor control.

Summary of generated evidence

In **Chapter 4**, we explored the usability and repeatability of the PowerJar device as an alternative to the handheld dynamometer. The PowerJar simulated jar-opening tasks and allowed quantification of grip and rotational forces. In addition to sixty healthy volunteers, we included twenty patients with neuromuscular diseases, who were expected to have impaired (rotational) grip strength. We observed minimal data loss and generally positive usability results. Furthermore, the maximum voluntary grip strength measured with the PowerJar was strongly correlated with the handheld dynamometer (Pearson correlation of 0.79). We assessed the repeatability of biomarkers obtained using the PowerJar in healthy volunteers for three tasks, each requiring both grip and rotational strength. We found moderate to good repeatability for both grip and rotational biomarkers. For example, the maximum grip during the prolonged grip strength task had an ICC of 0.87, and the number of openings in the rapid opening and closing task had an ICC of 0.79. Due to the study design, we could not assess the repeatability of these biomarkers in patients or its relation to disease severity. We concluded that the PowerJar was moderately repeatable in healthy volunteers under controlled conditions. The study demonstrated the PowerJar's usability in both healthy volunteers and patients. The PowerJar simulates the common daily task of opening a jar and quantifies it into multiple high-resolution endpoints, which have the potential to serve as reliable and sensitive biomarkers. Further research is needed to assess the PowerJar's sensitivity to disease progression, drug responses, and its correlation with clinical symptoms in daily tasks.

Summary of evidence and tier assignment

The maximum grip measured using the standardized handheld dynamometer is a widely used biomarker for general health[33] and interventions.[34,35] It has been shown to reflect clinically relevant changes in strength in various patient populations.[36] The task requires a short burst of strength from the hand muscles, particularly the fingers and thumb.[37] However, many daily tasks require sustained grip strength and/or fine motor control at different grip levels, such as brushing teeth, cleaning, or dressing. While the handheld dynamometer may reflect general well-being, its level of ecological validity to common daily tasks is less clear. The assessment provides a clinical correlation (tier 3) of an aspect of a daily task.

The maximum grip in the prolonged grip strength assessment of the PowerJar is a parameter similar to the maximum grip strength measured by the handheld dynamometer. However, it differs in two aspects that are important when assessing ecological validity: (1) the shape of the object resembles a real-life item (a typical jar), and (2) it requires prolonged strength, which is more representative of daily activities such as holding an object. The strong correlation between maximum grip strength measured by the handheld dynamometer and the PowerJar suggests that both assessments capture similar aspects of grip strength. However, the significantly lower grip strength measured with the PowerJar indicates a difference in clinical relevance. Due to the limited clinical evidence for the PowerJar, the ecological validity of maximum grip strength during the prolonged grip strength task remains classified as tier 4.

Evaluation

Among other parameters, the PowerJar's rapid opening and closing task measures the number of successful openings. This parameter reflects both rotational speed and the participant's strength in turning the PowerJar lid. Unlike grip strength, which had a validated comparative assessment, rotational strength lacking such validation in the study. Rotation is an important component of daily tasks, as seen in actions such as turning a key or opening a bottle. The current study does not provide enough information

to place the biomarker on a tier of the ecological validity scale, as the effects of interventions and/or disease remain unknown. There is a need to relate the findings to a quantification of daily life tasks. This can be done with a questionnaire such as the patient-rated wrist/hand evaluation,[38] or a known intervention resulting in a decrease of certain aspects of daily life activities. If this evidence is provided, the number of openings would correspond to a tier 4 biomarker. However, with further validation against real-world activities (such as opening a jar or a door with a twist knob) it could potentially reach tier 5.

BIOMARKERS FOR THE EMOTIONAL EXPERIENCE OF PAIN

Pain is an inherently subjective experience, making its clinical assessment challenging. Often, the clinical outcome for pain is assessed by using a numerical rating score (NRS), where patients rate their perceived pain from 1 (almost no pain) to 10 (worst imaginable pain). While this NRS is simple to assess, it fails to capture the complexity of pain perception, which is influenced by various factors such as emotional state, psychological conditions, and external stimuli (e.g., drugs). However, early-phase pain research often involves healthy volunteers who do not experience pain that the investigational medicinal product aims to alleviate, making the NRS unsuitable. Consequently, evoked pain tasks are used to study the analgesic efficacy in healthy volunteers. In such tasks, volunteers are subjected to short painful stimuli and report pain detection and tolerance thresholds. However, these tasks do not adequately simulate the prolonged and emotionally burdensome nature of chronic pain. Therefore, developing analgesics that target the emotional aspects of pain requires alternative methodologies that more accurately reflect the (chronic) pain experience in healthy volunteers.

Summary of generated evidence

There are currently no biomarkers other than pain characteristics questionnaires to assess the emotional pain experience in healthy volunteers. The evoked pain tasks do not (or only little) include this aspect of pain. This is primarily because pain experience requires an emotional connection to the painful stimulus which is difficult to induce in the setting of a clinical trial. There is a need for a pain task that immerse participants in a setting where emotions can be modulated. Virtual Reality (VR) has highly immersive characteristics and is known to modulate pain effectively as analgesic. In **Chapter 5**, we aimed to change pain perception by adding an affective component to painful stimulation using VR. We assessed the effect of a simulated wound in VR on the electrical pain detection (PDT) and tolerance (PTT) threshold in 24 healthy male study participants. The emotional experience of the simulation and the experienced pain were assessed with VAS-Questionnaires on unpleasantness and fear. We demonstrated that a virtual wound decreases the PDT compared to the neutral condition (ED: -18.4%, 95%CI: (-26.9%, -9.0%) p<.001). Additionally, study participants experienced the electrical stimulation as more unpleasant when shown the virtual wound (ED: 5.9, 95%CI: (2.1, 9.8) p=.0028). In conclusion, we demonstrated the potential of VR in combination with a pain task to provide a challenge model highlighting the affective component of pain.

This challenge model was further validated in **Chapter 6**, where we assessed the sensitivity of this model to a (pharmacological) intervention. Diazepam, a benzodiazepine used to treat anxiety, may affect the emotional processing of pain. We hypothesized that the VR-PainCart would lower the PDT in healthy participants, but that diazepam would inhibit this effect. The study was conducted with 24 healthy male participants who received diazepam and placebo in randomized order in a two-way crossover study. Indeed, diazepam increased the PDT in the VR-wound condition (ED: 6.0%, CI 2.4–53.2, p<0.05), while no significant differences were found for the VAS ratings for unpleasantness across any of the VR conditions or treatment effects. We concluded that a VR-simulated wound enhanced pain perception in an electrical nociceptive task and that diazepam increased the PDT only in the VR-wound condition, indicating pharmacological modulation of the pain experience.

Summary of evidence and tier assignment

The pain thresholds (i.e., PDT and PTT) obtained in an electrical pain task are frequently used biomarkers to study the effects of analgesics. The task, inducing a sharp pain sensation, evaluates primarily nociceptive pain. The pain thresholds have been validated with registered analgesics and found to be sensitive to a wide variety of drugs, such as ion channel blockers, opioids, and NMDA receptor blockers. Even though this indicates that the biomarker is not selective for one specific mechanism, it is easily obtained and sensitive to drug effects.[39] Depending on the (type of) analgesic, the biomarker can therefore provide proof of pharmacology and/or mechanistic insights. However, the translational evidence of electrical pain thresholds is limited. This can be partly explained by the fact that patients rarely

experience electrical stimulation in real life. Additionally, due to the influence of method and stimuli paradigms on pain thresholds, translational evidence must be provided per task. The task used in this thesis is mostly assumed to be translational for pain syndromes impacting the nociceptive system. Therefore, the pain thresholds recorded during electrical pain task as used in this thesis are classified as tier 1 on the scale of ecological validity. If the analgesic assessed has a specific mechanism of action, it can be considered as a tier 2 biomarker.

The pain detection thresholds recorded with the VR-PainCart, as described in this thesis, can be divided into three parts: 1) the thresholds recorded with the neutral simulation, 2) the thresholds recorded with the wound simulation and 3) the difference between the thresholds recorded with the neutral simulation and the wound simulation (delta). We propose the delta recorded with the VR-PainCart as a new biomarker where pain experience is expected to be a prominent factor of the analgesic effect. Diazepam is known to influence emotional processing and reduce anxiety. In our study, diazepam influenced the pain thresholds assessed with the VR-PainCart when presenting the virtual wound. This supports our hypothesis that emotional processing is modulated by the virtual wound paradigm.

Pain often includes (or warns against) tissue damage. One might think that by including the suggestion of tissue damage with the virtual wound, the pain thresholds would increase in ecological validity compared to the normal electrical pain thresholds. However, the level of evidence is limited. The delta thresholds of the VR-PainCart aim to measure a different type of pain. Therefore, the delta thresholds are classified on the same tiers as the electrical pain thresholds. With the evidence provided by this study for diazepam, the delta thresholds would result in tier 2, proof of mechanism. Advancing to a higher level of ecological validity should be feasible when the VR-PainCart is validated in a patient population with altered affective pain processing.

Evaluation

The biomarkers in this chapter highlight how important both standardisation and fit-for-purpose evidence are for the level of ecological validity. Even though the VR-PainCart pain thresholds might present a more realistic pain by emphasizing the possible consequences (e.g., tissue damage), the evidence relating these thresholds to pain syndromes is not (yet) available. Additionally, pain thresholds assessed with electrical stimulation are linked to pain syndromes based on logical reasoning and clinically evidence. However, the stimulus paradigm influences the type of nerves stimulated and consequently the type of pain mechanism involved. Therefore, each type of stimulus paradigm might result in new biomarkers and these new biomarkers need to be validated to potentially increase their ecological validity.

CONSIDERATIONS WHEN USING THIS FRAMEWORK

The current framework aims to evaluate biomarkers in terms of ecological validity. Besides ecological validity, biomarkers can have other characteristics that may be more important depending on the context. Other biomarker characteristics, such as the technical capabilities, quantifiability and variability, must also be considered. For example, if biomarker A has a lower ecological validity than biomarker B, but has a lower within-subject variability, then biomarker A could be considered a better choice for a first study in patients. This can be illustrated with the example of assessing changes in nerve excitability using threshold tracking techniques instead of the assessment of pain threshold changes.[40]

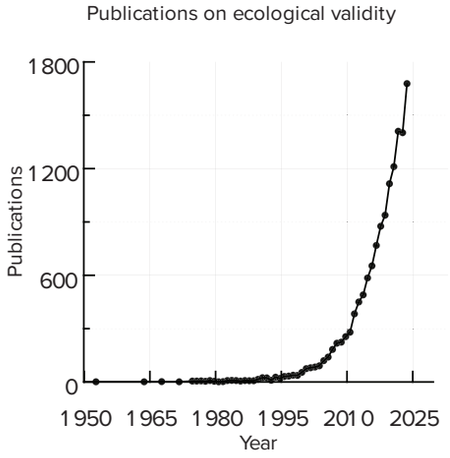
It is important to realise that a biomarker can be placed into a high tier while lacking evidence for the lower tiers. Studies based on real-world evidence (i.e., tier 6) or clinical data (i.e., tier 3) can result in the development of new biomarkers for drug development (i.e., tier 1). Additionally, a clinical rating score such as the Montreal Cognitive Assessment (MOCA), might be representative of disease severity (i.e., tier 4) without providing mechanistic insight (i.e., tier 2).

Additionally, the framework requires (scientific) evidence for a biomarker to obtain a high-level tier. This evidence can be costly, in both time and money. Such studies are long-lasting studies with multiple follow-up visits, epidemiological studies, and studies recording events in detail (e.g., accidents, falls, etc). When the link between the biomarker and the disease can logically be assumed, it will be less likely that these studies are conducted, and the evidence remains circumstantial. It might appear that this framework provides arguments against the use of the highest-tier ecological biomarkers. However, this framework only aims to encourage the incorporation of biomarkers with proven relationships to high level ecological biomarkers as early as possible in clinical trials for drug development.

CONCLUSIONS

To accelerate drug development, it is crucial that registration studies are as de-risked as possible. The selection of biomarkers plays a key role in this process. We identified ecological validity as an important factor in biomarker selection. Ecological validity determines how well a biomarker reflects real-world conditions that might impact clinical outcome assessments. Because this relationship can be complex and change over time, a framework is needed to standardize the assessment of ecological validity for biomarkers. Even though frameworks exist that address aspects of ecological validity, none were found to focus solely on biomarkers. We introduced a new framework which was applied to the biomarkers used in this thesis. We have drawn several important conclusions: reducing ecological validity can enhance study feasibility while still providing sufficient predictive power (Chapter 2); a well-validated biomarker does not always offer a complete picture of real-world living (Chapter 3); standardization of the biomarker is essential for increasing ecological validity (Chapters 4 and 5); additionally, a biomarker may have potential, but without a demonstrated intervention effect or clinical relevance, it lacks ecological validity (Chapter 6). The presented framework also has limitations, and it is important to be aware of these in order to use, refine and improve the framework for future applications. In conclusion, the framework presented here provides a better connection between clinical study results as part of early phase drug development and everyday life, which is relevant for medication registration studies.

FIGURE 1 **Scientific publications in the Pubmed database following a general search on 'ecological validity'.**



REFERENCES

1 S. Kieffer, U. B. Sangiorgi, and J. Vanderdonckt, 'ECOVAL: A framework for increasing the ecological validity in usability testing,' Proc. Annu. Hawaii Int. Conf. Syst. Sci., vol. 2015-March, no. 4, pp. 452–461, 2015, doi: 10.1109/HICSS.2015.61.

2 A. Kushniruk, C. Nohr, S. Jensen, and E. M. Borycki, 'From Usability Testing to Clinical Simulations: Bringing Context into the Design and Evaluation of Usable and Safe Health Information Technologies.,' Yearb. Med. Inform., vol. 8, pp. 78–85, 2013, doi: 10.1055/s-0038-1638836.

3 N. Van Atteveldt, M. T. R. Van Kesteren, and B. Braams, 'Europe PMC Funders Group Neuroimaging of learning and development : improving ecological validity,' Front. Learn Res, vol. 6, no. 3, pp. 186–203, 2018, doi: 10.14786/flr.v6i3.366. Neuroimaging.

4 S. Naumann et al., 'Assessing the Degree of Ecological Validity of Your Study: Introducing the Multidimensional Assessment of Research in Context (MARC) Tool,' Mind, Brain, Educ., vol. 16, no. 3, pp. 228–238, 2022, doi: 10.1111/mbe.12318.

5 W. E. Norton, K. Loudon, D. A. Chambers, and M. Zwarenstein, 'Designing provider-focused implementation trials with purpose and intent: introducing the PRECIS-2-PS tool,' Implement. Sci., vol. 16, no. 1, pp. 1–11, 2021, doi: 10.1186/s13012-020-01075-y.

6 B. J. Cummings, C. J. Pike, R. Shankle, and C. W. Cotman, 'β-amyloid deposition and other measures of neuropathology predict cognitive status in Alzheimer's disease,' Neurobiol. Aging, vol. 17, no. 6, pp. 921–933, 1996, doi: 10.1016/S0197-4580(96)00170-4.

7 C. C. Gispen-De Wied, M. Kritsidima, and A. J. A. Elferink, 'The validity of biomarkers as surrogate endpoints in Alzheimer's disease by means of the Quantitative Surrogate Validation Level of Evidence Scheme (QSVLES),' J. Nutr. Heal. Aging, vol. 13, no. 4, pp. 376–387, 2009, doi: 10.1007/s12603-009-0049-2.

8 S. Ackley, S. Zimmerman, W. Brenowitz, and E. Tchetgen, 'Effect of reduction in myloid levels on cognitive change in randomized trials: intrumental variable meta-analysis,' BMJ, vol. 372, no. 156, 2021.

9 E. Karran and B. De Strooper, 'The amyloid hypothesis in Alzheimer disease: new insights from new therapeutics,' Nat. Rev. Drug Discov., vol. 21, no. 4, pp. 306–318, 2022, doi: 10.1038/s41573-022-00391-w.

10 H. Barenholtz Levy, 'Accelerated Approval of Aducanumab: Where Do We Stand Now?,' Ann. Pharmacother., vol. 56, no. 6, pp. 736–739, 2022, doi: 10.1177/10600280211050405.

11 F. Riederer, 'Donanemab in early Alzheimer's Disease,' J. fur Neurol. Neurochir. und Psychiatr., vol. 22, no. 3, pp. 142–143, 2021, doi: 10.1056/nejmoa2212948.

12 S. M. Fu, C. Dai, Z. Zhao, and F. Gaskin, 'Anti-dsDNA Antibodies are one of the many autoantibodies in systemic lupus erythematosus,' F1000Research, vol. 4, 2015, doi: 10.12688/f1000research.6875.1.

13 R. F. Van Vollenhoven et al., 'Belimumab in the treatment of systemic lupus erythematosus: High disease activity predictors of response,' Ann. Rheum. Dis., vol. 71, no. 8, pp. 1343–1349, 2012, doi: 10.1136/annrheumdis-2011-200937.

14 R. S. Fisher, 'Serum prolactin in seizure diagnosis: Glass half-full or half-empty?,' Neurol. Clin. Pract., vol. 6, no. 2, pp. 100–101, 2016, [Online]. Available: http: //www.ncbi.nlm.nih.gov/pubmed/29377043%0Ahttp: //www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5720615

15 D. K. Chen, Y. T. So, and R. S. Fisher, 'Use of serum prolactin in diagnosing epileptic seizures: Report of the Therapeutics and Technology Assessment Subcommittee of the American Academy of Neurology,' Contin. Lifelong Learn. Neurol., vol. 13, no. 4 EPILEPSY, pp. 188–195, 2007, doi: 10.1212/01.CON.0000284527.28025.0d.

16 M. S. Momani et al., 'Effect of Age, Gender, Food Intake, Obesity, and Smoking on Serum Levels of Prolactin in Healthy Adults,' J. Pers. Med., vol. 14, no. 9, p. 905, 2024, doi: 10.3390/jpm14090905.

17 J. Bauer, 'Epilepsy and prolactin in adults: A clinical review,' Epilepsy Res., vol. 24, no. 1, pp. 1–7, 1996, doi: 10.1016/0920-1211(96)00009-5.

18 FDA-CDER, 'Evaluating Drug Effects on the Ability to Operate a Motor Vehicle – Guidance for Industry,' 2017. [Online]. Available: https: //www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/default.htm

19 T. Sawada et al., 'Reliability and validity of on-road driving tests in vulnerable adults: a systematic review,' Int. J. Rehabil. Res., vol. 42, no. 4, pp. 289–299, 2019, doi: 10.1097/MRR.0000000000000374.

20 M. Rosenfeld, Y. Goverover, and P. Weiss, 'Self-awareness predicts fitness to drive among adults referred to occupational therapy evaluation,' Front. Rehabil. Sci., vol. 3, no. November, pp. 1–8, 2022, doi: 10.3389/fresc.2022.1005025.

21 J. Verster and T. Roth, 'Standard operation procedures for conducting the on-the-road driving test, and measurement of the standard deviation of lateral position (SDLP),' Int. J. Gen. Med., p. 359, 2011, doi: 10.2147/ijgm.s19639.

22 J. Verster, D. Veldhuijzen, A. Patat, B. Olivier, and E. Volkerts, 'Hypnotics and Driving Safety: Meta-Analyses of Randomized Controlled Trials Applying the on-the-Road Driving Test,' Curr. Drug Saf., vol. 1, no. 1, pp. 63–71, 2008, doi: 10.2174/157488606775252674.

23 L. A. Talbot, R. J. Musiol, E. K. Witham, and E. J. Metter, 'Falls in young, middle-aged and older community dwelling adults: Perceived cause, environmental factors and injury,' BMC Public Health, vol. 5, no. 86, Aug. 2005, doi: 10.1186/1471-2458-5-86.

24 I. Na, J. Seo, E. Park, and J. Lee, 'Risk of Falls Associated with Long-Acting Benzodiazepines or Tricyclic Antidepressants Use in Community-Dwelling Older Adults: A Nationwide Population-Based Case–Crossover Study,' Int. J. Environ. Res. Public Health, vol. 19, no. 14, 2022, doi: 10.3390/ijerph19148564.

25 Y. Jiang et al., 'Insomnia, benzodiazepine use, and falls among residents in long-term care facilities,' Int. J. Environ. Res. Public Health, vol. 16, no. 23, pp. 1–11, 2019, doi: 10.3390/ijerph16234623.

26 T. Hortobágyi, L. A. Teixeira, J. Duysens, U. Granacher, J. Van Diëen, and Renato de Moraes, 'Is standing sway an accurate measure of fall risk and predictor of future falls in older adults?,' Brazilian J. Mot. Behav., vol. 14, no. 01, pp. 1–3, 2020, doi: 10.20338/bjmb.v14i01.171.

27 Y. H. Pua, P. H. Ong, R. A. Clark, D. B. Matcher, and E. C. W. Lim, 'Falls efficacy, postural balance, and risk for falls in older adults with falls-related emergency department visits: Prospective cohort study,' BMC Geriatr., vol. 17, no. 1, pp. 1–7, 2017, doi: 10.1186/s12877-017-0682-2.

28 H. Shimada, S. Obuchi, N. Kamide, Y. Shiba, M. Okamoto, and S. Kakurai, 'Relationship with Dynamic Balance Function During Standing and Walking,' Am. J. Phys. Med. Rehabil., vol. 82, no. 7, pp. 511–516, 2003, doi: 10.1097/01.phm.0000064726.59036.cb.

29 A. Gabell, M. A. Simons, and U. L. S. Nayak, 'Falls in the healthy elderly: Predisposing causes,' Ergonomics, vol. 28, no. 7, pp. 965–975, 1985, doi: 10.1080/00140138508963219.

30 G. J. Groeneveld, J. L. Hay, and J. M. Van Gerven, 'Measuring blood–brain barrier penetration using the NeuroCart, a CNS test battery,' Jun. 01, 2016, Elsevier Ltd. doi: 10.1016/j.ddtec.2016.07.004.

31 J. Kim and S. Choi, 'Association of timed up and go test results with future injurious falls among older adults by sex: a population-based cohort study,' BMC Geriatr., vol. 24, no. 1, 2024, doi: 10.1186/s12877-024-05588-9.

32 E. Barry, R. Galvin, C. Keogh, F. Horgan, and T. Fahey, 'Is the Timed Up and Go test a useful predictor of risk of falls in community dwelling older adults: A systematic review and meta-analysis,' BMC Geriatr., vol. 14, no. 1, 2014, doi: 10.1186/1471-2318-14-14.

33 U. S. L. Nayak and J. M. Queiroga, 'Pinch grip, power grip and wrist twisting strengths of healthy older adults,' Gerontechnology, vol. 3, no. 2, 2004, doi: 10.4017/gt.2004.03.02.003.00.

34 J. A. Allen et al., 'Safety, tolerability, and efficacy of subcutaneous efgartigimod in patients with chronic inflammatory demyelinating polyradiculoneuropathy (ADHERE): a multicentre, randomised-withdrawal, double-blind, placebo-controlled, phase 2 trial,' Lancet Neurol., vol. 23, no. 10, pp. 1013–1024, 2024, doi: 10.1016/S1474-4422(24)00309-0.

35 F. Louter et al., 'Instruments for measuring the neuromuscular function domain of vitality capacity in older persons: an umbrella review,' Eur. Geriatr. Med., vol. 15, no. 5, pp. 1191–1213, 2024, doi: 10.1007/s41999-024-01017-7.

36 S. C. Higgins, J. Adams, and R. Hughes, 'Measuring hand grip strength in rheumatoid arthritis,' Rheumatol. Int., vol. 38, no. 5, pp. 707–714, 2018, doi: 10.1007/s00296-018-4024-2.

37 H. C. Roberts et al., 'A review of the measurement of grip strength in clinical and epidemiological studies: Towards a standardised approach,' Age Ageing, vol. 40, no. 4, pp. 423–429, 2011, doi: 10.1093/ageing/afr051.

38 J. C. MacDermid and V. Tottenham, 'Responsiveness of the Disability of the Arm, Shoulder, and Hand (DASH) and Patient-Rated Wrist/Hand Evaluation (PRWHE) in Evaluating Change after Hand Therapy,' J. Hand Ther., vol. 17, no. 1, pp. 18–23, 2004, doi: 10.1197/j.jht.2003.10.003.

39 L. Arendt-Nielsen and H. C. Hoeck, 'Optimizing the early phase development of new analgesics by human pain biomarkers,' Expert Rev. Neurother., vol. 11, no. 11, pp. 1631–1651, 2011, [Online]. Available: https: //doi.org/10.1586/ern.11.147

40 K. Rietdijk and M. Claessens, 'Distinct pharmacodynamic effects of selective sodium channel blockers with different mechanisms of actions on Nerve Excitability Threshold Tracking: a randomized, double-blind, placebo-controlled study,' Poster presented at: 19th European Congress of Clinical Neurophysiology; Sep 2025; London

APPENDICES

ENGLISH SUMMARY

This thesis explores the integration of ecological validity in biomarkers used in early-phase clinical drug development. Ecological validity refers to the extent to which a biomarker reflects real-world clinical outcomes and patient-relevant functional capacities. While early clinical trials primarily assess safety and pharmacokinetics in healthy volunteers, the progression toward registration trials requires outcome measures that are both scientifically robust and clinically meaningful. Traditional biomarkers often lack the ability to predict clinical outcome assessments (COAs) needed for registration, creating a translational gap between proof-of-concept studies and real-world efficacy.

To bridge this gap, this research introduces a novel tiered framework named the Ecological Validity of Biomarkers (EVIB). This framework categorizes biomarkers based on their relevance and predictive value for real-world outcomes. The framework consists of six tiers, ranging from pharmacological proof (Tier 1) to real-world evidence (Tier 6). This provides a structured method for assessing biomarkers' applicability throughout the clinical development pipeline.

Besides the introduction, this thesis applies the framework across multiple case studies, each involving different functional domains: driving behaviour, fall risk, grip strength, and pain perception. These studies demonstrate that while high-tier biomarkers (e.g., on-the-road driving tests or real-life gait assessments) offer higher ecological validity, they are often costly, time-consuming, or logistically challenging. Intermediate biomarkers, such as those derived from driving simulators or dynamic walking tasks, can offer a balance between feasibility and real-world relevance. Similarly, novel tools like the PowerJar device and VR-PainCart show potential for increasing ecological validity when conventional measures fall short.

Biomarker selection for clinical trials should account not only for technical rigor but also for contextual relevance to daily life activities. Ecological validity, though dynamic and influenced by evolving scientific evidence, is essential for enhancing the predictive power of early-phase studies and optimizing drug development efficiency. By structuring the assessment of ecological validity, this work offers a practical strategy to improve translational reliability and reduce late-stage clinical trial failures.

NEDERLANDSE SAMENVATTING

Inleiding en onderzoeksvraag

De ontwikkeling van nieuwe geneesmiddelen is een langdurig, kostbaar en risicovol proces. Slechts een klein deel van de kandidaatmiddelen die de klinische fase bereiken, komt uiteindelijk op de markt. In vroege klinische studies, doorgaans uitgevoerd in gezonde vrijwilligers, ligt de nadruk op veiligheid, farmacokinetiek en de eerste farmacodynamische signalen. De overgang van deze onderzoeken naar grootschalige registratieonderzoeken verloopt echter vaak moeizaam. Een belangrijke oorzaak hiervoor lijkt te zijn dat de in vroege fasen gebruikte biomarkers onvoldoende aansluiten bij klinisch relevante uitkomstmaten (clinical outcome assessments, COA's) die door registratieautoriteiten worden geaccepteerd. Dit leidt tot een kloof tussen proof-of-concept studies en daadwerkelijke effectiviteit in de klinische praktijk.

Het doel van dit proefschrift is het identificeren en beoordelen van biomarkers met een hoge ecologische validiteit in vroege-fase geneesmiddelenonderzoek. Ecologische validiteit verwijst naar de mate waarin een biomarker aansluit bij het dagelijks functioneren van patiënten en daadwerkelijk voorspellend is voor klinische uitkomsten. Om dit systematisch te evalueren is een raamwerk ontwikkeld: de Ecological Validity of Biomarkers (EVIB). Deze score kent zes niveaus, variërend van puur farmacologisch bewijs (niveau 1) tot gegevens gebaseerd op de dagelijkse praktijk (niveau 6). Met behulp van de EVIB is de ecologische validiteit van verschillende biomarkers onderzocht binnen vier functionele domeinen: rijgedrag, valrisico, spierkracht en pijnbeleving.

Hoofdstuk 2 – Rijgedrag

Een belangrijk veiligheidsaspect bij geneesmiddelen die slaperigheid of motorische vertraging veroorzaken, is de invloed op de rijvaardigheid. Een verminderde rijvaardigheid kan zowel de verkeersveiligheid als de maatschappelijke acceptatie van een geneesmiddel ernstig ondermijnen. Daarom is het cruciaal dergelijke effecten al vroeg in het ontwikkelproces te identificeren.

In dit hoofdstuk werd onderzocht hoe slaaptekort de rijvaardigheid beïnvloedt en in welke mate verschillende meetmethoden dit effect kunnen vastleggen. Drie benaderingen werden vergeleken: een rijtest op de openbare weg, een rijsimulator en een batterij van psychomotorische taken. De belangrijkste uitkomstmaat was de standaarddeviatie van laterale positie (SDLP), een objectieve maat voor slingeren binnen de rijstrook.

De resultaten toonden dat slaaptekort leidde tot een duidelijke verslechtering van rijgedrag, zowel op de weg als in de simulator. Psychomotorische taken, zoals het volgen van een bewegende stip (pursuit tracking), bleken echter slechts beperkt voorspellend voor rijprestaties. Binnen het EVIB-raamwerk werd de openbare-wegtest geclassificeerd op niveau 5 (dicht bij dagelijks functioneren), de simulator op niveau 4–5 afhankelijk van de mate van realisme, en de psychomotortaken op niveau 3 (klinische correlatie zonder directe functionele relevantie).

Deze resultaten illustreren dat tests met een lagere ecologische validiteit waardevol zijn in vroege studies vanwege hun praktische toepasbaarheid, maar dat hogere niveaus noodzakelijk zijn om de klinische relevantie te bevestigen. Een gefaseerde aanpak, waarbij eenvoudige tests in vroege fasen worden gecombineerd met realistischere metingen in latere stadia, is het meest efficiënt.

Hoofdstuk 3 – Valrisico

Vallen is een van de belangrijkste veiligheidsrisico's bij het gebruik van geneesmiddelen, met name bij oudere patiënten. Slaapmiddelen zoals benzodiazepinen verhogen het valrisico aanzienlijk. Het daadwerkelijk meten van het aantal vallen in klinische studies is echter onpraktisch, omdat dit langdurige follow-up en grote cohorten vereist. Daarom is er behoefte aan betrouwbare surrogaatmaten die valrisico vroegtijdig kunnen voorspellen.

In dit hoofdstuk werden drie opties voor deze surrogaatmaat vergeleken: de klassieke maat body sway (het meten van lichaamszwaai tijdens stilstand), de Timed-Up-and-Go test (TUG, een eenvoudige mobiliteitstest), en de Interactive Walkway (IWW), waarin deelnemers dynamische taken zoals obstakelvermijding uitvoeren. In een crossover-studie bij ouderen werden placebo, zolpidem (bekend verhoogd valrisico) en suvorexant (een nieuw middel met mogelijk gunstiger profiel) vergeleken.

De resultaten toonden dat zolpidem leidde tot significante verslechteringen op alle maten, terwijl suvorexant veel minder effecten vertoonde. De IWW bleek het meest gevoelig voor verschillen en bootste bovendien realistisch gedrag na, zoals het ontwijken van obstakels. Binnen de EVIB werd body sway geclassificeerd op niveau 3, de TUG op niveau 4, en de IWW op niveau 4 met potentie tot 5.

Deze studie benadrukt dat traditionele maten zoals body sway te beperkt zijn om valrisico adequaat te voorspellen. Dynamische tests zoals de IWW bieden daarentegen een betrouwbaarder beeld doordat zij meerdere oorzaken van vallen (bijvoorbeeld obstakels of onverwachte situaties) integreren.

Hoofdstuk 4 – Spierkracht

Spierkracht is een belangrijke parameter in veel ziekten en een bekende voorspeller van functioneren. De standaardmethode is het meten van knijpkracht met een handdynamometer. Hoewel deze test eenvoudig en reproduceerbaar is, weerspiegelt zij slechts beperkt hoe spierkracht zich vertaalt naar dagelijkse activiteiten.

In dit hoofdstuk werd de PowerJar onderzocht, een nieuw instrument dat het openen van een potje nabootst en zowel knijp- als draaikracht meet. Dit apparaat werd getest bij gezonde vrijwilligers en patiënten met neuromusculaire aandoeningen. De PowerJar leverde stabiele en betrouwbare metingen, met een sterke correlatie met de dynamometer, maar voegde ook nieuwe informatie toe doordat een alledaagse taak werd nagebootst.

Binnen de EVIB werd de dynamometer geclassificeerd als niveau 3, omdat het een klinisch relevant verband toont zonder directe functionele betekenis. De PowerJar bereikte niveau 4, doordat deze een herkenbare dagelijkse activiteit weerspiegelt en daardoor dichter bij het dagelijks functioneren van patiënten staat.

Deze resultaten laten zien dat relatief kleine innovaties in meetinstrumenten kunnen leiden tot biomarkers die ecologisch meer valide zijn. Daarmee neemt hun voorspellende waarde voor klinische relevantie aanzienlijk toe.

Hoofdstukken 5 en 6 – Pijnbeleving

Pijn is een complex en subjectief verschijnsel waarbij zowel de sensorische als de affectieve-emotionele componenten een rol spelen. Klassieke pijnmodellen in gezonde vrijwilligers richten zich vooral op het meten van pijnprikkeldrempels, maar laten de emotionele dimensie grotendeels buiten beschouwing. Daardoor sluiten deze modellen beperkt aan bij de ervaring van klinische pijn bij patiënten.

Om dit probleem te adresseren werd de VR-PainCart ontwikkeld. In deze methode kregen deelnemers een elektrische pijnstimulus toegediend terwijl een virtuele wond op hun arm werd geprojecteerd. Dit leidde tot een hogere ervaren pijnintensiteit en onaangenaamheid. In een vervolgonderzoek werd onderzocht of diazepam, een angstremmend middel, dit effect kon moduleren. Inderdaad nam de invloed van de virtuele wond af onder diazepam, wat aangeeft dat ook de emotionele component van pijn farmacologisch te beïnvloeden is.

Binnen de EVIB werd de klassieke elektrische stimulatie ingedeeld op niveau 1–2 (farmacologisch bewijs en mechanistisch inzicht), terwijl de VR-PainCart werd geclassificeerd als niveau 2, met potentie om door te groeien naar hogere niveaus bij verdere validatie in patiëntenpopulaties.

Deze studies tonen dat het toevoegen van emotionele dimensies met behulp van virtual reality een veelbelovende weg is om pijn realistischer te meten. Dit kan leiden tot biomarkers die relevanter zijn voor de klinische praktijk, met name voor geneesmiddelen die inwerken op de emotionele component van pijn.

Hoofdstuk 7 – Algemene discussie en raamwerk

In de algemene discussie wordt de centrale onderzoeksvraag beantwoord: ecologische validiteit is een cruciaal criterium bij de keuze en ontwikkeling van biomarkers in geneesmiddelenonderzoek. Het EVIB-raamwerk biedt een systematische manier om biomarkers te plaatsen binnen een continuum dat loopt van puur farmacologisch bewijs tot real-world evidence.

De toepassing van het EVIB op de vier domeinen toont dat biomarkers met een lage ecologische validiteit waardevol zijn voor vroege signalering en kostenefficiëntie, maar dat biomarkers met een hogere ecologische validiteit essentieel zijn om de werkelijke impact voor patiënten vast te stellen. Innovatieve methoden zoals de Interactive Walkway, de PowerJar en de VR-PainCart laten zien dat ecologische validiteit daadwerkelijk kan worden verhoogd door tests dichter bij het dagelijks functioneren te brengen.

Daarnaast wordt benadrukt dat ecologische validiteit dynamisch is: de waarde van een biomarker kan veranderen met nieuw wetenschappelijk bewijs. Een voorbeeld is amyloïd als biomarker bij Alzheimer, waarvan de betekenis door de jaren heen sterk fluctueerde. Dit toont aan dat ecologische validiteit niet statisch is, maar meegroeit met de stand van de wetenschap.

De conclusie luidt dat selectie van biomarkers op basis van hun plaats binnen het EVIB-raamwerk kan bijdragen aan het verminderen van mislukkingen in latere klinische fasen en daarmee aan een efficiëntere, betrouwbaardere en patiëntgerichtere geneesmiddelenontwikkeling.

Conclusie

Dit proefschrift laat zien dat het structureel evalueren van ecologische validiteit essentieel is voor het verbeteren van de vertaalslag van vroege biomarkers naar klinisch relevante uitkomsten. Het EVIB-raamwerk vormt hiervoor een praktisch instrument en kan breed worden toegepast in geneesmiddelenonderzoek. Door biomarkers systematisch te plaatsen binnen dit raamwerk wordt de voorspellende waarde van vroege studies verhoogd, wordt de kans op discontinuatie van medicatie in late fase verkleind en kan de ontwikkeling van geneesmiddelen meer in lijn worden gebracht met de werkelijke behoeften van patiënten.

SAMENVATTING IN BEGRIJPELIJKE TAAL

Het ontwikkelen van nieuwe medicijnen is ingewikkeld, duur en gaat vaak mis. Van alle middelen die in mensen worden getest komt uiteindelijk maar een klein deel echt op de markt. In de eerste onderzoeken bij gezonde vrijwilligers kijken onderzoekers vooral of een middel veilig is, hoe het lichaam het verwerkt en of het middel werkt (dus doet wat het hoort te doen). Maar of een middel ook daadwerkelijk helpt in het dagelijks leven van patiënten, blijkt meestal pas veel later. Daardoor vallen veel medicijnen in een laat stadium alsnog af.

In dit proefschrift is onderzocht hoe we eerder en beter kunnen voorspellen of een medicijn echt helpt. Daarvoor hebben we gekeken naar zogenaamde *biomarkers*: meetbare signalen, zoals lichaamstesten of gedragsmetingen, die kunnen aangeven of een medicijn effect heeft. Het probleem is dat veel van die biomarkers weinig zeggen over hoe iemand functioneert in het dagelijks leven. Daarom staat in dit onderzoek de ecologische validiteit centraal: de vraag hoe goed een biomarker aansluit bij de echte wereld van de patiënt.

Om dit te meten hebben we een nieuw scoresysteem ontwikkeld: de Ecological Validity of Biomarkers (EVIB). Deze score loopt van laag (een meting die vooral iets zegt over hoe een middel in het lichaam werkt) tot hoog (een meting die direct laat zien hoe iemand functioneert in het dagelijks leven). Met dit systeem hebben we vier verschillende domeinen onderzocht: rijgedrag, valrisico, spierkracht en pijnbeleving.

Rijgedrag

Sommige medicijnen, zoals slaapmiddelen, maken mensen slaperig of trager, wat gevaarlijk kan zijn in het verkeer. In een studie hebben we deelnemers een nacht wakker gehouden zodat ze slaperig ware en daarna hun rijvaardigheid gemeten met drie methoden: een echte rijtest op de weg, een rijsimulator en korte computertaken. De echte rijtest gaf het meest realistische beeld. De simulator bleek ook bruikbaar, al iets minder nauwkeurig. De computertaken gaven vooral vroege aanwijzingen, maar konden het echte rijden niet goed voorspellen.

Valrisico

Oudere mensen hebben een groter risico om te vallen, en sommige slaapmiddelen vergroten dat risico nog eens. Het tellen van daadwerkelijke vallen in een onderzoek is bijna niet te doen, omdat het te veel tijd en deelnemers kost. Daarom hebben we verschillende testen vergeleken: stilstaan en de

lichaamszwaai meten, de Timed-Up-and-Go test (opstaan, een stukje lopen en weer gaan zitten), en een nieuwe methode: de *Interactive Walkway*. Daarbij moeten deelnemers tijdens het lopen obstakels vermijden. Uit de resultaten bleek dat de Interactive Walkway het meest realistische en gevoelige beeld gaf van valrisico, veel beter dan de eenvoudige evenwichtstest.

Spierkracht

Knijpkracht wordt vaak gemeten met een handdynamometer: een apparaat dat meet hoeveel kracht iemand in de handen heeft. Deze test is handig, maar zegt weinig over hoe iemand die kracht in het dagelijks leven gebruikt. Daarom hebben we de *PowerJar* ontwikkeld, een apparaat dat het openen van een potje nabootst. Bij gezonde vrijwilligers en patiënten met spierziekten bleek de PowerJar goed bruikbaar en stabiel. Bovendien liet het apparaat meer zien over hoe spierkracht nodig is in gewone handelingen. Dit maakt de PowerJar een realistischer meetinstrument.

Pijnbeleving

Pijn is niet alleen een lichamelijk, maar ook een emotioneel verschijnsel. Klassieke testen meten vooral de drempel waarop pijn wordt gevoeld, maar houden geen rekening met de gevoelens die pijn oproept. Om dit beter te onderzoeken ontwikkelden we de *VR-PainCart*. Tijdens een elektrische prikkel zagen deelnemers in een virtuele omgeving een wond op hun arm verschijnen. Dit maakte de pijnbeleving sterker en onaangenamer. Toen deelnemers een angstremmer kregen (diazepam), nam dit effect af. Dat laat zien dat geneesmiddeleffecten op de emotionele kant van pijn meetbaar zijn.

Conclusie

Dit proefschrift laat zien dat het belangrijk is om biomarkers te kiezen die iets zeggen over het dagelijks functioneren van patiënten. Hoe realistischer de test, hoe groter de kans dat het resultaat in vroege studies ook echt voorspellend is voor succes in latere stadia van onderzoek. Met de EVIB-score is er nu een praktisch systeem om de ecologische validiteit van biomarkers te beoordelen.

Door dit systeem te gebruiken, kunnen onderzoekers al vroeg zien welke biomarkers goed passen en dus welke medicijnen kansrijk zijn. Dat kan tijd en geld besparen, maar zorgt er ook voor dat de medicijnen die uiteindelijk de markt bereiken, beter aansluiten bij wat voor patiënten echt belangrijk is: functioneren in het dagelijks leven.

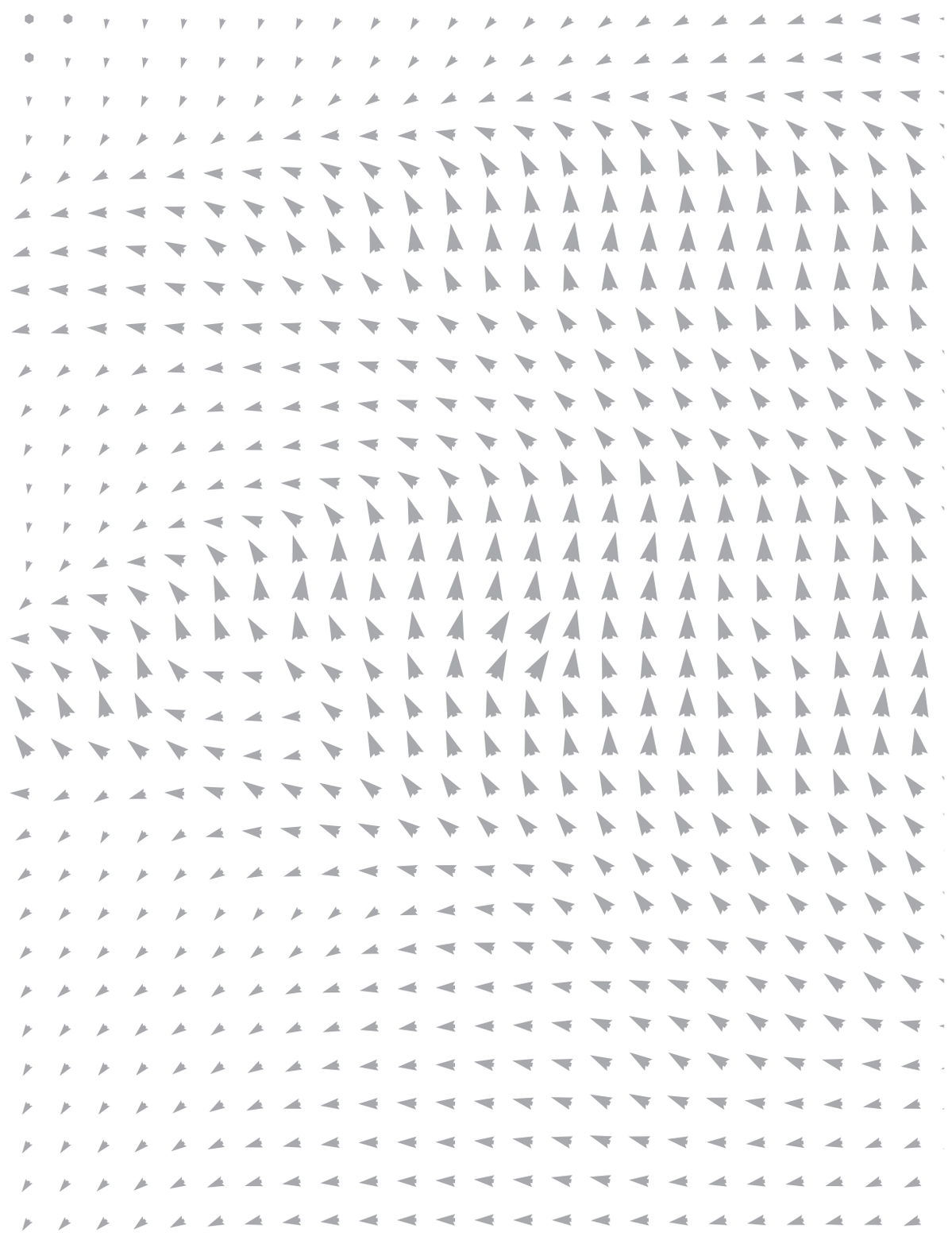
CURRICULUM VITAE

Ingrid Koopmans (born 1991 in Wageningen) grew up in Lochem, the Netherlands. She completed her secondary education at Staring College before enrolling at the University of Twente, where she studied Technical Medicine. During her studies, she was actively involved in student life and participated in various extracurricular activities alongside her academic work.

After graduating, Ingrid started her professional career in November 2017 at CHDR (Centre for Human Drug Research) in the Method Development group. There, she worked within a multidisciplinary team to unlock the true potential of biomarkers in clinical trials. Although initially uncertain about pursuing a PhD, she started her doctoral journey in 2021 under supervision of dr. ir. R.J. Doll and prof. dr. G.J. Groeneveld in 2021.

Ingrid's research initially covered a broad scope but became increasingly focused on pain-related studies after she joined the Pain research group as a Clinical Scientist in 2022. In 2023, she was appointed Experienced Clinical Scientist within the same group. Since July 2025, she continued as Senior Clinical Scientist. In this role, she guides research initiatives, oversees the project leaders working at clinical trials aimed at the understanding and treatment of pain, and supervises PhD students involved in related research.

Since 2016, Ingrid has lived in Leiden with her partner, Marc Hulsebosch. Outside of work, she enjoys travelling with Marc and renovating their home and garden.



SCIENTIFIC CONTRIBUTIONS

FULL PAPERS

The interactive walkway provides fit-for-purpose fall-risk biomarkers in the elderly: Comparison of zolpidem and suvorexant. **Ingrid Koopmans**, Daphne J. Geerse, Lara de Ridder, Melvyn Roerdink, Maria Joanna Juachon, Clemens Muehlan, Jasper Dingemanse, Joop van Gerven, Geert Jan Groeneveld, Rob Zuiker

Fit for purpose of on-the-road driving and simulated driving: A randomised crossover study using the effect of sleep deprivation. **Ingrid Koopmans**, Robert J. Doll, Hein E C van der Wall, Marieke de Kam, Geert Jan Groeneveld, Adam F. Cohen, Rob Zuiker

The impact of a virtual wound on pain sensitivity: insights into the affective dimension of pain. **Ingrid Koopmans**, Robert-Jan Doll, Maurice Hagemeijer, Robert van Barneveld, Marieke de Kam, Geert Jan Groeneveld

Virtual Reality in a nociceptive pain test battery: a randomized, placebo controlled two-way crossover study with diazepam. **Ingrid Koopmans**, Koen Rietdijk, Roman Bohoslavsky, Robert-Jan Doll, Geert Jan Groeneveld

Safety, pharmacokinetics, and pharmacodynamics of a clc-1 inhibitor - a first-in-class compound that enhances muscle excitability: a phase I, single- and multiple-ascending dose study. Titia Ruijs, Kaye de Cuba, Jules Heuberger, John Hutchison, Jane Bold, Thomas Grønnebæk, Klaus Jensen, Eva Chin, Jorge Quiroz, Thomas Petersen, Peter Flagstad, Marieke de Kam, Michiel van Esdonk, Erica Klaassen, Robert J. Doll, **Ingrid Koopmans**, Annika de Goede, Thomas Pedersen, Geert Jan Groeneveld

Objective monitoring of facioscapulohumeral dystrophy during clinical trials using a smartphone app and wearables: observational study. Ghobad Maleki*, Ahnjili Zhuparris*, **Ingrid Koopmans**, Robert J Doll, Nicoline Voet, Adam Cohen, Emilie van Brummelen, Geert Jan Groeneveld, Joris De Maeyer
**these authors contributed equally*

A smartphone- and wearable-based biomarker for the estimation of unipolar depression severity. Ahnjili Zhuparris, Ghobad Maleki, Liesbeth van Londen, **Ingrid Koopmans**, Vincent Aalten, Iris E. Yocarini, Vasileios Exadaktylos, A. van Hemert, Adam Cohen, Pim Gal, Robert J. Doll, Geert Jan Groeneveld, Gabriël Jacobs, Wessel Kraaij

Smartphone and Wearable Sensors for the Estimation of Facioscapulohumeral Muscular Dystrophy Disease Severity: Cross-sectional Study. Ahnjili Zhuparris, Ghobad Maleki, **Ingrid Koopmans**, Robert J. Doll, Nicoline Voet, Wessel Kraaij, Adam Cohen, Emilie van Brummelen, Joris De Maeyer, Geert Jan Groeneveld

Muscle velocity recovery cycles as pharmacodynamic biomarker: Effects of mexiletine in a randomized double-blind placebo-controlled cross-over study. Titia Ruijs, **Ingrid Koopmans**, Marieke de Kam, Martijn Tannemaat, Geert Jan Groeneveld, Jules Heuberger

Objective Monitoring of Facioscapulohumeral Dystrophy During Clinical Trials Using a Smartphone App and Wearables: Observational Study. Ghobad Maleki, Ahnjili Zhuparris, **Ingrid Koopmans**, Robert J. Doll, Nicoline Voet, Adam Cohen, Emilie van Brummelen, Geert Jan Groeneveld, Joris De Maeyer

A crossover study evaluating the sex-dependent and sensitizing effects of sleep deprivation using a nociceptive test battery in healthy subjects. Hemme J. Hijma, **Ingrid Koopmans**, Erica Klaassen, Robert J. Doll, Rob Zuiker, Geert Jan Groeneveld

Effects of Mexiletine and Lacosamide on Nerve Excitability in Healthy Subjects: A Randomized, Double-Blind, Placebo-Controlled, Crossover Study. Titia Q Ruijs, **Ingrid Koopmans**, Marieke de Kam, Michiel van Esdonk, Martin Koltzenburg, Geert Jan Groeneveld, Jules Heuberger

Simultaneous measurement of intra-epidermal electric detection thresholds and evoked potentials for observation of nociceptive processing following sleep deprivation. Boudewijn van den Berg, Hemme Hijma, **Ingrid Koopmans**, Robert J. Doll, Rob Zuiker, Geert Jan Groeneveld, Jan Buitenweg

Using machine learning techniques to characterize sleep-deprived driving behaviour. Hein van der Wall, Robert J. Doll, Gerard van Westen, **Ingrid Koopmans**, Rob Zuiker, Koos Burggraaf, Adam Cohen

The use of machine learning improves the assessment of drug-induced driving behaviour. Hein van der Wall, Robert J. Doll, Gerard van Westen, **Ingrid Koopmans**, Rob Zuiker, Koos Burggraaf, Adam Cohen

Nighttime safety of daridorexant: Evaluation of responsiveness to an external noise stimulus, postural stability, walking, and cognitive function. Massimo Magliocca, **Ingrid Koopmans**, Cedric Vaillant, Vincent Lemoine, Rob Zuiker, Jasper Dingemanse, Clemens Muehlan