Ecological validity of biomarkers in drug research

Koopmans, I.W.

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]

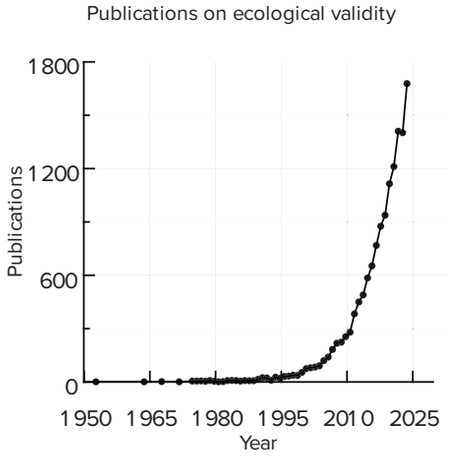
It is important to realise that a biomarker can be placed into a high tier while lacking evidence for the lower tiers. Studies based on real-world evidence (i.e., tier 6) or clinical data (i.e., tier 3) can result in the development of new biomarkers for drug development (i.e., tier 1). Additionally, 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