Aus dem
CharitéCentrum für Innere Medizin und Dermatologie
Medizinische Klinik mit Schwerpunkt Psychosomatik
Direktor: Prof. Dr. med. Matthias Rose

Habilitationsschrift

Der Nutzen von modernen, patientenzentrierten Erfassungsmethoden der Präzisionsmedizin bei neurobehavioralen Erkrankungen

zur Erlangung der Lehrbefähigung
für das Fach ***Innere Medizin***

vorgelegt dem Fakultätsrat der Medizinischen Fakultät
Charité-Universitätsmedizin Berlin

von
Dr. med. Alexander Obbarius

Eingereicht: Juni 2023
Dekan: Prof. Dr. Joachim Spranger
1. Gutachter/in: Prof. Dr. Bernd Löwe, Hamburg-Eppendorf
2. Gutachter/in: Prof. Dr. Tanja Stamm, Wien

Inhaltsverzeichnis

Abkürzungsverzeichnis

CAT Computer-adaptiver Test

DSEM Dynamic Structural Equation Modeling

EMA Ecological Momentary Assessment

EORTC European Organization for Research and Treatment of Cancer

HER2/neu Human epidermal growth factor receptor 2

ICHOM International Consortium for Health Outcomes Measurement

IMMPACT Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials

IRT Item Response Theory

iSD intraindividuelle Standardabweichung

JITAI Just-In-Time-Adaptive Intervention

KTT Klassische Testtheorie

MID Minimally Important Difference

OPD Operationalisierte Psychodynamische Diagnostik

PCT Precision Clinical Trial

PHQ Patient Health Questionnaire

PRO Patient-Reported Outcome

PROM Patient-Reported Outcome Measure

PROMIS Patient-Reported Outcome Measurement Information System

RCI Reliable Change Index

RCT Randomized Controlled Trial

VBC Value-based Healthcare

WHO World Health Organization

WHODAS 2.0 WHO Disability Assessment Schedule 2.0

WPD Within-Person Dynamic

YLD Years lived with disability

1. Einleitung

1.1 Krankheitslast neurobehavioraler Erkrankungen und die Notwendigkeit für Präzisionsmedizin

Neurobehaviorale Erkrankungen umfassen eine heterogene Gruppe an Erkrankungen, für die zwar pathophysiologische Veränderungen im Nervensystem und anderen Organen bekannt sind, diese Erkrankungen jedoch nicht im Sinne eines (mono-)kausalen Zusammenhangs bedingen. Bei den meisten dieser Erkrankungen wird mit zunehmender Sicherheit vermutet, dass diese multifaktoriell sowohl durch biologische als auch psychische und soziale Ursachen verursacht und aufrechterhalten werden [1]. Zu diesen Erkrankungen gehören beispielsweise Depressionen, Angsterkrankungen, psychotische Erkrankungen, Suchterkrankungen, Adipositas, Tinnitus oder chronische Schmerzen. Die hohe Bedeutung dieser Erkrankungen entsteht durch den hohen Beitrag zur weltweiten Morbidität und Mortalität. So befinden sich laut der “Global Burden of Disease Study“ Schmerzerkrankungen, Angsterkrankungen und Depressionen unter den 10 Erkrankungen, die für die meisten mit Einschränkungen gelebten Lebensjahre („Years lived with disability“, YLD) verantwortlich sind [2].

Die hohe Krankheitslast durch neurobehaviorale Erkrankungen wie auch anderer chronischer Erkrankungen ist unter anderem eine Folge der gestiegenen Lebenserwartung, die vor allem durch eine gebesserte Gesundheitsversorgung und Lebensstilveränderungen in den letzten Jahrzehnten bedingt ist. Weil die Menschen immer älter werden, ist auch die globale Morbidität angestiegen, die sich in einer längeren Erkrankungsdauer ausdrückt. Die Herausforderung für die Gesundheitssysteme besteht deshalb darin, die Morbidität zu verringern, indem Menschen länger gesund bleiben (z.B. durch verbesserte Prävention) und auch im Verlauf chronischer Erkrankungen eine möglichst gute Lebensqualität aufrechtzuerhalten („Kompression“ der Morbidität) [3]. Eine Verbesserung der Lebensqualität bei neurobehavioralen Erkrankungen, wie z.B. Schmerz- oder depressiven Erkrankungen, ist nur möglich, wenn bestehende Behandlungsmöglichkeiten weiterentwickelt werden, um eine höhere Effektivität der Therapien zu erreichen. Bei vielen neurobehavioralen Erkrankungen existiert aktuell das Problem, dass generell wirksame Behandlungen nicht verfügbar sind. Das bedeutet, dass Behandlungen nur bei einem Teil der Patient:innen oder nur in begrenztem Maße wirksam sind [4-6]. Das könnte daran liegen, dass pathophysiologische Zusammenhänge nicht ausreichend bekannt sind. Bei chronischen Erkrankungen, deren Pathophysiologie besser verstanden ist, wie z.B. bei einigen kardiovaskulären und onkologischen Erkrankungen, konnten in den letzten Jahrzehnten immer bessere zielgerichtete Präventionsmaßnahmen oder Therapien entwickelt werden. Eines der prominentesten Beispiele ist die rasante Entwicklung im Bereich der Biologika, welche aufgrund eines genaueren Verständnisses von immunologischen Prozessen bei Tumoren oder Autoimmunerkrankungen möglich ist und welche die Behandlung vieler Erkrankungen revolutioniert

hat. So können beispielsweise unterstützt durch die neuen zielgerichteten Therapien knapp 90% der Patientinnen mit Mamma-Karzinom [7] und praktisch alle Patient:innen mit malignen Lymphomen in frühen Krankheitsstadien geheilt werden [8].

In randomisierten, kontrollierten Studien („Randomized Controlled Trials“, RCTs) zu neurobehavioralen Erkrankungen findet sich immer wieder eine hohe Variabilität der Behandlungsergebnisse. Während einzelne Patient:innen gut von einer bestimmten Behandlung profitieren, profitieren andere Patient:innen von der gleichen Behandlung gar nicht, oder verschlechtern sich darunter sogar [4, 5]. So finden sich in Placebo-kontrollierten Medikamentenstudien bei Patient:innen mit Depressionen, Angst- oder Schmerzerkrankungen im Durchschnitt sehr kleine Effektstärken [9-11].

Eine mögliche Erklärung für die große Variabilität der Behandlungsergebnisse ist, dass sich hinter einer bestimmten Diagnose, wie einer Depression oder chronischen Rückenschmerzen nicht jeweils eine Erkrankung, sondern unterschiedliche Erkrankungen verbergen oder dass es bestimmte Kombinationen von Merkmalen gibt, die Subgruppen voneinander unterscheiden. Daher gibt es im Bereich der neurobehavioralen Erkrankungen einen zunehmenden Wunsch nach einem Paradigmenwechsel im Sinne der Präzisionsmedizin [4, 5].

Präzisionsmedizin ist ein Behandlungsansatz, der die individuellen Unterschiede berücksichtigt, indem die richtige Behandlung der/dem richtigen Patient:in zum richtigen Zeitpunkt zukommt, in der Erwartung einer besseren Gesundheit bei geringeren Kosten [12]. Im Gegensatz zu personalisierter Medizin, die häufig als individuelle Behandlung einzelner Patient:innen verstanden wird, hat die Präzisionsmedizin das Ziel, auf der Basis umfangreicher Daten bestimmte Subgruppen zu identifizieren, die besonders gut auf bestimmte Behandlungen ansprechen. Neben genetischen Daten („Genomics“) können in Zukunft mutmaßlich z.B. Daten über Proteine („Proteomics“), Stoffwechsel-Eigenschaften („Metabolomics“) oder auch mobile Gesundheitsdaten für die Charakterisierung von Patient:innen oder Gruppen von Patient:innen genutzt werden. Die Verfügbarkeit umfangreicher Informationen aus Datenbanken, elektronischen Gesundheitsakten und Daten von mobilen Endgeräten („Big Data“) ist einer der zahlreichen Faktoren, die es jetzt ermöglichen, große Datenmengen zu integrieren und auszuwerten.

1.2 Marker für die Stratifizierung bei neurobehavioralen Erkrankungen

Eine Herausforderung für die Präzisionsmedizin bei neurobehavioralen Erkrankungen ist bisher das Fehlen von einheitlichen Markern, welche die Stratifizierung für unterschiedliche Therapien erlauben. Eine zunehmende Anzahl von Studien beschäftigt sich damit, *Biomarker* zu identifizieren, die „einfach“, d.h. direkt messbar sind, vergleichbar beispielsweise mit der Bestimmung der genetischen

Überexpression von HER2/neu bei Patient:innen mit Mammakarzinom [7]. Zu den Biomarkern gehören neben genetischen Markern auch Neuroimaging, elektrophysiologische Verfahren wie Transkranielle Magnetstimulation und Elektroenzephalografie oder auch psychophysische Verfahren wie die Quantitative Sensorische Testung [4, 5].

Aufgrund ihrer Komplexität sind viele pathophysiologische Prozesse, die neurobehaviorale Erkrankungen verursachen oder aufrechterhalten (noch) nicht bekannt. Die Komplexität ergibt sich aus der Tatsache, dass viele dieser Krankheitsbilder keine einzelnen Erkrankungen darstellen, sondern vielmehr heterogene Syndrome mit multiplen neurobiologischen, verhaltensbezogenen und umweltbedingten Ursachen sind. Hinzu kommt, dass der Behandlungsverlauf bei diesen Erkrankungen durch ein komplexes, dynamisches Zusammenspiel zwischen Patient:innen, der Behandlung, den Behandler:innen, sowie den Umgebungsfaktoren gekennzeichnet ist.

Aus diesen Voraussetzungen ergeben sich mehrere Bedingungen, die Marker für die Stratifizierung bei neurobehavioralen Erkrankungen mutmaßlich erfüllen sollten, um relevante Patient:innen-Subgruppen zu identifizieren:

1) Die Stratifizierung sollte auf den messbaren behavioralen und interaktionellen Faktoren beruhen, zumindest so lange die zugrundeliegenden pathophysiologischen Prozesse für Entwicklung, Aufrechterhaltung der Erkrankungen, oder für die Auslösung von Episoden nicht besser verstanden sind. Es ist naheliegend, dass behaviorale und interaktionelle Faktoren Folge der zugrundeliegenden pathophysiologischen Prozesse sind. Somit können diese Faktoren möglicherweise auch dazu beitragen, in einem „Top-Down“ Forschungsansatz zugrundeliegende Prozesse besser zu verstehen.
2) Bei der Stratifizierung sollten die unterschiedlichen biologischen und psychosozialen Faktoren mit einbezogen werden, die mutmaßlich zur Entstehung und Aufrechterhaltung neurobehavioraler Erkrankungen beitragen.
3) Aufgrund der dynamischen Veränderungen während der Behandlung sind Variablen, die z.B. zu Beginn einer Behandlung einmalig erfasst wurden wahrscheinlich keine ausreichenden Prädiktoren für den Verlauf der Behandlung. Im zeitlichen Verlauf könnten zusätzliche Informationen liegen, z.B. wie Patient:innen auf bestimmte Stimuli während der Behandlung reagieren. Deshalb sollten bei der Stratifizierung dynamische Muster mit einbezogen werden. Zusätzlich ist in vielen Fällen mutmaßlich eine wiederholte Stratifizierung im Verlauf der Behandlung notwendig, da sich die individuellen biologischen und psychosozialen Bedingungen und damit die Behandlungsanforderungen verändern können.

Da viele der oben genannten Faktoren nicht direkt messbar sind, sind indirekte Messungen notwendig, durch die man auf die Zielkonstrukte schließen kann. Im Falle neurobehavioraler Erkrankungen handelt es sich häufig um „latente“ (=nicht direkt beobachtbare) Konstrukte, wie zum Beispiel Depressivität, körperliche Funktionsfähigkeit, soziale Teilhabe oder die Auswirkung von Schmerzen. Für die Messung dieser latenten Konstrukte hat sich der Einsatz patientenzentrierter Methoden bewährt [13]. Zur Bezeichnung patient:innen-berichteter Konstrukte hat sich der Begriff „Patient-Reported Outcome“ (PRO) etabliert, während die Instrumente, die zur Erfassung verwendet werden, als „Patient-Reported Outcome Measure“ (PROM) bezeichnet werden. Als PROMs werden eine große Bandbreite an herkömmlichen und modernen Fragebögen bezeichnet, die den Patient:innen in unterschiedlicher Weise präsentiert werden können, beispielsweise als Papierfragebogen oder als Computer-adaptiver Test (CAT).

1.3 Probleme herkömmlicher Erfassungsmethoden für den Einsatz in der Präzisionsmedizin

Variablen, die mit herkömmlichen PROMs erfasst wurden, weisen jedoch einige Schwächen auf, die diese in der Regel für den Einsatz als Prädiktor oder Outcome in der Präzisionsmedizin ungeeignet machen. Mit herkömmlichen PROMs sind dabei Instrumente gemeint, die überwiegend auf der Basis klassischer Testtheorie (KTT) [14] entwickelt wurden und die zu einem bestimmten Zeitpunkt einmalig erhoben werden. In der KTT wird der gemessene Testwert als eine fehlerbehaftete Messung des unbeobachteten „wahren“ Wertes operationalisiert. Als Folge ist der Testwert direkt an das jeweilige Instrument gebunden, wodurch Nachteile bei der Messung und Vergleichbarkeit von Ergebnissen entstehen.

1.3.1 Genauigkeit individueller Messungen

Für die Bestimmung individueller Unterschiede wie sie zur Gruppenzuteilung in der Präzisionsmedizin notwendig ist, ist ein besonders hohes Maß an Messpräzision („Precision“) bzw. Reliabilität notwendig [15, 16]. Messpräzision und Reliabilität liefern einander ergänzende Informationen und sind direkt ineinander umrechenbar. Die Messpräzision beschreibt die Unsicherheit bei der Messung (Schätzung) der Veränderung einer Person, während die Reliabilität auch die Heterogenität des Merkmals in der Population widerspiegelt [16]. Die folgenden Abschnitte beziehen sich primär auf die Reliabilität, da dieser Begriff weiterverbreitet ist. Formal lässt sich der Reliabilitätskoeffizient R als Verhältnis der wahren und beobachteten Unterschiede in den Testwerten beschreiben:

$$R = \frac{Variance_{true}}{Variance_{observed}}$$

Ein hohes Maß an Reliabilität bedeutet, dass ein großer Anteil der beobachteten Unterschiede auf echte Unterschiede zurückzuführen ist, während eine niedrige Reliabilität bedeutet, dass die beobachteten Unterschiede erhebliche Messfehler aufweisen [17]. Für Gruppenstudien gilt häufig eine Reliabilität von ≥0,7 als ausreichend, ≥0,8 als moderat und ≥0,9 als hoch [16]. Auf der individuellen Ebene sind jedoch Reliabilitäten von mindestens 0,9 vonnöten, um mit ausreichender Sicherheit davon ausgehen zu können, dass der gemessene Wert auch dem wahren Wert entspricht [16]. Der Unterschied zwischen dem wahren Wert und dem beobachteten Wert lässt sich anhand des Konfidenzintervalls einer Messung quantifizieren. Das Konfidenzintervall beschreibt den Bereich, in dem der wahre Wert eines Individuums mit einer bestimmten Wahrscheinlichkeit liegt, wenn der beobachtete Wert (Y), die Reliabilität (R) und die Standardabweichung (SD) des Tests bekannt sind. Das 95%-Konfidenzintervall (CI) einer Messung kann dann anhand der folgenden Formel geschätzt werden:

$$CI = Y \pm 1{,}96 \times SD \times \sqrt{1 - R}$$

Wenn beispielsweise der gemessene Wert der Depressivität 60 wäre bei einer SD = 10, dann läge der wahre Wert im Falle einer Reliabilität von 0,9 mit einer Wahrscheinlichkeit von 95% zwischen 54 und 66, falls die Reliabilität aber nur 0,7 betragen würde, läge der wahre Wert mit einer Wahrscheinlichkeit von 95% in einem viel größeren Bereich, nämlich zwischen 49 und 71.

Anders als einige Biomarker wie z.B. Laborparameter ermöglichen viele herkömmliche PROMs keine hoch-präzise Erfassung auf der Ebene der einzelnen Person [18]. In RCTs werden meist nur Vergleiche auf Gruppenebene evaluiert. Dazu werden häufig Outcome-Instrumente eingesetzt, für die akzeptable Reliabilitäten vorbeschrieben sind. Das hat zur Folge, dass die Aussagen, die in diesen klinischen Studien getroffen werden, z.B. ob eine neue Therapie wirksam ist oder nicht, nicht auf die Individualebene übertragbar sind [5]. Hohe Reliabilitäten sind nicht nur für Outcome-Messungen relevant, sondern ebenso für die Messung von Prädiktoren oder von Mediatoren, die für die Bestimmung von kausalen Zusammenhängen auf der Individualebene unabdingbar sind [15]. Eine Strategie, die individuelle Messpräzision zu steigern, ist, die Anzahl der Fragen („Items“) pro Outcome zu erhöhen, wodurch allerdings auch die Belastung des/der Patient:in steigt und die Compliance abnimmt [16]. Bei der Beantwortung hoher Itemzahlen, wie dies dann in Studien mit multiplen Outcomes erforderlich wäre, kommt es mit höherer Wahrscheinlichkeit zu unbedachter Beantwortung von Fragen („Careless Responding“) und zu höheren zusätzlichen Kosten, weshalb nicht mehr alle Fragen korrekt beantwortet werden [19]. Eine weitere Möglichkeit, um die Reliabilität zu erhöhen, ist die Formulierung von besseren Items [15].

Das Problem der niedrigen Messpräzision auf individueller Ebene und das Problem des hohen Itemburden lässt sich durch den Einsatz von Item Response Theorie (IRT) bei der Konstruktion neuer Instrumente adressieren [20]. IRT, auch „probabilistische Testtheorie“ unterscheidet sich von der KTT, insbesondere darin, dass latente, nicht direkt beobachtbare Konstrukte (z.b. Depressivität oder körperliche Funktionsfähigkeit) mit unterschiedlichen Items gemessen werden können [20]. Dabei wird durch mathematische Modelle der Zusammenhang zwischen einzelnen Itemantworten und dem latenten Konstrukt dargestellt. Dies geschieht in IRT-Modellen, die im PROM-Bereich üblicherweise verwendet werden, anhand von zwei Parametern. Während der Diskriminationsparameter („Slope“) beschreibt, wie gut ein Item in einem bestimmten Bereich des latenten Konstruktes misst, wird durch einen oder mehrere Thresholds die Lage der Itemantworten auf dem Konstrukt festgelegt. Abhängig von der Ausprägung der latenten Variable beim getesteten Individuum und abhängig von den spezifischen Itemeigenschaften (Slope und Thresholds) wird die Wahrscheinlichkeit für die Wahl einer Antwortkategorie durch das Modell vorgegeben [20]. Mit sukzessiver Beantwortung mehrerer Items wird der Bereich um den wahren Wert auf dem latenten Konstrukt beim Individuum immer weiter eingegrenzt. Ein sehr häufiges 2-Parameter IRT Modell im PROM-Bereich ist das Graded-Response-Modell [21]. Im Gegensatz zur Entwicklung von Instrumenten auf Basis der KTT müssen für die Schätzung eines IRT-Modells die Items bestimmte Voraussetzungen erfüllen. Dazu gehört, dass das Zielkonstrukt (hinreichend) unidimensional sein muss und dass die Items lokal unabhängig sein müssen. Lokale Unabhängigkeit bedeutet, dass es nur eine geringe gemeinsame Varianz zwischen Items gibt, die nicht durch die latente Variable erklärt wird, oder anders gesagt, die nicht durch das latente Konstrukt erklärte Fehlervarianz einzelner Items sollte nicht hoch mit der Fehlervarianz anderer Items korrelieren. Als Ergebnis der Entwicklung eines Instrumentes auf Basis der IRT entsteht ein Pool an Items („Itembank“), die unterschiedlich „schwierig“ sind, d.h. die in unterschiedlichen Bereichen des latenten Konstruktes am besten messen [22]. Während beispielsweise das weniger schwierige Item „Waren Sie in den letzten 7 Tagen traurig?“ in einem niedrigeren Bereich des latenten Konstruktes Depressivität die besten Messeigenschaften aufweist, misst das schwierigere Item „Hatten Sie in den letzten 7 Tagen lebensmüde Gedanken?“ in einem höheren Bereich des Konstruktes besser.

Aus einer Itembank können je nach Zweck der Erhebung Subsets von Items ausgewählt werden. Fixierte Kurzformen mit z.B. 4-8 Items können an bestimmte klinische Populationen oder Studienpopulationen angepasst werden, damit die Items den zu erwartenden Messbereich möglichst gut abdecken. Außerdem kann die Erhebung als Computer-adaptiver Test (CAT) an individuelle Personen angepasst werden, indem nach einem vorgegebenen Algorithmus Items auf der Basis der vorhergehenden Antworten ausgewählt werden. Durch CATs ist eine noch effizientere Messung mit

weniger Items bei gleicher Messpräzision möglich, da nur die für ein bestimmtes Individuum in der bestimmten Situation sinnvollen Items angeboten werden [22]. Am Beispiel der Depressivität konnte gezeigt werden, dass ein CAT mit 2-4 Items eine ähnliche diagnostische Treffsicherheit erreichen kann wie der weit verbreitete „Gesundheitsfragebogen für Patienten“ (PHQ-9) mit 9 Items [23]. In den letzten Jahren wurden CATs für viele Gesundheitsdomänen entwickelt, z.B. zur Messung von Depressivität, Angst oder Stress [24-26].

In den letzten 20 Jahren haben die National Institutes of Health in den USA über 180 Millionen Dollar investiert um es der Patient-Reported Outcomes Measurement Information System (PROMIS) Initiative zu ermöglichen, auf der Basis von IRT PROMs zu entwickeln, die eine höhere Messpräzision bzw. einen niedrigeren Messfehler als herkömmliche Instrumente aufweisen [27]. Die beteiligten Wissenschaftler:innen an der PROMIS Initiative haben dazu einen standardisierten Leitfaden entwickelt, die einen strukturierten Prozess zur Entwicklung von Itembanken vorgibt [18]. Der sorgfältige Entwicklungsprozess beginnt mit der Definition des unidimensionalen Zielkonstruktes und beinhaltet mehrere repetitive Stufen, um die Items möglichst ideal zu formulieren und die Items mit den besten Eigenschaften auszuwählen [28]. Die Domänen, die aktuell gemessen werden können, decken viele Aspekte der körperlichen, psychischen und sozialen Gesundheit ab. Bisher wurden über 70 Itembanken zur Erfassung von PROMIS Gesundheitsdomänen bei Erwachsenen entwickelt, die Itembanken zu den Haupt-Domänen wurden bereits in viele andere Sprachen übersetzt und dort validiert (www.healthmeasures.net). In Deutsch sind aktuell 16 Itembanken verfügbar und Kurzformen für viele weitere Domänen (promis-germany.de).

1.3.2 Genauigkeit von Veränderungsmessungen

Eine weitere Herausforderung für die Präzisionsmedizin ist die Erfassung individueller Veränderungen, z.B. im Verlauf einer Behandlung [15, 29]. In RCTs wird die Veränderung üblicherweise auf Gruppenebene als signifikanter Mittelwertsunterschied erfasst. Um Responder auf der Gruppenebene zu identifizieren kann der „Minimally Important Difference“ (MID) eingesetzt werden [30]. Anhand eines „Ankers“ wird die minimal klinisch bedeutsame Veränderung eingeschätzt [31]. Wenn allerdings Verbesserung auf der Gruppenebene dazu eingesetzt werden würde, um Responder zu identifizieren, würde das zu einer falschen Klassifizierung von Patient:innen als Responder führen, die sich tatsächlich gar nicht verändert haben. Auf der individuellen Ebene ist eine sehr viel größere Veränderung eines individuellen Testwerten notwendig, um statistisch signifikant zu sein, als auf der Gruppenebene, da der Messfehler auf der individuellen Ebene deutlich größer ist [32]. Daher müssen Responder auf der Grundlage individuell signifikanter Unterschiede identifiziert werden. Eine mögliche Methode bietet

zum Beispiel die Berechnung des „Reliable Change Index“ (RCI) [33]. Der RCI entspricht der minimalen Differenz von Testwerten, die notwendig ist, um eine Veränderung mit einer Sicherheit von 95% anzuzeigen:

$$RCI = 1{,}96 \times SD \times \sqrt{2} \times \sqrt{(1 - R)}$$

Anhand dieser Formel wird wieder die hohe Bedeutung der Reliabilität deutlich. Für das bereits erwähnte Beispiel zur Depressivität von 60 (SD = 10) würde eine verlässliche Veränderung bei einer Reliabilität von 0,9 eine minimale Veränderung von 9 Punkten erfordern und bei einer Reliabilität von 0,7 eine minimale Veränderung von 15 Punkten, also deutlich mehr als eine Standardabweichung.

Rodebaugh et al. [15] haben exemplarisch gezeigt, welchen Einfluss Testwerte mit niedriger Reliabilität (in ihrem Fall Testwerte zum Konstrukt der Aufmerksamkeitsverzerrung) auf die Beurteilung von Verläufen von Subgruppen haben können. Bei geringer Reliabilität von zwei aufeinanderfolgenden Messungen können sich die Verlaufsmessungen deutlich von den ursprünglichen Messwerten unterscheiden. Hinzu kommt das Phänomen der „Regression zur Mitte“. So verändern sich in dem Beispiel von Rodebaugh et al. die Testwerte zur Aufmerksamkeitsverzerrung im Verlauf deutlich. Dazu wurden aus einer Stichprobe die Fälle mit den besonders hohen und besonders niedrigen Werten herausgenommen – wie es bei der Subgruppenzuteilung auf der Basis dieses Wertes passieren könnte. Beide Mittelwerte der Testwerte von den zwei Gruppen, die bei der ersten Messung besonders hoch oder besonders niedrig waren, veränderten sich bei der zweiten Messung in Richtung des Mittelwertes und darüber hinaus, jedoch ohne dass in der Zwischenzeit eine Intervention erfolgt war. Ein ähnliches Muster in den Verlaufsmessungen von Subgruppen wurde in anderen Studien beschrieben, obwohl dort eine Intervention erfolgte [34]. Wenn also unreliable Testwerte genutzt werden, um Subgruppenverläufe zu beschreiben, kann der Anteil der gemessenen Veränderung, der durch eine wahre Veränderung entstanden ist, nicht von der Veränderung unterschieden werden, die zufällig („Fehlervarianz“) entsteht.

Theoretisch kann die Reliabilität von Veränderungsmessungen nicht nur durch die Erhöhung der Messgenauigkeit der einzelnen Messungen, sondern auch durch häufigere Erhebungen gesteigert werden [35]. Auf Basis von Daten aus der Praxis konnten Stone et al. [36] zeigen, dass die Reliabilität von Schmerzintensität-Messungen zunimmt, wenn die Anzahl der Erhebungen gesteigert wird. Dazu wurden Daten von Tagebuch-Studien ausgewertet, bei denen Patient:innen täglich nach ihrer Schmerzstärke gefragt wurden. Während die Reliabilitäten auf der Basis von zwei Messungen in den meisten Studien unter 0,9 lagen, waren diese bei Messungen, die täglich innerhalb von einer Woche durchgeführt wurden am höchsten und in allen Studien über 0,9 [36]. Erwartungsgemäß war die

Reliabilität bei zwei Messungen umso geringer, je weiter die Messzeitpunkte auseinanderlagen. Nichtsdestotrotz ergab sich in einer anderen Studie ein gegenteiliges Bild. Jensen et al. [37] berichteten Ergebnisse aus einer Studie, bei der zwei Erfassungen der Schmerzstärke, die 13 Tage auseinanderlagen, eine Reliabilität von 0,938 erreichten. Dies entspricht einer sehr hohen Übereinstimmung der beiden Messungen. Eine Erklärung dieses ungewöhnlichen Ergebnisses war jedoch bisher nicht möglich.

Zur Reliabilität von Messungen individueller Veränderungen ist bisher nur sehr wenig publiziert worden. Die Studie von Moinpour et al. [38] verdeutlicht die große Herausforderung, individuelle Veränderungen verlässlich zu messen. Am Beispiel von monatlichen Fatigue-Messungen bei Patient:innen mit Tumor-assoziierter Fatigue evaluierten die Autor:innen die Reliabilität von individuellen Veränderungsmessungen nach 3 und nach 6 Monaten. Es wurden somit bis zu 4 Messungen (3-Monats-Verlauf) oder bis zu 7 Messungen (6-Monats-Verlauf) in die Modellberechnungen mit einbezogen. Obwohl auch PROMIS Instrumente mit hoher Messpräzision verwendet wurden, darunter auch der PROMIS Fatigue CAT, waren die Ergebnisse ernüchternd, da die erreichten Reliabilitäten zwischen 0,342 und 0,533 lagen. Anders als erwartet waren die Reliabilitäten des PROMIS CAT nicht höher, sondern sogar etwas niedriger als bei der 7-Item PROMIS Kurzform. Die Autoren schlugen unterschiedliche Ansätze vor, um die Reliabilität der Verlaufsmessungen zu steigern, z.B. durch häufigere Messungen. Es konnte berechnet werden, dass ca. 15 Erhebungen über 6 Monate notwendig wären, um eine Reliabilität von 0,9 zu erreichen. Idealerweise sollten die Abstände zwischen den Erhebungen zu Beginn und zum Ende hin geringer sein. Die Autoren vermuteten, dass durch eine höhere Item-Anzahl oder durch die Identifikation von besseren Items, die Veränderungen sensitiver erfassen, ebenfalls eine Erhöhung der Reliabilität möglich sein könnte. Zuletzt könnte auch die Nutzung anderer statistischer Modelle zur Berechnung der Testwerte (z.B. Empirische Bayes-Methode) zu besserer Messgenauigkeit beitragen [38].

1.3.3 Erfassung dynamischer Prozesse

Bei vielen neurobehavioralen Erkrankungen besteht eine hohe Dynamik der Symptome und Verhaltensweisen [39, 40]. Das stellt unter anderem eine Herausforderung für die Erfassung individueller „Trends“ dar, also der tatsächlichen Veränderungen über die Zeit ohne das kurzfristige „Rauschen“ in den Symptomen. In Abbildung 1 ist dieses Phänomen am Beispiel von Schmerzratings dargestellt. Hier sieht man, dass der Trend einer Schmerzreduktion über die Zeit entspricht, wenn alle Erfassungen beachtet werden. Würden jedoch zufällig einzelne Erfassungen herausgenommen, könnte es in Extremfällen dazu kommen, dass fälschlicherweise keine Verbesserung oder sogar eine

Verschlechterung über die Zeit angenommen wird. Wenn einzelne Werte aus den komplexen Verläufen über die Zeit herausgenommen werden - wie dies bei herkömmlicher Messung üblich ist - sind diese nicht besonders reliabel und spiegeln daher ggf. auch nicht den tatsächlichen Trend eines Individuums wider [15, 40].

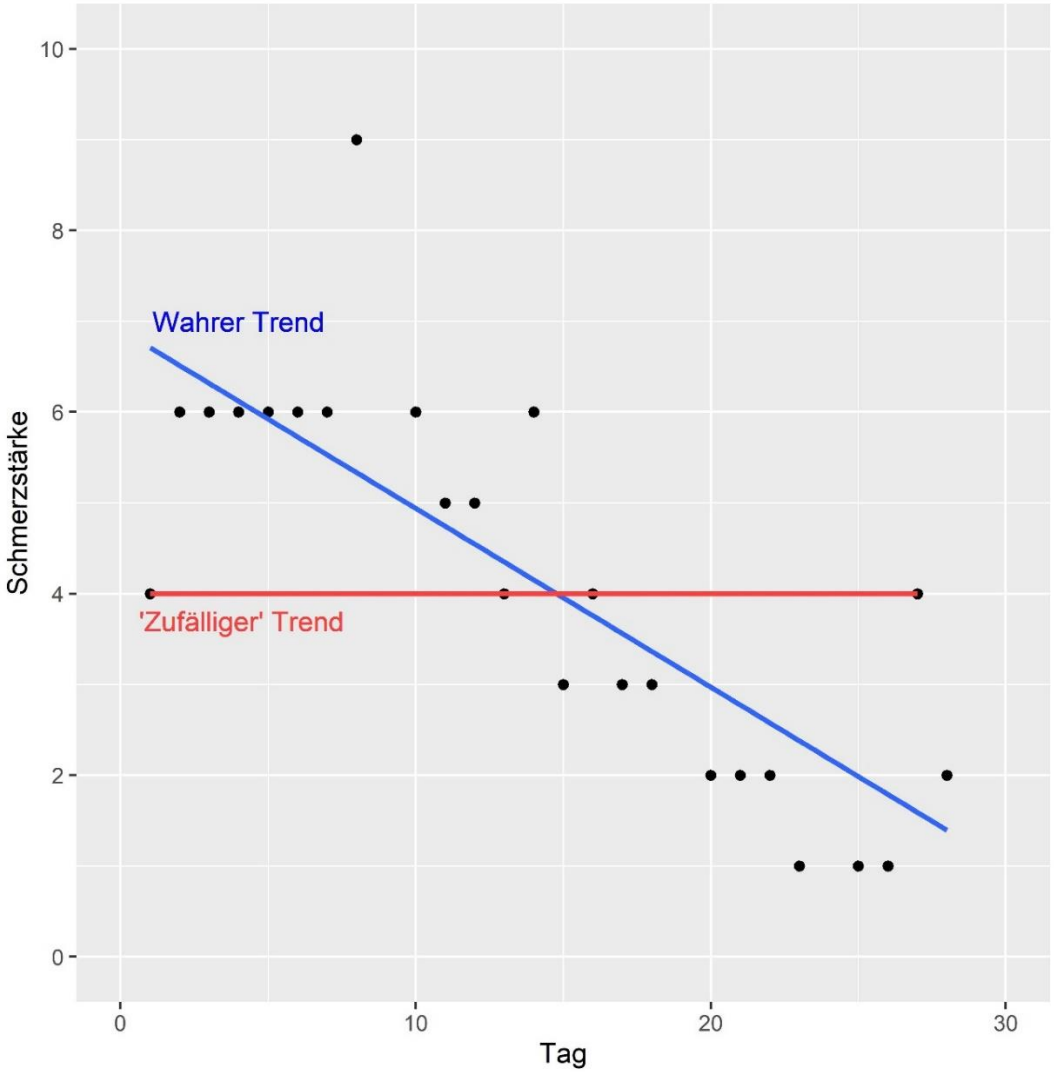

Abbildung 1: Der wahre Trend basierend auf täglichen Erfassungen der Schmerzstärke (blaue Linie) weist auf eine Schmerzreduktion hin. Dahingegen würde ein Trend, der auf einigen zufällig ausgewählten Erfassungen beruht, auf einen Trend hinweisen, bei der sich die Schmerzstärke über die Zeit scheinbar nicht verändert (rote Linie). Die Werte entstammen einer Erhebung bei einem Patienten, der an Tag 8 eine Hernien-OP erhalten hat [41, 42].

Darüber hinaus bestehen bei neurobehavioralen Symptomen wie Schmerz, Fatigue und Depressivität häufig tageszeitliche Schwankungen, die sich z.B. repetitiv wiederholen können und relevant für Diagnostik und Therapie sein können [43]. Neben der Dynamik einzelner Symptome über die Zeit spielen dynamische Wechselbeziehungen mit Umgebungsfaktoren eine bedeutende Rolle, z.B. da die

Symptome abhängig davon sein können, in welcher Situation sich eine Person befindet, was sie/er gerade tut und mit welchen anderen Menschen sie/er gerade interagiert [40]. Herkömmliche Instrumente können kurzfristige Veränderungen und Wechselbeziehungen jedoch nicht erfassen, da sie nach dem Zustand in den letzten Tagen oder sogar Wochen fragen [44]. Hinzu kommt, dass bei der Nutzung herkömmlicher Instrumente kognitive Verzerrungseffekte entstehen, wenn Personen Zustände retrospektiv bewerten (sog. „Recall Bias“). Beispielsweise kommt es zu „Peak-and-End“ Effekten, das heißt besonders bedeutsame Phasen (z.B. mit hohen Schmerzen) oder Phasen die noch nicht so weit zurückliegen fließen stärker in die retrospektive Bewertung mit ein [45]. Bei retrospektiven Erhebungen können beispielsweise Symptome als zu hoch eingeschätzt werden [41]. Aus diesen genannten Gründen haben sich insbesondere für wissenschaftliche Zwecke Methoden etabliert, um Variablen von Personen häufig und im täglichen Leben zu erfassen. Dazu gehören z.B. Tagebuch-Methoden, bei denen Patient:innen täglich zu ihren Beschwerden befragt werden oder auch „Ecological Momentary Assessments“ (EMAs), bei denen die Befragungen mehrmals täglich – meist zu zufälligen Zeitpunkten – erfolgen, um möglichst wenig verzerrte Einschätzungen zu erhalten [40, 45].

Entwicklungen in den letzten Jahren haben das Potenzial zur Nutzung von EMAs für die Präzisionsmedizin deutlich verbessert. So wird gerade durch die Weiterentwicklung und Verbreitung mobiler Technologie die Umsetzung von EMAs in Smartphone-Applikationen einfacher und diese Selbstbefragungs-Daten können mit anderen gemessenen Daten wie Bewegungs-, Standortdaten oder Daten zum Nutzer:innen-Verhalten kombiniert werden. Darüber hinaus machen kürzlich vorgeschlagene statistische Methoden wie dynamische Strukturgleichungsmodelle („Dynamic Structural Equation Models“, DSEM) oder maschinelles Lernen die Modellierung komplexer und zeitlich variabler Zusammenhänge auf individueller Ebene einfacher und generalisierbarer [46, 47].

Intensive longitudinale Erfassungsmethoden wie EMAs, Tagebücher oder Ambulantes Assessment von physiologischen- und Verhaltensdaten bieten auch die Möglichkeit „within-person“ Dynamiken (WPD) zu bestimmen [48, 49]. Diese Indices ermöglichen eine Zusammenfassung komplexer Verläufe einzelner Personen über die Zeit und bieten so die Möglichkeit vereinfachte Werte zur Variabilität von Symptomen zu erhalten. Einer der bekanntesten Indices ist zum Beispiel die intraindividuelle Standardabweichung, also ein Wert, der die Streuung von Werten über die Zeit um den Mittelwert einer Person darstellt [50]. Dieser und andere Indices wurden z.B. im Rahmen von dynamischen Veränderungen von Emotionen, Persönlichkeit oder Schmerz untersucht [50-52]. Diese Indices könnten vielfältig eingesetzt werden. Schmerzindices wurden beispielsweise als vielversprechende Alternativen vorgeschlagen, um Behandlungseffekte auf Gruppenebene zu detektieren [53]. Viele der Indices werden jedoch vor allem als Möglichkeit für eine bessere Charakterisierung von Individuen in der Präzisionsmedizin gesehen. So wird beispielsweise die intraindividuelle Standardabweichung als einer der Marker zur Phänotypisierung bei

Schmerzpatient:innen empfohlen [4, 54]. Im Schmerzbereich wurde die Einführung dieser Indices als Paradigmenwechsel für die Schmerzforschung bezeichnet [55].

1.3.4 Standardisierung von Messungen

Die Marker, auf denen die Stratifizierung und Verlaufsbeurteilung von Patient:innen beruht, sollten möglichst standardisiert sein, also von Klinik zu Klinik und von Land zu Land vergleichbar [56]. Andernfalls wären Vergleiche der identifizierten Subgruppen und deren Behandlungsergebnisse über unterschiedliche Settings hinweg nicht gut möglich. Eine Vergleichbarkeit ist bei vielen biologischen Markern wie z.B. genetischen Profilen, Laborwerten, oder elektrophysiologischen Messungen natürlicherweise gegeben. Dies ist jedoch bei PROs nicht der Fall. Für viele Konstrukte haben sich eine Vielzahl unterschiedlicher Instrumente etabliert, je nach Setting und Anwendungszweck, Sprache, persönlichen Vorlieben etc. [57]. Zur Erfassung von Depressivität konnten wir in einer Studie beispielsweise 29 Instrumente identifizieren, die zur Messung bei Erwachsenen allein im englischen Sprachraum eingesetzt werden [58]. Momentan gibt es eine Vielzahl von Bemühungen PRO Erhebungen zu standardisieren. Die vielleicht prominenteste Initiative ist das bereits oben beschriebene PROMIS System, welches effiziente Instrumente für die Erhebung von Gesundheitsdomänen bereitstellt, die krankheitsunabhängig sind [27]. Dies ermöglicht eine vergleichbare Erfassung von Domänen über unterschiedliche Erkrankungen hinweg. Andere Initiativen wie die Quality of Life Group der European Organisation for Research and Treatment of Cancer (EORTC) entwickeln krankheitsspezifische Instrumente, die Aspekte erfassen, die für Patient:innen mit malignen Erkrankungen relevant sind. Krankheitsspezifische Instrumente sind spezifischer, lassen sich jedoch nur innerhalb einer Erkrankung oder einer Gruppe von Erkrankungen einsetzten und vergleichen [59]. Im Gegensatz dazu bemühen sich andere Initiativen wie beispielsweise das International Consortium for Health Outcomes Measurement (ICHOM) darum, Kombinationen von bereits validierten Instrumenten (sogenannte „Standard Sets“) vorzuschlagen, die krankheitsspezifisch sind [60]. Die Grundidee zur Standardisierung von PROs bei ICHOM beruht auf dem System des „Value-based Healthcare“ (VBHC) [61]. VBHC wurde durch den Gesundheitsökonom Michael Porter vorgeschlagen und soll die Grundlage für eine Restrukturierung von Gesundheitssystemen sein, bei dem nicht die Kosten einer Behandlung im Vordergrund stehen, sondern der Wert („Value“), der sich durch die Formel $Value = \frac{Gesundheits-Outcomes}{Kosten}$ beschreiben lässt [60]. Das Ziel ist, einen größtmöglichen Value durch bessere Behandlungsergebnisse bei möglichst geringen Kosten zu erreichen. Eine Kernvoraussetzung hierfür ist die kontinuierliche Messung von Outcomes, die zum großen Teil durch PROs erfasst werden können. Bis zum ersten Quartal 2023 wurden bereits 40 Standard-Sets publiziert, die viele der Erkrankungen mit einer hohen Krankheitslast abdecken.

Hierunter befinden sich auch Standard-Sets für viele neurobehaviorale Erkrankungen wie Depressionen, Angsterkrankungen, Rückenschmerzen und Abhängigkeitserkrankungen (www.ichom.org).

1.4 Forschungsfragen

Zusammenfassend lässt sich sagen, dass herkömmliche Ansätze zur Messung latenter Konstrukte wie Depressivität oder Schmerz nicht die Voraussetzungen erfüllen, die für die Präzisionsmedizin neurobehavioraler Erkrankungen erforderlich sind. Allerdings gibt es wie bereits beschrieben schon unterschiedliche Ansätze, um einzelne Aspekte – wie beispielsweise die Genauigkeit individueller Messungen – zu adressieren. Die Arbeiten in dieser Habilitationsschrift beschäftigen sich mit unterschiedlichen Forschungsfragen zu diesen Aspekten. In den Arbeiten wurde untersucht,

1) ob auf der Basis herkömmlicher Fragebögen mithilfe von IRT Instrumente entwickelt werden können, die zu einer genaueren und effizienteren Messung von Individuen führen,
2) welche Faktoren die Reliabilität von Veränderungsmessungen beeinflussen können,
3) ob auf der Basis herkömmlicher PROs klinische Phänotypen identifiziert werden können, die prädiktiv für den Verlauf der Behandlung sind,
4) ob hochauflösende EMA-Daten zur Identifizierung von Subgruppen geeignet sind, die prädiktiv für den Verlauf der Behandlung sind,
5) ob und wie die standardisierte Erfassung von PROs bei neurobehavioralen Erkrankungen gelingen kann.

2. Eigene Arbeiten

2.1 Applying Item Response Theory to the OPD Structure Questionnaire: Identification of a Unidimensional Core Construct and Feasibility of Computer Adaptive Testing

Obbarius, A., Ehrenthal, J. C., Fischer, F., Liegl, G., Obbarius, N., Sarrar, L., & Rose, M. (2021). Applying Item Response Theory to the OPD Structure Questionnaire: Identification of a Unidimensional Core Construct and Feasibility of Computer Adaptive Testing. *Journal of Personality Assessment*, *103*(5), 645-658. https://doi.org/10.1080/00223891.2020.1828435

Die Operationalisierte Psychodynamische Diagnostik (OPD) ist ein im deutschsprachigen Raum verbreitetes Klassifikationssystem für psychische Störungen. Zur OPD gehört auch die Strukturachse, welche eine hohe Ähnlichkeit mit dem Funktionsniveau der Persönlichkeit aufweist, das im DSM-5 bzw. ICD-11 eingeführt wurde. Beide beschreiben Fähigkeiten und Einschränkungen in Bezug auf das Selbst (z.B. Selbstwahrnehmung oder Selbstregulierung) und im zwischenmenschlichen Bereich (z.B. Objektwahrnehmung oder Beziehungsregulation). Die Ausprägung des Funktionsniveaus ist ein wichtiger Prädiktor für den Behandlungsverlauf von psychischen Erkrankungen und wird zur Auswahl der Behandlungsform oder des Behandlungsfokus verwendet. Das Ziel dieser Arbeit war die Entwicklung eines effizienten Messinstrumentes auf der Basis des OPD-Strukturfragebogens, einem herkömmlichen 95-Item Instrument zur Erfassung des Funktionsniveaus der Persönlichkeit.

Der methodische Prozess orientierte sich an Empfehlungen der PROMIS Initiative und der Food and Drug Administration. Ca. 1200 Patient:innen füllten den Strukturfragebogen aus. Zunächst wurden die Unidimensionalität und lokale Unabhängigkeit als Voraussetzungen für IRT-Modelle geprüft und es wurden Graded-Response IRT Modelle geschätzt. In einem iterativen Prozess wurden die Items ausgeschlossen, die für eine standardisierte Messung des Zielkonstruktes nicht geeignet waren. Die Messgenauigkeit der resultierenden Itembank wurde evaluiert und mit einem Computer-Adaptiven-Test und einer 12-Item-Kurzform verglichen, die auf der 95-Item-Originalversion beruht.

Von den ursprünglich 95 Items erfüllten 36 Items hinreichende Gütekriterien und wurden in die resultierende Itembank übernommen. Die Messgenauigkeit der Itembank war auf einer großen Breite des latenten Konstruktes sehr gut (Reliabilität > 0,95 zwischen $-1{,}7 < \Theta < 2.4$). Ein CAT, bei dem die Patient:innen im Schnitt 6,8 Items beantworteten (Range 6-12 Items) erreichte die gleiche Messgenauigkeit (Reliabilität > 0,90 zwischen $-1{,}5 < \Theta < 2.0$) wie der herkömmliche 12-Item-Kurzfragebogen.

Zusammenfassend konnte durch die Anwendung des IRT-Modells die Messung des Strukturniveaus der Persönlichkeit nach OPD effizienter gemacht werden. Wir konnten zeigen, dass das Strukturniveau bei individuellen Patient:innen mit hoher Messgenauigkeit bzw. Reliabilität erfasst

wird [62]. Während bei dieser Arbeit die Messgenauigkeit einzelner Messungen untersucht wurde, bezieht sich die folgende Arbeit auf die Untersuchung von Veränderungsmessungen.

2.2 Achieving reliable pain change scores for individuals in the postoperative phase: carefully choose sampling density, test length, and administration mode

Obbarius, A., Schneider, S., Junghaenel, D. U., & Stone, A. A. (2022). Achieving reliable pain change scores for individuals in the postoperative phase: carefully choose sampling density, test length, and administration mode. *Pain*, *163*(1), 170-179. https://doi.org/10.1097/j.pain.0000000000002328

Das Ziel dieser Arbeit war es, herauszufinden, welche Faktoren in welchem Ausmaß Einfluss auf die Reliabilität von Veränderungsmessungen haben. Relevant ist die Reliabilität bei jeder Verlaufsmessung von Symptomen über die Zeit, wie sie beispielsweise im postoperativen Verlauf durchgeführt wird. Abhängig von der Verbesserung der Schmerzen wird die analgetische Therapie angepasst und der Entlassungstermin festgelegt. Anhaltend hohe Schmerzen können beispielsweise ein Hinweis auf Komplikationen wie Entzündungen oder eine Unterdosierung von Analgetika sein. Obwohl durch die PROMIS Initiative bereits große Anstrengungen unternommen wurden, die individuelle Messung effizienter und reliabler zu machen, wurde die Reliabilität von Veränderungsmessungen nur wenig untersucht.

Es erfolgten post-hoc Simulationen auf der Basis von Daten einer Tagebuch-Studie, bei der Patienten nach einer Hernien-Operation zu ihren Schmerzen befragt wurden. Über knapp 3 Wochen erfolgte dazu eine tägliche Befragung vmit PROMIS Items zur Auswirkung von Schmerzen („Pain Interference“) und schmerzbezogenem Verhalten („Pain Behavior“). Für die Post-hoc Simulationen wurden sehr viele unterschiedliche Szenarien verglichen. Neben der Testlänge (Anzahl der Items) und der Häufigkeit der Befragungen wurde auch die Art der Befragung (statische Kurzform versus CAT) untersucht.

Es konnte gezeigt werden, dass alle beschriebenen Faktoren einen relevanten Einfluss auf die Reliabilität von Veränderungsmessungen hatten. Für ein Szenario in welchem mit allen verfügbaren Items (6-7) innerhalb von knapp 3 Wochen 5-6 Erhebungen erfolgen, konnte eine ausreichende Reliabilität für individuelle Veränderungsmessungen (≥0,90) erreicht werden. Ein weiteres Ergebnis war, dass CATs den statischen Kurzformen in Bezug auf die Reliabilität von Veränderungsmessungen überlegen sind.

Die Ergebnisse dieser Studie weisen darauf hin, dass für eine verlässliche Messung von Verläufen eine höherfrequente Erfassung notwendig ist als bisher in klinischen Settings üblich. Weitere Studien sind dringend notwendig, um die Ergebnisse bei anderen Konstrukten zu replizieren und um die Reliabilität von Veränderungsmessungen bei komplexeren Verläufen zu untersuchen [42].

2.3 A Step Towards a Better Understanding of Pain Phenotypes: Latent Class Analysis in Chronic Pain Patients Receiving Multimodal Inpatient Treatment

Obbarius, A., Fischer, F., Liegl, G., Obbarius, N., van Bebber, J., Hofmann, T., & Rose, M. (2020). A Step Towards a Better Understanding of Pain Phenotypes: Latent Class Analysis in Chronic Pain Patients Receiving Multimodal Inpatient Treatment. *Journal of pain research*, *13*, 1023. https://doi.org/10.2147/JPR.S223092

Ziel der Studie war es, Subgruppen von Patient:innen mit chronischen Schmerzen anhand von herkömmlichen PROMs zu identifizieren, Gruppenunterschiede zu beschreiben. Zudem wurde untersucht, ob die Unterschiede der Gruppen prädiktiv für den Verlauf während einer stationären multimodalen Schmerztherapie waren.

Zur Identifikation von Subgruppen führten wir Latente Klassenanalysen auf der Basis von selbstberichteten Daten zu Schmerzaspekten, emotionaler Belastung und körperlicher Funktionsfähigkeit durch. Im Gegensatz zu anderen ähnlichen Studien wurden nur PROMs genutzt, jedoch keine soziodemographischen oder klinischen Daten.

Ein Modell mit vier Klassen erreichte die beste Kombination aus Modell-Fit und klinischer Interpretierbarkeit. Die Klassen umfassten Patient:innen mit hoher (54,7%), extremer (17,0%), moderater (15,6%) und niedriger (12,7%) Schmerzbelastung. Patient:innen in der Klasse mit niedriger Schmerzbelastung zeigten eine hohe emotionale Belastung, während die emotionale Belastung in den anderen Klassen den Schmerzbelastungsniveaus entsprach. Während sich im Verlauf der multimodalen Behandlung in der Gruppe mit den extremen Schmerzen die Schmerzstärke sowie physische und psychische Gesundheit verbesserten, verbesserten sich in den anderen Gruppen nur die Depressivität und Angst.

Diese Ergebnisse weisen darauf hin, dass Subgruppen von Patient:innen mit chronischen Schmerzen mutmaßlich von individualisierten Behandlungen profitieren könnten, die die spezifischen Bedürfnisse berücksichtigen. Beispielsweise könnten Patient:innen in den Klassen mit hoher emotionaler Belastung von zusätzlichen psychologischen Interventionen profitieren, während solche in den Klassen mit hoher Schmerzbelastung möglicherweise mehr körperliche Rehabilitation benötigen. Die Generalisierbarkeit der Ergebnisse ist jedoch durch das retrospektive Studiendesign und das im Verhältnis zu anderen ähnlichen Studien relativ kleine Sample eingeschränkt, weshalb eine Replikation in prospektiven Studien dringend erforderlich scheint [63].

Während in dieser Arbeit einmalige Erhebungen verwendet wurden, um Subgruppen von Schmerzpatient:innen zu identifizieren, wurden in der folgenden Arbeit Messungen dynamischer Veränderung für die Identifikation von Subgruppen verwendet.

2.4 A combination of pain indices based on momentary assessments can predict placebo response in patients with fibromyalgia syndrome

Obbarius, A., Schneider, S., & Stone, A. A. (2021). A combination of pain indices based on momentary assessments can predict placebo response in patients with fibromyalgia syndrome. *Pain*, *162*(2), 543-551. https://doi.org/10.1097/j.pain.0000000000002025

Das Ziel dieser Studie war es, herauszufinden, ob kürzlich für den Schmerzbereich vorgeschlagene WPDs, die auf hochauflösenden EMA Daten basieren, dazu geeignet sind, relevante Subgruppen von Schmerzpatient:innen zu identifizieren. In bisherigen Studien bei Schmerzpatient:innen wurden nur die prädiktiven Eigenschaften einzelner WPDs – vor allem der intraindividuellen Standardabweichung (iSD) – evaluiert. Aufbauend darauf wurden in dieser Arbeit die prädiktiven Eigenschaften einer Kombination von drei WPDs untersucht.

Die WPDs wurden so ausgewählt, dass Diese unterschiedliche Aspekte der Schmerzwahrnehmung reflektieren. Dazu gehörten neben der iSD der individuelle Mittelwert über die Zeit und ein Index zur Konsistenz, der die Abhängigkeit späterer Messungen von früheren Messungen darstellt (Autokorrelationen 1. Ordnung). Als Stichprobe dienten N = 2084 Patient:innen mit Fibromyalgie-Syndrom aus zwei klinischen Studien, bei denen die Wirksamkeit eines Antidepressivums untersucht wurde. Auf der Basis der drei WPDs erfolgte eine Latente Profilanalyse (LPA; =LCA mit kontinuierlichen Variablen) und die Behandlungseffekte der resultierenden Subgruppen wurden anhand einer Varianzanalyse mit Messwiederholungen verglichen.

Die beste Lösung der LPA ergab 3 Subgruppen. Patient:innen in der kleinsten Gruppe (5% der Stichprobe) wiesen die höchste Schmerzintensität, die niedrigste Konsistenz und die geringste Variabilität auf. Diese Gruppe zeigte die größte Reduktion von Schmerzen in Reaktion auf eine Placebo-Behandlung. Überraschenderweise sprachen diese Patient:innen sogar besser auf die Placebo-Behandlung als auf die Behandlung mit dem Verum an.

Zusammenfassend konnte hier erstmalig für den Schmerzbereich gezeigt werden, dass mit Hilfe der WPDs relevante Subgruppen von Patient:innen identifiziert werden können. Ein möglicher Nutzen besteht darin, die Assay Sensitivität für klinische Studien zu erhöhen. D.h., dass die Anzahl der Studienteilnehmer:innen und dadurch die Kosten gesenkt werden, indem die Patient:innen mit der hohen Response auf die Placebo-Behandlung ausgeschlossen werden. Im Sinne der Präzisionsmedizin könnten diese Erkenntnisse jedoch auch genutzt werden, um spezifische Behandlungen für die Subgruppen zu entwickeln [64].

2.5 Standardization of health outcomes assessment for depression and anxiety: recommendations from the ICHOM Depression and Anxiety Working Group

Obbarius, A., van Maasakkers, L., Baer, L., Clark, D. M., Crocker, A. G., de Beurs, E., Emmelkamp, P. M. G., Furukawa, T. A., Hedman-Lagerlof, E., Kangas, M., Langford, L., Lesage, A., Mwesigire, D. M., Nolte, S., Patel, V., Pilkonis, P. A., Pincus, H. A., Reis, R. A., Rojas, G., Sherbourne, C., Smithson, D., Stowell, C., Woolaway-Bickel, K., Rose, M. (2017). Standardization of health outcomes assessment for depression and anxiety: recommendations from the ICHOM Depression and Anxiety Working Group. *Quality of Life Research*. https://doi.org/10.1007/s11136-017-1659-5

Wie bereits beschrieben ist die Standardisierung von PRO Messungen ein Baustein, um Präzisionsmedizin für neurobehaviorale Erkrankungen zu ermöglichen, zu denen auch Depressionen und Angsterkrankungen gehören. Ziel dieser Studie war es deshalb ein Standard Outcome Set für Depressionen und Angsterkrankungen vorzuschlagen, welches global eingesetzt werden kann.

Dazu wurde eine Arbeitsgruppe aus 24 Expert:innen zusammengestellt, zu denen Outcome-Forscher:innen, klinische Expert:innen, Patientenvertreter:innen und ICHOM-Koordinator:innen gehörten. Über 7 Monate erfolgte ein adaptierter Delphi-Prozess nach dem Vorbild ähnlicher Studien. Während eines iterativen Prozesses wurden Literaturrecherchen durchgeführt und die Ergebnisse bei monatlichen Treffen diskutiert. Die Festlegung der Domänen und Instrumente erfolgte dann auf der Basis von Online-Umfragen der Arbeitsgruppenteilnehmer:innen.

Die vorgeschlagenen Outcome-Domänen umfassen die Symptomlast, Funktionsfähigkeit, Krankheitsaktivität, Nachhaltigkeit von Behandlungen, sowie Nebenwirkungen von Behandlungen. Zur Erfassung von Depressivität wurde der PHQ-9 vorgeschlagen, zur Erfassung von Angst der GAD-7. Beide sind Module aus dem Patient Health Questionnaire (PHQ). Zur Erfassung der Funktionsfähigkeit wurde der WHO Disability Assessment Schedule 2.0 (WHODAS 2.0) empfohlen. Alle drei Instrumente sind wissenschaftlich gut untersucht und in vielen Sprachen verfügbar. Zur Erfassung der Symptomlast wurde auch die Möglichkeit beschrieben, eigene etablierte Instrumente zur Erfassung zu benutzen und diese in PHQ-9 und GAD-7 Testwerte umzurechnen. Dies ist möglich, da beide Instrumente und viele andere häufig eingesetzte Instrumente an die PROMIS-Metrik gelinkt wurden und dadurch in standardisierte PROMIS-Testwerte oder ineinander umgerechnet werden können (www.common-metrics.org).

Zusammenfassend wurde ein umfangreiches Outcome Set für die Erfassung von Depression und Angst vorgeschlagen. Die einheitliche Erfassung von Outcomes bei Patient:innen kann dabei helfen, herauszufinden, welche Behandlung für welche Personen am besten geeignet ist und somit

mittelfristig die Gesundheitsversorgung verbessern. Eine große Herausforderung ist die Implementierung der Outcome Sets in den unterschiedlichen Gesundheitssystemen [58].

3. Diskussion

„What Works For Whom?“ ist nicht mehr nur eine hypothetische Frage, die verdeutlichen soll, dass wir zu wenig über die Wirksamkeit von Behandlungen bei Individuen verstehen [65]. Durch digitale und statistische Weiterentwicklungen wird nun die tatsächliche Beantwortung dieser Frage auch für Erkrankungen wie Depressionen, Schmerzerkrankungen oder andere heterogene Krankheitsbilder möglich. In den Studien, die in dieser Habilitationsschrift zusammengefasst sind, wurden einige Ansatzmöglichkeiten untersucht, die einen Fortschritt auf dem Weg hin zu Präzisionsmedizin bei neurobehavioralen Erkrankungen bedeuten können. Obwohl sich die einzelnen Studien nur jeweils mit kleinen Ausschnitten aus den beschriebenen Bereichen der individuellen Messgenauigkeit, Messgenauigkeit von Veränderungsmessungen, Erfassung dynamischer Prozesse und Standardisierung von Messungen beschäftigen, so weisen die Ergebnisse doch auf ein hohes Potenzial für die Präzisionsmedizin hin.

Insbesondere wenn mehrere der beschriebenen Methoden für zukünftige Anwendungen kombiniert werden, könnten leistungsstarke Werkzeuge für die Stratifizierung und Verlaufsbeurteilung entstehen. Naheliegend ist dabei zum Beispiel die Kombination von CAT-Anwendungen mit hochfrequenter Erfassung wie Tagebüchern oder EMAs. Eine der größten Herausforderungen bei der EMA-Erfassung ist, die Belastung für die Befragten in alltäglichen Situationen möglichst gering zu halten, also nur die minimal notwendige Anzahl von Items zu erheben [66]. So sind längere EMA-Befragungen in Beobachtungsstudien mit niedrigerer Compliance und höheren Abbruchraten verbunden [67]. Wie oben beschrieben und in Arbeit 2.1 gezeigt kann der Einsatz von IRT und CATs Befragungen effizienter machen. In Arbeit 2.1 konnte die Anzahl der Items eines Screening-Instrumentes von 12 auf durchschnittlich unter 7 Items reduziert werden - bei gleichbleibender Messpräzision. Ähnliche Verbesserungen der Effizienz konnten bereits für viele Konstrukte im Gesundheits-, Beratungs- und Bildungsbereich gezeigt werden [68-70]. Neben der Minimierung der Itemanzahl bei EMA-Befragungen ist selbstverständlich auch dort eine hohe Messpräzision für die Messungen von Zuständen und Veränderungen vonnöten [71, 72]. Herkömmliche EMA-Items wurden bisher jedoch häufig nicht im Rahmen eines standardisierten Entwicklungsprozesses, inklusive einer psychometrischen Evaluation entwickelt, zu der die Untersuchung der Validität und Reliabilität gehört. Vielmehr wurden häufig ad-hoc Items eingesetzt, ohne zuvor die psychometrischen Eigenschaften zu überprüfen [66]. Dies und die Notwendigkeiten der geringen Itemanzahlen haben EMA-Fragebögen entstehen lassen, die nicht die für die Erfassung bei Individuen erforderliche Reliabilität von ≥ 0,90 erreichen. Beispielsweise erfassen verbreitete EMA-Instrumente zur Messung von Emotionen mit 3-5 Items Fluktuationen innerhalb einer Person nur mit Reliabilitäten von 0,60 bis 0,80 [73, 74]. Der Einsatz

von kurzen CATs mit hoher Messpräzision kann mutmaßlich dabei helfen, die erforderlichen Reliabilitäten zu erreichen, ohne dabei eine zu große Belastung für die Befragten darzustellen.

CATs nutzen Informationen vorangegangener Itemantworten, um die Folgefragen auszuwählen. Dieses Prinzip kann im Rahmen von EMA-Befragungen zusätzlich genutzt werden, um Befragungen noch effizienter zu machen. Bei CATs die nur einmalig oder unregelmäßig bei einzelnen Personen angewendet werden, wird häufig ein bestimmtes Start-Item vorgegeben, welches im mittleren Messbereich der gesamtem Population liegt ($Theta_{Start} = 0$) [75]. Alternativ können jedoch auch Vorinformationen genutzt werden, um das erste Item für den zu erwartenden Messbereich der/des Befragten ideal auszuwählen. Möglich wäre aufgrund der zeitlich nahe zusammenliegenden EMA-Befragung beispielsweise, den Testwert der letzten Befragung als Startwert der Folgebefragung zu verwenden. Dieses Prinzip wurde in Arbeit 2.2 für die CAT-Simulationen genutzt. In dieser Arbeit konnte der Vorteil des Einsatzes von CATs bei der reliablen Messung von Veränderungen gezeigt werden. Denkbar wäre auch die Nutzung von komplexeren Vorinformationen, also z.B. des Durchschnittswerts über mehrere Vorbefragungen oder aber der zu erwartende Wert, der auf einer individuellen zirkadianen Rythmik besteht. So konnten eine erste Studie die zirkadianen Veränderungen von Fatigue-Symptomen erfolgreich nutzen, um Startwerte für CAT-Befragungen effizienter zu schätzen [76]. Eine weitere Herausforderung herkömmlicher EMA-Befragungen ist die wiederholte Präsentation der identischen Items bei jeder Befragung, die wahrscheinlich zu einem Ermüdungseffekt bei den Befragten führt, durch den es zu einem Bias bei den Ergebnissen kommen könnte [72, 77]. Die CAT-Regeln könnten für EMA-Anwendungen so vorgegeben werden, dass sich die Items nicht bei jeder Befragung wiederholen und dadurch auch kein Ermüdungseffekt entsteht [76]. Über die beschriebenen Möglichkeiten hinaus wäre auch eine multivariate Erweiterung der CATs in Kombination mit EMAs denkbar, um die Effizienz noch weiter zu erhöhen. Multivariate CATs nutzen Informationen von zusammenhängenden Konstrukten. Die Items werden durch den multivariaten CAT-Algorithmus so vorgegeben, dass möglichst schnell die Information für alle Dimensionen gleichzeitig maximiert wird [75]. Aufgrund der herausfordernden Entwicklung multivariater CATs gibt es im medizinischen Bereich bisher wenige Studien dazu. In einer Studie wurde beispielsweise der Einsatz unidimensionaler CATs zur Erfassung von Angst, Depressivität und Ärger mit einem multivariaten CAT hinsichtlich der Effizienz verglichen [78]. Die Autoren fanden heraus, dass mithilfe eines multivariaten CATs zur Messung von Angst, Depressivität und Ärger die Testlänge um 42% (von 12 Items auf 7 Items) im Vergleich zu unidimensionalen CATs reduziert werden kann [78]. Im Vergleich zu konventionellen Fragebögen ist durch beide CAT-Anwendungen eine enorme Verbesserung der Effizienz möglich. Während der Einsatz klassischer, statischer Fragebögen einen zeitlichen Aufwand von über 10 Minuten für die drei PROs bedeutet, kann diese Zeit auf ca. 2 Minuten bei der Anwendung

des unidimensionalen CATs bzw. auf etwas über eine Minute im Falle des multivariaten CATs reduziert werden. In Tabelle 1 sind einige Vorteile aktuell im medizinischen Bereich üblicherweise eingesetzten CATs sowie mögliche Weiterentwicklungen gezeigt, die für hochfrequente Erhebungen im Rahmen der Präzisionsmedizin vielversprechend sind.

Tabelle 1: Vorteile aktueller Computer-adaptiver Tests und mögliche Weiterentwicklungen

Aktuelle CATs	**Mögliche Weiterentwicklungen**
Hohe Messgenauigkeit	
Items passend für den/die Patient:in	Items und Dimensionen passend für den/die Patient:in und passend zur Situation
Konstanter Itempool (Patient:innen könnten bei jeder Befragung gleiche Items erhalten)	Wechselnder Itempool (Patient:innen erhalten wechselnde Items, geringerer Ermüdungseffekt)
Start-Item: mittlere Symptomausprägung	Start-Item: basierend auf Vorinformationen, z.B. letzte Erhebung(en), zirkadiane Rhythmik, ähnliche Zeit/Situation
Univariat (1 Dimension)	Multivariat (gleichzeitige Erfassung multipler Dimensionen)
Nutzung von PRO Daten	Nutzung von PRO Daten + anderen patientenbezogenen Daten
Hohe Effizienz im Vergleich zu herkömmlichen Instrumenten (zeitsparend)	Noch höhere Effizienz im Vergleich zu aktuellen CATs

Wichtige Fragen für die Stratifizierung von Subgruppen sind die Art und der Umfang der Indikatoren, die dafür eingesetzt werden müssen. Natürlich wäre ein kleines Set von zu erfassenden Konstrukten vor dem Beginn der Behandlung einfacher in der Handhabung und Implementierung. Bisherige Empfehlungen und Studien haben jedoch häufig eine große Bandbreite unterschiedlicher Faktoren empfohlen oder verwendet, um Subgruppen zu identifizieren. So schlägt beispielsweise die Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials (IMMPACT) ein größeres Set mit über 10 Instrumenten zur Phänotypisierung von Schmerzpatient:innen vor. Dazu gehört neben PROMs auch ein psychophysisches Testverfahren, die Quantitative Sensorische Testung [54]. Auch in den bisher durchgeführten Studien zur Identifikation von Phänotypen bei Schmerzerkrankungen wurden eine große Anzahl an Indikatoren verwendet [79, 80]. Die Ergebnisse der Arbeiten 2.3 und 2.4 weisen darauf hin, dass die ausschließliche Verwendung von Selbstberichten prädiktiv relevante Phänotypen identifizieren kann und dass in manchen Fällen möglicherweise auch eine sehr kleine Anzahl an

Indikatoren ausreicht, um klinisch bedeutsame Phänotypen zu identifizieren. Möglicherweise können dabei WPDs eine zentrale Rolle spielen. Es wurde bereits eine große Anzahl an Indizes vorgeschlagen, von denen bisher die wenigsten im Kontext der Präzisionsmedizin weiter untersucht wurden [55, 64]. Besonders vielversprechend könnten Indices sein, die auf „Regime-Switching Models“ beruhen [81]. Regime-Switching Models beschreiben die Intensität, die Dauer, sowie den Wechsel zwischen unterschiedlichen Zuständen. Diese Modelle werden seit Jahrzehnten in den Wirtschaftswissenschaften angewendet. Erst kürzlich wurden diese Modelle auch für die Beschreibung von Krankheitszuständen z.B. bei Patient:innen mit depressiven und bipolaren Erkrankungen eingesetzt [82, 83]. Auch die Nutzung der bisher eingesetzten WPDs ist jedoch nicht ausgeschöpft. So beruhen die in Arbeit 2.4 beschriebenen WPDs alle auf dem gleichen Konstrukt, nämlich auf der momentanen Erfassung der Schmerzstärke. Bei vielen neurobehavioralen Erkrankungen spielen jedoch viele unterschiedliche Konstrukte eine Rolle, für die ebenfalls eine hohe Variabilität beschrieben wurde. Dazu gehören beispielsweise emotionale Belastung, Coping-Versuche oder Katastrophisierung [84-86]. Zusätzlich zur Untersuchung anderer Arten von WPDs scheint es daher sinnvoll, die Forschung auf weitere relevante Konstrukte auszuweiten.

Viele der Methoden, die in den vorgestellten Arbeiten beschriebenen wurden, könnten dabei helfen, sogenannte „Precision Clinical Trials“ (PCTs) zu entwickeln und durchzuführen [5]. Wie zuvor ausführlich beschrieben können herkömmliche RCTs die Effektivität von Interventionen für einzelne Personen nicht gut detektieren. Die Effektivität von Interventionen für Subgruppen kann in RCTs nur unter der Voraussetzung geprüft werden, dass die Studie für diesen Zweck eine ausreichende Power aufweist. Deshalb wurde kürzlich ein PCT-Framework vorgeschlagen, welches den Anforderungen der Präzisionsmedizin besser gerecht wird. Darin werden drei Kernpunkte für das Design von PCTs empfohlen: 1) Verbesserung der Vorauswahl von Patient:innen: Dazu wird zunächst das Ansprechen der Patient:innen auf eine kurze Probeintervention gemessen. Danach werden Patient:innen für (unterschiedliche) vollständige Behandlungen randomisiert. Das Ansprechen auf die Probeintervention kann zur Vorhersage des Outcomes nach der Randomisierung genutzt werden. 2) Anpassung der Behandlungsparameter (z.B. Dosis, Zeitpunkt, etc.) während der Behandlung, um diese individuell zu optimieren. Dies geschieht auf der Basis von regelmäßigen Outcome-Messungen. 3) Präzise und reliable Messung der Prädiktor- und Outcomevariablen mit Techniken wie EMA [5]. Insbesondere für die Punkte 2) und 3) sind die methodischen Anpassungen relevant, die in den Arbeiten 2.1, 2.2 und 2.4 beschrieben wurden. Kürzlich Untersuchungen weisen auch darauf hin, dass WPDs nützlich bei der Evaluation von Behandlungseffekten sein können. Die Indices „Anteil an maximalem Schmerz“ und iSD konnten in einer meta-analytischen Studie zusätzliche Informationen im Vergleich zu dem üblicherweise verwendeten Mittelwert liefern [53]. Viele WPDs wurden jedoch noch

nicht in Bezug auf ihre Eigenschaft untersucht, Behandlungseffekte detektieren zu können, weshalb hier weitere Studien sinnvoll erscheinen.

PCTs könnten dafür genutzt werden, pharmakologische oder psychotherapeutische Behandlungsoptionen zu evaluieren, die je nach Ergebnis der Verlaufsmessungen während der Behandlung angepasst werden. Eine andere Art der Behandlung, die von sich aus schon dazu konzipiert ist, um die Behandlung zu individualisieren, ist die Just-In-Time-Adaptive Intervention (JITAI). JITAIs haben das Ziel, Patient:innen die richtige Art und Dosis einer Behandlung zum richtigen Zeitpunkt zukommen zu lassen und dabei sich ändernde innere und äußere Bedingungen mit einzubeziehen [87]. Hierbei werden unterschiedliche Echtzeitdaten wie Bewegungs-, Nutzungs- oder EMA-Daten genutzt, um den richtigen Zeitpunkt für eine „Mikrointervention“ zu bestimmen. Viele wiederholte Mikrointerventionen können dann zu dem erwünschten langfristigen Outcome führen (Abbildung 2).

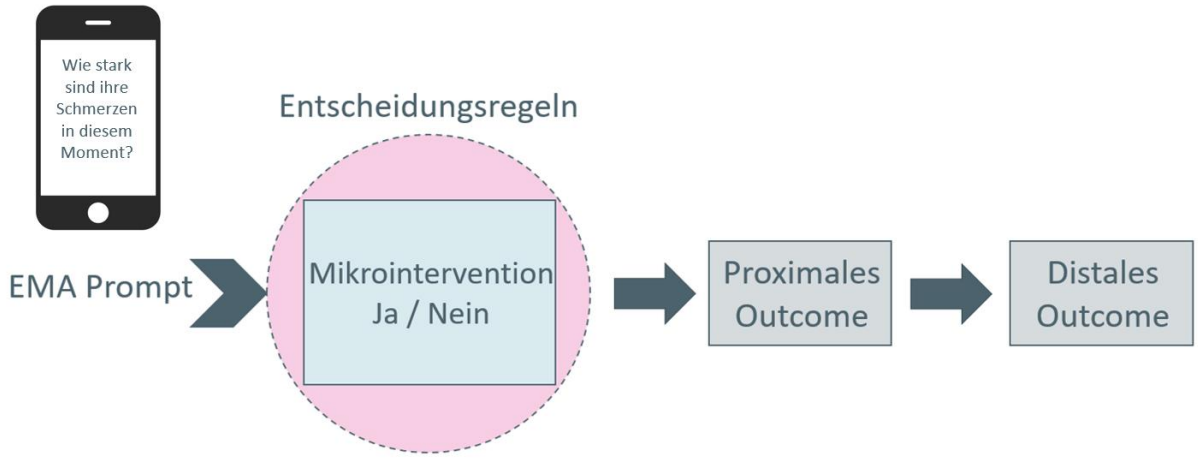

Abbildung 2: Komponenten einer Just-In-Time-Adaptiven Intervention, welche auf Ecological Momentary Assessments (EMA) Erhebungen beruht.

JITAIs wurden schon für einige neurobehaviorale Erkrankungen entwickelt, wie z.B. für Abhängigkeitserkrankungen, Adipositas, Angsterkrankungen, Depressionen und Bipolare Störungen [88]. Dabei erreichen diese Interventionen mittlere bis große Effekte. Trotz des großen Potenzials von JITAIs befindet sich die Entwicklung und Umsetzung dieser Interventionen noch im Anfangsstadium und erfordert multidisziplinäre Anstrengungen zur Weiterentwicklung der Theorien zu dynamischen Veränderungen [89], des Studiendesigns [90], der Interventionstechniken [87], und der Messmethoden [91]. Bei der Weiterentwicklung der Messmethoden könnten die in den Arbeiten 2.1 und 2.2 beschriebenen methodischen Prinzipien von hoher Bedeutung sein. In einer kürzlich publizierten Arbeit wurden Empfehlungen zur Messung von Zuständen in JITAIs beschrieben, welche auch IRT-Methoden und CAT-Anwendungen mit einbeziehen [76]. Hier konnten die Autoren die

Überlegenheit eines Instrumentes mit CAT-Komponente bei der Klassifizierung von veränderten Fatigue-Zuständen zeigen.

3.1 Zukünftige Forschung und Ausblick

Obwohl es einige vielversprechende Ansätze für die Erfassungsmethoden in der Präzisionsmedizin bei neurobehavioralen Erkrankungen gibt, so bedarf es noch einer besseren Evidenzlage in vielen Bereichen, die nur durch multidisziplinäre Zusammenarbeit zwischen Kliniker:innen und Wissenschaftler:innen aus der digitalen Medizin, Outcome-Forschung und Medizin-Informatik erreichbar scheint. Einige Möglichkeiten zur Kombination unterschiedlicher methodischer Ansätze wie EMA- und CAT-Anwendungen wurden bereits skizziert. Diese könnten für die reliable Messung von Markern bei der Stratifizierung und zur Verlaufsmessung von hoher Bedeutung sein. Hochauflösende Erfassungsmethoden wie EMA könnten auch bei der Identifizierung von Markern zur Stratifizierung von großer Relevanz sein. Aktuell kommt der Erforschung direkt messbarer Marker zur Stratifizierung, die auf Neuroimaging (z.B. funktionelle Magnetresonanztomographie), elektrophysiologischen Verfahren (z.B. Elektroenzephalografie) oder genetischen Profilen beruhen eine hohe Aufmerksamkeit zu [4, 6]. Viele dieser Ansätze sind jedoch noch in einer frühen Phase der Forschung und weit von der klinischen Anwendung entfernt. Möglicherweise ist es deshalb aktuell naheliegender, einen Forschungsschwerpunkt auf die einfacher messbaren, selbstberichteten Faktoren und deren Dynamik zu legen, die mutmaßlich Ausdruck der zugrundeliegenden pathophysiologischen Prozesse sind. Die Messung dieser (selbstberichteten) Symptome wird durch die zunehmende Verbreitung von Smartphones und Smartwatches immer besser und einfacher auch im Alltag möglich. Eine sinnvolle Forschungsrichtung scheint daher die Identifikation von phänomenologischen Markern, die auf diesen einfach zugänglichen Informationen beruht.

Um zukünftig die Zustände von Individuen im Verlauf von Behandlungen noch effizienter zu erfassen könnten z.B. Methoden des Maschinellen Lernens oder andere Bayes-basierte Modelle eingesetzt werden, die kontinuierlich Daten während der Behandlung verarbeiten. Beim Maschinellen Lernen werden üblicherweise hunderte von Eigenschaften oder Variablen in ein Modell mit aufgenommen, um präzise Vorhersagen zu einem Outcome zu machen [92]. Maschinelles Lernen wurde bereits erfolgreich für Vorhersagen im klinischen Bereich eingesetzt, wie für die Bildverarbeitung in der Radiologie und Pathologie oder zur Vorhersage von klinischen Outcomes (z.B. Früherkennung von Sepsis oder stationärer Wiederaufnahme nach Entlassung). In einem Scoping Review konnten über 300 Artikel identifiziert werden, die maschinelles Lernen zur Charakterisierung von psychischen Erkrankungen wie Depressionen oder Schizophrenie eingesetzt haben [93]. In den hier beschriebenen

Analysen beruhen Anwendungen des Maschinellen Lernens weitestgehend auf Sekundärdaten von klinischen Studien, bei denen auch strukturiert Outcomes erfasst wurden. Dadurch sind robuste Vorhersagen möglich. Dies wäre in klinischen Routine-Settings nicht der Fall. Zudem ist es sehr wahrscheinlich, dass sich die besten Prädiktoren von Person zu Person unterscheiden. Deshalb scheint die Weiterentwicklungen der Methoden des Maschinellen Lernens für dynamische Settings, bei denen richtige Vorhersagen für Individuen getroffen werden müssen, ein vielversprechendes Forschungsziel [46].

4. Zusammenfassung

Neurobehaviorale Erkrankungen wie Depressionen oder chronische Schmerzen sind verantwortlich für einen beachtlichen Anteil der globalen Krankheitslast. Eine große Herausforderung bei der Behandlung dieser Erkrankungen ist die hohe Variabilität der Behandlungsergebnisse. Während einige Patient:innen gut auf bestimmte Behandlungen ansprechen, ist dieselbe Behandlung bei anderen Patient:innen mit ähnlichen Beschwerden wenig oder gar nicht wirksam. Deshalb wird zunehmend davon ausgegangen, dass sich hinter der jeweiligen Diagnose eine Vielzahl von Subgruppen verbirgt, die mutmaßlich von unterschiedlichen Behandlungsansätzen profitieren würden. Die Identifikation dieser Subgruppen würde eine bessere Stratifizierung erlauben. Dieses Konzept der Präzisionsmedizin hat sich bei einigen kardialen oder onkologischen Erkrankungen bereits etabliert. Bei den neurobehavioralen Erkrankungen bestehen jedoch zusätzliche Hürden. Unter anderem sind viele der Konstrukte, die zur Stratifizierung eingesetzt werden könnten, nicht direkt messbar.

Die Arbeiten in dieser Habilitationsschrift beschäftigen sich deshalb mit ausgewählten Aspekten der Messung von Konstrukten, welche die Stratifizierung neurobehavioraler Erkrankungen erleichtern könnten. Während sich eine Arbeit mit der Verbesserung der Genauigkeit von Messungen individueller Zustände am Beispiel der Persönlichkeitsstruktur beschäftigt, werden in einer weiteren Arbeit Faktoren untersucht, welche die Genauigkeit von Veränderungsmessungen beeinflussen können. Zwei weitere Arbeiten beschäftigen sich damit, wie auf der Basis von Patient-Reported Outcomes (PROs) und Ecological Momentary Assessments (EMAs) Subgruppen identifiziert werden können, die relevant für den Verlauf der Behandlungen sind. Die letzte Arbeit beschäftigt sich mit der internationalen Standardisierung von PROs, die eine weitere Voraussetzung für die Implementierung von Präzisionsmedizin bei neurobehavioralen Erkrankungen darstellt.

Die Ergebnisse der Arbeiten bieten Ansatzpunkte für die Verbesserung der Messungen als Voraussetzung für Präzisionsmedizin bei neurobehavioralen Erkrankungen. Insbesondere die Kombination unterschiedlicher Ansätze, wie dynamische Messungen von Symptomen und Computer-adaptives Testen könnten zukünftig zur besseren Stratifizierung, Erfassung der Behandlungsverläufe und Vorhersage der Outcomes beitragen.

5. Literaturangaben

1. Engel, G. L. (1980). The clinical application of the biopsychosocial model. *American Journal of Psychiatry*, 137(5), 535-544.https://doi.org/10.1176/ajp.137.5.535
2. Vos, T., Abajobir, A. A., Abate, K. H., Abbafati, C., Abbas, K. M., Abd-Allah, F., Abdulkader, R. S., Abdulle, A. M., Abebo, T. A., & Abera, S. F. (2017). Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet*, 390(10100), 1211-1259.https://doi.org/10.1016/S0140-6736(17)32154-2
3. Fries, J. F. (1980). Aging, Natural Death, and the Compression of Morbidity. *New England Journal of Medicine*, 303(3), 130-135.https://doi.org/10.1056/nejm198007173030304
4. Edwards, R. R., Schreiber, K. L., Dworkin, R. H., Turk, D. C., Baron, R., Freeman, R., Jensen, T. S., Latremoliere, A., Markman, J. D., Rice, A. S. C., Rowbotham, M., Staud, R., Tate, S., Woolf, C. J., Andrews, N. A., Carr, D. B., Colloca, L., Cosma-Roman, D., Cowan, P., . . . Wesselmann, U. (2023). Optimizing and Accelerating the Development of Precision Pain Treatments for Chronic Pain: IMMPACT Review and Recommendations. *The Journal of Pain*, 24(2), 204-225.https://doi.org/10.1016/j.jpain.2022.08.010
5. Lenze, E. J., Nicol, G. E., Barbour, D. L., Kannampallil, T., Wong, A. W. K., Piccirillo, J., Drysdale, A. T., Sylvester, C. M., Haddad, R., Miller, J. P., Low, C. A., Lenze, S. N., Freedland, K. E., & Rodebaugh, T. L. (2021). Precision clinical trials: A framework for getting to precision medicine for neurobehavioural disorders. *Journal of Psychiatry & Neuroscience*, 46, E97-E110.https://doi.org/10.1503/jpn.200042
6. Schumann, G., Binder, E. B., Holte, A., de Kloet, E. R., Oedegaard, K. J., Robbins, T. W., Walker-Tilley, T. R., Bitter, I., Brown, V. J., Buitelaar, J., Ciccocioppo, R., Cools, R., Escera, C., Fleischhacker, W., Flor, H., Frith, C. D., Heinz, A., Johnsen, E., Kirschbaum, C., . . . Wittchen, H. U. (2014). Stratified medicine for mental disorders. *European Neuropsychopharmacology*, 24(1), 5-50.https://doi.org/https://doi.org/10.1016/j.euroneuro.2013.09.010
7. Waks, A. G., & Winer, E. P. (2019). Breast Cancer Treatment: A Review. *Journal of the American Medical Association*, 321(3), 288-300.https://doi.org/10.1001/jama.2018.19323
8. Armitage, J. O., Gascoyne, R. D., Lunning, M. A., & Cavalli, F. (2017). Non-Hodgkin lymphoma. *The Lancet*, 390(10091), 298-310.https://doi.org/https://doi.org/10.1016/S0140-6736(16)32407-2
9. Khan, A., Mar, K. F., & Brown, W. A. (2018). The conundrum of depression clinical trials: one size does not fit all. *International Clinical Psychopharmacology*, 33(5), 239-248.https://doi.org/10.1097/yic.0000000000000229
10. Bokma, W. A., Wetzer, G. A. A. M., Gehrels, J. B., Penninx, B. W. J. H., Batelaan, N. M., & van Balkom, A. L. J. M. (2019). Aligning the many definitions of treatment resistance in anxiety disorders: A systematic review. *Depression and Anxiety*, 36(9), 801-812.https://doi.org/https://doi.org/10.1002/da.22895
11. Dickenson, A. H., & Patel, R. (2020). Translational issues in precision medicine in neuropathic pain. *Canadian Journal of Pain*, 4(1), 30-38.https://doi.org/10.1080/24740527.2020.1720502
12. Collins, F. S., & Varmus, H. (2015). A New Initiative on Precision Medicine. *New England Journal of Medicine*, 372(9), 793-795.https://doi.org/10.1056/NEJMp1500523
13. Obbarius, A., Fischer, K. I., Fischer, F., Liegl, G., Obbarius, N., Nolte, S., & Rose, M. (2018). Empirische Erfassung subjektiver Gesundheitsmerkmale am Beispiel der gesundheitsbezogenen Lebensqualität. [Empirical Assessment of Patient-Reported Outcomes and Exemplary Introduction to Health-Related Quality of Life] *Psychother Psychosom Med Psychol*, 68(12), 534-547.https://doi.org/10.1055/a-0764-4691
14. Bühner, M. (2011). *Einführung in die Test-und Fragebogenkonstruktion*: Pearson Studium
15. Rodebaugh, T. L., Scullin, R. B., Langer, J. K., Dixon, D. J., Huppert, J. D., Bernstein, A., Zvielli, A., & Lenze, E. J. (2016). Unreliability as a threat to understanding psychopathology: The

cautionary tale of attentional bias. *Journal of Abnormal Psychology*, 125, 840-851.https://doi.org/10.1037/abn0000184

16. Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill
17. Danner, D. (2015). *Reliabilität – die Genauigkeit einer Messung*. Mannheim: GESIS – Leibniz Institut für Sozialwissenschaften (GESIS Survey Guidelines).
18. Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., Thissen, D., Revicki, D. A., Weiss, D. J., Hambleton, R. K., Liu, H., Gershon, R., Reise, S. P., Lai, J. S., Cella, D., & Promis Cooperative Group. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care*, 45(5 Suppl 1), S22-31.https://doi.org/10.1097/01.mlr.0000250483.85507.04
19. Ward, M. K., & Meade, A. W. (2023). Dealing with Careless Responding in Survey Data: Prevention, Identification, and Recommended Best Practices. *Annual Review of Psychology*, 74(1), 577-596.https://doi.org/10.1146/annurev-psych-040422-045007
20. Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Mahway, NJ: Lawrence Erlbaum Associates
21. Samejima, F. (1997). Graded Response Model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 85-100). New York, NY: Springer New York.https://doi.org/10.1007/978-1-4757-2691-6_5
22. Bjorner, J. B., Chang, C.-H., Thissen, D., & Reeve, B. B. (2007). Developing tailored instruments: item banking and computerized adaptive assessment. *Quality of Life Research*, 16(1), 95-108.https://doi.org/10.1007/s11136-007-9168-6
23. Fischer, H. F., Klug, C., Roeper, K., Blozik, E., Edelmann, F., Eisele, M., Störk, S., Wachter, R., Scherer, M., Rose, M., & Herrmann-Lingen, C. (2014). Screening for mental disorders in heart failure patients using computer-adaptive tests. *Quality of Life Research*, 23(5), 1609-1618.https://doi.org/10.1007/s11136-013-0599-y
24. Kocalevent, R. D., Rose, M., Becker, J., Walter, O. B., Fliege, H., Bjorner, J. B., Kleiber, D., & Klapp, B. F. (2009). An evaluation of patient-reported outcomes found computerized adaptive testing was efficient in assessing stress perception. *Journal of Clinical Epidemiology*, 62(3), 278-287, 287 e271-273.https://doi.org/10.1016/j.jclinepi.2008.03.003
25. Fliege, H., Becker, J., Walter, O. B., Bjorner, J. B., Klapp, B. F., & Rose, M. (2005). Development of a Computer-adaptive Test for Depression (D-CAT). *Quality of Life Research*, 14(10), 2277-2291.https://doi.org/10.1007/s11136-005-6651-9
26. Walter, O. B., Becker, J., Bjorner, J. B., Fliege, H., Klapp, B. F., & Rose, M. (2007). Development and evaluation of a computer adaptive test for 'Anxiety' (Anxiety-CAT). *Quality of Life Research*, 16(1), 143-155.https://doi.org/10.1007/s11136-007-9191-7
27. Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., Amtmann, D., Bode, R., Buysse, D., Choi, S., Cook, K., Devellis, R., DeWalt, D., Fries, J. F., Gershon, R., Hahn, E. A., Lai, J. S., Pilkonis, P., Revicki, D., . . . Promis Cooperative Group. (2010). The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. *Journal of Clinical Epidemiology*, 63(11), 1179-1194.https://doi.org/10.1016/j.jclinepi.2010.04.011
28. Patient Reported Outcome Information System (PROMIS®) (2013). *Instrument Development and Validation, Scientific Standards Version 2.0 (revised May 2013)*. Abgerufen 08.09.2023, https://www.healthmeasures.net/images/PROMIS/PROMISStandards_Vers2.0_Final.pdf
29. Rogosa, D. (1988). Myths about longitudinal research. In *Methodological issues in aging research.* (pp. 171-209). New York, NY, US: Springer Publishing Company
30. Coon, C. D., & Cook, K. F. (2018). Moving from significance to real-world meaning: methods for interpreting change in clinical outcome assessment scores. *Quality of Life Research*, 27(1), 33-40.https://doi.org/10.1007/s11136-017-1616-3
31. McLeod, L. D., Coon, C. D., Martin, S. A., Fehnel, S. E., & Hays, R. D. (2011). Interpreting patient-reported outcome results: US FDA guidance and emerging methods. *Expert Review of Pharmacoeconomics & Outcomes Research*, 11(2), 163-169.https://doi.org/10.1586/erp.11.12

32. Hays, R. D., Brodsky, M., Johnston, M. F., Spritzer, K. L., & Hui, K.-K. (2005). Evaluating the Statistical Significance of Health-Related Quality-Of-Life Change in Individual Patients. *Evaluation & the Health Professions*, 28(2), 160-171.https://doi.org/10.1177/0163278705275339
33. Jacobson, N. S., & Truax, P. (1991). Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, 59(1), 12-19.https://doi.org/10.1037//0022-006x.59.1.12
34. Calamaras, M. R., Tone, E. B., & Anderson, P. L. (2012). A Pilot Study of Attention Bias Subtypes: Examining Their Relation to Cognitive Bias and Their Change following Cognitive Behavioral Therapy. *Journal of Clinical Psychology*, 68(7), 745-754.https://doi.org/https://doi.org/10.1002/jclp.21875
35. Rogosa, D. R., & Willett, J. B. (1983). Demonstrating the Reliability of the Difference Score in the Measurement of Change. *Journal of Educational Measurement*, 20(4), 335-343
36. Stone, A. A., Schneider, S., Broderick, J. E., & Schwartz, J. E. (2014). Single-day Pain Assessments as Clinical Outcomes: Not So Fast. *The Clinical Journal of Pain*, 30(9), 739-743.https://doi.org/10.1097/ajp.0000000000000030
37. Jensen, M. P., Hu, X., Potts, S. L., & Gould, E. M. (2013). Single vs composite measures of pain intensity: Relative sensitivity for detecting treatment effects. *Pain*, 154(4), 534-538.https://doi.org/https://doi.org/10.1016/j.pain.2012.12.017
38. Moinpour, C. M., Donaldson, G. W., Davis, K. M., Potosky, A. L., Jensen, R. E., Gralow, J. R., Back, A. L., Hwang, J. J., Yoon, J., Bernard, D. L., Loeffler, D. R., Rothrock, N. E., Hays, R. D., Reeve, B. B., Smith, A. W., Hahn, E. A., & Cella, D. (2017). The challenge of measuring intra-individual change in fatigue during cancer treatment. *Quality of Life Research*, 26(2), 259-271.https://doi.org/10.1007/s11136-016-1372-9
39. Wichers, M. (2014). The dynamic nature of depression: a new micro-level perspective of mental disorder that meets current challenges. *Psychological Medicine*, 44(7), 1349-1360.https://doi.org/10.1017/S0033291713001979
40. Stone, A. A., Obbarius, A., Junghaenel, D. U., Wen, C. K. F., & Schneider, S. (2021). High-resolution, field approaches for assessing pain: Ecological Momentary Assessment. *Pain*, 162(1), 4-9.https://doi.org/10.1097/j.pain.0000000000002049
41. Stone, A. A., Broderick, J. E., Junghaenel, D. U., Schneider, S., & Schwartz, J. E. (2016). PROMIS fatigue, pain intensity, pain interference, pain behavior, physical function, depression, anxiety, and anger scales demonstrate ecological validity. *Journal of Clinical Epidemiology*, 74, 194-206.https://doi.org/10.1016/j.jclinepi.2015.08.029
42. Obbarius, A., Schneider, S., Junghaenel, D. U., & Stone, A. A. (2022). Achieving reliable pain change scores for individuals in the postoperative phase: carefully choose sampling density, test length, and administration mode. *Pain*, 163(1), 170-179.https://doi.org/10.1097/j.pain.0000000000002328
43. Schneider, S., Junghaenel, D. U., Keefe, F. J., Schwartz, J. E., Stone, A. A., & Broderick, J. E. (2012). Individual differences in the day-to-day variability of pain, fatigue, and well-being in patients with rheumatic disease: Associations with psychological variables. *Pain*, 153(4), 813-822.https://doi.org/10.1016/j.pain.2012.01.001
44. Myin-Germeys, I., Kasanova, Z., Vaessen, T., Vachon, H., Kirtley, O., Viechtbauer, W., & Reininghaus, U. (2018). Experience sampling methodology in mental health research: new insights and technical developments. *World Psychiatry*, 17(2), 123-132.https://doi.org/https://doi.org/10.1002/wps.20513
45. Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology*, 4, 1-32.https://doi.org/10.1146/annurev.clinpsy.3.022806.091415
46. Bzdok, D., & Meyer-Lindenberg, A. (2018). Machine Learning for Precision Psychiatry: Opportunities and Challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 3(3), 223-230.https://doi.org/https://doi.org/10.1016/j.bpsc.2017.11.007

47. Asparouhov, T., Hamaker, E. L., & Muthén, B. (2018). Dynamic Structural Equation Models. *Structural Equation Modeling: A Multidisciplinary Journal*, 25(3), 359-388.https://doi.org/10.1080/10705511.2017.1406803
48. Hamaker, E. L., & Wichers, M. (2017). No Time Like the Present:Discovering the Hidden Dynamics in Intensive Longitudinal Data. *Current Directions in Psychological Science*, 26(1), 10-15.https://doi.org/10.1177/0963721416666518
49. Molenaar, P. C. M., & Campbell, C. G. (2009). The New Person-Specific Paradigm in Psychology. *Current Directions in Psychological Science*, 18(2), 112-117.https://doi.org/10.1111/j.1467-8721.2009.01619.x
50. Mun, C. J., Suk, H. W., Davis, M. C., Karoly, P., Finan, P., Tennen, H., & Jensen, M. P. (2019). Investigating intraindividual pain variability: methods, applications, issues, and directions. *Pain*, 160(11), 2415-2429.https://doi.org/10.1097/j.pain.0000000000001626
51. Schneider, S., & Stone, A. A. (2015). Mixed emotions across the adult life span in the United States. *Psychology and Aging*, 30, 369-382.https://doi.org/10.1037/pag0000018
52. Kuppens, P., Allen, N. B., & Sheeber, L. B. (2010). Emotional Inertia and Psychological Maladjustment. *Psychological Science*, 21(7), 984-991.https://doi.org/10.1177/0956797610372634
53. Schneider, S., Junghaenel, D. U., Ono, M., Broderick, J. E., & Stone, A. A. (2021). III. Detecting Treatment Effects in Clinical Trials With Different Indices of Pain Intensity Derived From Ecological Momentary Assessment. *The Journal of Pain*, 22(4), 386-399.https://doi.org/https://doi.org/10.1016/j.jpain.2020.10.003
54. Edwards, R. R., Dworkin, R. H., Turk, D. C., Angst, M. S., Dionne, R., Freeman, R., Hansson, P., Haroutounian, S., Arendt-Nielsen, L., Attal, N., Baron, R., Brell, J., Bujanover, S., Burke, L. B., Carr, D., Chappell, A. S., Cowan, P., Etropolski, M., Fillingim, R. B., . . . Yarnitsky, D. (2016). Patient phenotyping in clinical trials of chronic pain treatments: IMMPACT recommendations. *Pain*, 157(9), 1851-1871.https://doi.org/10.1097/j.pain.0000000000000602
55. Winger, J. G., Plumb Vilardaga, J. C., & Keefe, F. J. (2019). Indices of pain variability: a paradigm shift. *Pain*, 160(11), 2411-2412.https://doi.org/10.1097/j.pain.0000000000001627
56. Huckvale, K., Venkatesh, S., & Christensen, H. (2019). Toward clinical digital phenotyping: a timely opportunity to consider purpose, quality, and safety. *npj Digital Medicine*, 2(1), 88.https://doi.org/10.1038/s41746-019-0166-1
57. Döhmen, A., Kock, M., Fischer, F., Rose, M., Obbarius, A., & Klapproth, C. P. (2022). Are OMERACT recommendations followed in clinical trials on fibromyalgia? A systematic review of patient-reported outcomes and their measures. *Quality of Life Research*.https://doi.org/10.1007/s11136-022-03261-5
58. Obbarius, A., van Maasakkers, L., Baer, L., Clark, D. M., Crocker, A. G., de Beurs, E., Emmelkamp, P. M. G., Furukawa, T. A., Hedman-Lagerlof, E., Kangas, M., Langford, L., Lesage, A., Mwesigire, D. M., Nolte, S., Patel, V., Pilkonis, P. A., Pincus, H. A., Reis, R. A., Rojas, G., . . . Rose, M. (2017). Standardization of health outcomes assessment for depression and anxiety: recommendations from the ICHOM Depression and Anxiety Working Group. *Quality of Life Research*, 26(12), 3211-3225.https://doi.org/10.1007/s11136-017-1659-5
59. Fayers, P., & Bottomley, A. (2002). Quality of life research within the EORTC—the EORTC QLQ-C30. *European Journal of Cancer*, 38, 125-133.https://doi.org/https://doi.org/10.1016/S0959-8049(01)00448-8
60. Porter, M. E., Larsson, S., & Lee, T. H. (2016). Standardizing Patient Outcomes Measurement. *New England Journal of Medicine*, 374(6), 504-506.https://doi.org/10.1056/NEJMp1511701
61. Hartmann, C., Obbarius, A., Haneke, H., Klapproth, C. P., & Rose, M. (2021). Value-based Healthcare – Gesundheitsversorgung neu denken. *OP-Management up2date*, 01(01), 87-105.https://doi.org/10.1055/a-1325-7371
62. Obbarius, A., Ehrenthal, J. C., Fischer, F., Liegl, G., Obbarius, N., Sarrar, L., & Rose, M. (2021). Applying Item Response Theory to the OPD Structure Questionnaire: Identification of a Unidimensional Core Construct and Feasibility of Computer Adaptive Testing. *Journal of Personality Assessment*, 103(5), 645-658.https://doi.org/10.1080/00223891.2020.1828435

63. Obbarius, A., Fischer, F., Liegl, G., Obbarius, N., van Bebber, J., Hofmann, T., & Rose, M. (2020). A Step Towards a Better Understanding of Pain Phenotypes: Latent Class Analysis in Chronic Pain Patients Receiving Multimodal Inpatient Treatment. *Journal of Pain Research*, 13, 1023.https://doi.org/https://doi.org/10.2147/JPR.S223092
64. Obbarius, A., Schneider, S., & Stone, A. A. (2021). A combination of pain indices based on momentary assessments can predict placebo response in patients with fibromyalgia syndrome. *Pain*, 162(2), 543-551.https://doi.org/10.1097/j.pain.0000000000002025
65. Fonagy, P. (2010). Psychotherapy research: do we know what works for whom? *British Journal of Psychiatry*, 197(2), 83-85.https://doi.org/10.1192/bjp.bp.110.079657
66. Stone, A. A., Schneider, S., & Smyth, J. M. (2023). Evaluation of Pressing Issues in Ecological Momentary Assessment. *Annual Review of Clinical Psychology*, 19(1), null.https://doi.org/10.1146/annurev-clinpsy-080921-083128
67. Eisele, G., Vachon, H., Lafit, G., Kuppens, P., Houben, M., Myin-Germeys, I., & Viechtbauer, W. (2022). The Effects of Sampling Frequency and Questionnaire Length on Perceived Burden, Compliance, and Careless Responding in Experience Sampling Data in a Student Population. *Assessment*, 29(2), 136-151.https://doi.org/10.1177/1073191120957102
68. Weiss, D. J. (2004). Computerized Adaptive Testing for Effective and Efficient Measurement in Counseling and Education. *Measurement and Evaluation in Counseling and Development*, 37(2), 70-84.https://doi.org/10.1080/07481756.2004.11909751
69. Choi, S. W., Reise, S. P., Pilkonis, P. A., Hays, R. D., & Cella, D. (2010). Efficiency of static and computer adaptive short forms compared to full-length measures of depressive symptoms. *Quality of Life Research*, 19(1), 125-136.https://doi.org/10.1007/s11136-009-9560-5
70. Segawa, E., Schalet, B., & Cella, D. (2020). A comparison of computer adaptive tests (CATs) and short forms in terms of accuracy and number of items administrated using PROMIS profile. *Quality of Life Research*, 29(1), 213-221.https://doi.org/10.1007/s11136-019-02312-8
71. Gibbons, C. J. (2017). Turning the Page on Pen-and-Paper Questionnaires: Combining Ecological Momentary Assessment and Computer Adaptive Testing to Transform Psychological Assessment in the 21st Century. *Frontiers in Psychology*, 7.https://doi.org/10.3389/fpsyg.2016.01933
72. Rose, M., Bjorner, J. B., Fischer, F., Anatchkova, M., Gandek, B., Klapp, B. F., & Ware, J. E. (2012). Computerized adaptive testing--ready for ambulatory monitoring? *Psychosomatic Medicine*, 74(4), 338-348.https://doi.org/10.1097/PSY.0b013e3182547392
73. Cranford, J. A., Shrout, P. E., Iida, M., Rafaeli, E., Yip, T., & Bolger, N. (2006). A Procedure for Evaluating Sensitivity to Within-Person Change: Can Mood Measures in Diary Studies Detect Change Reliably? *Personality and Social Psychology Bulletin*, 32(7), 917-929.https://doi.org/10.1177/0146167206287721
74. Scott, S. B., Sliwinski, M. J., Zawadzki, M., Stawski, R. S., Kim, J., Marcusson-Clavertz, D., Lanza, S. T., Conroy, D. E., Buxton, O., Almeida, D. M., & Smyth, J. M. (2020). A Coordinated Analysis of Variance in Affect in Daily Life. *Assessment*, 27(8), 1683-1698.https://doi.org/10.1177/1073191118799460
75. Chalmers, R. P. (2016). Generating Adaptive and Non-Adaptive Test Interfaces for Multidimensional Item Response Theory Applications. *Journal of Statistical Software*, 71(5), 38.https://doi.org/10.18637/jss.v071.i05
76. Schneider, S., Junghaenel, D. U., Smyth, J. M., Fred Wen, C. K., & Stone, A. A. (2023). Just-in-time adaptive ecological momentary assessment (JITA-EMA). *Behavior Research Methods*.https://doi.org/10.3758/s13428-023-02083-8
77. Silvia, P. J., Kwapil, T. R., Walsh, M. A., & Myin-Germeys, I. (2014). Planned missing-data designs in experience-sampling research: Monte Carlo simulations of efficient designs for assessing within-person constructs. *Behavior Research Methods*, 46, 41-54.https://doi.org/10.3758/s13428-013-0353-y
78. Morris, S., Bass, M., Lee, M., & Neapolitan, R. E. (2017). Advancing the efficiency and efficacy of patient reported outcomes with multivariate computer adaptive testing. *Journal of the*

American Medical Informatics Association, 24(5), 897-902.https://doi.org/10.1093/jamia/ocx003

79. Hassan, S., Nesovic, K., Babineau, J., Furlan, A. D., Kumbhare, D., & Carlesso, L. C. (9900). Identifying chronic low back pain phenotypic domains and characteristics accounting for individual variation: a systematic review. *Pain*, 10.1097/j.pain.0000000000002911.https://doi.org/10.1097/j.pain.0000000000002911
80. Diatchenko, L., Fillingim, R. B., Smith, S. B., & Maixner, W. (2013). The phenotypic and genetic signatures of common musculoskeletal pain conditions. *Nature Reviews Rheumatology*, 9(6), 340-350.https://doi.org/10.1038/nrrheum.2013.43
81. Schneider, S., Junghaenel, D. U., Ono, M., & Stone, A. A. (2018). Temporal dynamics of pain: an application of regime-switching models to ecological momentary assessments in patients with rheumatic diseases. *Pain*, 159(7), 1346-1358.https://doi.org/10.1097/j.pain.0000000000001215
82. Hamaker, E. L., Grasman, R. P. P. P., & Kamphuis, J. H. (2016). Modeling BAS Dysregulation in Bipolar Disorder:Illustrating the Potential of Time Series Analysis. *Assessment*, 23(4), 436-446.https://doi.org/10.1177/1073191116632339
83. Hamaker, E. L., Grasman, R. P. P. P., & Kamphuis, J. H. (2010). Regime-switching models to study psychological processes.
84. Chiros, C., & O'Brien, W. H. (2011). Acceptance, appraisals, and coping in relation to migraine headache: an evaluation of interrelationships using daily diary methods. *Journal of Behavioral Medicine*, 34(4), 307-320.https://doi.org/10.1007/s10865-011-9313-0
85. Holtzman, S., & DeLongis, A. (2007). One day at a time: The impact of daily satisfaction with spouse responses on pain, negative affect and catastrophizing among individuals with rheumatoid arthritis. *Pain*, 131(1), 202-213.https://doi.org/10.1016/j.pain.2007.04.005
86. Turner, J. A., Mancl, L., & Aaron, L. A. (2004). Pain-related catastrophizing: a daily process study. *Pain*, 110(1), 103-111.https://doi.org/10.1016/j.pain.2004.03.014
87. Nahum-Shani, I., Smith, S. N., Spring, B. J., Collins, L. M., Witkiewitz, K., Tewari, A., & Murphy, S. A. (2017). Just-in-Time Adaptive Interventions (JITAIs) in Mobile Health: Key Components and Design Principles for Ongoing Health Behavior Support. *Annals of Behavioral Medicine*, 52(6), 446-462.https://doi.org/10.1007/s12160-016-9830-8
88. Wang, L., & Miller, L. C. (2020). Just-in-the-Moment Adaptive Interventions (JITAI): A Meta-Analytical Review. *Health Communication*, 35(12), 1531-1544.https://doi.org/10.1080/10410236.2019.1652388
89. Spruijt-Metz, D., & Nilsen, W. (2014). Dynamic Models of Behavior for Just-in-Time Adaptive Interventions. *IEEE Pervasive Computing*, 13(3), 13-17.https://doi.org/10.1109/MPRV.2014.46
90. Klasnja, P., Hekler, E. B., Shiffman, S., Boruvka, A., Almirall, D., Tewari, A., & Murphy, S. A. (2015). Microrandomized trials: An experimental design for developing just-in-time adaptive interventions. *Health Psychology*, 34(Suppl), 1220-1228.https://doi.org/10.1037/hea0000305
91. Collins, L. M., Murphy, S. A., & Bierman, K. L. (2004). A Conceptual Framework for Adaptive Preventive Interventions. *Prevention Science*, 5(3), 185-196.https://doi.org/10.1023/B:PREV.0000037641.26017.00
92. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine Learning in Medicine. *New England Journal of Medicine*, 380(14), 1347-1358.https://doi.org/10.1056/NEJMra1814259
93. Shatte, A. B. R., Hutchinson, D. M., & Teague, S. J. (2019). Machine learning in mental health: a scoping review of methods and applications. *Psychological Medicine*, 49(9), 1426-1448.https://doi.org/10.1017/S0033291719000151

Danksagung

Zunächst einmal möchte ich mich bei meinen wissenschaftlichen Mentor:innen bedanken, die mir in den letzten Jahren bei meiner akademischen Arbeit und Forschung zur Seite standen. Vor allem gilt mein tiefer Dank Prof. Matthias Rose für seine fortdauernde Unterstützung und Wertschätzung in den letzten Jahren. Seine Motivation, Inspiration und Ermutigung haben mir immer wieder geholfen meinen eigenen Weg zu finden und die jeweils nächsten Schritte zu gehen. Dankbar bin ich auch Prof. Arthur Stone, der mich während und nach meinem Postdoc-Aufenthalt in Los Angeles sehr bei meinen Forschungsideen unterstützt und dazu motiviert, noch einen Schritt weiter zu denken. Außerdem danke ich Prof. Ralph Grabhorn für seine Betreuung und Unterstützung bei meiner Promotion, durch die mein Interesse an psychometrischer Forschung geweckt wurde.

Neben meinen akademischen Lehrern haben mich in den letzten Jahren viele Kolleg:innen auf dem Weg zur Habilitation unterstützt, denen ich sehr dankbar bin. Dazu gehören Tobias Hofmann, Eva Winter, Barbara Voigt, Sandra Nolte, Felix Fischer, Gregor Liegl, Paul Klapproth, Kathrin Fischer, Claudia Hartmann, Annett Mierke, Lea Sarrar, Andreas Stengel, Stefan Schneider, Doerte Junghaenel und viele mehr.

Mein größter Dank gilt aber Jonas, Noah, Lia und Nina. Sie waren meine Stütze und mein Anker, wenn ich Zweifel hatte oder mich überwältigt fühlte. Nina hat mir unglaublich oft den Rücken freigehalten, mich durch die schwierigen Zeiten getragen und mich dazu inspiriert, weiterzumachen.

Erklärung

§ 4 Abs. 3 (k) der HabOMed der Charité

Hiermit erkläre ich, dass

- weder früher noch gleichzeitig ein Habilitationsverfahren durchgeführt oder angemeldet wurde,
- die vorgelegte Habilitationsschrift ohne fremde Hilfe verfasst, die beschriebenen Ergebnisse selbst gewonnen sowie die verwendeten Hilfsmittel, die Zusammenarbeit mit anderen Wissenschaftlern/Wissenschaftlerinnen und mit technischen Hilfskräften sowie die verwendete Literatur vollständig in der Habilitationsschrift angegeben wurden,
- mir die geltende Habilitationsordnung bekannt ist.

Ich erkläre ferner, dass mir die Satzung der Charité – Universitätsmedizin Berlin zur Sicherung Guter Wissenschaftlicher Praxis bekannt ist und ich mich zur Einhaltung dieser Satzung verpflichte.

…… ……

Datum Unterschrift