ProHTA Advisors in Healthcare
Schwabachanlage 6
91052 Erlangen
Germany

Tel: +49 (0) 163-8535855
Email: peter.kolominsky@prohta.com
URL: www.prohta.com

Erlangen, 22.07.2021

Gutachten

Bewertung von digitalen Gesundheitsanwendungen (DIGA) auf evidenz-basierter Grundlage

Auftraggeber: Vorstand der Kassenärztlichen Vereinigung Bayern (KVB)

Verfasser: Prof. Dr. med. Peter Kolominsky-Rabas, MBA

Laufzeit: 03.06.2021 - 22.07.2021 (7 Wochen)

Inhalt

Gender-Hinweis

Aus Gründen der besseren Lesbarkeit wird auf die gleichzeitige Verwendung der Sprachformen männlich, weiblich und divers (m/w/d) verzichtet.

Sämtliche Personenbezeichnungen gelten gleichermaßen für alle Geschlechter.

1. Gesundheitspolitischer Hintergrund und Fragestellung des Gutachtens

Im nachfolgenden Text wird der gesetzgeberische Begriff „Digitale Gesundheitsanwendungen“ (DiGA) und die in der internationalen Literatur gängigen Begriffe „mobile Gesundheitsapplikationen“, „mobile Health Apps, „mHealth Apps“ und „Gesundheits-Apps“ synonym verwendet.

Seit Oktober 2020 stehen die ersten „Digitalen Gesundheitsanwendungen“ (DiGA) zur Verfügung und können zu Lasten der GKV verschrieben werden. DiGA dürfen entweder auf Verordnung des behandelnden Arztes oder des behandelnden Psychotherapeuten oder mit Genehmigung der Krankenkasse angewendet werden. Voraussetzung für eine Genehmigung durch die Krankenkasse ist der Nachweis über das Vorliegen der medizinischen Indikation für die die DIGA bestimmt ist. Deutschland ist damit das erste Land weltweit, das DIGA auf Rezept verschreibt und erstattet. Grundlage für die Erstattungsfähigkeit der DiGA ist das Digitale-Versorgung-Gesetz [1] und die Digitale Gesundheitsanwendungen-Verordnung [2]. Das Bundesgesundheitsministerium führte dazu aus: „Künftig können solche Apps vom Arzt verschrieben werden. Die Kosten dafür zahlt die gesetzliche Krankenversicherung (GKV). Damit das möglichst unbürokratisch möglich ist, wird der Zugang für die Hersteller erleichtert“ [3]. Dieser erleichterte Zugang sieht vor, dass zwischen dem Antrag des Herstellers der DIGA und der Prüfungsentscheidung des BfArM laut DVG maximal drei Monate liegen dürfen [4]. Danach erfolgt bei positiver Entscheidung eine „vorläufige Aufnahme“ in das Register. Eine Aussage über den medizinischen Nutzen und die Qualität der Software trifft das BfArM zu diesem Zeitpunkt nicht. Diese Angaben müssen die Produkthersteller innerhalb des ersten Jahres nachliefern, damit eine dauerhafte Aufnahme erfolgen kann.
Anfang des 4. Quartals 2020 standen zwei Apps in dem DiGA-Verzeichnis, das beim Bundesinstitut für Arzneimittel und Medizinprodukte (BfArM) angesiedelt ist. Mit Stand zum 12.07.2021 sind 5 DIGA dauerhaft und 14 DIGA bislang vorläufig in das DIGA-Verzeichnis aufgenommen worden.

Die Beurteilung des medizinischen Nutzens und der Qualität von DIGA, bzw. von sog. „Gesundheits-Apps“ ist unerlässlich, da die Anzahl der verfügbaren DIGA in den App-Stores stetig wächst. Mittlerweile können über 100.000 Anwendungen mit Bezug zur Gesundheit in den Kategorien Health and Fitness und Medizin aus dem Google Play Store und dem Apple App Store heruntergeladen werden [5].
Patienten wünschen sich mehr Informationen für ihre am Nutzen orientierte Entscheidung, Ärzte, Psychotherapeuten und weitere Akteure im Gesundheitssystem benötigen ein solides Fundament für ihre Empfehlung bzw. Verordnung von DiGA. Die Qualitätstransparenz ist deshalb von grundlegender Bedeutung, da sich der Vergütungsbetrag einer DiGA insbesondere 1.) an dem Ausmaß der positiven Versorgungseffekte orientiert; 2.) in Relation zum Kosten-Nutzen-Verhältnis der DiGA und zu den

Preisen bereits bestehender Leistungen gesetzt wird (z. B. der GOÄ); 3.) die Ärzte als gesetzliche Verordner von DIGA für deren Qualität, Nutzen aber auch Schaden haften.

Zielsetzung des vorliegenden Gutachtens ist

1.) die Identifikation und Beschreibung der aktuell existierenden, interanationalen Instrumente für die Bewertung von DIGA,

2.) die Bewertung der wissenschaftlich-methodischen Qualität der den DIGA zugrundeliegenden Studien.

2. International verfügbare Instrumente für die Bewertung von DIGA

2.1. Methoden

Zur Identifikation der relevanten Studien wurde eine systematische Übersichtsarbeit, auch als ‚systematischer Review' bezeichnet, durchgeführt. Das Vorgehen bei der Erstellung folgte den PRISMA-Empfehlungen (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) [6].

Suchstrategien: Folgende Datenbanken wurden bei der systematischen Suche einbezogen: PubMed, EMBASE, CINAHL, PsychINFO und Cochrane Library. Folgende Suchbegriffe waren Ausgangspunkt der Recherche: [mobile medical application] OR [smartphone app] AND [framework] OR [evaluation] OR [criteria] OR [rating].

Auswahlkriterien: Studien wurden gemäß der folgenden Einschlusskriterien eingeschlossen: Bewertungsinstrumente für digitale Gesundheitsanwendungen (mobile medical applications). ‚Mobile medical applications' umfassten definitorisch Apps auf Smartphones, Tablets, Smart watches ect., die einen diagnostischen oder therapeutischen Zweck erfüllten. Die Bewertungsinstrumente richteten sich an Ärzte, Patienten bzw. an Endnutzer der digitalen Gesundheitsanwendung.

Ausgeschlossen wurden Studien, die nur Bewertungen einer einzelnen DIGA-Anwendung beschreiben, kein neues Bewertungsinstrumente darstellen, sich das Bewertungsinstrument auf die App-Entwickler und nicht auf potenzielle Endnutzer fokussiert, und Bewertungsinstrumente, die sich nicht in den Rahmen von Gesundheitsanwendungen einordnen ließen.

Studienauswahl: Die identifizierten Studien wurden von zwei Reviewern unabhängig voneinander gemäß der o. g. Einschlusskriterien begutachtet. Bei unterschiedlicher Einschätzung über den Einschluss wurde ein dritter Reviewer hinzugezogen und ein Konsens herbeigeführt.

2.2. Ergebnisse der systematischen Literaturrecherche

Der Prozess der Identifikation, Auswahl und Bewertung der Studien ist der Grafik 1. abgebildet.

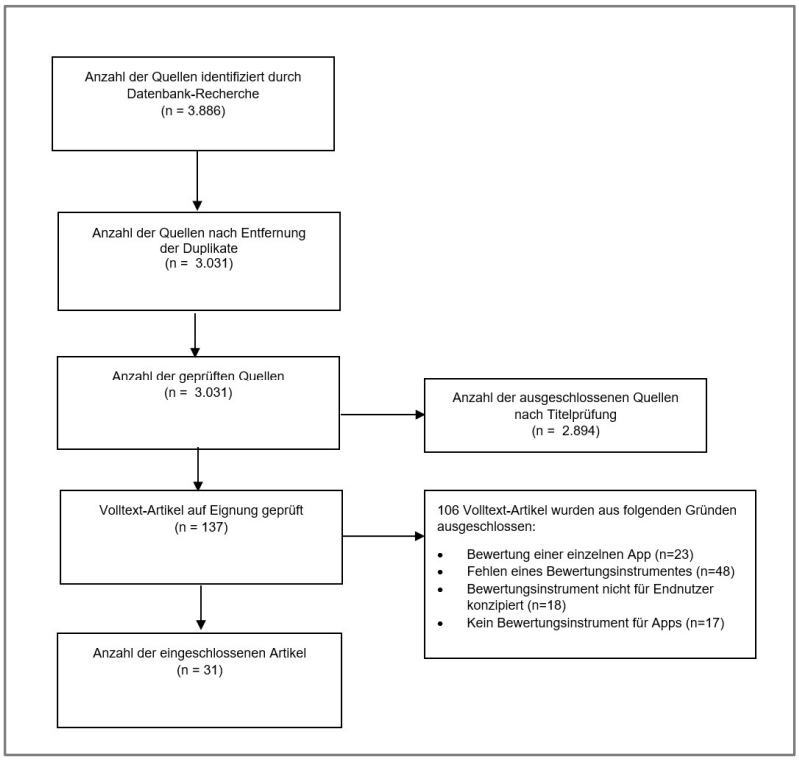
Ausgehend von den im Jahre 2018 von Moshi et al. [7] vorgestellten 45 Bewertungsinstrumenten wurde ein Review für den Zeitraum Januar 2018 bis Mai 2021 durchgeführt und 31 Studien zu Bewertungsinstrumenten identifiziert.

Die Charakteristika der in das Review eingeschlossenen Studien finden sich in Tabelle 1. Von den in Studien identifizierten Bewertungsinstrumenten wurden nur diejenigen betrachtet, die eine breite Anwendung finden. Der Begriff der „breiten Anwendung" wurde definiert, indem die Bewertungsinstrumente über eine angeschlossene Datenbank verfügen, die für Nutzer jederzeit frei zugänglich ist und die Bewertungsergebnisse der digitalen Gesundheitsanwendungen dort für den Nutzer abrufbar sind.

Tabelle 1: Übersicht über die eingeschlossenen Studien

ID	Framework	Autor	Datum	Land	Internetreferenz / Publikation
1	/	Anxiety and Depression Association of America (ADAA)	2016	USA	https://adaa.org/finding-help/mobile-apps
2	Usability Evaluation of Mobile Applications for Diabetics	Arnhold, Quade & Kirch	2014	Germany	PMID: 24718852
3	PIS: graphical classification tool for mHealth apps	Basilico et al.	2016	Italy	PMID: 26897071
4	Framework for Evaluating Mobile Applications for Cardiac Rehabilitation	Beatty, Fukuoka & Whooley	2013	USA	PMID: 24185949
5	/	Brooks et al.	2015	USA	PMID: 26283292
6	Guidelines for Mental Health App Evaluation Framework	Chan et al.	2015	USA	PMID: 26171663
7	mHealth App Evaluation for HIV	Cortes et al.	2014	Spain	PMID: 24849001
8	Evidence-based Mobile Medical Applications in Diabetes	Drincic et al.	2013	USA	PMID: 27823614
9	Apps and Eating Disorders	Fairburn & Rothwell	2015	UK	PMID: 25728705
10	/	Gautham & Johnson	2015	UK	PMID: 24621929
11	Assessing Mobile Health App Quality	Grundy et al.	2016	Australia	PMID: 27659122
12	/	Hacking Medicine Institute (HMi)	2016	USA	http://www.rankedhealth.com/approach/
13	/	Hoppe & Carter	2016	UK	PMID: 27747955
14	Mobile Apps for Asthma	Huckvale et al.	2015	UK	PMID: 25857569
15	IMS 2013: Patient Apps for Improved Healthcare	IMS Institute for Healthcare Informatics	2013	USA	https://www.iqvia.com/-/media/iqvia/pdfs/institute-reports/patient-adoption-of-mhealth.pdf
16	Evaluation Tool for Healthcare SmartphoneApplications	Jin & Kim	2015	Republic of Korea	PMID: 26431261
17	/	Kassianos et al.	2015	UK	PMID: 25600815
18	Concussion App Evaluation, Apps for Pain Management	Lalloo et al.	2015	Canada	PMID: 25370138
19	/	McMillan et al.	2016	UK	PMID: 26607787
20	mHIMSS	mHIMSS App Usability Work Group	2012	USA	https://www.himss.org/sites/hde/files/d7/HIMSSorg/Content/files/ApplicationSecurityv2.3.pdf
21	Evaluation of Mobile Clinical Applications	Murfin	2013	USA	PMID: 24261171
22	/	Pandey et al.	2013	USA	PMID: 23275239
23	/	Portelli & Eldred	2016	UK	PMID: 27583140
24	/	Powell et al.	2016	USA	PMID: 26863986
25	PsyberGuide	Psyber Guide	n.a.	USA	https://onemindpsyberguide.org/
26	/	Reynoldson et al.	2014	UK	PMID: 24422990
27	mHealth App Evaluation for HIV	Schnall et al.	2015	USA	PMID: 26385783
28	/	Shah & Castro	2014	USA	PMID: 24512633
29	APPLICATIONS scoring system	Shaia et al.	2016	USA	PMID: 27483365
30	Mobile Health Evaluation Framework	Singh et al.	2016	USA	PMID: 26934758
31	Mobile App Rating Scale (MARS)	Stoyanvo et al.	2015	Australia	PMID: 25760773

Grafik 1: Ablauf der systematischen Literatursuche



2.3. Darstellung der Bewertungsinstrumente

2.3.1. Mobile App Rating Scale (MARS)

Die Mobile App Rating Scale (MARS) ist ein einfaches Werkzeug zur Klassifizierung und Bewertung der Qualität von DIGA. Die MARS-German stellt die deutschsprachige Version dar und ist in drei Ebenen aufgebaut: Zuerst wird die Anwendung anhand von den im App Store oder der Herstellerwebseite verfügbaren Informationen klassifiziert. Im zweiten Schritt wird die App über 23 Bewertungsfragen in den Bereichen: *Engagement, Funktionalität, Ästhetik, Information* und *subjektive Qualität* evaluiert. Aus den Antworten berechnet sich für jeden Bereich eine Punktzahl, aus denen anschließend die Gesamtbewertung ermittelt wird. Die Anwendungen sind daher untereinander vergleichbar.

Die Ergebnisse der Bewertung durch die MARS finden sich in der *mobile Health App Datanbank (mHAD)* und sind dort abrufbar (siehe unten). Die mHAD bietet die technischen Funktionen und die Oberfläche für eine systematische Nutzung der MARS wobei die bewerteten Apps nach der Gesamtpunktzahl dort in einer Liste sortiert und ohne Anmeldung einsehbar sind.

Zielsetzung

Das Ziel der MARS war es, eine mehrdimensionale Skala zur Klassifizierung und Bewertung der Qualität mobiler Gesundheits-Apps, sog. mHealth-Anwendungen, zu entwickeln. Die 2015 von Stoyanov et al. [8] veröffentlichte MARS-Skala ist nach Angaben der Autoren das erste multidimensionale App-Bewertungsinstrument für mHealth-Anwendungen und findet weltweiten Einsatz. Die bis zu diesem Zeitpunkt veröffentlichten Bewertungsinstrumente fokussierten sich auf die technischen Aspekte der Anwendungen oder waren ursprünglich für Webanwendungen entwickelt worden.

Die Autoren betonen, dass die bislang existierenden Instrumente keine Kriterien zur Informationsqualität verwendeten. Auf Kriterien basierende Bewertungsinstrumente sind dagegen über die errechnete Punktzahl der Bewertungsfragen untereinander vergleichbar. Dasselbe gilt auch für die Ergebnisse von unterschiedlichen Prüfern auf Grund des Fragendesigns [8]. Das Bewertungsinstrument eignet sich zudem ebenfalls für Softwareentwickler zum Erstellen von Checklisten für das Design neuer Anwendungen. Durch die von Experten durchgeführten MARS-Ratings wird eine höhere Validität in der Qualitätsbewertung im Vergleich zu den Benutzerbewertungen im App-Store angestrebt. Eine Validierungsprüfung im Vergleich zur ursprünglichen MARS wurde im Rahmen der Entwicklung von den Autoren durchgeführt [9]. Zudem zeigt sich durch eine von verschiedenen internationalen Forschern veröffentlichte Validierungsstudie, dass sich das Bewertungsschema sehr gut eignet, um die Qualität der mHealth Anwendungen für Interessensgruppen und Patienten transparent zu machen [10]. Die MARS-G Bewertungsskala ist ebenfalls die Grundlage für die Bewertungen in der Mobile Health App Database (MHAD) [11].

Wissenschaftliche Grundlage

Die deutsche Version des Mobile App Rating Scale (MARS-G) wurde 2020 von Messner et al. [9] veröffentlicht, um Anwendungen auf diesem Weg einfacher und besser bewerten und vergleichen zu können.

Die methodische Grundlage der MARS beruht auf insgesamt 25 zwischen Januar 2000 und Januar 2013 publizierte Artikeln, die sich mit Qualitätskriterien für mobile Anwendungen beschäftigen und im Rahmen einer systematischen Suche identifiziert wurden. Dabei sind lediglich Modelle und Skalen mit quantitativen Kriterien aufgenommen worden. Qualitative Bewertungspunkte sowie Kriterien, die irrelevant für mobile Anwendungen sind, wurden ausgeschlossen. Auf diese Weise wurden 372 Bewertungskriterien identifiziert, die sich in 23 Unterkategorien unterteilen, welche jeweils eine zu beantwortende Frage (MARS-Punkt) beinhaltet. Die Gütekriterien Konstruktvalidität, Reabilität und Objektivität wurden auf der Grundlage von 1299 mHealth Anwendungen geprüft und positiv bewertet [10].

Struktur des Bewertungsinstrumentes

Bei MARS-G werden zu Beginn technische Informationen über die App gesammelt, die aus der Beschreibung im App-Store, der Webseite des Entwicklers und der App selbst hervorgehen. Es werden auch Daten zum Ziel bzw. dem theoretischen Hintergrund der Intervention erhoben, welche dann zur Einordnung der Anwendung herangezogen werden. Die in diesem Teil erhobenen Daten dienen der Klassifizierung der App. Die ursprünglich 23 Bewertungsfragen sind dabei in fünf Dimensionen gegliedert:

- Engagement (A): Interesse, Interaktivität (z.B. Senden von Nachrichten, Erinnerungen), Zielgruppenspezifität.
- Funktionalität (B): Funktionen, Benutzerfreundlichkeit und Handhabung, Navigation.
- Ästhetik (C): Graphisches Design, visueller Anreiz, farbliche Gestaltung.
- Information (D): Informationsqualität, Glaubwürdigkeit der Quellen, Literaturhinweise.
- Subjektive Qualität (E): Weiternutzung, Weiterempfehlung.

Die deutschsprachige Version der Skala beinhaltet die folgenden zusätzlichen Dimensionen und wird dadurch um 10 Fragen erweitert:

- Psychotherapie (PT): Gütekriterien im Hinblick auf die Patientensicherheit und des therapeutischen Angebots.
- Zusätzliche Angaben (F): Wahrgenommene Auswirkungen durch App-Nutzung.

Jede dieser Punkte erhält eine Bewertung auf einer 5-Punkte Skala (1: Inakzeptabel, 2: Schlecht, 3: Akzeptabel, 4: Gut, 5: Exzellent) sowie die Auswahlmöglichkeit „nicht anwendbar“, falls der Punkt nicht auf die Anwendung zutrifft. Für jeden dieser Bewertungspunkte sind fragenspezifische Anforderungen festgelegt, um die Bewertung einheitlich zu gestalten. Im Anschluss wird für jede Kategorie einzeln der Mittelwert berechnet. Der Punkt „nicht anwendbar“ wird bei der Berechnung nicht mit einbezogen. Die abschließenden Werte der einzelnen Kategorien sind somit mit denen anderer Anwendungen vergleichbar.

Angaben zu Interessenskonflikten und Finanzierung

Die Erstpublikation wurde von Mitarbeitern des *Institute of Health & Biomedical Innovation* der Queensland University of Technology (QUT) in Brisbane veröffentlicht. Die Autoren um Hides geben an, dass kein Interessenskonflikt vorliegt. Dieses Projekt wurde durch das *Young and Well Cooperative Research Center* finanziert, welches versichert keinen Einfluss auf die veröffentlichten Ergebnisse zu haben [8]. Die Autoren Messner et al. der Universität Ulm, geben ebenfalls an, dass kein Interessenskonflikt in der Publikation der deutschen Version des Bewertungsschemas vorliegt [9].

2.3.2. Anwendung: Mobile Health App Datenbank (mHAD)

Die *Mobile Health App Database* ist ein interdisziplinäres non-profit Projekt der Universitäten Ulm, Würzburg und Freiburg mit dem Ziel der Transparenz und Qualitätssicherung mobiler Gesundheits-Apps [11]. Die mobile Health App Datanbank (mHAD) bietet die technischen Funktionen und die Oberfläche für eine systematische Nutzung der MARS. Sie schafft eine höhere Transparenz der Bewertungsergebnisse der von der MARS evaluierten mobilen Gesundheits-Apps. Die Datenbank beinhaltet zurzeit 1180 Bewertungen für Apps aus dem Google-Play Store und dem Apple App-Store zurzeit (Stand 12.07.2021). Die Anwendungen sind den folgenden Themenbereiche Achtsamkeit, Angst, Depression, Gastrointestinale Erkrankungen, Kinder und Jugendliche, Krebs, Post traumatische Belastungsstörung (PTBS), Schmerz, Senioren, und Sport zugeordnet worden

Der erste Schritt bei der Aktualisierung der Datenbank ist die Identifikationsphase. In dieser wird ein sog. Web-Crawler, ein Computerprogramm, welches automatisch das World Wide Web durchsucht, auf die App-Stores angesetzt, um potenziell relevante Anwendungen zu identifizieren. Die Software wurde speziell für diesen Anwendungsfall entwickelt, da keine zugänglichen Schnittstellen zu den App-Stores bestehen [11]. Im zweiten Schritt werden die über den Suchbegriff gefundenen Anwendungen manuell von einem Expertenteam überprüft, ob sie für die gewünschte Zielgruppe entwickelt wurden und ob eine Zuordnung in eine der Kategorien erfolgte. Im dritten Schritt werden im Rahmen eines internen Review-Verfahrens die identifizierten Apps von zwei unabhängigen Prüfern mit dem MARS Bewertungsinstrument evaluiert. Nach der Überprüfung und Freigabe durch einen dritten Gutachter wird das Ergebnis auf der Plattform veröffentlicht. Liegen die Ergebnisse der Prüfer zu weit auseinander (IRR <0,75) erfolgt eine erneute Bewertung durch einen weiteren vierten Gutachter [9]. Inzwischen existiert für die Prüfer eine Weboberfläche, sodass die Apps direkt dort bewertet werden können [12]. Der Nutzer der Datenbank hat aktuell lediglich die Möglichkeit über das Setzen des Auswahlfilters zum Themenbereichs Anwendungen zu finden. Die Ergebnisse werden in Form einer Rankingliste, nach der MARS Gesamtpunktzahl mit der besten Anwendung oben sortiert, dargestellt. Auf den ersten Blick werden neben der Gesamtbewertung die entsprechende Kategorie und der Preis im App Store sowie allgemeine Informationen dargestellt. Über den Reiter ‚Bewertungen' können die Ergebnisse der Bewertungen getrennt für in eine der insgesamt fünf MARS-Dimensionen abgerufen werden. Im letzten Reiter kann der Nutzer weitere Angaben zur Klassifikation einer Anwendung, wie den App Fokus, die theoretischen Hintergründe und Informationen über den Entwickler einsehen. Auch ist ein direkter Link zu der App im jeweiligen App Store eingebunden [11].

2.3.3. Bewertungsinstrument der American Psychiatric Association (APA)

Das Bewertungsinstrument der American Psychiatric Association (APA) stellt eine Möglichkeit dar, systematisch Informationen über mHealth Anwendungen zu erfassen. Da unterschiedliche Endnutzer auch unterschiedliche Anforderungen an die Anwendungen haben, versucht das Bewertungsinstrument nicht, die beste Anwendung anhand einer Gesamtpunktzahl zu finden, sondern gibt dem Benutzer die Möglichkeit Prioritäten festzulegen. Das Ziel der Bewertungen liegt darin, den Nutzern ausreichende Informationen zur Verfügung zu stellen, anhand derer sie eine fundierte Entscheidung für ihre spezielle Situation treffen können. Während das ursprüngliche APA-Bewertungsinstrument aus dem Jahr 2019 den Fokus überwiegend auf medizinisches Personal gelegt hat [13], werden in dem aktuellen Modell andere Berufsgruppen, wie z.B. Sozialarbeiter, Psychologen und Informatiker, stärker berücksichtigt. Das Bewertungsinstrument bildet die Grundlage für App Bewertungen in der *mHealth Index and Navigation Database (MIND)* (siehe unten).

Zielsetzung

Das Ziel von APA war es, bereits veröffentlichte Bewertungsinstrumente zu identifizieren, deren Inhalte zu vergleichen, Gemeinsamkeiten herauszuarbeiten und in einem neuen Bewertungsinstrument zusammenzuführen. Dabei verzichtet das Instrument vollständig auf qualitativen Fragen, um die Bewertung durch ausgebildete Prüfer zu möglichen und vergleichbare Ergebnisse zu generieren. Die Ergebnisse der Bewertungen sollen den interessierten Gruppen in Form einer Online-Datenbank zugänglich gemacht werden. Weiteres Ziel ist es, mit Hilfe dieses Bewertungsinstruments eine kostenlose und nach der Registrierung zugängliche Datenbank für App-Bewertungen zu schaffen, die fundierte Unterstützung anhand von klinisch relevanten Kriterien bei der Auswahl der richtigen mHealth-Anwendung für Patient und Kliniker liefert. Benutzer können dort die einzelnen Bewertungen einsehen und somit Anwendungen miteinander vergleichen. Die Ergebnisse sollen für jeden Anwender selbsterklärend sein. Die veröffentlichte Datenbank enthält sowohl Informations- als auch Schulungsmodule für die Bewertung der Anwendungen, sodass über die Plattform eine Weiterbildung zum Prüfer stattfindet.

Im Unterschied zu anderen bekannten Bewertungsinstrumenten, wie z.B. der mobile App Rating Scale (MARS), vergibt der Prüfer keine vergleichenden Gesamtpunkte bei der Bewertung. Der Grund hierfür liegt darin, dass unterschiedliche Bevölkerungsgruppen, wie Jugendliche und Erwachsene auch unterschiedliche Bedürfnisse an eine App haben. Die Eingrenzung der Anwendungen erfolgt über das Setzen von Filtern in den einzelnen Kategorien, die beliebig miteinander kombiniert werden können [14].

Wissenschaftliche Grundlage

Bei diesem Instrument handelt es sich um eine Weiterentwicklung des ursprünglichen Bewertungsinstruments der American Psychiatric Association, welches erstmals 2019 von Henson et al. in der Zeitschrift ‚The Lancet Digital Health' veröffentlicht wurde [13]. Das Bewertungsinstrument wurde von Lagan et al. weiterentwickelt und 2021 in der Fachzeitschrift British Medical Journal (BMJ) publiziert [14].

Struktur des Bewertungsinstrumentes

Das Bewertungsinstrument besteht aus insgesamt 105 zu beantwortenden Fragen, die entweder ein binäres Ergebnis (ja / nein) oder eine numerische Eingabe als Antwort liefern. Diese einzelnen Bewertungspunkte gehen aus einer Häufigkeitsanalyse der identifizierten Bewertungsinstrumente hervor. Die Fragen gliedern sich in die sechs folgenden Kategorien:

- App-Ursprung und Funktionalität (1): Kosten, Funktionen (Medizinische Ansprüche), Verfügbarkeit, Plattform, Zahl der Downloads, Offline-Funktionalität.
- In- und Output (2): Kompatibilität mit externen Geräten, Benachrichtigungen und Erinnerungen, Datennutzung.
- Datenschutz und Datensicherheit (3): Datenschutz-Bestimmungen, Verschlüsselung, Informationen zu gesammelten und gespeicherten Daten.
- Klinische Grundlage (4): Klinische Gültigkeit, Evidenz des Wirksamkeitsnachweises, Patientennutzen, Erfüllung der Ansprüche.
- Funktionen (5): Zusammenarbeit mit dem Anbieter, „Gamification / Seriouse Gaming".
- Interoperabilität und Datenaustausch (6): Dateneigentum, Export der Daten.

Subjektive und qualitative Fragen z.B. zur Benutzerfreundlichkeit, Layout und grafische Darstellung deckt die Bewertungsskala nicht ab. Als Ergebnis liefert das Instrument keine abschließende numerische Gesamtbewertung, die einen Vergleich der Anwendungen untereinander ermöglicht. Stattdessen steht die Flexibilität der Interpretation im Vordergrund. Der Endnutzer bzw. Kliniker kann durch das Setzen von Filtern, die für den speziellen Anwendungsbereich in Frage kommenden Anwendungen mit deren entsprechenden Bewertungen herausfinden [15].

Angabe zu Interessenskonflikten und Finanzierung

Die genannten Publikationen wurden von Mitarbeitern der *Division of Digital Psychiatry Collaborative Research Group, Beth Israel Deaconess Medical Center* der Harvard Medical School in Boston veröffentlicht. Die Autoren geben zudem an, dass kein Interessenskonflikt vorliegt. Die öffentlich zugängliche Datenbank MIND wird durch die *Argosy-Foundation* unterstützt, welche wiederum ebenfalls angibt, keinen Einfluss auf die Forschungsergebnisse zu haben [16].

2.3.4. Anwendung: mHealth Index & Navigation Database (MIND)

Die *mHealth Index and Navigation Database (MIND)* bietet eine Plattform und die technische Umsetzung, um Informationen über mHealth Anwendungen systematisch zu erfassen. Als Grundlage dient das aktuelle Bewertungsinstrument der American Psychiatric Association (APA). In diesem sind Fragen aus über 70 unterschiedlichen Bewertungsinstrumenten auf Basis einer Häufigkeitsanalyse harmoniert. Diese ist nach der Registrierung verwendbar und stellt alle Inhalte kostenlos zur Verfügung. Der Fokus der Entwickler lag darin, ein Instrument und eine Plattform zu schaffen, die für unterschiedliche Interessensgruppen Informationen anbietet, was einen Mehrwert darstellt [15]. Der Nutzer kann im Suchfeld Anwendungen über den Namen eine bestimmte Funktion oder die Plattform finden. Des Weiteren stellt die Plattform sehr gute Filtermöglichkeiten zur Verfügung [16]. Es können 84 einzelne Filter aus neun Kategorien beliebig miteinander kombiniert werden, wodurch eine sehr genaue Eingrenzung der Anwendungen möglich ist. Wird eine App identifiziert, können in der Detailansicht alle Antworten der 105 Bewertungsfragen eingesehen werden. Diese sind entweder binär (ja/ nein) oder numerisch zu beantworten, um die maximale Vergleichbarkeit der Ergebnisse verschiedener Gutachter zu erreichen [14]. Zudem sind alle Bewertungen, die zu der App auf der Plattform veröffentlicht wurden, frei zugänglich. Dadurch wird ein potenzieller Verlauf der Bewertungen und Veränderungen bei unterschiedlichen Versionen oder Prüfern sichtbar. Ebenfalls wird automatisch der aktuelle Informationstext aus dem entsprechenden App Store angezeigt. Auch kann jeder Gutachter ein kurzes qualitatives Review zu der Anwendung erstellen, welches ebenfalls in der Detailansicht angezeigt wird.

Die Plattform bietet zudem zwei für jeden registrierten Nutzer zugängliche Schulungsmodule an, um sich als Gutachter qualifizieren zu können. Dieses dezentrale Verfahren zur Rekrutierung von neuen Prüfern führt zu einer aktuellen Datenbank. Die weltweit verteilten Gutachter können z.B. Anwendungen, die nur in bestimmten Ländern in den App Stores verfügbar sind, zur Datenbank hinzufügen. Ermöglicht wird dies durch den Aufbau der Bewertungsfragen, die ein standardisiertes Verfahren zur Erhebung der Daten vorgeben [16].

2.3.5. AppQ-Gütekriterien-Kernset der Bertelsmann-Stiftung

Im deutschsprachigen Raum entwickelten Thranberend und Bittner, beide Mitarbeiter der Bertelsmann Stiftung, den Kriterienkatalog „AppQ“ mit neun Themenbereichen (=Qualitätsanforderungen), der von DiGA-Entwicklern zur strukturellen Qualitätsberichtserstattung verwendet werden soll [17]. Diese Informationen sollen für mehr Transparenz im digitalen Gesundheitsmarkt sorgen. Der Themenkatalog ist in drei Bereiche gegliedert: Evidenzbasierung (z.B. Einhaltung klinischer Leitlinien), Vertrauenswürdigkeit (z.B. Datenschutz und Datensicherheit) und Nutzerperspektive (z.B. Benutzerfreundlichkeit). Das Sammeln der Informationen aus Eigenrecherche und freiwilligen Herstellerangaben bildet die Grundlage der Ergebnisse, die auf der Online Plattform „Weisse Liste“ der Öffentlichkeit zur Verfügung gestellt werden. Die Angaben sollen sowohl dem Patienten als auch dem Arzt dabei helfen, ein Fundament für eine Empfehlung bzw. Nutzenentscheidung zu schaffen. Aufgrund der schnellen Veränderungen im Gesundheitsmarkt wird der Kriterienkatalog dynamisch weiterentwickelt.

Zielsetzung

Die aktuelle Version des App Bewertungsmodells wurde am 15.06.2020 mit dem Titel „AppQ 1.1 – Gütekriterien-Kernset für mehr Qualitätstransparenz bei digitalen Gesundheitsanwendungen“ von Thranberend und Bittner [18] veröffentlicht. Die Bertelsmann Stiftung befasst sich seit 2015 mit der Analyse von DiGA, um Qualitätskriterien zur Bewertung mobiler Anwendungen im deutschsprachigen Raum zu erarbeiten. Das Ziel dabei ist, eine Qualitätseinschätzung der Anwendungen zu erreichen, um einen standardisierten Vergleich aller als Medizinprodukt zertifizierter DiGA, unabhängig der Risikoklasse, durchführen zu können. Patienten wünschen sich mehr fundierte Informationen für ihre Nutzenentscheidung und behandelnde Ärzte suchen ein Fundament für ihre Empfehlungen und Verordnungen. Das dynamisch entwickelte Bewertungsschema dient Softwareherstellern auch zu einer strukturierten und zentralen Qualitätsberichtserstattung [17]. Das Bewertungsschema ist ebenfalls die Grundlage für die Bewertungen im App-Verzeichnis „Weisse Liste gemeinnützige GmbH“, einer Tochtergesellschaft der Bertelsmann Stiftung, und ist Teil des Projektes mit dem Titel „Trusted Health Apps“. Das übergeordnete Ziel des App-Verzeichnisses ist es, die Transparenz im Feld der DiGA zu erhöhen. In die Liste werden dabei Anwendungen aufgenommen, die in Deutschland als Medizinprodukt nach der Medical Device Regulation (MDR) oder im Rahmen der Übergangsvorschrift nach der Medizinprodukterichtlinie (MDD) zertifiziert sind. Die Bewertung erfolgt unabhängig von der Risikoklasse. Es handelt sich dabei um Apps, Webanwendungen oder andere digitale Produkte, die von Patienten genutzt werden können, um Gesundheitsziele zu erreichen. Die App-Liste ist öffentlich und ohne Anmeldung zugänglich [21].

Wissenschaftliche Grundlage

Das Framework baut auf dem im Juni 2018 veröffentlichten Meta-Kriterienkatalog APPKRI des Fraunhofer-Institutes für Offene Kommunikationssysteme (FOKUS) auf [19]. Ziel war es ein Werkzeug

zu entwickeln, um spezifischere Kriterienkataloge für bestimmte Klassen von GesundheitsApps erstellen zu können [20]. In einer Expertenrunde wurde aus dem APPKRI-Katalog ein Entwurf des AppQ erstellt. Aus dieser Grundlage entstand zusammen mit der Analyse „Transfer von Digitalen-Health-Anwendungen in den Versorgungsalltag“ der Bertelsmann Stiftung ein entsprechender Referenzentwurf. Die Weiterentwicklung erfolgte durch Fachgespräche mit stationär und ambulant tätigen Ärzten, Vertretern von Krankenkassen, medizinischen Fachgesellschaften und Datenschutzbehörden, um verschiede Interessengruppen an der Entwicklung teilhaben zu lassen. In der Version 1.1 sind ebenfalls „Hinweise auf unerwünschte Wirkungen“ aufgenommen, da bisher nur Untersuchungen zum Nutzen und nicht zum potenziellen Schaden gemacht wurden [18]. Im Anschluss wurden in einer ersten Analyse 79 deutschsprachige DiGAs identifiziert. Damit eine Anwendung in das App-Verzeichnis aufgenommen wird, müssen folgende Einschlusskriterien erfüllt sein: Es muss eine deutschsprachige Digitale Anwendung (mHealth Anwendung, Webanwendung oder Sprachanwendung für Sprachassistenten) sein, deren Hauptfunktionen in der App selbst beinhaltet sind (z.B. keine externen Sensoren notwendig) und deren Gestaltung der Benutzeroberfläche auf den Patienten ausgelegt ist.

Struktur des Bewertungsinstrumentes

Das Bewertungsinstrument besteht aus zwei Teilen: dem Metadatenmodell zur Beschreibung und Klassifizierung einer DiGA und dem Gütekriterien-Kernset, über welches die strukturelle Qualitätstransparenz angegeben wird. Mit den Informationen des Metamodells werden die DiGA selbst, deren Ziel und die entsprechende Nutzergruppe beschrieben. Dabei wird zwischen den Stammdaten und den klassifizierenden Metadaten unterschieden. Die Stammdaten haben einen beschreibenden Charakter und umfassen Angaben zur Anwendung, der Plattform, zum Hersteller, über den Medizinproduktstatus, zur Zielgruppe, zum Preismodell und zu Hard- und Softwareanforderungen. Jede dieser Kategorien besteht aus spezifischen Fragen, um die Anwendung genauer zu beschreiben. Bei den klassifizierenden Metadaten wird nochmal in Funktionalität und Versorgungseffekt unterschieden. Anhand dieser Angaben kann der Hersteller den Funktionsumfang beschreiben. Die Funktionen sind wiederum ein wesentlicher Bestandteil zur Erfüllung des Versorgungseffektes, der die beabsichtigte Wirksamkeit der DiGA beschreibt. Im Rahmen des Metamodells werden hier nur Hypothesen aus Angaben des Herstellers erfasst.

Die Überprüfung erfolgt im Rahmen des Gütekriterien-Kernsets. Dieses soll zu einer größeren strukturellen Qualitätstransparenz beitragen und ist in drei Hierarchieebenen gegliedert. Die erste Ebene enthält neun Themen zur inhaltlichen Struktur, die aus aktuellen Erkenntnissen wissenschaftlicher Studien, Leitlinien und Expertenwissen hervorgehen. Die Fragen sind in die folgenden neun Themenbereiche gegliedert:

- Medizinische Qualität (1):
Medizinisch-fachliche Fundierung, Berücksichtigung von Erkenntnissen aus wissenschaftlichen Studien und Leitlinien, Maßnahmen zur Reduzierung des Nutzenrisikos
- Positiver Versorgungseffekt (2):
Effekte der medizinischen Wirksamkeit und des medizinischen Nutzens, Verfahrens- und Strukturverbesserungen in der Gesundheitsversorgung, Hinweise auf unerwünschte Wirkungen
- Datenschutz (3):
Allgemeine Einhaltung der Datenschutz Grundverordnung (DSGVO), Abfrage von Einwilligungen zur Datenverarbeitung, Schutz der Privatspähre des Nutzers, Umsetzung der Datenminimierung und Zweckbindung
- Informationssicherheit (4):
Maßnahmen zur Absicherung der Vertraulichkeit, Integrität, Verfügbarkeit und Belastbarkeit der über die DiGA verarbeiteten Daten
- Technische Qualität (5):
Bezeichnung als Medizinprodukt, Erfüllung der je nach Risikoklasse verschiedenen Anforderungen (Prüfzertifikate, usw.)
- Verbraucherschutz und Fairness (6):
Verhältnis zwischen dem Hersteller und dem Endnutzer, Verbraucherfreundlichkeit, Kommunikation und Unterstützungsleistungen des Herstellers
- Interoperabilität (7):
Export und Import von Daten, Standardschnittstellen, Interaktion mit anderen Anwendungen und Diensten
- Nutzerfreundlichkeit und Motivation (8):
Gebrauchstauglichkeit (Maßnahmen zur Usability), Personalisierte Nutzung, Maßnahmen zur Förderung der Nutzermotivation und Nutzertreue, Verlässlichkeit der Anwendung, Verwendung geeigneter Gesundheitsinformationen
- Anbindung an das Gesundheitssystem (9):
Unterstützung von Angehörigen von Gesundheitsberufen, Anbindung an nationale E-Health-Dienste der Telematikinfrastruktur

Jedes dieser Themen hat mehrere Kriterien (Ebene 2), die als repräsentative Fragestellungen formuliert sind. Auf der dritten Hierarchieebene stehen 187 Indikatoren, die eindeutig binär (ja/ nein) zu beantworten sind. Unter den Indikatoren befinden sich sowohl qualitative als auch quantitative Bewertungsfragen. Einzelne Indikatoren sind nicht auf alle DiGA anwendbar, hier muss der Hersteller in einem Freitextfeld den Grund dafür angeben [17].

Angabe zu Interessenskonflikten und Finanzierung

Der Studienbericht zu AppQ wurde von der Bertelsmann Stiftung veröffentlicht. Das Projekt wird vom Bundesministerium für Gesundheit gefördert und entsteht in Zusammenarbeit mit dem Fraunhofer-Institut für Offene Kommunikationssysteme (FOKUS) [19]. Die Bertelsmann Stiftung hält die Mehrheit der Anteile am Bertelsmann Konzern, einem der weltweit größten Medienunternehmen, zu dem unter anderem auch die Fernsehersender der RTL Group (z.B. RTL, VOX, n-tv) gehören. Der Journalist Thomas Schuler kritisiert dabei, dass die kommerziellen und auch politischen Interessen im Kontrast zu den Tätigkeiten der Stiftung stehen und wirft in seinem Buch *„Bertelsmannrepublik Deutschland – eine Stiftung macht Politik“* dem Konzern mangelnde Transparenz vor [22].

2.3.6. Anwendung: App Verzeichnis „Weisse Liste“ der Bertelsmann-Stiftung

Die „Weisse Liste“ ist eine gemeinnützige GmbH der Bertelsmann Stiftung, welche zusammen mit den Dachverbänden der wichtigsten Patientenvertretungen und Verbraucherorganisationen 2008 ins Leben gerufen wurde. Zu den Angeboten zählen die Unterstützung von Patienten, Angehörigen und Gesundheitsexperten bei der Suche nach einem passenden Arzt, einem Krankenhaus oder einer Pflegeeinrichtung sowie bei der Auswahl einer Gesundheitsanwendung. Das Projekt „Trusted Health Apps“ soll Transparenz im Bereich der DiGA schaffen und dem Nutzer Informationen über das Angebot und die Qualität von Gesundheitsanwendungen zur Verfügung stellen. Diese stehen auf der Webseite ohne Anmeldung der Öffentlichkeit zur Verfügung und bieten dem Nutzer die Möglichkeiten sich zu informieren [21].

Damit eine mHealth Anwendung in der Datenbank gelistet wird, müssen definierte Einschlusskriterien erfüllt sein. Diese weichen von den Anforderungen im Fast-Track-Verfahren der DiGA ab. Es muss sich um eine deutschsprachige mHealth Anwendung, Webanwendung oder Sprachanwendung für einen Sprachassistenten handeln. Die Benutzeroberfläche muss dabei auf den Patienten ausgelegt sein, die Nutzung soll vom Patienten selbst oder gemeinsam durch den Patienten mit Angehörigen oder professioneller Hilfe erfolgen.
Ebenfalls muss die Anwendung ein zertifiziertes Medizinprodukt nach Medicine Device Regulation (MDR) oder im Rahmen der Übergangsvorschrift (bis 2021) nach Medicine Device Directive (MDD) sein. Die Risikoklasse hat dabei keinen Einfluss. In regelmäßigen Abständen überprüft das Team, ob neue Anwendungen die Zertifizierung erlangen und in die Liste aufgenommen werden können [21].

Die Grundlage der Datenbank bilden Informationen, die zum einen aus der Eigenrecherche des Gutachterteams, welches aus Medizinern, Datenschutz- und Informationssicherheitsspezialisten besteht und zum anderen aus freiwilligen Herstellerangaben, die strukturiert über die AppQ-Kriterien erfasst werden. Bestimmte Angaben müssen dabei über ein Zertifikat oder mit Studienergebnissen nachgewiesen werden [17]. Die Anwendungen sind zu einer der 25 verfügbaren Kategorien im Bereich Prävention, Diagnostik und Therapie zugeordnet, diese reichen von unspezifischen

Zustandsbeschreibung (z.B. Alter und Pflege) bis hin zu konkreten klinischen Diagnosen (z.B. Diabetes). Zu jeder gelisteten App wird vom Expertenteam eine kurze Zusammenfassung mit den wichtigsten Informationen über den intendierten Versorgungseffekt, die Hauptfunktionen für den Patienten und über die Nachweise zu medizinischer Wirksamkeit, zum medizinischen Nutzen sowie zu Verfahrens- und Strukturverbesserungen in der Gesundheitsversorgung. Der Nutzer kann Anwendung über das Setzen des Kategorienfilters oder über die Suchleiste finden. Weitere Informationen können in der Detailansicht abgerufen werden. Im ersten Reiter sind die Plattformen dargestellt auf denen die Anwendung verfügbar sind. Im mittleren Reiter sind Basisinformationen wie beispielsweise die Risikoklasse oder der Preis angegeben und im letzten Reiter können eigene Erfahrungen mit der App dem Entwicklerteam mitgeteilt werden. Unter den für die Anwendung repräsentativen Abbildungen kann der Nutzer weitere Details zu einzelnen Bereichen der AppQ-Kriterien einsehen [21].

2.4. Fazit

Bewertung: Mobile App Rating Scale – German (MARS-G)

Der Wirksamkeitsnachweis, den die Hersteller erbringen, wird dabei in der Kategorie 4: *Informationsqualität* („Beinhaltet Informationen mit hoher Güte und Qualität aus einer glaubwürdigen Quelle?“) Die sieben Fragen beziehen sich auf Informationen, ob eine App dem beschriebenen Anwendungsgebiet entspricht sowie auf Aussagen zu spezifischen, messbaren und erreichbaren Zielen. Die Punkte zur Glaubwürdigkeit der Informationen „Kommt die App aus einer legitimen Quelle?“ und zur Evidenzbasierung „Wurde die App geprüft/ getestet?“ beziehen sich auf das Vorhandensein wissenschaftlicher Belege zur Wirksamkeit der Anwendung. Die Frage nach deren methodischer Qualität bzw der Evidenz zur Wirksamkeit der DIGA wird nicht hinterfragt.

Bewertung: American Psychiatric Association (APA)

Von den insgesamt 105 Bewertungsfragen fallen lediglich 9 in die Kategorie der Klinischen Grundlage, das sind ca. 9% der Gesamtfragen. Die Fragen versuchen dabei darzustellen, ob und in welcher Form ein Wirksamkeitsnachweis vom jeweiligen Hersteller einer Anwendung erbracht wurde. Dies wird in Form von Machbarkeits- bzw. evidenzbasierten Studien, die durchgeführt bzw. veröffentlich wurden und mit der Frage, ob die Anwendung das tut, was sie vorgibt, überprüft. Hier wird jeweils der höchste Impact-Factor, der den Einfluss der wissenschaftlichen Fachzeitschrift angibt, angegeben. Auch werden potenzielle Schäden, die durch die Anwendung entstehen können, und die entsprechenden Warnungen bzw. Warnhinweisen dazu analysiert. Ein grundlegendes Problem ist jedoch, dass keine Angaben zur Qualität der Studien bzw. zu den Ergebnissen gemacht werden, sondern die Fragen lediglich das Vorhandensein von Studien abbilden. Abschließend sei bemerkt, dass die Höhe eines sog. Impact Factors (Wirkfaktor) einer Fachzeitschrift allein keine Aussage über die Qualität einer Studie oder gar über Wirksamkeitsbeleg einer DIGA darstellt.

Bewertung: AppQ-Gütekriterien-Kernset der Bertelsmann-Stiftung

Zur Wirksamkeit bzw. zum Nutzen führt die AppQ folgendes aus: „Als positive Versorgungseffekte werden im Sinne von AppQ zum einen solche Effekte bezeichnet, welche die medizinische Wirksamkeit oder den medizinischen Nutzen von digitalen Gesundheitsanwendungen (DiGA) betreffen (AppQ-VE-01). Zum anderen umfasst der Begriff Verfahrens- und Strukturverbesserungen in der Gesundheitsversorgung (AppQ-VE-02). Die in beiden Kriterien erfragten Nachweise beziehen sich ausschließlich auf solche Effekte, die für die jeweilige DiGA im Speziellen erbracht worden sind oder werden sollen." Grundlage dieser dort zusammengetragenen Informationen zur Evidenz im Hinblick auf Nutzen und Verbesserungen ist die Selbstauskunft des Herstellers der DIGA.

Als Fallbeispiel sei hier die Aufnahme der DIGA *Demenz-Screening Test* (DST) in die „Weisse Liste" im September 2020 genannt. Unter dem Gütekriterium M*edizinischer Wirksamkeit/ Medizinischer Nutzen* (s.o.) gibt die „Weisse Liste" folgende Bewertung ab: „DST soll Beschwerden und Komplikationen beim Nutzer verringern. Aus den Angaben des Herstellers lässt sich ableiten, dass dieser Effekt für DST gut belegt ist. Dieser postulierte Effekt hinsichtlich Wirksamkeit oder Nutzen von DST ist durch prospektive, parallel kontrollierte klinische Studien (Evidenzgrad II) belegt." (https://www.trustedhealthapps.org/de/suche/DST%20-%20Demenz%20Screening%20Test/dst-demenz-screening-test-1).

Eine Suche in den internationalen Datenbanken fand keine Belege für das Vorliegen einer Wirksamkeitsstudie in der welcher der namentlich gennannte Screening –Test DST evaluiert wurde. Auf Anfrage des Verfassers an den Betreiber der DIGA, wurden insgesamt drei Dokumente übersendet. Bei einem der Dokumente handelte es sich um eine klinische Studie, jedoch ohne Bezug zum DST, bzw. ohne Hinweis auf eine erfolgte Evaluation dieser DIGA. Bei den anderen beiden Unterlagen handelte es sich um sog. Konferenzabstracts einer US-amerikanischen Arbeitsgruppe.
Trotz der durchgeführten Recherchen und Rückfragen bleibt es unklar, ob für diese DIGA eine wissenschaftliche Evidenz auf klinischer Grundlage existiert.

3. Studienqualität der im deutschen GKV-System vergüteten DIGA

3.1. Methoden

Damit die Wirksamkeit einer medizinischen Behandlung, also beispielsweise eines Medikaments oder einer (digitalen) Gesundheitsanwendung, eingeordnet werden kann, ist es von zentraler Bedeutung, die Studiendurchführung zu betrachten. Dabei kann im Rahmen der Studiendurchführung ein Verzerrungsrisiko bestehen. Von einem Verzerrungsrisiko in Studien spricht man dann, wenn die Ergebnisse aufgrund des Studiendesigns oder der methodischen Durchführung der Studie von dem „tatsächlichen" Ergebnis abweichen können. Randomisiert kontrollierte Studien (RCTs) gelten in der medizinischen Forschung als Goldstandard, da bei ihnen das Verzerrungsrisiko verglichen mit anderen Studienformen am geringsten ist [23]. Nichtsdestotrotz können auch die Ergebnisse in RCTs verzerrt sein, etwa wenn die strengen methodischen Kriterien zur Studiendurchführung nicht eingehalten werden.

Gegenstand der Untersuchung waren die dauerhaft in das DIGA-Verzeichnis aufgenommen DIGA, für die wissenschaftliche Belege vorlagen (Stand 12.07.2021). Das Verzerrungsrisiko der Studien zu den hier aufgeführten digitalen Gesundheitsanwendungen (DiGA) wurde mit dem *Revised Cochrane risk-of-bias tool for randomized trials (RoB 2),* Version of 22 August 2019 bewertet [23].
Die Bewertungsergebnisse der insgesamt 15 untersuchten Studien werden detailliert im Anhang dargestellt. Das Risk of Bias tool ist eine Checkliste zur Bewertung des Verzerrungsrisikos von randomisiert kontrollierten Studien (RCT). Die ursprüngliche Checkliste wurde 2008 von Autoren des Cochrane Netzwerkes entwickelt und 2011 aktualisiert. Bei Cochrane handelt es sich um ein internationales Netzwerk, das wissenschaftliche Grundlagen für Entscheidungen im Gesundheitswesen, in der Regeln in Form von systematischen Übersichtsarbeiten, bereitstellt. Diese systematischen Übersichtsarbeiten fassen die gesamte wissenschaftliche Evidenz zu einer konkreten Fragestellung aus der Medizin oder anderen Gesundheitswissenschaften auf Basis einer strengen wissenschaftlichen Methodik und frei von kommerziellen Interessenkonflikten zusammen. Die aktualisierte Checkliste ermöglicht eine Bewertung des Verzerrungsrisikos einzelner RCTs bezogen auf den 1.) Randomisierungsprozess, 2.) Abweichungen von der beabsichtigten Behandlungsdurchführung (hier: der digitalen Gesundheitsanwendung), 3.) fehlende Werte, die 4.) Bewertung des Endpunkts oder 4.) ein selektives Berichten bestimmter Ergebnisse.

Bei dem Randomisierungsprozess geht es darum, ob die Verteilung der Studienteilnehmer auf die verschiedenen Gruppen (Behandlungs- und Kontrollgruppe) zufällig und verdeckt erfolgte, also ohne das Wissen des Studienpersonals um die Verteilung. Zudem ist zu prüfen, ob sich die Gruppen zu Beginn der Studie unterscheiden.

Abweichungen von der beabsichtigten Behandlungsdurchführung können entstehen, indem bspw. zusätzliche, nicht im Vorhinein geplante Behandlungsmaßnahmen ergriffen werden oder, wenn die Behandlung nicht wie geplant durchgeführt werden kann. Sie können jedoch auch dann entstehen, wenn sich die Studienteilnehmenden nicht an die geplante Behandlung halten, bspw. wenn Medikamente nicht (zeitgerecht) eingenommen werden oder die digitale Anwendung nicht verwendet wird. Um den Effekt einer Behandlung im Rahmen der Datenauswertung wissenschaftlich und statistisch beurteilen zu können, gibt es grundsätzlich zwei Möglichkeiten.
Zum einen können die Daten aller Studienteilnehmer, die ursprünglich auf einer der beiden Gruppen (Behandlungs- und Kontrollgruppe) verteilt wurden, entsprechend ihrer ursprünglichen Gruppenzuteilung ausgewertet werden. Die Auswertung erfolgt in diesem Fall bespielweise auch bei solchen Teilnehmern, die die Studie abbrechen, die Behandlung nicht wie vorgesehen durchführen oder eine andere Behandlung erhalten, als ursprünglich geplant. Man bezeichnet diese Form der Auswertung als Intention-to-Treat (ITT).
Zum anderen können ausschließlich die Daten solcher Teilnehmenden ausgewertet werden, die die Studie streng nach Studienprotokoll durchgeführt haben. Hierbei werden Personen ausgeschlossen, die die Studie beispielsweise nicht bis zum Ende durchgeführt oder die Anwendung der Behandlung zwischenzeitlich vergessen haben. Diese Form der Auswertung nennt man Per-Protocol-Analysis (PP).

Für gesundheitspolitische Entscheidungen gilt der Intention-to-Treat Effekt als bedeutsamer, da die gleichmäßige Verteilung zufälliger Störgrößen, also beispielsweise Alter, Geschlecht, etc., auf beide Gruppen erhalten bleibt. Er wird deshalb im Rahmen dieser Evidenzbewertung dem Per-Protocol Effekt vorgezogen. Das RoB 2 Tool ermöglicht die Untersuchung von Verzerrungen in diesem Punkt, beispielsweise durch eine fehlende Verblindung. Eine fehlende Verblindung besteht dann, wenn die Teilnehmenden und/oder die Prüfer im Vorhinein oder im Nachgang der Randomisierung erfahren, welche Gruppenzuteilung besteht.

Fehlende Werte, etwa bezogen auf den zu untersuchenden Endpunkt, können die Studie ebenfalls verzerren, insbesondere, wenn es in einer der beiden Gruppen eine höhere Anzahl fehlender Werte gibt. Zudem können fehlende Werte die statistische Aussagekraft verringern. Fehlende Werte können beispielsweise entstehen, wenn Teilnehmende aus der Studie ausscheiden (Drop-Out) oder nicht mehr auffindbar sind (Loss-to-Follow-Up). Obwohl Intention-to-Treat Analysen alle Teilnehmenden einbeziehen sollten, können auch in diesem Fall Teilnehmende ausgeschlossen werden, bei denen zu wenige Daten vorliegen. Man spricht dann von einer modifizierten Intention-to-Treat Analyse (mITT). Als Richtwert für kontinuierliche Ergebnisvariablen schlägt das RoB 2 Tool die Verfügbarkeit von 95% der Daten nach der abschließenden Erhebung des Endpunkts vor, andernfalls kann es zu Verzerrungen kommen. Das Verzerrungsrisiko in diesem Punkt kann allerdings beispielsweise durch Sensitivitätsanalysen verringert werden. Sensitivitätsanalysen untersuchen, wie sich Ergebnisse verändern, wenn sich bestimmte Paramater in der Untersuchung ändern, beispielweise durch

fehlende Werte. So könnte man etwa beobachten, wie sich die Ergebnisse verändern, wenn man bestimmte Personen aus der Analyse ausschließt, bei denen Angaben fehlen. Unterscheiden sich die Ergebnisse verschiedener Sensitivitätsanalysen nicht oder nur geringfügig, gelten die Ergebnisse als robust.

Hinsichtlich einer Verzerrung im Rahmen der Bewertung des Endpunkts wird überprüft, ob die Messinstrumente angemessen waren, ob die Messmethodik in beiden Gruppen (Behandlungs- und Kontrollgruppe) identisch war, also mittels der gleichen Instrumente und zu den gleichen Zeitpunkten erfolgte, und ob die Prüfer über die Gruppenzuteilung der Teilnehmenden Bescheid wussten.

Eine Verzerrung bezogen auf die Wirksamkeit der Behandlung kann zudem auftreten, weil nur bestimmte Studienergebnisse berichtet werden. Dies ist dann relevant, sofern ein Endpunkt auf mehrere verschiedenen Arten gemessen und analysiert werden kann. Dabei könnte es vorkommen, dass die Messung auf verschiedene Weisen erfolgte, jedoch nur die Ergebnisse veröffentlicht wurden, die das gewünschte Ergebnis, also einen Mehrwert der untersuchten Behandlung, bestätigten. Um ein Verzerrungsrisiko in diesem Punkt bewerten zu können, bedarf es vorab veröffentlichter Informationen zur geplanten Studiendurchführung. Diese können im Rahmen der Registrierung der Studie in einem anerkannten Studienregister und in Form eines Studienprotokolls oder Analyse-Plans bereitgestellt werden (siehe unten). Anschließend ist zu prüfen, ob die vorab definierte Studiendurchführung von der abschließenden Veröffentlichung abweicht.

3.2. Bewertung der Studienqualität der dauerhaft aufgenommenen DIGA

3.2.1. DIGA deprexis

Deprexis ist eine digitale Gesundheitsanwendung für Menschen mit Depressionen oder depressiven Stimmungen. Es ist als onlinebasiertes Selbsthilfeprogramm zur Therapieunterstützung konzipiert und soll in Ergänzung zur medizinischen Behandlung durch Allgemein- oder Fachärzt*innen erfolgen. Die Anwendung basiert auf der kognitiven Verhaltenstherapie (KVT), bei der es darum geht, bestimmte, für den Menschen schädliche Verhaltensweisen zu erkennen und zu verändern.
Die Wirksamkeit wurde in elf randomisiert kontrollierten Studien untersucht [24-34]. Zudem wurde eine systematische Übersichtsarbeit und Meta-Analyse basierend auf den veröffentlichten RCTs publiziert [35].
In der Meta-Analyse zeigte sich eine signifikante Verbesserung der depressiven Symptomatik bei den Nutzer*innen von deprexis verglichen mit den Teilnehmenden der jeweiligen Kontrollgruppe, die in den meisten Fällen auf einer Warteliste standen und die übliche medizinische Versorgung erhielten (I^2 = 26%, 95 % CI: 0.40-0.62, p = <0.001, Effektstärke Hedge's g = 0.25). Eine signifikante Verbesserung in den Studien bedeutet, dass die Ergebnisse auch für einen dahinterliegende Bevölkerung gelten und keine (zufälligen) studienspezifischen Ergebnisse waren. Zudem zeigte sich

in den Studien eine geringe Heterogenität. Das spricht dafür, dass alle bzw. ein Großteil der Studien (tendenziell) dieselben Ergebnisse ausweisen. Die elf randomisiert kontrollierten Studien wurden im Rahmen dieses Gutachtens anhand des RoB2 Tools auf ihre methodische Qualität untersucht (siehe Anhang: RoB 2 Listen deprexis).

Das Verzerrungsrisiko bezogen auf den Prozess der Randomisierung war bei der Mehrheit der Studien gering. In vier Studien wurden jedoch keine Angaben dazu gemacht, ob die Randomisierung verdeckt stattgefunden hat, sodass in diesen Studien Bedenken hinsichtlich möglicher Verzerrung bestehen. Zudem zeigten sich in einer Studie wesentlichere Gruppenunterschiede nach der zufälligen Zuteilung der Teilnehmenden, was ebenfalls ein Hinweis auf eine mögliche Verzerrung in diesem Bereich sein könnte.

In allen elf beurteilten Studien [24-34] bestehen Bedenken hinsichtlich des Verzerrungsrisikos durch Abweichungen von der Anwendung der geplanten Behandlung. Dies liegt darin begründet, dass keine der Studien verblindet war, d.h. die Teilnehmenden hatten Kenntnis von ihrer Gruppenzuteilung (Behandlungs- oder Kontrollgruppe). Gleichwohl fand die Auswertung in allen elf Studien [24-34] entsprechend der Intention-to-Treat Methodik statt, was als angemessene Methode zur Reduzierung des Verzerrungsrisikos in diesem Bereich gilt.

In drei der elf Studien [26, 28, 30] war das Verzerrungsrisiko hinsichtlich einer Verzerrung durch fehlende Werte gering. In sechs der elf Studien [25, 27, 29, 31-33] bestehen aufgrund fehlender Werte Bedenken hinsichtlich eines Verzerrungsrisikos, da keine Maßnahmen zur Reduzierung des Verzerrungsrisikos getroffen wurden. Gleichwohl war der Anteil fehlender Werte in beiden Gruppen (Behandlungs- und Kontrollgruppe) in diesen sechs Studien gleich verteilt. In zwei der elf Studien [24,34] ist das Risiko einer Verzerrung durch fehlende Werte als hoch einzustufen, da es zwischen den Gruppen deutliche Unterschiede im Anteil der fehlenden Werte gibt. Hier kann folglich nicht ausgeschlossen werden, dass der Anteil an fehlenden Werten mit der Behandlung selbst zusammenhängt.

In allen elf Studien [24-34] wurden angemessene und validierte Messinstrumente zur Erfassung des Endpunktes eingesetzt. Zudem gibt es keine Hinweise, dass die Messung zwischen den Gruppen unterschiedliche erfolgt ist. Dennoch muss das Risiko einer Verzerrung hinsichtlich der Messung des Endpunkts in zehn Studien [24-27, 29-34] als hoch eingestuft werden. Dies liegt darin begründet, dass diejenigen, die den Endpunkt tatsächlich messen – in den meisten Fällen waren das die Teilnehmenden selbst im Rahmen von Selbstbewertungsbögen – von der Gruppenzuteilung Kenntnis hatten. Lediglich in einer Studie wurde angegeben, dass diejenigen, die den Endpunkt gemessen haben, verblindet waren, sodass das Verzerrungsrisiko in dieser Studie als gering bewertet werden kann.

In drei der elf Studien [25, 26, 28] kann das Verzerrungsrisikos durch eine Selektion der veröffentlichten Ergebnisse als gering bewertet werden. In diesen drei Studien wurden vorab ein Studienprotokoll veröffentlicht, dass einen Vergleich der Methodik zwischen Studienprotokoll und Studie möglich machte. In allen drei Studien gibt es keine Hinweise darauf, dass die Endpunkte

und/oder statistische Auswertung entsprechend gewünschter Ergebnisse verändert wurden. In den acht weiteren Studien [24, 27, 2,5, 29-34] bestehen Bedenken hinsichtlich des Verzerrungsrisikos in diesem Bereich. Zwar konnte auf Basis der Registrierungen der Studien in einer Datenbank ausgeschlossen werden, dass nur bestimmte Ergebnisse der Endpunktuntersuchung veröffentlicht wurden. Gleichwohl lassen sich keine Rückschlüsse auf Veränderungen in der Auswertungsmethodik ziehen.

3.2.2. DIGA elevida

Die digitale Gesundheitsanwendung *elevida* wurde für Menschen mit multipler Sklerose entwickelt, bei denen ein krankheitsbedingter, dauerhafter Müdigkeits- oder Erschöpfungszustand (Fatigue) vorliegt. Ziel der Anwendung ist es, diesen Erschöpfungszustand zu reduzieren. Die Anwendung basiert im Wesentlichen auf der kognitiven Verhaltenstherapie (KVT). Sie soll in Ergänzung zu einer ärztlichen Behandlung eingesetzt werden. Dabei wird die Anwendung von den Nutzern selbst über einen Zeitraum von 180 Tagen angewendet. Die Wirksamkeit wurde in einer randomisiert kontrollierten Studie untersucht [36].

An der Studie nahmen 275 Personen mit einer Multiplen Sklerose und einer zusätzlichen Fatigue teil, von denen 139 der Behandlungsgruppe (mit elevida) und 136 der Kontrollgruppe zugeordnet wurden. Die Personen in der Kontrollgruppe erhielten die übliche medizinische Versorgung und konnten elevida nach Abschluss der Studie ebenfalls nutzen. Gemessen wurde der physische und mentale Ausprägungsgrad der Fatigue mit einer standardisierten Skala (Chalder Fatigue Scale (CFS) [37, 38], bei dem ein abschließender numerischer Gesamtwert ermittelt wurde. In der Behandlungsgruppe war der Ausprägungsgrad der Fatigue nach 12 Wochen signifikant niedriger als in der Kontrollgruppe (ITT-Analyse: Zwischengruppendifferenz: 2.74 Punkte; 95 % CI: 1.16-4.32; p = 0.0007; Effektstärke Cohen's d = 0.53). Die Unterschiede waren auch nach 24 Wochen nachweisbar (ITT-Analyse: Zwischengruppendifferenz: 2.19 Punkte; 95 % CI: 0.57-3.82; p = 0.0080). Die Studie wurde im Rahmen dieser schriftlichen Stellungnahme anhand des RoB2 Tools auf ihre methodische Qualität untersucht (siehe Anhang: RoB 2 Liste elevida).

In der Studie zeigte sich ein geringes Verzerrungsrisiko hinsichtlich des Prozesses der Randomisierung (siehe Anhang: RoB 2 Liste elevida). Dabei erfolgte die Randomisierung zufällig (1:1 Block-Randomisierung mit 10 Teilnehmenden pro Block) und vollautomatisch mittels eines Computer-Algorithmus, sodass eine verdeckte Zuteilung gewährleistet werden konnte.

Hinsichtlich eines Verzerrungsrisikos durch Abweichungen von der Anwendung der geplanten Behandlung wurde der Intention-to-Treat Effekt beurteilt, was als angemessene Methode zur Vermeidung von Verzerrungen gilt. Gleichwohl bestehen einige Bedenken bezüglich eines Verzerrungsrisikos. Die Bedenken liegen darin begründet, dass die Teilnehmenden im Nachgang der Randomisierung von der Zuordnung zu Behandlungs- bzw. Kontrollgruppe Kenntnis hatten (fehlende

Verblindung). Inwiefern dadurch mögliche Abweichungen entstanden sind, wurde seitens der Autoren nicht berichtet.

Zudem bestehen einige Bedenken hinsichtlich einer Verzerrung durch fehlende Werte. Der Anteil der Studienteilnehmer, die im Laufe der Studie ausgeschieden sind, betrug am Ende der Erhebung (24 Wochen) 19% (n=51). In diesem Fall wird das Verzerrungsrisiko durch den Einsatz verschiedener Sensitivitätsanalysen gemindert.

Das Verzerrungsrisiko hinsichtlich der Messung des Endpunktes muss ebenfalls als hoch eingestuft werden. Wiederum liegt dies in der fehlenden Verblindung begründet, sodass eine Verzerrung der Endpunkterfassung durch das Wissen der Teilnehmenden um ihre Gruppenzugehörigkeit nicht ausgeschlossen werden kann. Hierbei ist festzuhalten, dass die Teilnehmenden den Endpunkt auf Basis eines Online-Fragebogen als Selbstbericht beurteilten.

Hinsichtlich eines Verzerrungsrisikos durch eine Selektion der veröffentlichten Ergebnisse bestehen einige Bedenken. Zwar werden im Rahmen der Registrierung in der Registerdatenbank ISRCTN die Instrumente (inklusive der Quellen) zur Erhebung der primären und sekundären Endpunkte angegeben, sodass ein selektives Berichten der Ergebnisse bestimmter Messinstrumente ausgeschlossen werden kann. Gleichwohl werden die Erhebungszeitpunkte explizit nur für die Messung des primären Endpunkts beschrieben. Zudem gibt es keine Angaben zu der geplanten statistischen Auswertung. Hier wäre die vorab Veröffentlichung eines Studienprotokolls notwendig gewesen, um das Verzerrungsrisiko besser einschätzen zu können.

3.2.3. DIGA somnio

somnio ist eine digitale Gesundheitsanwendung, die sich an Menschen mit Ein- und Durchschlafstörungen (Insomnie) richtet. Sie basiert auf der kognitiven Verhaltenstherapie für Insomnie (KVT-I). Die Nutzer lernen beispielsweise, ihre Schlafzeiten zu optimieren, wie sie mit schlafhindernden Gedanken umgehen oder Entspannungstechniken zum Einschlafen einsetzen können. Die Wirksamkeit wurde in einer randomisiert kontrollierten Studie untersucht [39].
An der Studie nahmen 56 Personen teil, die unter einer Insomnie litten. Von den 56 Personen wurden 29 der Behandlungsgruppe (mit somnio) und 27 Personen der Kontrollgruppe zugeordnet. Die Teilnehmenden in der Kontrollgruppe erhielten die übliche medizinische Versorgung und Zugang zur Nutzung von somnio nach Abschluss der Studie. Die Schwere der Symptomatik wurde mit einem standardisierten Fragebogen (Insomnie Schweregrad Index (ISI) [40] gemessen, bei dem ein numerischer Gesamtwert ermittelt wurde. Bei der Studie zeigte sich, dass somnio die Symptome in der Behandlungsgruppe im Vergleich zu Kontrollgruppe signifikant reduzieren konnte ($p < 0.001$, Effektstärke Cohen´s $d = 1.79$). Die Reduktion blieb auch nach 12 Monaten stabil. Die Studie wurde

im Rahmen dieser schriftlichen Stellungnahme anhand des RoB2 Tools auf ihre methodische Qualität untersucht (siehe Anhang: RoB 2 Liste somnio).

Im Falle der Studie zu somnio wurde eine zufällige und verdeckte 1:1 Randomisierung durch eine unabhängige Person ohne Kontakt zu den Teilnehmenden durchgeführt. Dieses Vorgehen minimiert das Verzerrungsrisiko hinsichtlich des Randomisierungsprozesses. Bei der anschließenden Eingangsuntersuchung gab es jedoch trotz Randomisierung signifikante Unterschiede zwischen der Behandlungs- und der Kontrollgruppe. Dabei zeigten sich Unterschiede bezogen auf die Schwere der Symptomatik der Insomnie, gemessen mit dem Insomnia Severity Index (ISI) [40] sowie den Grad einer Depression, gemessen mittels Beck Depression Inventory revised (BDI-II) [41]. In beiden Fällen waren die Ausgangswerte bei der Behandlungsgruppe signifikant höher als bei der Kontrollgruppe.

Weiterhin war die Studie nicht verblindet, sodass die Teilnehmenden nach der Randomisierung von der Gruppenzugehörigkeit Kenntnis hatten. Entsprechend ist das Verzerrungsrisiko bezogen auf den Intention-to-Treat Effekt als hoch einzustufen.
Zu möglichen Abweichungen von der geplanten Anwendung der Behandlung machten die Autoren keine Angaben. Zudem wurden in der veröffentlichten Studie auch keine Angaben über die Auswertungsmethode (Intention-to-Treat oder Per-Protocol) gemacht, sodass nicht eingeschätzt werden kann, ob hier zusätzliche Verzerrungsrisiken bestehen.

Das Risiko einer Verzerrung durch fehlende Werte ist hingegen als gering einzustufen, da die Rate der Studienteilnehmer, die aus der Studie ausgeschieden sind, lediglich bei 7% lag.
Die Untersuchungsinstrumente können als angemessen bewertet werden, auch gibt es keine Hinweise auf unterschiedliche Messmethoden zwischen den Gruppen. Zudem waren die Interviewer zur Erhebung der Daten hinsichtlich der Gruppenzuteilung der Teilnehmenden verblindet, sodass das Verzerrungsrisiko hinsichtlich der Messung des Outcomes ebenfalls als gering bewertet werden kann.

Hinsichtlich eines Verzerrungsrisikos durch eine Selektion der veröffentlichten Ergebnisse bestehen jedoch Bedenken. Im Rahmen der Registrierung in der Registerdatenbank ClinicalTrials.gov wurden Angaben zu den Erhebungsinstrumenten und den Erhebungszeitpunkten gemacht. Quellen zum Beleg der Validität der Instrumente wurden jedoch nicht angegeben. Auf Basis der Ergebnisse kann ein selektives Berichten bestimmter Ergebnisse ausgeschlossen werden. Nichtsdestotrotz finden sich keine Informationen über die geplante statistische Auswertung, sodass hier ein Verzerrungsrisiko besteht. Zur Beurteilung der Verzerrung in diesem Punkt wäre der Zugriff auf ein vorab veröffentlichtes Studienprotokoll notwendig.

3.2.4. DIGA velibra

Die digitale Gesundheitsanwendung *velibra* wurde für Menschen mit einer generalisierten Angststörung, einer Panikstörung mit oder ohne Agoraphobie oder einer sozialen Angststörung entwickelt. Die Anwendung basiert auf der kognitiven Verhaltenstherapie (KVT) und soll in Ergänzung zu einer ärztlichen Behandlung eingesetzt werden. Die Anwendung wird von den Nutzern selbst über einen Zeitraum von 180 Tagen angewendet. Die Wirksamkeit wurde in einer randomisiert kontrollierten Studie untersucht [42].

An der Studie zu velibra nahmen 139 Personen mit einer generalisierten Angststörung, einer Panikstörung mit oder ohne Agoraphobie oder einer sozialen Angststörung teil. 70 der Teilnehmenden wurden der Behandlungsgruppe (übliche medizinische Versorgung + velibra), 69 der Teilnehmenden wurden der Kontrollgruppe zugeordnet. Die Kontrollgruppe erhielt ausschließlich die übliche medizinische Behandlung und konnte velibra nach Abschluss der Studie nutzen. Gemessen wurden eine Reduktion der Angstsymptomatik und depressiver Beschwerden mittels folgender standardisierter Fragebögen: Beck Depression Inventory-II (BDI-II) [41]; Depression Anxiety Stress Scales – Short Form (DASS-21) [43]; Beck Anxiety Inventory (BAI) [44]; Brief Symptom Inventory (BSI) [45] und Short-Form Health Survey-12 (SF-12) [46].

In der Behandlungsgruppe konnte nach 9 Wochen eine signifikante Verringerung der Angst- und depressiven Symptome sowie signifikant bessere Werte bezogen auf die Lebensqualität im Vergleich zur Kontrollgruppe festgestellt werden. Die Therapieeffekte waren auch nach 6 Monaten nachweisbar. Die Studie wurde im Rahmen dieser schriftlichen Stellungnahme anhand des RoB2 Tools auf ihre methodische Qualität untersucht (siehe Anhang: RoB 2 Liste velibra).

Bezogen auf die Studie zu velibra kann das Risiko einer Verzerrung im Rahmen des Randomisierungsprozesses als gering eingestuft werden. Es wurde eine stratifizierte Randomisierung unter Berücksichtigung der Diagnose, der medikamentösen sowie der psychotherapeutischen Behandlung vorgenommen. Eine stratifizierte Randomisierung bedeutet, dass die Teilnehmenden anhand bekannter Störgrößen vor der Randomisierung in verschiedene Gruppen (Strata) eingeordnet werden. Bei der Randomisierung wird anschließend darauf geachtet, dass die Behandlungs- und die Kontrollgruppe hinsichtlich der Diagnose sowie der medikamentösen und psychotherapeutischen Behandlung ausgewogen besetzt sind. Die Randomisierung wurde verdeckt mittels eines computerbasierten Generators durchgeführt. Nach der Randomisierung wurden keine signifikanten Unterschiede bei den Ausgangswerten zwischen den Gruppen festgestellt.

Die hier eingesetzte Auswertungsmethodik entsprechend Intention-to-Treat gilt als angemessen, um das Verzerrungsrisiko gering zu halten. Gleichwohl bestehen Bedenken hinsichtlich möglicher Abweichungen von der Anwendung der geplanten Behandlung durch eine fehlende Verblindung der Teilnehmenden.

Die Drop-Out-Rate in der Studie betrug 14% zum Ende der Studie, sodass hinsichtlich einer Verzerrung durch fehlende Werte Bedenken geäußert werden müssen.

Das Risiko einer Verzerrung bezogen auf die Messung des Outcomes muss als hoch eingestuft werden, da sowohl die Teilnehmenden als auch die Interviewer, die den Endpunkt erheben, nicht verblindet werden konnten.

Im Falle der Studie zu velibra war kein vorab veröffentlichtes Studienprotokoll verfügbar. Informationen zur den vorab definierten Erhebungsinstrumenten und Erhebungszeitpunkten finden sich jedoch bei der Registrierung in der Registerdatenbank ISRCTN. Angaben zu den Quellen der einzelnen Instrumente wurden nicht gemacht. Dennoch kann ein selektives Berichten der Ergebnisse bestimmter Messinstrumente ausgeschlossen werden. Im Gegensatz dazu bestehen Bedenken hinsichtlich der selektiven Berichterstattung von Ergebnissen bestimmter Auswertungsmethoden. Zur statistischen Auswertung finden sich keine vorab veröffentlichten Angaben.

3.2.5. DIGA vorvida

*Vorvida i*st eine digitale Gesundheitsanwendung für Patienten mit schädlichem Alkoholkonsum oder Alkoholabhängigkeit. Basierend auf der kognitiven Verhaltenstherapie (KVT) und weiteren psychotherapeutischen Ansätzen und Verfahren sollen Nutzer beim Management ihres gesundheitsschädlichen Trinkverhaltens und der Reduzierung ihrer Trinkmenge unterstützt werden. Die Anwendung soll in Ergänzung zu einer ärztlichen Behandlung eingesetzt und von dem Nutzer selbst über einen Zeitraum von 180 Tagen angewendet werden. Die Wirksamkeit wurde in einer randomisiert kontrollierten Studie untersucht [47].

An der Studie beteiligten sich 608 Personen mit problematischem Alkoholkonsum, von denen 306 der Behandlungsgruppe (mit vorvida) und 302 der Kontrollgruppe zugeordnet wurden. Die Personen in der Kontrollgruppe erhielten die übliche medizinische Versorgung und Zugang zur Nutzung von vorvida nach Abschluss der Studie. Der tägliche durchschnittliche Alkoholkonsum wurde mittels standardisierter Selbstbeurteilungsfragebögen (Quantity-Frequency-Index (Mengenangaben in Alkohol in g) in den letzten 30 Tagen und Timeline-Follow-Back (Mengenangaben in Alkohol in g) in den letzten 7 Tagen [48-50] gemessen. Der Alkoholkonsum in der Behandlungsgruppe reduzierte sich im Vergleich zu der Kontrollgruppe innerhalb von drei und sechs Monaten signifikant. Die Studie wurde im Rahmen dieser schriftlichen Stellungnahme anhand des RoB2 Tools auf ihre methodische Qualität untersucht (siehe Anhang: RoB 2 Liste vorvida).

Die Studie zu vorvida weist ein niedriges Verzerrungsrisiko bezogen auf den Prozess der Randomisierung auf. Dabei fand eine 1:1 Randomisierung statt. Die verdeckte Zuteilung wurde durch den Einsatz eines zentralisierten, software- und computerbasierten Randomisierungsprozesses

gewährleistet. Zudem konnten keine signifikanten Gruppenunterschiede nach der Eingangsbeurteilung festgestellt werden.

Die Studie selbst war nicht verblindet, sodass die Teilnehmenden von ihrer Gruppenzuordnung Kenntnis hatten. Gleichwohl fand die Auswertung entsprechend Intention-to-Treat statt, was das Verzerrungsrisiko hinsichtlich möglicher Abweichungen von der Anwendung der beabsichtigten Behandlung verringert. Insgesamt bestehen in diesem Punkt auf Grund der fehlenden Verblindung dennoch einige Bedenken bezüglich einer möglichen Verzerrung.
Die Rate der ausgeschiedenen Studienteilnehmer von 30% birgt zudem das Risiko einer Verzerrung durch fehlende Werte. Sensitivitätsanalysen wurden hier eingesetzt, um das Risiko in diesem Punkt zu minimieren.

Das Risiko einer Verzerrung hinsichtlich der Messung des Endpunktes ist als hoch einzustufen. Dies liegt darin begründet, dass die Teilnehmenden, die den Endpunkt auf Basis von Online-Fragebögen erhoben, von ihrer Gruppenzugehörigkeit wussten.
Positiv ist hervorzuheben, dass ein im Vorfeld veröffentlichtes Studienprotokoll verfügbar war. Auf Basis des Studienprotokolls konnten Abweichungen hinsichtlich einer selektiven Veröffentlichung von Ergebnissen ausgeschlossen werden. Lediglich anzumerken ist, dass einige der im Studienprotokoll angefügten sekundären Outcomes im Rahmen der Durchführung der Studien nicht erhoben oder nicht berichtet wurden.

3.2.6. DIGA ohne wissenschaftliche Belege zur Wirksamkeit (Stand: 12.07.2021)

Vierzehn digitale Gesundheitsanwendungen wurden vorläufig in das DiGA-Verzeichnis aufgenommen (Stand 12.07.2021). Eine vorläufige Aufnahme ist möglich, sofern der Hersteller noch keine Studie zum Nachweis eines „positiven Versorgungseffekts“ vorlegen kann. Ein positiver Versorgungseffekt kann entweder in Form eines medizinischen Nutzens oder einer patientenrelevanten Verfahrens- und Strukturverbesserungen nachgewiesen werden. Sofern ein Antrag auf eine vorläufige Aufnahme in das DiGA-Verzeichnis gestellt wird, müssen bereits bei Antragstellung alle Anforderungen hinsichtlich Sicherheit, Funktionstauglichkeit, Qualität, Datenschutz und Informationssicherheit erfüllt sein. Nach Antragstellung hat der Hersteller 12 Monate Zeit, einen positiven Versorgungseffekt nachzuweisen. In dieser Zeit ist die Anwendung zu einem vom Hersteller festgelegten Preis flächendeckend erstattungsfähig, allerdings können Höchstpreise für bestimmte Gruppen von DiGAs festgelegt werden. Anschließend entscheidet das Bundesinstitut für Arzneimittel und Medizinprodukte (BfArM) über eine dauerhafte Aufnahme der Anwendung in das DiGA-Verzeichnis [51].Unabhängig von der dauerhaften Aufnahme in das Verzeichnis können digitale Anwendungen somit für einen Zeitraum von 12 Monaten voll erstattungsfähig sein – und dass, ohne bis dato den Nachweis eines Nutzens erbringen zu müssen.

Aktuell wurden die 14 Gesundheitsanwendungen CANKADO (Brustkrebs), ESYSTA (Diabetes), Invirto (Angststörungen), Kalmeda (Tinnitus), M-sense Migräne (Migräne), Mika (Krebs im Mund-Zungenbereich), Mindable (Angststörungen), NichtraucherHelden-App (Suchterkrankungen), Rehappy (Schlaganfall), Selfapy (Depression), Selfapy (Angststörung), Selfapy (Panikstörung), Vivira (Arthrose) und Zanadio (Adipositas) ohne vorhandene Evidenz vorläufig in das DiGA-Verzeichnis aufgenommen. Eine Einschätzung des Nutzens sowie der methodischen Vorgehensweise und des Verzerrungsrisikos sind somit nicht möglich. Bei der digitalen Gesundheitsanwendung Selfapy Online-Kurs bei Depression liegen bisher lediglich vorläufige Daten auf der Homepage des Herstellers vor. Die Studie selbst wurde bisher nicht veröffentlicht, sodass ebenfalls keine Nachvollziehbarkeit der Methodik und Ergebnisse möglich ist.

3.3. Fazit

Gegenstand der Untersuchung in Abschnitt 3. waren die dauerhaft in das DIGA-Verzeichnis aufgenommen DIGA, für die wissenschaftlichen Belege vorlagen (Stand 12.07.2021). Das Verzerrungsrisiko der Studien zu den hier aufgeführten digitalen Gesundheitsanwendungen (DiGA) wurde mit dem *Revised Cochrane risk-of-bias tool for randomized trials (RoB 2),* Version of 22 August 2019 bewertet [23]. Die Bewertungsergebnisse der insgesamt fünf digitalen Gesundheitsanwendungen zeigten bei vier der fünf untersuchten DIGA ein beträchtliches Verzerrungspotential. Dies insbesondere deshalb, da eine Überprüfung der Ergebnisse anhand eines vorab veröffentlichen Studienprotokolls oder Analyseplans nicht möglich war. Um eine angemessene Bewertung des Verzerrungsrisikos gewährleisten zu können, sollte neben der Registrierung demnach auch die Publikation und öffentliche Bereitstellung eines Studienprotokolls oder Analyseplans vor der Durchführung der Studie verbindlich sein. Dies beinhaltet insbesondere auch die Erklärung, nach welcher Auswertungsmethodik (Intention-to-Treat oder Per-Protocol) die abschließenden Auswertungen vorgenommen werden sollen.
Im Sinne der Transparenz und Nachvollziehbarkeit schreibt die Digitale-Gesundheitsanwendungen-Verordnung (DiGAV) die Registrierung in einem deutschem und zusätzlich in einem internationalen Studienregister vor Studiendurchführung verbindlich vor. Digitale Anwendungen, die in das DiGA-Verzeichnis aufgenommen werden möchten, sollten ihre Studien prioritär im Deutschen Register Klinischer Studien (DRKS) registrieren. Für internationale Registrierungsmöglichkeiten bieten sich die international anerkannten Register ClinicalTrials.gov oder ISRCTN an.

Gleichwohl sind die Angaben, die bei der Registrierung der DIGA in einem klinischen Register gemacht werden, zur Bewertung des Verzerrungsrisikos nicht ausreichend. Im Vergleich mit der international anerkannten SPIRIT-Checkliste zur Erstellung von Studienprotokollen von RCTs [58, 59] zeigte sich, dass die Angaben der drei zugelassen DIGA elevida, somnio und velibra deutlich unter den geltenden internationalen Anforderungskriterien liegen (Tabelle 2).

Tabelle 2: Vergleich der Registrierungsdatenbanken der DIGA elevida, somnio uns velibra und der SPIRIT-Checkliste zur Erstellung von Studienprotokollen

Kategorien der SPIRIT-Checkliste		Angaben aus der Registrierung in der Registerdatenbank		
		elevida	somnio	velibra
Title	1			
Trial registration	2a			
	2b	Separate Checkliste WHO	Separate Checkliste WHO	Separate Checkliste WHO
Protocol version	3			
Funding	4			
Roles and responsibilities	5a	Nur Kontaktdaten eines Ansprechpartners	Nur Name der Universität	Nur Kontaktdaten eines Ansprechpartners
	5b		Nur Name	
	5c			
	5d			
Background and rationale	6a			
	6b			
Objectives	7			
Trial design	8	Nur Studientyp, keine Randomisierung	Nur Studientyp, keine Randomisierung	Nur Studientyp, keine Randomisierung
Study setting	9			
Eligibility criteria	10	Nur Einschlusskriterien		Nur Einschlusskriterien
Interventions	11a			
	11b			
	11c			
	11d			
Outcomes	12	Nur Outcome, Instrument und Zeitpunkt	Nur Outcome, Instrument und Zeitpunkt	Nur Outcome, Instrument und Zeitpunkt
Participant timeline	13			
Sample size	14	Nur Anzahl	Nur Anzahl	Nur Anzahl
Recruitment	15			
Sequence generation	16a			
Allocation concealment mechanism	16b			
Implementation	16c			
Blinding (masking)	17a			
	17b			
Data collection methods	18a	Nur Instrumente und Quellen, Erhebungs-zeitpunkte bei primärem Outcome	Nur Instrumente und Erhebungs-zeitpunkte	Nur Instrumente und Erhebungs-zeitpunkte
	18b			
Data management	19			
Statistical methods	20a			
	20b			
	20c			
Data monitoring	21a			
	21b			
Harms	22			
Auditing	23			
Research ethics approval	24			
Protocol amendments	25			
Consent or assent	26a			
	26b			
Confidentiality	27			
Declaration of interests	28			
Access to data	29			
Ancillary and post-trial care	30	Nicht anwendbar	Nicht anwendbar	Nicht anwendbar
Dissemination policy	31a			
	31b			
	31c			
Informed consent materials	32	Nur auf Anfrage		Nur auf Anfrage
Biological specimens	33			

Grün = Übereinstimmung zwischen Registrierungsangaben und SPIRIT-Checkliste
Gelb = Teilweise Übereinstimmung zwischen Registrierungsangaben und SPIRIT-Checkliste
Rot = Keine Übereinstimmung zwischen Registrierungsangaben und SPIRIT-Checkliste

Zudem zeigte sich in allen Studien zu den digitalen Gesundheitsanwendungen ein Verzerrungsrisiko durch eine fehlende Verblindung der Teilnehmenden und Prüfer. Dies lag darin begründet, dass die Kontrollgruppe die übliche medizinische Behandlung erhielt und die Gruppenzuteilung dahingehend nicht geheim gehalten werden konnte. Eine Verblindung aller beteiligten ist dann möglich, wenn sich ohne weiteres nicht unterscheiden lässt, welche Art der Behandlung die Teilnehmenden erhalten. Dies ist beispielsweise häufig bei Arzneimittelstudien der Fall, bei denen jede teilnehmende Person eine optisch identisch aussehende Tablette erhält, bei der es sich entweder um das Arzneimittel oder ein Placebo handeln kann.

Auch im Falle digitaler Gesundheitsanwendungen ist eine Verblindung möglich. In einer Studie von Fiatarone et al. [60] wurde die Behandlung (computerbasiertes kognitives Training) in der Behandlungsgruppe mit einer alternativen Behandlung (Anschauen von 5 National Geographic Videos mit anschließenden Gedächtnisfragen) in der Kontrollgruppe verglichen. Man spricht in so einem Fall von einer *Sham-Behandlung* oder *Sham-Intervention*. Dadurch kann eine Verblindung der Studie gewährleistet werden. Hier ist deshalb zu fordern, dass zukünftige Studien zur Wirksamkeit von DIGA die Behandlung mit der digitalen Gesundheitsanwendung mit einer digitalen Sham-Behandlung vergleichen.

4. Neue Studiendesigns zur Bewertung digitaler Gesundheitsanwendungen

Wie zuvor dargestellt, sind die Wahl des Studiendesign und der Auswertungsmethoden maßgeblich für die Höhe des Verzerrungspotentials und damit für die Aussage zur Wirksamkeit der DIGA. Das BfArM empfiehlt in seinem Leitfaden für die Bewertung von DIGA weitere, sogenannte „alternative Studiendesigns“ [4] . Dazu führt das BfArM auf Seite 104 folgendes aus: „*Neben den zuvor genannten Studiendesigns können in Abhängigkeit des Versorgungskontexts der DiGA sowie der angestrebten Nachweise auch andere alternative Studiendesigns und -methoden wie beispielsweise Pragmatic Clinical Trials (PCT), Sequential Multiple Assignment Randomized Trial (SMART) oder Multiphase Optimization Strategy (MOST) sinnvoll sein. Auch der Einbezug weiterer Datenquellen im Sinne von Real-World-Data kann im Beleg der pVE nutzbringend sein.*“ (Seite 104, ‚Das Fast Track Verfahren für digitale Gesundheitsanwendungen (DiGA) nach § 139e SGB V‘).

Die vom BfArM empfohlenen „alternativen Studiendesigns“ werden hier im Detail beschrieben und in Hinblick auf die international geltenden wissenschaftlichen Standards des Health-Technology Assessments bewertet.

4.1. Continuous Evaluation of Evolving Behavioral Intervention Technologies (CEEBIT)

Die Methode *Continuous Evaluation of Evolving Behavioral Intervention Technologies‘ (CEEBIT)* wurde erstmalig 2013 von Mohr et al. berichtet [52]. Die Methode fokussiert sich auf die kontinuierliche Bewertung von webbasierten und mobilen Anwendungen innerhalb eines Systems (bspw. einer Klinik), die Nutzern bei der Änderung von (gesundheitsbezogenen) Verhaltensweisen unterstützen sollen (sogenannte Behavioral Intervention Technologies (BITs)). Dabei folgt die Methode einem mehrstufigen Verfahren, bei dem verschiedene Anwendungen oder im Zeitverlauf angepasste Anwendungen in Echtzeit miteinander verglichen werden.

Wird beispielsweise bei einer digitalen Anwendung eine Anpassung vorgenommen, zum Beispiel in Form eines Updates, oder wird eine weitere Anwendung eingeführt, werden die verschiedenen Versionen bzw. Anwendungen miteinander verglichen. Die bei diesem Vergleich schlechter bewertete Version wird anschließend verworfen. Dies gelingt, indem kontinuierlich Nutzungsdaten erhoben werden. Diese Daten können sich auf vordefinierte gesundheitsbezogene Endpunkte, wie Verringerung des Gewichts oder depressiver Symptome, oder auch Nutzungswerte, wie die Dauer oder Häufigkeit der Nutzung, beziehen. Für den Vergleich der Anwendungen können Nutzer, ähnlich wie bei RCTs, im Rahmen der Datenerhebung zufällig einer Gruppe zugeordnet werden. Allerdings gilt dies nur für solche Nutzer, die keine Präferenz für die eine oder andere Anwendung äußern. Sofern Nutzer lieber eine der zu vergleichenden Anwendungen nutzen wollen, können sie der entsprechenden Gruppe zugeordnet werden [52]. In diesem Fall handelt es sich jedoch um keine Randomisierung. Generell werden auch andere nicht-randomisierte Zuteilungsmethoden

vorgeschlagen. Da sich die digitalen Systeme und Anwendungen stetig weiterentwickeln können, ist ein zeitlicher Endpunkt für die Bewertung nicht vorgesehen [52].

Als Vorteil dieser Methodik gilt die dynamische Optimierung von Nutzerpräferenzen und Wirksamkeit und somit auch des Nutzens. Dadurch soll die Qualität der Anwendungen verbessert werden und Nutzern nur Zugang zu den effektivsten Anwendungen bekommen. Zudem gelten eine kurze Entwicklungs- und Untersuchungszeit sowie niedrige Entwicklungskosten als Vorteil [53]. Die letztgenannten Punkte sind jedoch in erster Linie für die Entwickler der Anwendungen von Vorteil, da auf diese Weise Zeit und Kosten in der Produktentwicklung gespart werden können.

Dem stehen jedoch gravierende methodische Problematiken entgegen [54].
Als Signifikanzschwelle wird ein Wert von 0,5 vorgeschlagen, womit die Fehlertoleranz bei 50% liegt. Die Signifikanz beschreibt die Wahrscheinlichkeit, dass man von den Ergebnissen der Studie auf die gesamte Bevölkerung schließen kann. Je höher der Signifikanzwert ist, desto höher ist die Wahrscheinlichkeit, dass sich die Ergebnisse nur auf die Personen aus der Studie beziehen und keine Rückschlüsse auf die Bevölkerung möglich sind. Bei der genannten Methode wäre selbst eine Wahrscheinlichkeit von fast 50%, dass die Ergebnisse nur für die Personen aus der Studie gelten, noch akzeptabel. Zum Vergleich: in der Regel gilt eine Wahrscheinlichkeit von 5% als akzeptabel, alle darüber liegenden Werte gelten als nicht signifikant. Ein Effekt dieser hohen Fehlertoleranz liegt darin, dass dadurch die Anzahl an Personen in der Studie wesentlich geringer sein kann. Der Signifikanzwert hängt wesentlich mit der Anzahl an Personen in einer Studie zusammen. Je höher die Personenanzahl, desto niedriger wird das Ergebnis der Signifikanzprüfung ausfallen. Für die Entwickler hat dies den Vorteil, dass sie durch die niedrigere Personenanzahl Zeit und Geld sparen. Dies geht jedoch mit einem erheblichen Verzerrungspotenzial und Unsicherheiten im Hinblick auf die festgestellten Behandlungseffekte einher [54].
Eine weitere Problematik sind die sich ständig verändernden Personenkreise durch die kontinuierliche Bewertung der Anwendungen. Üblicherweise wird in wissenschaftlichen Studien ein fester Personenkreis, die sogenannte Studienpopulation, über einen festen Zeitraum untersucht. Wie zuvor beschrieben wird für gesundheitspolitische Entscheidungen der Intention-to-Treat Effekt bevorzugt, da die gleichmäßige Verteilung zufälliger Störgrößen, also beispielsweise Alter, Geschlecht, etc., auf beide Gruppen erhalten bleibt. Bei sich verändernden Studienpopulationen ist diese Untersuchung nicht möglich, auch, wenn die Personen jedes Mal zufällig zugeteilt werden. Zudem wurde die Methode bisher sehr selten genutzt.

4.2. Multiphase Optimization Strategy (MOST)

Die Methode *Multiphase Optimization Strategy (MOST)* wurde erstmalig 2007 von Collins et al. vorgestellt [55]. Bei MOST geht es um die Entwicklung, Optimierung und Bewertung digitaler gesundheitsbezogener Behandlungsmethoden. Die Methode basiert auf einem dreistufigen

Verfahren von der Entwicklung bis zur Wirksamkeitsprüfung mit den Phasen Vorauswahl (screening), Weiterentwicklung (refinement) und Auswertung (confirmation).

In der ersten Phase geht es im Rahmen der Entwicklung der Behandlung darum zu entscheiden, welche Komponenten die Behandlung beinhalten und welche verworfen werden sollen. Auf Basis dieser Auswahl an Behandlungskomponenten werden im zweiten Schritt die optimale Dosis bzw. Intensität der einzelnen Komponenten ermittelt. In der Auswertungsphase wird die Wirksamkeit der Anwendung schließlich anhand einer RCT untersucht [55]. Dieses Vorgehen soll anhand eines Beispiels dargestellt werden: einer digitalen Anwendung zur Beendigung des Rauchens durch regelmäßige Nachrichten. Im Rahmen der ersten Phase werden vier positive Komponenten in der Entwicklung der Anwendung (bzw. der Nachrichten) identifiziert. Dabei handelt es sich um die Erwartungen der Nutzer, was nach Beendigung des Rauchens passiert; Hindernisse für das Durchhalten bei der Rauchentwöhnung; Erfahrungsberichte, sog. Testimonials, von Rauchern, die aufgehört haben; sowie die Frage, ob einzelne längere oder mehrere kürzere Nachrichten eingesetzt werden sollen. Als unwichtig wurden zwei Komponenten erachtet: das Framing der Nachrichten, also ob diese positiv oder negativ formuliert werden sollen, sowie die Quelle der Nachrichten, also ob sie bspw. von einem Arzt oder einer Gesundheitsorganisation bereitgestellt werden sollen. In der zweiten Phase wird dann bspw. bewertet, was die optimale Anzahl an Nachrichten ist, die den Nutzern zur Verfügung gestellt werden. Die im Ergebnis entwickelte Anwendung wird schließlich in einer RCT bspw. mit der Standardtherapie zur Rauchentwöhnung verglichen. Die Bewertungen und Entscheidungen innerhalb der ersten beiden Phasen findet auf Basis des sog. „randomisierten Experimentierens“ [55] statt. Dies könnte beispielweise in Form eines Vergleichs verschiedener Gruppen stattfinden, die alle eine unterschiedliche Anzahl an Nachrichten bekommen. Für diese randomisierten Experimente werden bestimmte Schwellenwerte für die Signifikanz (Fehlertoleranz von 50%) oder den Effekt festgelegt [55].

Im Rahmen der Entwicklung der Anwendungen kann aufgrund der höheren Fehlertoleranz oder der Festlegung großzügiger Schwellenwerte für die Effekte eine zeit- und kostensparende Entwicklung der Anwendung erfolgen [53]. Dies geht jedoch in zweierlei Hinsicht zulasten der methodischen Qualität [54]. Zum einen können sich zu niedrig angelegte Schwellenwerte für die Effekte auf die Qualität der einzelnen Anwendungskomponenten auswirken. Zum anderen lässt eine hohe Fehlertoleranz wiederum Zweifel daran, dass die Anwendung auch für die Bevölkerung wirksam und hilfreich ist, für die sie eingesetzt werden soll. Somit kann die Anwendung von MOST an sich keinen methodisch einwandfreien Wirksamkeitsnachweis der digitalen Anwendung hervorbringen. Dieses Fehlen eines Wirksamkeitsnachweises war den Autoren bewusst, sodass sie nach Abschluss der ersten beiden Phasen zur Bewertung der Wirksamkeit eine klassische RCT empfehlen, die höchsten methodischen Standards entspricht [55]. Dadurch kann der gesamte Zyklus der MOST-Methodik allerdings sehr lang werden, was wiederum die Zeit- und Kosteneinsparungen in den ersten Phasen konterkariert.

4.3. Sequential Multiple Assignment Randomized Trial (SMART)

Die Methode *Sequential Multiple Assignment Randomized Trial (SMART)* wurde erstmalig 2007 von Collins et al. berichtet [55]. Bei SMART handelt es sich um eine Methode zur Entwicklung und Optimierung sich stetig anpassender (digitaler) Anwendungen. Dier Methodik kann in den MOST-Prozess integriert werden. Es handelt sich um ein mehrstufiges Verfahren mit dem Ziel, die optimale Abfolge von Behandlungskomponenten zu identifizieren. Dazu finden in mehreren Schritten Gruppenvergleiche mit einer zufälligen Zuteilung der Personen statt (Randomisierung), auf deren Basis Entscheidungen für oder gegen bestimmte Behandlungsabfolgen getroffen werden. Die Entscheidungen erfolgen auf Basis vordefinierter Endpunkte. Die Personengruppen werden dabei in Responder und Non-Responder unterschieden, wobei Responder einen Behandlungserfolg charakterisieren und Non-Responder einen Misserfolg. Wichtig ist dabei, dass nur bei den Non-Responder in den folgenden Schritten eine Anpassung der Behandlungsfolge stattfindet oder diese zufällig anderen Behandlungsabfolgen zugeordnet werden. Responder fahren so lange mit der Behandlung fort, bis der Behandlungserfolg nachlässt [55, 56]. Die Methode soll wiederum anhand der zuvor genannten App zur Rauchentwöhnung beispielhaft dargestellt werden.

Im Rahmen der Entwicklung der App sollen verschiedene Behandlungsbestandteile und deren optimale Abfolge untersucht werden. Bewertet wird die App dahingehend, ob Nutzer nach der Nutzung mit dem Rauchen aufhören oder den Zigarettenkonsum reduzieren. In einem ersten Schritt werden nun zwei Gruppen verglichen, bei der eine Gruppe positive und eine Gruppe negativ formulierte Nachrichten zum Rauchverhalten bekommt. Die untersuchten Personen werden den Gruppen zufällig zugeordnet. Nach dieser ersten Untersuchung wird verglichen, in welcher der beiden Gruppen der Zigarettenkonsum stärker reduziert wurde, in dem Beispiel sei es die Gruppe mit den positiv formulierten Nachrichten. In einem zweiten Schritt erhält die Gruppe mit den negativen formulierten Nachrichten nun zusätzlich Erfahrungsberichte, sog. Testimonials, von Personen, die mit dem Rauchen aufgehört haben, währen die Gruppe mit den positiv formulierten Nachrichten mit der Behandlung fortfährt. Wiederum wird anhand des Behandlungserfolgs (Reduzierung des Zigarettenkonsums) die wirksamste Variante identifiziert. Auf diese Weise können viele verschiedene Abfolgen von Behandlungskomponenten untersucht werden. Abschließend kann die Behandlung entsprechend der optimalen Abfolge finalisiert werden. Im Rahmen dieser Methode können sowohl Daten aus der klinischen Forschung als auch der Versorgungsforschung einbezogen werden [55].

Zur Bewertung der Gruppenvergleiche in den einzelnen Phasen wird eine Fehlertoleranz von 50% vorgeschlagen. Auf diese Weise lässt sich die Methode zeit- und kostensparend mit geringeren Personenanzahlen durchführen [53]. Gleichwohl wird dadurch die statistische Aussagekraft geringer und somit steigt das Verzerrungspotenzial und Unsicherheit im Hinblick auf die festgestellten Behandlungseffekte an [54]. Problematisch ist zudem das Verfahren der Randomisierung in den einzelnen Phasen. Da die Randomisierung auf dem Behandlungserfolg beruht, werden nicht alle

Personen zu jedem Zeitpunkt zufällig einer Gruppe zugeordnet. Auch dies kann zu Verzerrungen hinsichtlich der Bewertung des Behandlungseffekts führen [54]. Aufgrund der methodischen Schwächen schlagen die Autoren nach erfolgreicher Entwicklung der Intervention die Durchführung einer randomisiert kontrollierten Studie zur Überprüfung der Wirksamkeit vor [55, 56].

4.4. Micro-Randomized Trials (MRT)

Die Methode *Micro-Randomized Trials (MRT)* wurde erstmalig 2015 von Klasnja et al. beschrieben [57]. Die MRT-Methode wurde mit dem Ziel entwickelt, die Optimierung von sogenannten Just-in-time adaptive interventions (JITAIs) zu unterstützen. Dabei handelt es sich um digitale Gesundheitsanwendungen, bei denen es darum geht, die richtigen Behandlungskomponenten zur richtigen Zeit einzusetzen, um Nutzer in ihrem Gesundheitsverhalten zu unterstützen. Im Rahmen der Entwicklung der digitalen Anwendung werden für die einzelnen Anwendungskomponenten randomisierte Experimente durchgeführt, um deren Wirksamkeit zu beurteilen [35]. Dies soll am Beispiel einer digitalen Anwendung zur physischen Aktivierung illustriert werden. Die Anwendung beinhaltet zwei Komponenten: eine Planung der täglichen Aktivitäten sowie kontextbezogene Empfehlungen für eine physische Aktivierung (bspw. an der Arbeit, beim Einkaufen, nach dem Abendessen, etc.). Zudem erfasst die Anwendung Bewegungsdaten. Für die Untersuchung der Wirksamkeit wird nun für jede der beiden Komponenten ein randomisiertes Experiment durchgeführt. Dabei werden Nutzer zufällig einer Gruppe zugeordnet. Bei der ersten Komponente erhält die eine Gruppe den Hinweis, einen Plan für die täglichen Aktivitäten zu erstellen und die andere Gruppe erhält keine Hinweise. Bezogen auf die zweite Komponente erhält eine Gruppe über den Tag verteilt Empfehlungen zur physischen Aktivierung und die andere Gruppe nicht. Auch eine unterschiedliche Anzahl an Empfehlungen pro Gruppe oder zu Empfehlungen zu unterschiedlichen Zeitpunkten könnten untersucht werden. Auf diese Weise können beliebig viele Bestandteile einer Anwendung oder Behandlung untersucht werden. Diese werden dann nach vorab festgelegten Endpunkten, bspw. der Anzahl an Schritten, verglichen. Durch die vielen Mikro-RCTs können auch bei kurzen Studienzeiträumen und einer geringen Personenanzahl viele Beobachtungspunkte gewonnen werden, zudem können auch zeitliche Variationen in den Effekten festgestellt werden [57].

Positiv hervorzuheben ist, dass die MRT Methode eine in wissenschaftlichen Studien übliche Fehlertoleranz von 5% angibt. Damit unterscheidet sich die Methode von den anderen genannten Verfahren wesentlich. Durch die zahlreichen Mikro RCTs kann die Methode dennoch mit niedrigeren Personenzahlen durchgeführt werden. Dies geht jedoch trotz der niedrigen Fehlertoleranz zu Lasten der methodischen Qualität [54]. Beispielsweise könnte sich bei Betrachtung ein und derselben Person im Zeitablauf vor allem die zeitliche Variation nicht beobachteter patientenindividueller Charakteristika verzerrend auf die Ergebnisse auswirken. Außerdem könnten beobachtete Behandlungseffekte im Zeitverlauf durch die Auswahl von Behandlungskomponenten vorangegebener Mikro RCTs verzerrt werden. Zudem bleibt fraglich, inwieweit die geringere Anzahl

an Studienteilnehmern es möglich macht, Ergebnisse zu erzielen, die repräsentativ für die dahinterliegende Bevölkerung sind [54]. Darüber hinaus berichten die Autoren über Einschränkungen des Einsatzes bei bestimmten medizinischen Fragestellungen. Als konkretes Beispiel geben die Autoren an, dass beim Vorkommen seltener Ereignisse, wie im Falle manischer Episoden bei bipolaren Störungen, die Anwendung der Methodik ungeeignet sei [57].

4.5. Fazit

Für den Nachweis des positiven Versorgungseffekts akzeptiert das BfArM beobachtende analytische Studien, experimentelle Interventionsstudien sowie Meta-Analysen. Neben diesen bekannten Studiendesigns werden unter bestimmten Voraussetzungen jedoch auch die neuartigen wissenschaftlichen Methoden MOST und SMART anerkannt [4]. Hierbei ist festzuhalten, dass diese genannten Methoden in erster Linie auf die Entwicklung und Weiterentwicklung/Optimierung von (digitalen) Anwendungen abzielen. Die Bewertung ihrer Wirksamkeit und ihres Nutzens steht dabei im Hintergrund [54]. Aufgrund der zuvor diskutierten methodischen Charakteristika, wie einer höheren Fehlertoleranz, spezieller Maßnahmen der Randomisierung oder sich verändernden Studienpopulationen, gehen die Methoden mit schwerwiegenden Verzerrungspotenzialen und Unsicherheiten im Hinblick auf die Behandlungseffekte einher und sollten demnach nicht empfohlen werden.

Zudem sehen alle dargestellten Methoden die Verwendung von Versorgungsdaten aus dem Alltag (sogenannte Real World Data) vor. Diese gehen, verglichen mit Studiendaten in experimentellen Designs, die für RCTs verwendet werden, ebenfalls mit einem höheren Verzerrungsrisiko einher. Im Falle von MOST und SMART wird als abschließende Wirksamkeitsprüfung eine klassische randomisiert kontrollierte Studie empfohlen. Die Durchführung einer solchen Studienart sollte allerdings auch bei der Verwendung von MOST und SMART verbindlich sein, um eine entsprechende methodische Qualität bei der Überprüfung der Wirksamkeit zu gewährleisten. In diesem Fall würden jedoch die angegebenen Vorteile der Methoden, in erster Linie die kürzeren Entwicklungszeiten, konterkariert.

5. Literaturverzeichnis

1. Gesetz für eine bessere Versorgung durch Digitalisierung und Innovation(Digitale-Versorgung-Gesetz–DVG), Bundesgesetzblatt Jahrgang 2019, Teil I, Nr. 49, Bonn am 18. Dezember 2019. https://www.bgbl.de/xaver/bgbl/start.xav?startbk=Bundesanzeiger_BGBl&jumpTo=bgbl119s2562.pdf#__bgbl__%2F%2F*%5B%40attr_id%3D%27bgbl119s2562.pdf%27%5D__1626432689273. Abgerufen am 30.06.2021.
2. DIGAV 2020. Verordnung über das Verfahren und die Anforderungen zur Prüfung der Erstattungsfähigkeit digitaler Gesundheitsanwendungen in der gesetzlichen Krankenversicherung (Digitale Gesundheitsanwendungen-Verordnung–DiGAV). Bundesgesetzblatt Jahrgang 2020 Teil I Nr. 18, ausgegeben zu Bonn am 20. April 2020. https://www.bgbl.de/xaver/bgbl/start.xav?startbk=Bundesanzeiger_BGBl&jumpTo=bgbl120s0768.pdf#__bgbl__%2F%2F*%5B%40attr_id%3D%27bgbl120s0768.pdf%27%5D__1626432819279. Abgerufen am 30.06.2021.
3. Ärzte sollen Apps verschreiben können. Bundesministerium für Gesundheit. https://www.bundesgesundheitsministerium.de/digitale-versorgung-gesetz.html. Abgerufen am 30.06.2021.
4. Das Fast Track Verfahren für digitale Gesundheitsanwendungen (DiGA) nach § 139e SGB V. Bundesinstitut für Arzneimittel und Medizinprodukte (BfArM) 2021. file:///C:/Users/kolomipr/AppData/Local/Temp/diga_leitfaden.pdf%3bjsessionid=CD2D89065A34F71DCF986FEB85F2543F.pdf. Abgerufen am 30.06.2021.
5. Most popular Google Play app categories as of 1st quarter 2021 by share of available apps. Statista Research Department, 06.07.2021. https://www.statista.com/statistics/279286/google-play-android-app-categories/. Abgerufen am 10.07.2021.
6. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P., and Moher, D. (2021). The prisma 2020 statement: an updated guideline for reporting systematic reviews. BMJ (Clinical research ed.), 372:n71.
7. Moshi, M., Tooher, R., & Merlin, T. (2018). SUITABILITY OF CURRENT EVALUATION FRAMEWORKS FOR USE IN THE HEALTH TECHNOLOGY ASSESSMENT OF MOBILE MEDICAL APPLICATIONS: A SYSTEMATIC REVIEW. *International Journal of Technology Assessment in Health Care, 34*(5), 464-475. doi:10.1017/S026646231800051X.

8. Stoyanov, S. R., Hides, L., Kavanagh, D. J., Zelenko, O., Tjondronegoro, D., and Mani, M. (2015). Mobile app rating scale: A new tool for assessing the quality of health mobile apps. JMIR mHealth uHealth, 3(1):e27.

9. Messner, E.-M., Terhorst, Y., Barke, A., Baumeister, H., Stoyanov, S., Hides, L., Kavanagh, D., Pryss, R., Sander, L., and Probst, T. (2020). The german version of the mobile app rating scale (mars-g): Development and validation study. JMIR mHealth uHealth, 8(3):e14479.

10. Terhorst, Y., Philippi, P., Sander, L. B., Schultchen, D., Paganini, S., Bardus, M., Santo, K., Knitza, J., Machado, G. C., Schoeppe, S., Bauereiß, N., Portenhauser, A., Domhardt, M., Walter, B., Krusche, M., Baumeister, H., and Messner, E.-M. (2020). Validation of the mobile application rating scale (mars). PLOS ONE, 15(11):e0241480.

11. Stach, M., Kraft, R., Probst, T., Messner, E.-M., Terhorst, Y., Baumeister, H., Schickler, M., Reichert, M., Sander, L. B., and Pryss, R. (2020). Mobile Health App Database – A Repository for Quality Ratings of mHealth Apps. 33rd IEEE CBMS International Symposium on Computer-Based Medical Systems (CBMS 2020). IEEE Computer Society Press.

12. Baumeister H. (2021). Mhad - mobilde health app database. http://www.mhad.science/, Abgerufen am 10.07.2021.

13. Henson, P., David, G., Albright, K., and Torous, J. (2019). Deriving a practical framework for the evaluation of health apps. The Lancet Digital Health, 1(2):e52–e54.

14. Lagan, S., Sandler, L., and Torous, J. (2021). Evaluating evaluation frameworks: a scoping review of frameworks for assessing health apps. BMJ open, 11(3):e047001.

15. Lagan, S., Aquino, P., Emerson, M. R., Fortuna, K., Walker, R., and Torous, J. (2020). Actionable health app evaluation: translating expert frameworks into objective metrics. NPJ Digital Medicine, 3:100.

16. Lagan S, Emerson MR, King D, Matwin S, Chan SR, Proctor S, Tartaglia J, Fortuna KL, Aquino P, Walker R, Dirst M, Benson N, Myrick KJ, Tatro N, Gratzer D, Torous J. (2021). Mental Health App Evaluation: Updating the American Psychiatric Association's Framework Through a Stakeholder-Engaged Workshop. Psychiatr Serv. 2:appips202000663. doi: 10.1176/appi.ps.202000663

17. Thranberend T., Bittner J. (2019). Appq: Gütekriterien-Kernset für mehr Qualitätstransparenz bei digitalen Gesundheitsanwendungen. Hrsgb. Bertelsmann Stiftung, 29.10.2019. https://www.bertelsmann-stiftung.de/de/publikationen/publikation/did/appq/ Abgerufen am 30.06.2021.

18. Thranberend T., Bittner J. (2020). Appq 1.1: Gütekriterien-Kernset für mehr Qualitätstransparenz bei digitalen Gesundheitsanwendungen. Hrsgb. Bertelsmann Stiftung,

15.06.2020.
https://www.bertelsmann-stiftung.de/de/publikationen/publikation/did/appq-1-1
Abgerufen am 30.06.2021.

19. Das Fraunhofer-Institut für Offene Kommunikationssysteme FOKUS (2019). AppKri Bedienungsanleitung.
https://websites.fraunhofer.de/appkri/index.php/AppKri_Bedienungsanleitung
Abgerufen am 30.06.2021.
20. Milne-Ives, M., Lam, C., van Velthoven, M., and Meinert, E. (2020). Mobile fitness and weight management apps: Protocol for a quality evaluation. JMIR Res Protoc, 9(9):e17685.
21. Schwenk U. (2021). Projekt „Weisse Liste“ - Wegweiser im Gesundheitswesen. Hrsgb. Weisse Liste gemeinnützige GmbH, Gütersloh. https://www.weisse-liste.de/projekt/. Abgerufen am 30.06.2021.
22. Schuler, T. (2010). Bertelsmannrepublik Deutschland: Eine Stiftung macht Politik. Business 2010. Campus, Frankfurt am Main and New York.
23. Higgins JPT, Savović J, Page MJ, Sterne JAC on behalf of the RoB2 Development Group. Revised Cochrane risk-of-bias tool for randomized trials (RoB 2). 22 August 2019. Verfügbar unter: https://methods.cochrane.org/bias/resources/rob-2-revised-cochrane-risk-bias-tool-randomized-trials.
24. Bücker L, Bierbrodt J, Hand I, Wittekind C, Moritz S. Effects of a depression-focused internet intervention in slot machine gamblers: A randomized controlled trial. PLoS One. 2018 Jun 8;13(6):e0198859. doi: 10.1371/journal.pone.0198859. Erratum in: PLoS One. 2018 Aug 23;13(8):e0203145. PMID: 29883479; PMCID: PMC5993308.
25. Zwerenz R, Becker J, Knickenberg RJ, Siepmann M, Hagen K, Beutel ME. Online Self-Help as an Add-On to Inpatient Psychotherapy: Efficacy of a New Blended Treatment Approach. Psychother Psychosom. 2017;86(6):341-350. doi: 10.1159/000481177. Epub 2017 Nov 3. PMID: 29131090.
26. Krieger T, Meyer B, Sude K, Urech A, Maercker A, Berger T. Evaluating an e-mental health program ("deprexis") as adjunctive treatment tool in psychotherapy for depression: design of a pragmatic randomized controlled trial. BMC Psychiatry. 2014;14:285. Published 2014 Oct 8. doi:10.1186/s12888-014-0285-9
27. Beevers CG, Pearson R, Hoffman JS, Foulser AA, Shumake J, Meyer B. Effectiveness of an internet intervention (Deprexis) for depression in a united states adult sample: A parallel-group pragmatic randomized controlled trial. J Consult Clin Psychol. 2017 Apr;85(4):367-380. doi: 10.1037/ccp0000171. Epub 2017 Feb 23. PMID: 28230390.

28. Klein JP, Berger T, Schröder J, Späth C, Meyer B, Caspar F, Lutz W, Arndt A, Greiner W, Gräfe V, Hautzinger M, Fuhr K, Rose M, Nolte S, Löwe B, Anderssoni G, Vettorazzi E, Moritz S, Hohagen F. Effects of a Psychological Internet Intervention in the Treatment of Mild to Moderate Depressive Symptoms: Results of the EVIDENT Study, a Randomized Controlled Trial. Psychother Psychosom. 2016;85(4):218-28. doi: 10.1159/000445355. Epub 2016 May 27. PMID: 27230863; PMCID: PMC8117387.

29. Fischer A, Schröder J, Vettorazzi E, Wolf OT, Pöttgen J, Lau S, Heesen C, Moritz S, Gold SM. An online programme to reduce depression in patients with multiple sclerosis: a randomised controlled trial. Lancet Psychiatry. 2015 Mar;2(3):217-23. doi: 10.1016/S2215-0366(14)00049-2. Epub 2015 Feb 25. PMID: 26359900.

30. Schröder J, Brückner K, Fischer A, Lindenau M, Köther U, Vettorazzi E, Moritz S. Efficacy of a psychological online intervention for depression in people with epilepsy: a randomized controlled trial. Epilepsia. 2014 Dec;55(12):2069-76. doi: 10.1111/epi.12833. Epub 2014 Nov 19. PMID: 25410633.

31. Meyer B. Bierbrodt J, Schröder J, Berger T, Beevers CG, Weiss M, Jacob G, Späth C, Andersson G, Lutz W, Hautzinger M, Löwe B, Rose M, Hohagen F, Caspar F, Greiner W, Moritz S, Klein JP. Effects of an Internet intervention (Deprexis) on severe depression symptoms: Randomized controlled trial. Internet Interventions. 2015;2(1): 48-59. http://dx.doi.org/10.1016/j.invent.2014.12.003

32. Moritz S, Schilling L, Hauschildt M, Schröder J, Treszl A. A randomized controlled trial of internet-based therapy in depression. Behav Res Ther. 2012 Aug;50(7-8):513-21. doi: 10.1016/j.brat.2012.04.006. Epub 2012 May 3. PMID: 22677231.

33. Berger T, Hämmerli K, Gubser N, Andersson G, Caspar F. Internet-based treatment of depression: a randomized controlled trial comparing guided with unguided self-help. Cogn Behav Ther. 2011;40(4):251-66. doi: 10.1080/16506073.2011.616531. PMID: 22060248.

34. Meyer B, Berger T, Caspar F, Beevers CG, Andersson G, Weiss M. Effectiveness of a novel integrative online treatment for depression (Deprexis): randomized controlled trial. J Med Internet Res. 2009 May 11;11(2):e15. doi: 10.2196/jmir.1151. PMID: 19632969; PMCID: PMC2762808.

35. Twomey C, O'Reilly G, Bültmann O, Meyer B. Effectiveness of a tailored, integrative Internet intervention (deprexis) for depression: Updated meta-analysis. PLoS One. 2020 Jan 30;15(1):e0228100. doi: 10.1371/journal.pone.0228100. PMID: 31999743; PMCID: PMC6992171.

36. Pöttgen J, Moss-Morris R, Wendebourg JM, Feddersen L, Lau S, Köpke S, Meyer B, Friede T, Penner IK, Heesen C, Gold SM. Randomised controlled trial of a self-guided online fatigue

intervention in multiple sclerosis. J Neurol Neurosurg Psychiatry. 2018 Sep;89(9):970-976. doi: 10.1136/jnnp-2017-317463. Epub 2018 Mar 16. PMID: 29549193.

37. Cella M, Chalder T. Measuring fatigue in clinical and community settings. J Psychosom Res 2010;69:17–22.

38. Chilcot J, Norton S, Kelly ME, Moss-Morris R. The Chalder Fatigue Questionnaire is a valid and reliable measure of perceived fatigue severity in multiple sclerosis. Mult Scler. 2016 Apr;22(5):677-84. doi: 10.1177/1352458515598019. Epub 2015 Jul 31. PMID: 26232100.

39. Lorenz N, Heim E, Roetger A, Birrer E, Maercker A. Randomized Controlled Trial to Test the Efficacy of an Unguided Online Intervention with Automated Feedback for the Treatment of Insomnia. Behav Cogn Psychother. 2019 May;47(3):287-302. doi: 10.1017/S1352465818000486. Epub 2018 Sep 6. PMID: 30185239.

40. Bastien C, Vallières A, Morin C. Validation of the Insomnia Severity Index as an outcome measure for insomnia research. Sleep Medicine 2001, 2, 297–307. http://doi.org/doi:10.1016/S1389-9457(00)00065-4

41. Beck AT, Steer RA, Brown GK. Beck Depression Inventory-II. San Antonio, 2016. 12–15. http://doi.org/10.1037/t00742-00

42. Berger T, Urech A, Krieger T, Stolz T, Schulz A, Vincent A, Moser CT, Moritz S, Meyer B. Effects of a transdiagnostic unguided Internet intervention ('velibra') for anxiety disorders in primary care: results of a randomized controlled trial. Psychol Med. 2017 Jan;47(1):67-80. doi: 10.1017/S0033291716002270. Epub 2016 Sep 22. PMID: 27655039.

43. Lovibond SH, Lovibond PF. Manual for the Depression Anxiety Stress Scales. Psychology Foundation, 1995: Sydney.

44. Beck AT, Epstein N, Brown G, Steer RA. An inventory for measuring clinical anxiety: psychometric properties. Journal of Consulting and Clinical Psychology 1988. 56, 893–897.

45. Derogatis LR. Brief Symptom Inventory (BSI): Administration, Scoring and Procedures Manual. National Computer Systems, 1993: Minneapolis, MN.

46. Ware Jr. J, Kosinski M, Keller SD. A 12-Item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity. Medical Care 1996. 34, 220–233

47. Zill JM, Christalle E, Meyer B, Härter M, Dirmaier J: The effectiveness of an internet intervention aimed at reducing alcohol consumption in adults: results of a randomized controlled trial (Vorvida). Dtsch Arztebl Int 2019; 116: 127–33. DOI: 10.3238/arztebl.2019.0127

48. Bloomfield K, Hope A, Kraus L. Alcohol survey measures for Europe: a literature review. Drugs Educ Prev Polic 2013; 20: 348–60

49. Kraus L, Piontek D, Pabst A, Gomes de Matos E. Studiendesign und Methodik des Epidemiologischen Suchtsurveys 2012. Sucht 2013;59:309–20.

50. Sobell LC, Sobell MB: Timeline Follow-Back: a technique for assessing self-reported alcohol consumption. In: Litten RZ, Allen JP (eds.): Measuring alcohol consumption: psychosocial and biochemical methods. Humana Press 1992.

51. Bundesinstitut für Arzneimittel und Medizinprodukte (BfArM). Das Fast Track Verfahren für digitale Gesundheitsanwendungen (DiGA) nach § 139e SGB V. Ein Leitfaden für Hersteller, Leistungserbringer und Anwender. Bonn 2020.

52. Mohr DC, Cheung K, Schueller SM, Hendricks Brown C, Duan N. Continuous evaluation of evolving behavioral intervention technologies. Am. J. Prev. Med 2013;45:517–23, http://dx.doi.org/10.1016/j.amepre.2013.06.006.

53. Knöppler K, Hesse S, Ex P. Transfer von Digital-Health-Anwendungen in den Versorgungsalltag. Teil 4: Wirksamkeitsnachweis und Nutzenbewertung Kontext, Methoden und Integration in die agile Produktentwicklung. Bertelsmann Stiftung 2018. URL: https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/DH-Transfer_Wirksamkeitsnachweis_Nutzenbewertung.pdf.

54. Gensorowsky D, Lampe D, Hasemann L, Düvel J, Greiner W. „Alternative Studiendesigns“ zur Bewertung digitaler Gesundheitsanwendungen – eine echte Alternative? ["Alternative study designs" for the evaluation of digital health applications- a real alternative?]. Z Evid Fortbild Qual Gesundhwes. 2021 Apr;161:33-41. German. doi: 10.1016/j.zefq.2021.01.006. Epub 2021 Feb 26. PMID: 33642251.

55. Collins LM, Murphy SA, Strecher V. The multiphase optimization strategy (MOST) and the sequential multiple assignment randomized trial (SMART): new methods for more potent eHealth interventions. Am J Prev Med. 2007 May;32(5 Suppl):S112-8. doi: 10.1016/j.amepre.2007.01.022. PMID: 17466815; PMCID: PMC2062525.

56. Almirall D, Nahum-Shani I, Sherwood NE, Murphy SA. Introduction to SMART designs for the development of adaptive interventions: with application to weight loss research. Transl Behav Med. 2014 Sep;4(3):260-74. doi: 10.1007/s13142-014-0265-0. PMID: 25264466; PMCID: PMC4167891.

57. Klasnja P, Hekler EB, Shiffman S, Boruvka A, Almirall D, Tewari A, Murphy SA. Microrandomized trials: An experimental design for developing just-in-time adaptive interventions. Health Psychol. 2015 Dec;34S(0):1220-8. doi: 10.1037/hea0000305. PMID: 26651463; PMCID: PMC4732571.

58. Chan A-W, Tetzlaff JM, Altman DG, Laupacis A, Gøtzsche PC, Krleža-Jerić K, Hróbjartsson A, Mann H, Dickersin K, Berlin J, Doré C, Parulekar W, Summerskill W, Groves T, Schulz K, Sox H, Rockhold FW, Rennie D, Moher D. SPIRIT 2013 Statement: Defining standard protocol items for clinical trials. Ann Intern Med 2013;158:200-207.

59. Chan A-W, Tetzlaff JM, Gøtzsche PC, Altman DG, Mann H, Berlin J, Dickersin K, Hróbjartsson A, Schulz KF, Parulekar WR, Krleža-Jerić K, Laupacis A, Moher D. SPIRIT 2013 Explanation and Elaboration: Guidance for protocols of clinical trials. BMJ 2013;346:e7586.

60. Fiatarone Singh MA, Gates N, Saigal N, Wilson GC, Meiklejohn J, Brodaty H, Wen W, Singh N, Baune BT, Suo C, Baker MK, Foroughi N, Wang Y, Sachdev PS, Valenzuela M. The Study of Mental and Resistance Training (SMART) study—resistance training and/or cognitive training in mild cognitive impairment: a randomized, double-blind, double-sham controlled trial. J Am Med Dir Assoc. 2014 Dec;15(12):873-80. doi: 10.1016/j.jamda.2014.09.010. Epub 2014 Oct 23. Erratum in: J Am Med Dir Assoc. 2021 Feb;22(2):479-481. PMID: 25444575.

6. Anlage (151 Seiten, separat)

6.1. Revised Cochrane risk-of-bias tool for randomized trials (RoB 2)

6.2. Bewertung der Studienqualität von 15 Studien mit dem RoB 2