

Unterstützt von DALL-E 3


# Künstliche Intelligenz und Psychotherapie

Prof. Dr. Harald Baumeister  
Abteilung für Klinische Psychologie und  
Psychotherapie

6. Oberberg Online-Vortragsreihe  
09. Januar 2024



Unterstützt von DALL-E 3



**Universität Ulm –  
Fakultät Ingenieurwissenschaften,  
Informatik und Psychologie**

**u.a. fachbereichsübergreifende  
Forschungsschwerpunkte:**

- **Kognitive Systeme &  
Mensch-Technik Interaktion**
- **Smart Sensing Systems**

**Forschungsschwerpunkte  
Psychologie:**

- **Mensch & Gesundheit**
- **Menschliches Verhalten  
in sozialen Interaktionen**

**Abteilung für Klinische Psychologie und Psychotherapie:**

- **Digital Behavioral and Mental Health**
- **Psychotherapieprozessforschung, Somato-Psychologie,  
Versorgungsforschung**
- **Teil der Deutschen Zentren für Psychische Gesundheit (DZP)  
und Kinder- und Jugendgesundheit (DZKJ)**

# Psychische Gesundheit, Digitalisierung und KI

## Grundlagenforschung

- Ätiologie und Pathogenese
- Nosologie und Klassifikation
- Epidemiologie

## Diagnostik- und Interventionsforschung / Versorgungsforschung

### Diagnostik

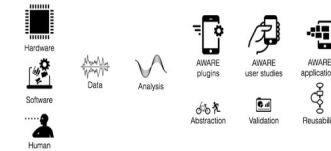
- Patient:innen-berichtete Outcomes (PROM)
- Screening psychischer Störungen
- Lebensqualität
- Motivation und Volition
- Psycho-soziale State/Trait-Konstrukte

### Interventionen

- Psychische Belastungen und Störungen
- Verhaltensmedizinische Interventionen
- Prävention und Gesundheitsförderung (Verhaltensänderung)
- Kompetenz- / Skilltraining
- Komplexe Interventionen



## Digital Phenotyping / EMA / Smart Sensing



## Server-basiertes PROM/PREM Computer-adaptive Testing (CAT) Clinical Decision Support Systems (CDSS)

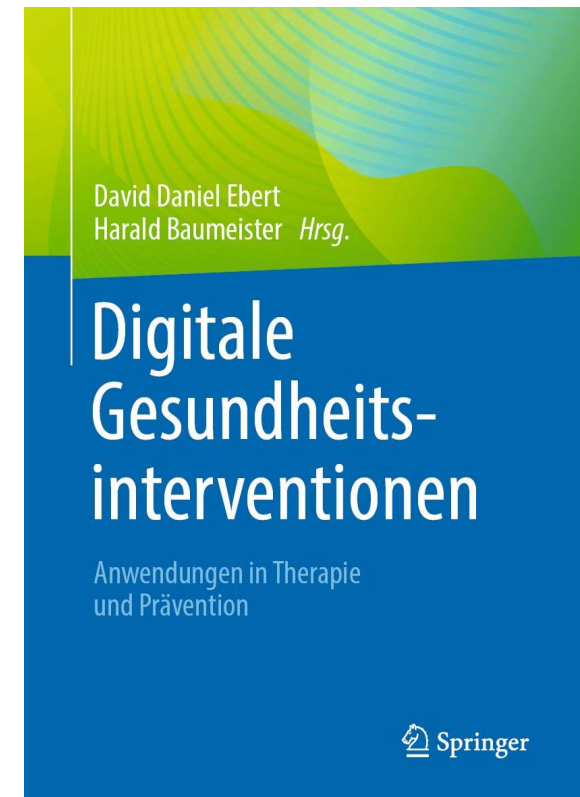


## Internet- und mobile-basiert (IMI) Blended Care Virtual Reality KI-basierte Innovationen Videokonferenz-basierte Therapie



## Digitale Gesundheitsinterventionen

- Potentiell akzeptabel, wirksam und effektiv für eine Vielzahl an psychischen und körperlichen Erkrankungen sowie in der Prävention und Gesundheitsförderung
- Neue, ergänzende Versorgungswege
- Versorgungsmöglichkeit von Menschen, die die vor-Ort Behandlung nicht in Anspruch nehmen (können)
- Skalierbarer Ansatz für ein Mehr an Versorgung
- Digitale Gesundheitsinterventionen aktuell erstentwicklungs- und pflegeaufwändige „Stangenwaren“-Interventionen mit begrenzten Personalisierungsmöglichkeiten



# Digitale Gesundheitsinterventionen

## Implementierungsvarianten und Innovationen

- Blended Therapy<sup>1,2</sup>
  
- (KI-basierte) Flexibilisierung und Personalisierung
  - Virtual Assistant / Chatbot<sup>3,4</sup>
  - Clinical Decision Support Systeme (CDSS)<sup>5</sup>
  - Just-in-Time adaptive Interventions (JITAI) / EMI<sup>6,7</sup>



# Blended Therapy

**PSYCHOnlineTHERAPIE für den verzahnten Einsatz in der ambulanten Psychotherapie – Durchführung u.a. gemeinsam mit Krankenkassen, gefördert vom Innovationsausschuss des Gemeinsamen Bundesausschusses (GBA)**



**Website:**  
→ <https://www.psychonlinetherapie.de>



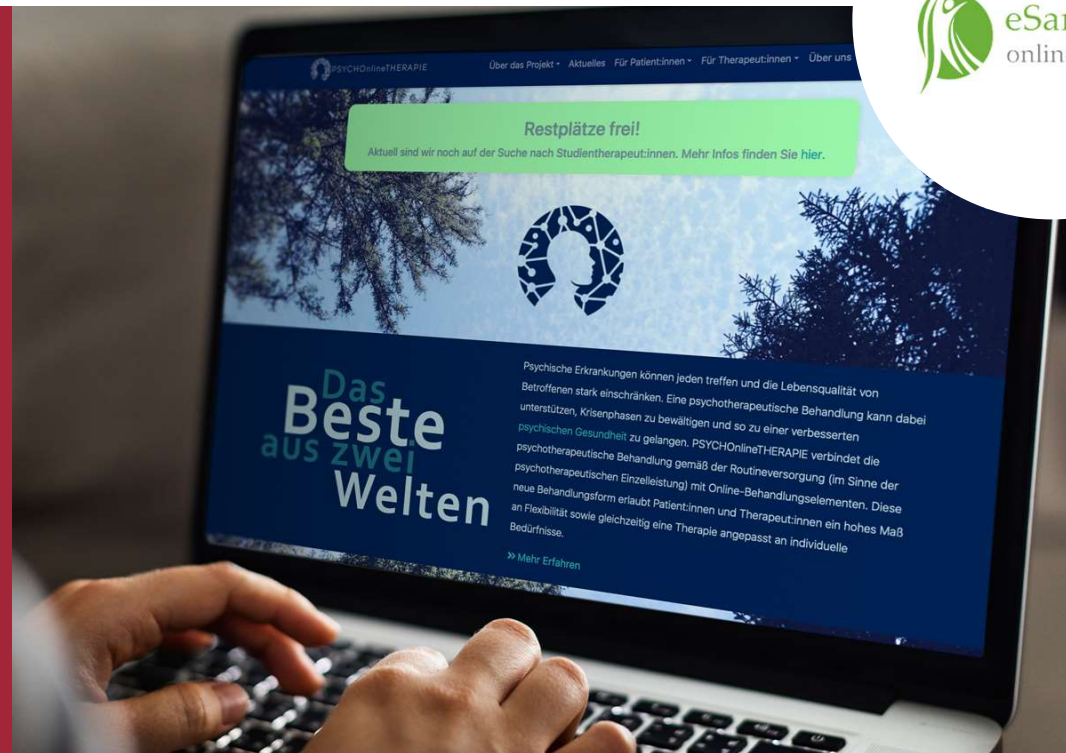
**eSano Online-Interventionen:**  
PSYCHOnlineTHERAPIE  
Aktuell für Depressionen, Angststörungen  
Weitgehend finalisiert für: ICD-10 F1-F6



**Social Media:**  
Facebook/Instagram  
@psychonlinetherapie



Twitter  
@psychONtherapie



Baumeister et al. Frontiers Psychiatry 2021; <https://esano.klips-ulm.de/de/it-plattform/>; Kraft et al. EMBC 2021; Idrees et al. WiMob 2022

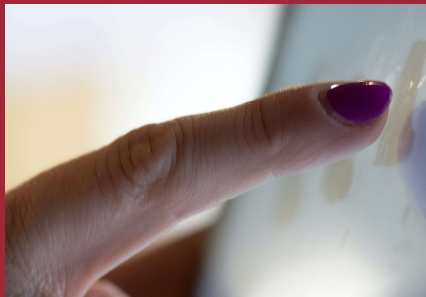
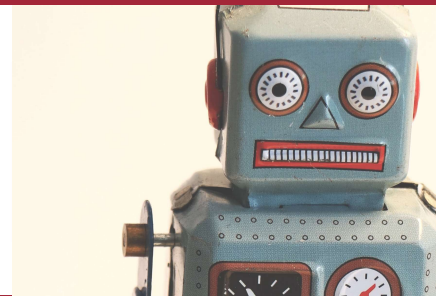
# Digitale Gesundheitsinterventionen

## Implementierungsvarianten und Innovationen

- Blended Therapy<sup>1,2</sup>
  
- (KI-basierte) Flexibilisierung und Personalisierung
  - Virtual Assistant / Chatbot<sup>3,4</sup>
  - Clinical Decision Support Systeme (CDSS)<sup>5</sup>
  - Just-in-Time adaptive Interventions (JITAI) / EMI<sup>6,7</sup>

## Next Generation Interventions

**SISU Chatbot:** Ein Chatbot zur Steigerung des psychischen Wohlbefindens



better  
care

**Better Care JITAI:** Eine auf ACTonCANCER aufbauende Just-in-Time-Intervention unter Nutzung von EMA und Smart Sensing



## Künstliche Intelligenz (KI): Definition

**Intelligenz:** Fähigkeit [des Menschen], abstrakt und vernünftig zu denken und daraus zweckvolles Handeln abzuleiten (Duden, 2023)

**Künstlich:** nicht natürlich, sondern mit chemischen und technischen Mitteln nachgebildet, nach einem natürlichen Vorbild angelegt, gefertigt, geschaffen (Duden, 2023)

„**Schwache**“ vs. „**Starke KI**“:<sup>1</sup> „Schwache KI (Artificial Narrow Intelligence (ANI))“ wird nur für spezielle und definierte Aufgaben eingesetzt, „Starke KI (Artificial General Intelligence (AGI))“ ist eine (derzeit) hypothetische Form der KI, die die menschliche Intelligenz erreichen (= AGI) oder übersteigen (= Artificial Superintelligence (ASI) ) und diese Problemlösungsfähigkeit auf beliebige Problemstellungen anwenden kann

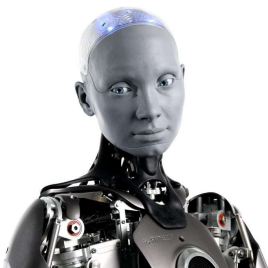
<sup>1</sup> Flowers ceur-WS.org, 2019; Scherk et al. 2017

- 1950er: KI-Begriff wird geprägt
- Dartmouth Konferenz 1956
  - Turing Test (Grundlage 1936)

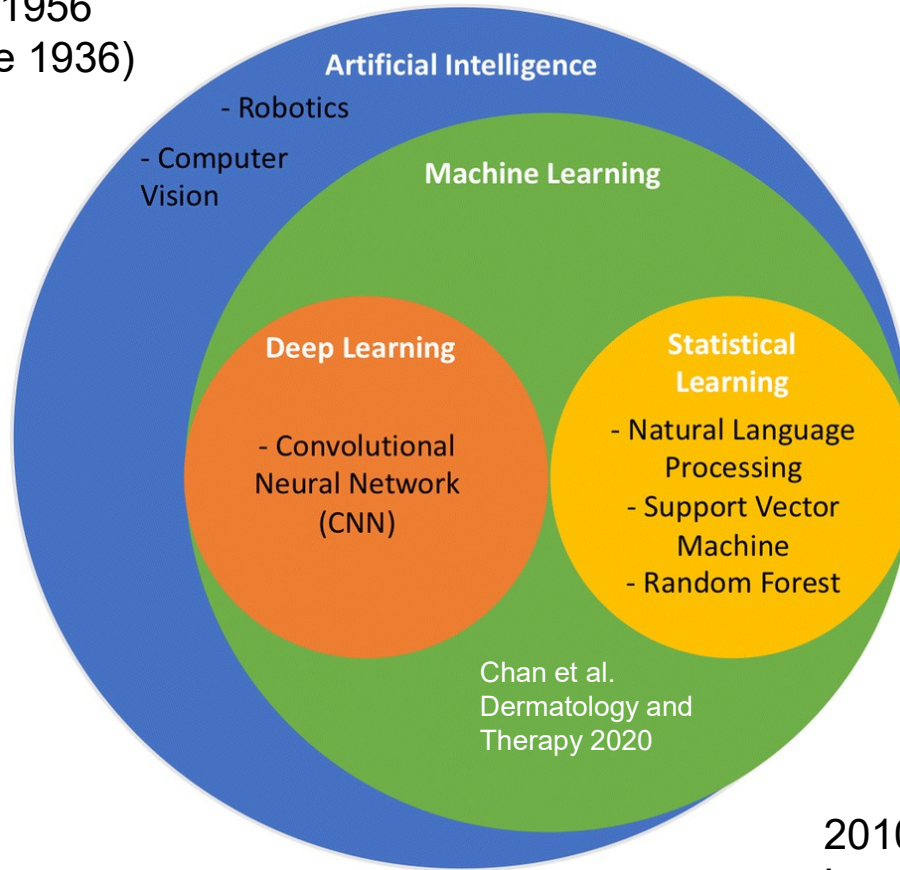
2022: ChatGPT  
Öffentlichkeit



2017: Deepfakes



2016: z.B. Sophia und (hier abgebildet) Ameca



1966



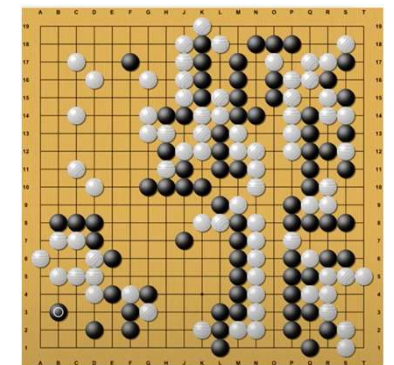
1980er: Machine Learning



1996 Deep Blue

3361 (= 1,7 · 10<sup>172</sup>) Möglichkeiten

2010er: Deep Learning  
2016 AlphaGo  
2017 AlphaGoZero  
2017 AlphaZero



# Artificial Intelligence Index Report 2023

## AI Journal Publications (% of World Total) by Geographic Area, 2010–21

Source: Center for Security and Emerging Technology, 2022 | Chart: 2023 AI Index Report

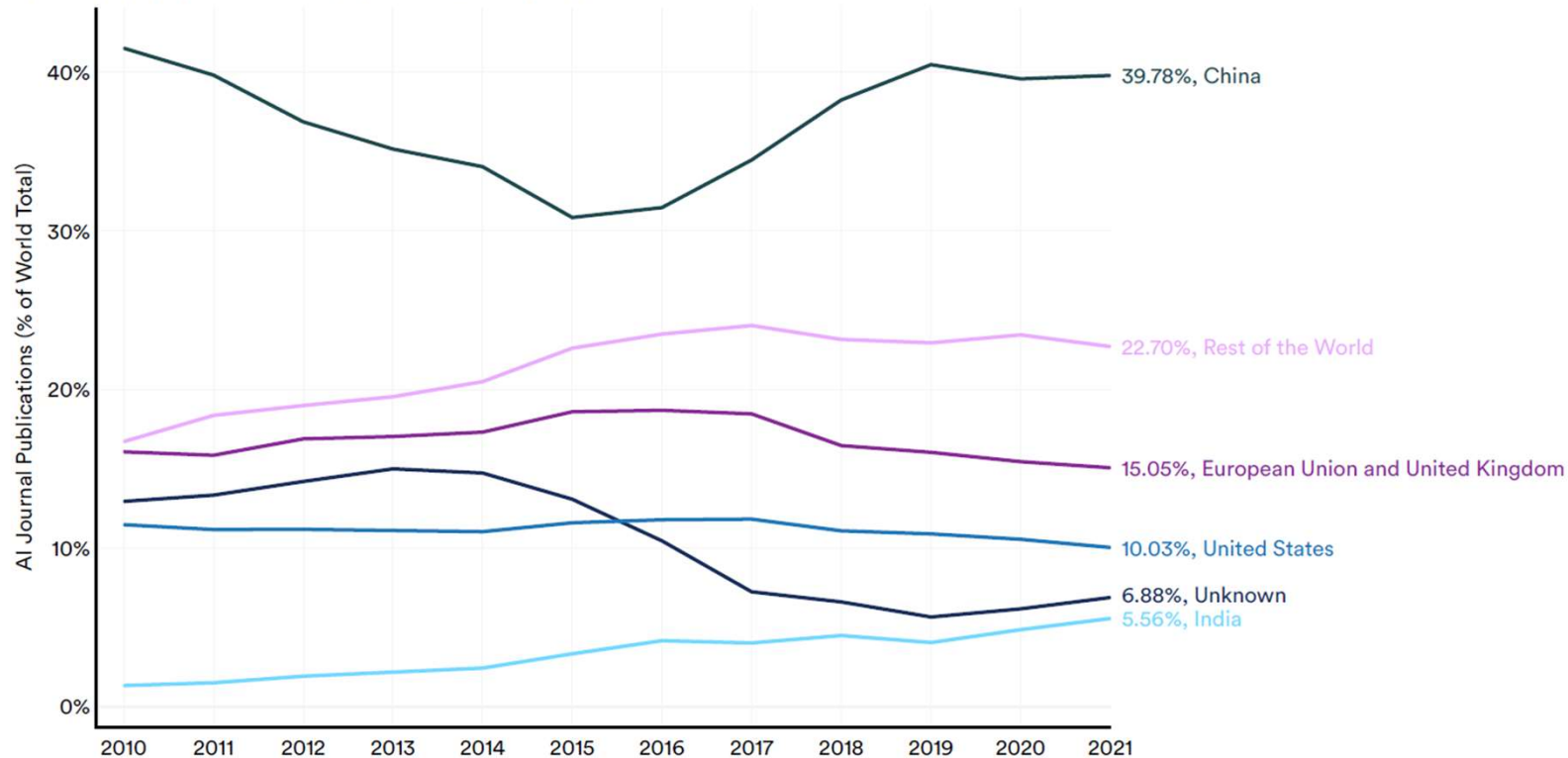


Figure 1.1.11

<https://aiindex.stanford.edu/report/>

# Artificial Intelligence Index Report 2023

## Private Investment in AI by Geographic Area, 2022

Source: NetBase Quid, 2022 | Chart: 2023 AI Index Report

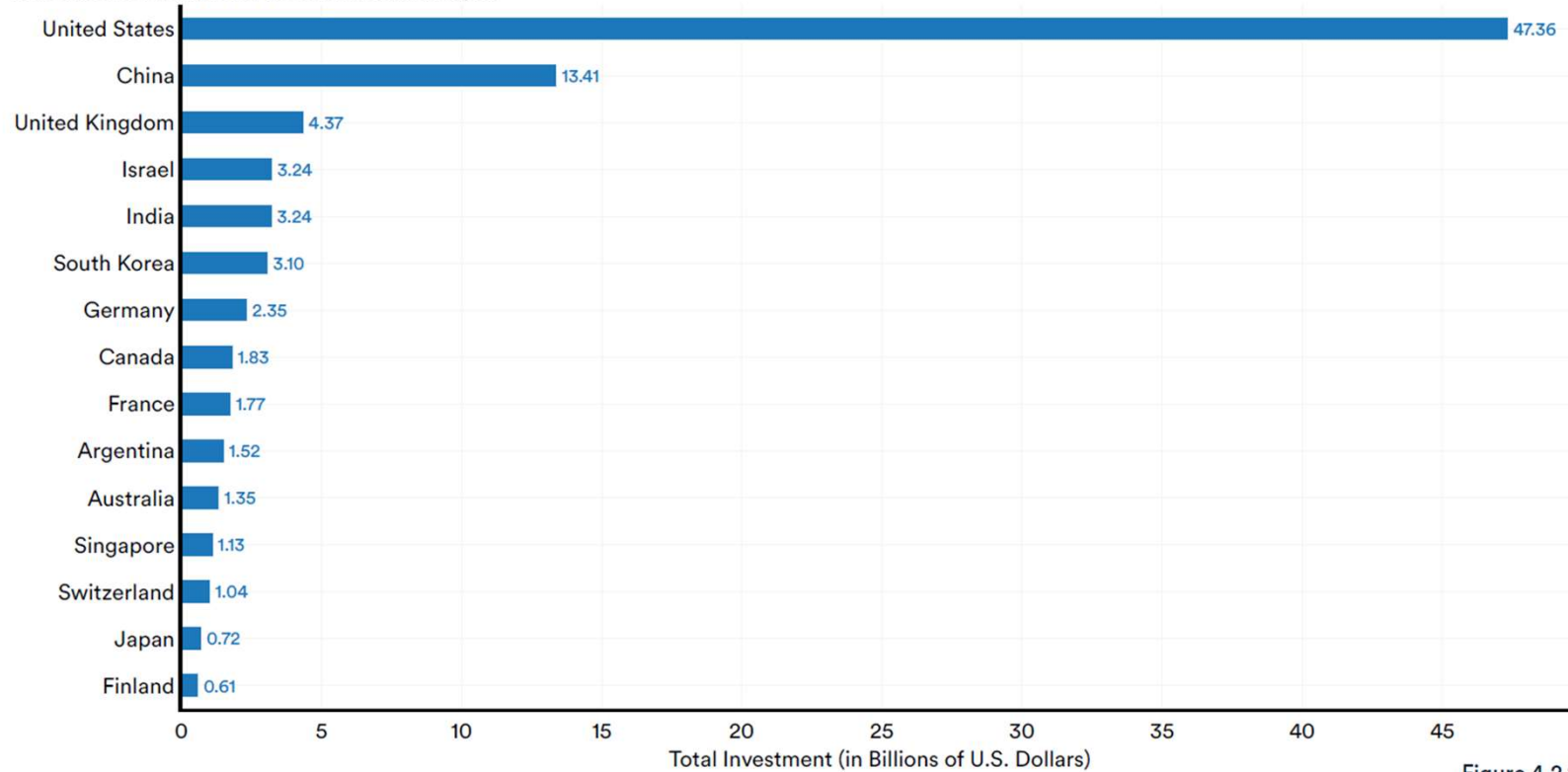


Figure 4.2.10

<https://aiindex.stanford.edu/report/>

# Artificial Intelligence Index Report 2023

## Number of AI Incidents and Controversies, 2012–21

Source: AIAAIC Repository, 2022 | Chart: 2023 AI Index Report

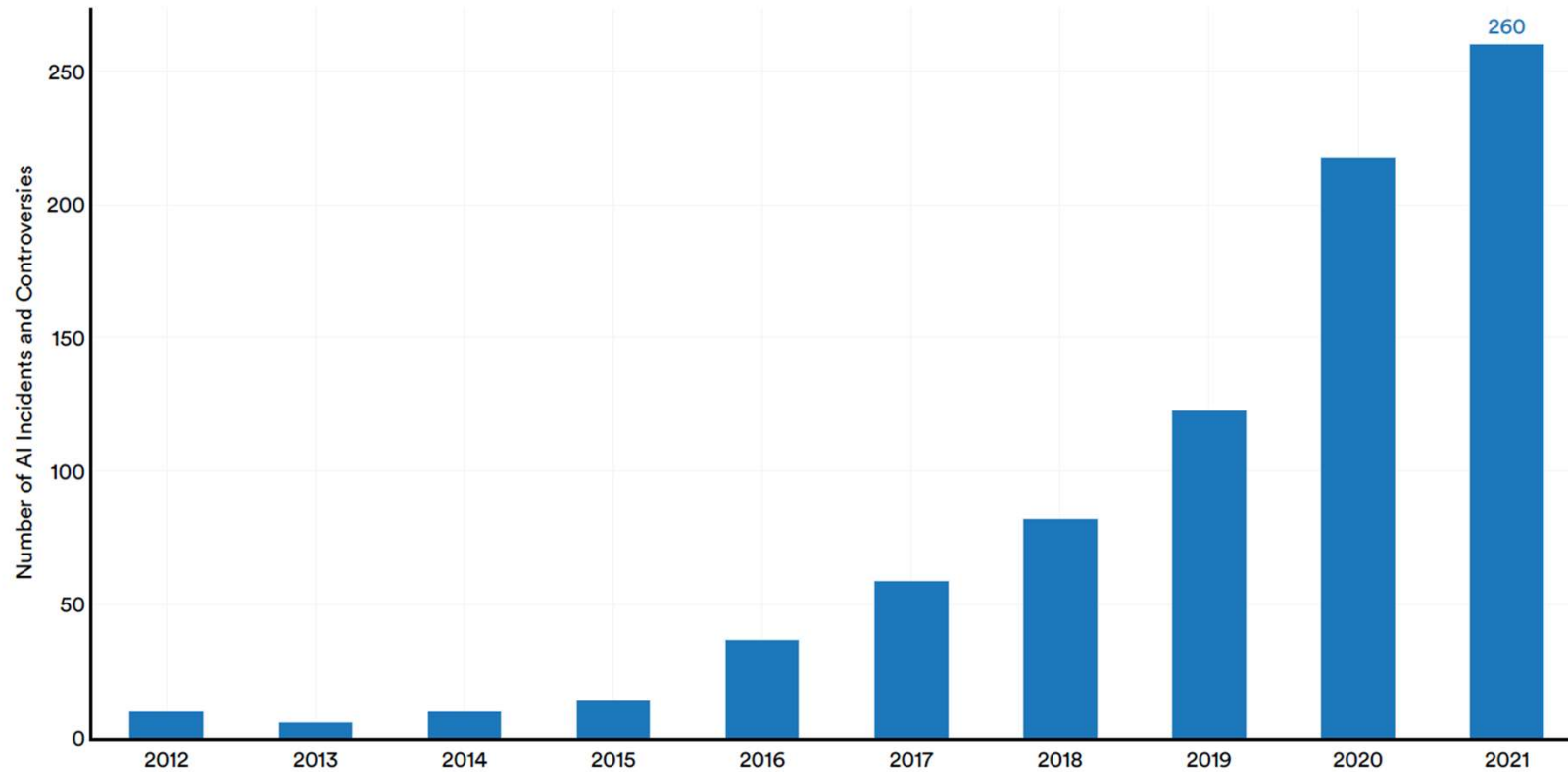


Figure 3.2.1

<https://aiindex.stanford.edu/report/>

# Artificial Intelligence Index Report 2023

## Private Investment in AI by Focus Area, 2021 Vs. 2022

Source: NetBase Quid, 2022 | Chart: 2023 AI Index Report

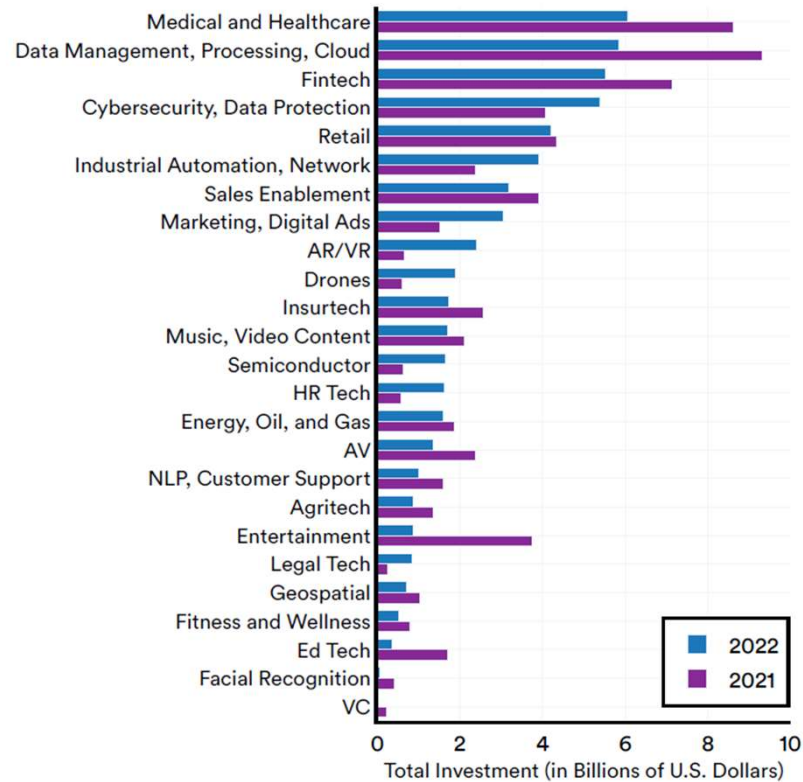


Figure 4.2.19

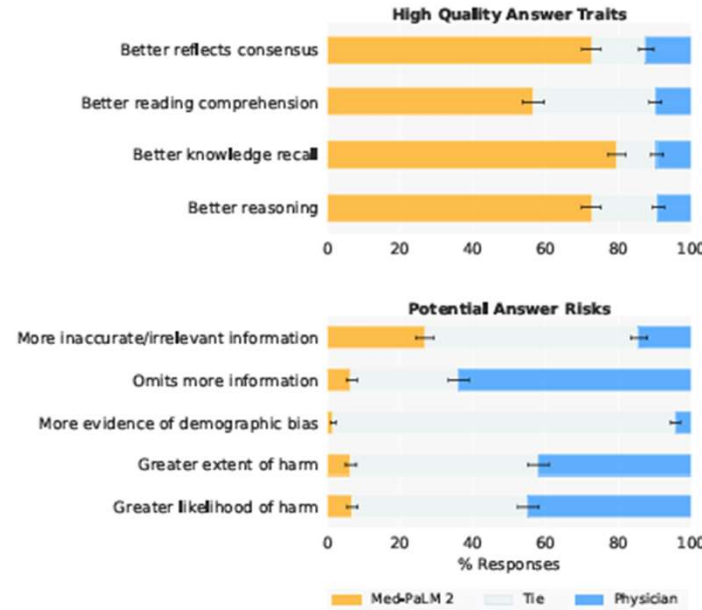
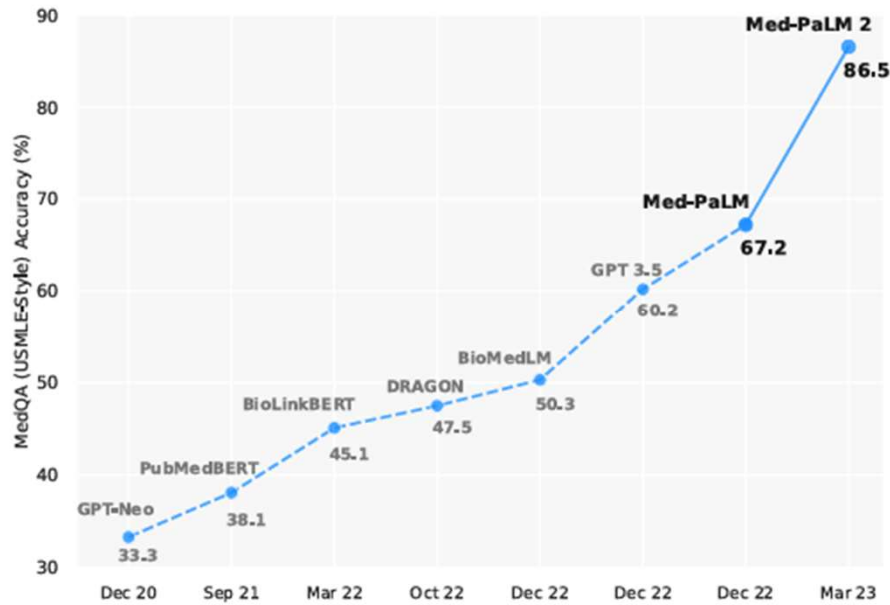
**Der Bereich mit dem höchsten  
Investment aus dem privaten Sektor:**

**Medizin und  
Gesundheitsversorgung**

<https://aiindex.stanford.edu/report/>



# MultiMedQA – Benchmark für medizinische Prüfungsfragen



## Towards Expert-Level Medical Question Answering with Large Language Models

Karan Singhal<sup>1,2</sup>, Tan Yu<sup>1</sup>, Jara Gotrova<sup>1</sup>, Rory Stryer<sup>1,3</sup>, Elley Wuzyu<sup>1</sup>, Le Hou<sup>1</sup>, Kevin Clark<sup>1</sup>, Stephen Pohl<sup>1</sup>, Heather Cole-Lewis<sup>1</sup>, Darlene Neal<sup>1</sup>, Mike Schachermann<sup>1</sup>, Amy Wang<sup>1</sup>, Mohamed Amin<sup>1</sup>, Sami Lachgar<sup>1</sup>, Philip Mandillo<sup>1</sup>, Sushant Prakash<sup>1</sup>, Bradley Green<sup>1</sup>, Ewa Dominiak<sup>1</sup>, Blake Agnew<sup>1</sup> & Arsal<sup>1</sup>, Nand Bhanushali<sup>1</sup>, Yun Liu<sup>1</sup>, Hense Wong<sup>1</sup>, Christopher Semur<sup>1</sup>, S. Sara Mahdavi<sup>1</sup>, Joelle Barral<sup>1</sup>, Dale Webster<sup>1</sup>, Greg S. Corrado<sup>1</sup>, Yossi Matias<sup>1</sup>, Shobkoth Azaria<sup>1</sup>, Alan Karthikesalingam<sup>1,2</sup> and Vivek Natarajan<sup>1,2</sup>

<sup>1</sup>Google Research, <sup>2</sup>DeepMind,

Recent artificial intelligence (AI) systems have reached milestones in "grand challenges" ranging from GPT to protein-folding. The capability to retrieve medical knowledge, reason over it, and answer medical questions comparably to physicians has long been viewed as one such grand challenge. Large language models (LLMs) have exhibited significant progress in medical question answering. Med-PaLM was the first model to exceed a "passing" score in US Medical Licensing Examination (USMLE) style questions with a score of 67.2% on the MedQA dataset. However, this and other prior work suggested significant room for improvement, especially when models' answers were compared to clinicians' answers. Here we present Med-PaLM 2, which bridges these gaps by leveraging a combination of base LLM improvements (PaLM 2), medical domain fine-tuning, and prompting strategies including a novel ensemble refinement approach. Med-PaLM 2 scored up to 86.5% on the MedQA dataset, improving upon Med-PaLM 1 over 19% and setting a new state-of-the-art. We also observed performance approaching or exceeding state-of-the-art across MedMCQA, PubMedQA, and MMLU clinical topics datasets. We performed detailed human evaluation on 1066 consumer medical questions including a novel ensemble refinement approach. In pairwise comparative ranking of 1066 consumer medical questions, physicians preferred Med-PaLM 2 answers to those produced by physicians on eight of nine axes pertaining to clinical utility ( $p < 0.001$ ). We also observed significant improvements compared to Med-PaLM 1 on every evaluation axis ( $p < 0.001$ ) on newly introduced datasets of 200 long-form "adversarial" questions to probe LLM limitations. While further studies are necessary to validate the efficacy of these models in real-world settings, these results highlight rapid progress towards physician-level performance in medical question answering.

## Article Large language models encode clinical knowledge

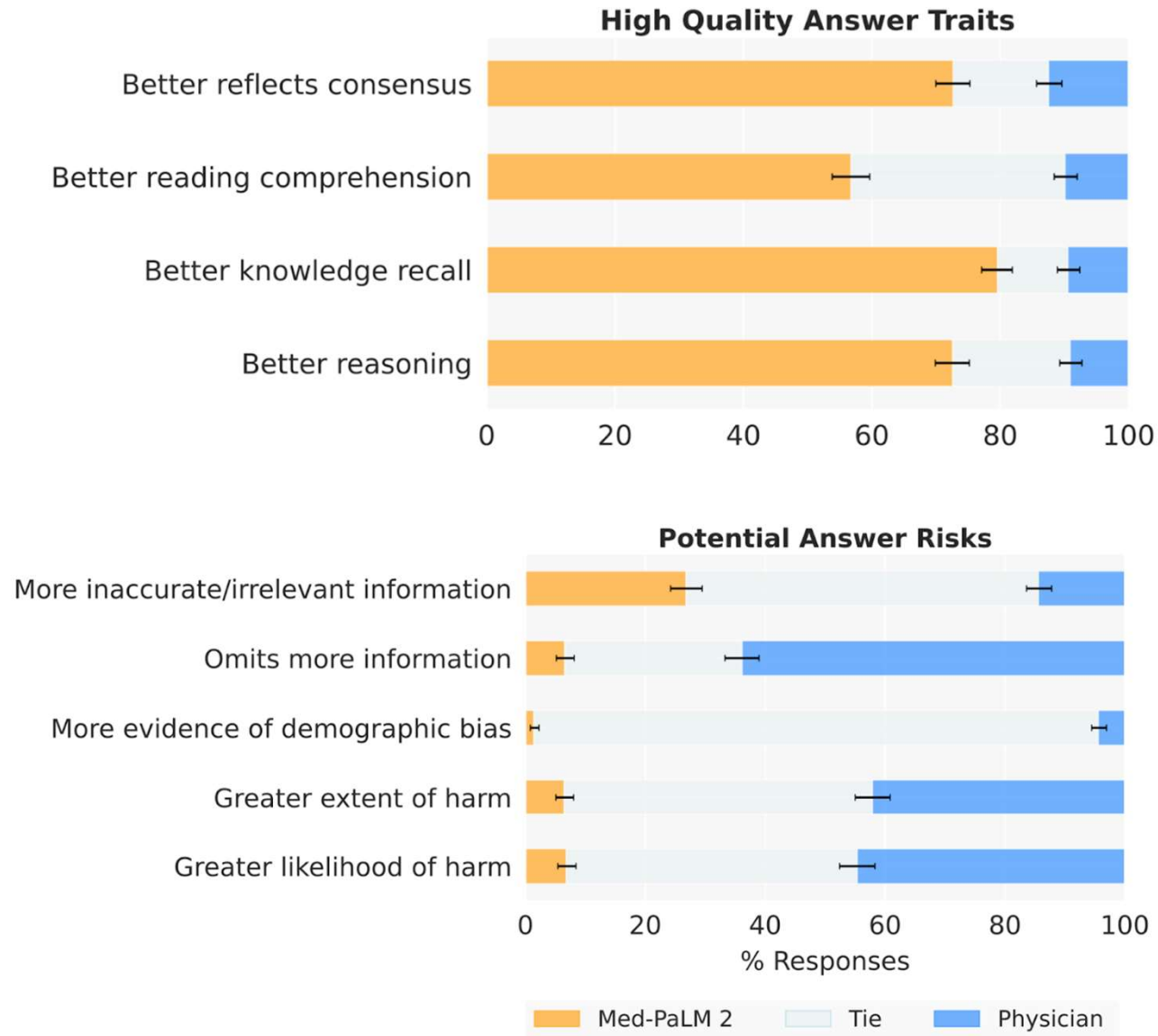
Karan Singhal<sup>1,2</sup>, Tan Yu<sup>1</sup>, Jara Gotrova<sup>1</sup>, Rory Stryer<sup>1,3</sup>, Elley Wuzyu<sup>1</sup>, Le Hou<sup>1</sup>, Kevin Clark<sup>1</sup>, Stephen Pohl<sup>1</sup>, Heather Cole-Lewis<sup>1</sup>, Darlene Neal<sup>1</sup>, Mike Schachermann<sup>1</sup>, Amy Wang<sup>1</sup>, Mohamed Amin<sup>1</sup>, Sami Lachgar<sup>1</sup>, Philip Mandillo<sup>1</sup>, Sushant Prakash<sup>1</sup>, Bradley Green<sup>1</sup>, Ewa Dominiak<sup>1</sup>, Blake Agnew<sup>1</sup> & Arsal<sup>1</sup>, Nand Bhanushali<sup>1</sup>, Yun Liu<sup>1</sup>, Hense Wong<sup>1</sup>, Christopher Semur<sup>1</sup>, S. Sara Mahdavi<sup>1</sup>, Joelle Barral<sup>1</sup>, Dale Webster<sup>1</sup>, Greg S. Corrado<sup>1</sup>, Yossi Matias<sup>1</sup>, Shobkoth Azaria<sup>1</sup>, Alan Karthikesalingam<sup>1,2</sup> and Vivek Natarajan<sup>1,2</sup>

Large language models (LLMs) have demonstrated impressive capabilities, but the bar for clinical applications is high. Attempts to assess the clinical knowledge of models typically rely on automated evaluations based on limited benchmarks. Here, to address these limitations, we present MultiMedQA, a benchmark combining six existing medical question answering datasets with newly constructed adversarial research and consumer queries and a new data set of medical questions created online. HealthMedQA. We propose a human evaluation framework for model answers along multiple axes including factuality, comprehension, reasoning, possible harm and bias. In addition, we evaluated Pathway Language Model (PaLM) at its full parameter LLM and instruction-tuned variant, Flu-PaLM, on MultiMedQA using a combination of prompting strategies. Flu-PaLM achieves state-of-the-art accuracy on every MultiMedQA multiple-choice question (MedMCQA, PubMedQA, PubMedQA) and answering Multiple-Choice Language Understanding (MMLU clinical topics), including 67.2% accuracy on MedQA US Medical Licensing Examination questions, surpassing the prior state of the art by more than 17%. However, human evaluation reveals key gaps. For example, we introduce instruction prompting, a parameter-efficient approach for aligning LLMs to new domains using a few exemplars. The resulting model, Med-PaLM 2, performs encouragingly, but remains inferior to clinicians. We show that comprehension, knowledge recall and reasoning improve with model scale and instruction prompting, suggesting the potential utility of LLMs in medicine. Our human evaluation reveals limitations of today's models, underscoring the importance of both evaluation frameworks and method development in creating safe, helpful LLMs for clinical applications.

Medicine is a human endeavor in which language enables us to interpret and understand the world. The knowledge encoded in medical corpora, at scale, offers several potential applications of such models in medicine, including but not limited to: assisting in diagnosis, providing patient education, streamlining administrative tasks, and supporting research. However, the safety-critical nature of the domain necessitates a thoughtful approach to evaluation, assessment, and deployment of such models in medicine. This is especially important for LLMs, since these models are trained on vast amounts of text, including medical literature, which may contain errors and biases. This paper reports on the results of a study that compares the performance of LLMs to that of human experts in answering medical questions. The study shows that LLMs can perform well on some tasks, but struggle on others, particularly those requiring deep reasoning and clinical judgment. These findings highlight the need for careful evaluation and deployment of LLMs in medicine, and suggest that LLMs may be most useful as tools to support human decision-making, rather than as standalone diagnostic or treatment tools.

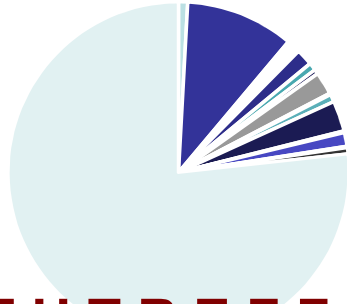
**Figure 1 | Med-PaLM 2 performance on MultiMedQA** Left: Med-PaLM 2 achieved an accuracy of 86.5% on USMLE-style questions in the MedQA dataset. Right: In a pairwise ranking study on 1066 consumer medical questions, Med-PaLM 2 answers were preferred over physician answers by a panel of physicians across eight of nine axes in our evaluation framework.

Singhal et al. Nature 2023; Singhal et al. arXiv preprint 2023



In a pairwise study, Med-PaLM 2 answers were preferred to physician answers across eight of nine axes considered.

FDA AI/ML-Enabled Medical Devices

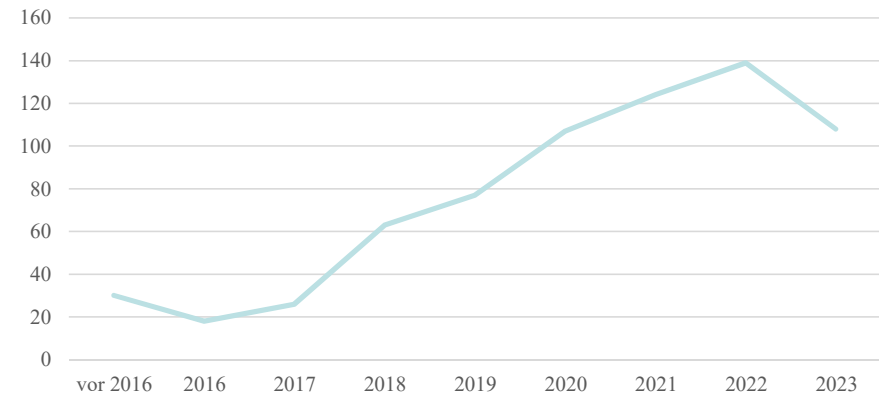


## NICHT ZUTREFFEND:

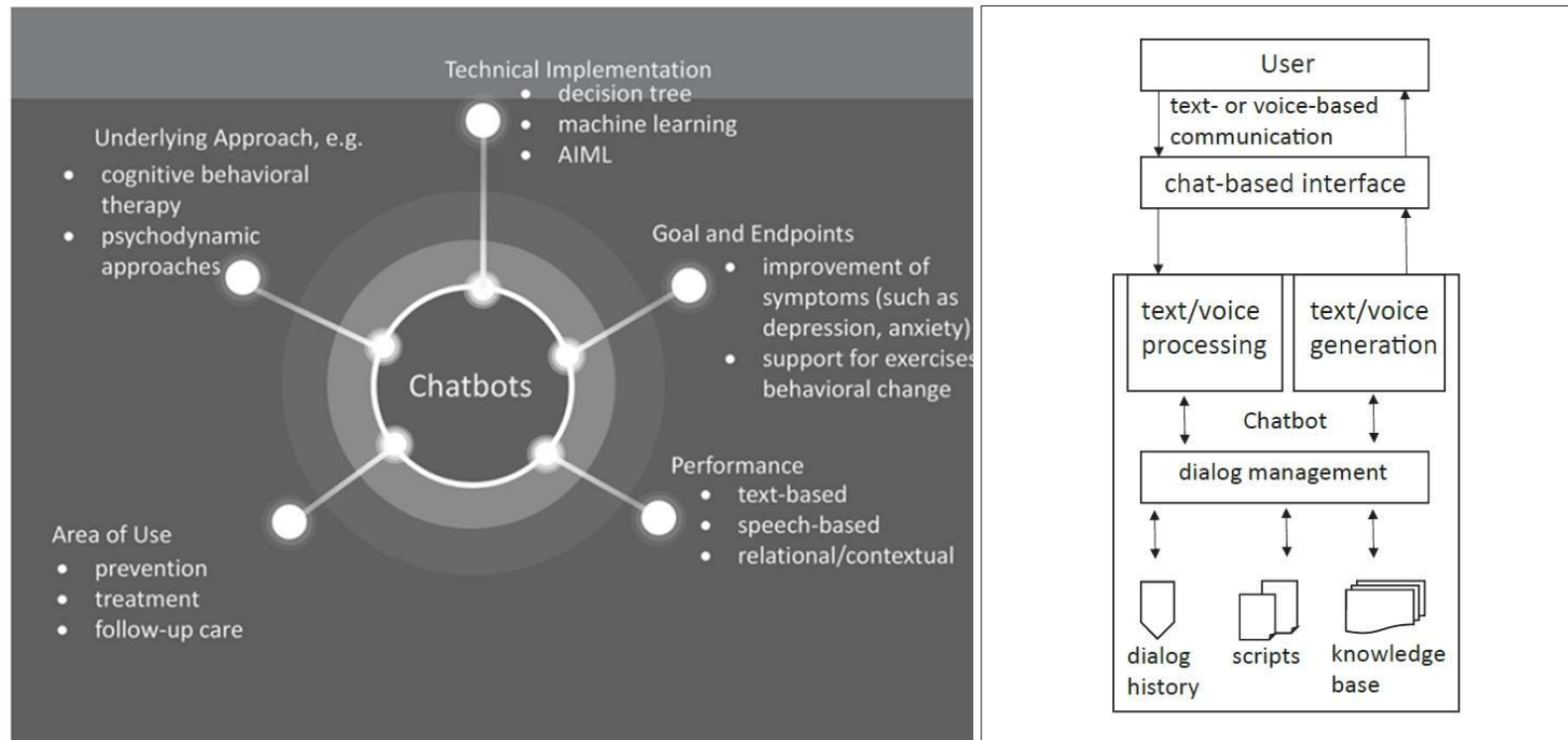
„Einer der wenigen medizinischen Handlungsbereiche, in denen KI-basierte Systeme für einzelne bereits ärztliches bzw. anderes Gesundheitspersonal – jedenfalls de facto – weitgehend oder vollständig ersetzen, ist die Psychotherapie. Hier sind seit einigen Jahren Instrumente in Entwicklung und Nutzung, meist in Form von bildschirmbasierten Apps, die auf algorithmischer Basis eine Art von Therapie anbieten und vielfach frei erhältlich sind. (S. 34/35)“

[www.fda.gov/Medical-devices](http://www.fda.gov/Medical-devices); Deutscher Ethikrat Mensch und Maschine – Herausforderungen durch KI 2023

FDA AI/ML Enabled Medical Devices



## Chatbots in der Psychotherapie: Gegenstand



**Fig. 2.** Graphic representation of the technical implementation of chatbots.

# ChatGPT

**Ich**

Hi, weißt du was, letzte Woche war ich im Urlaub

**ChatGPT**

Hallo! Es klingt großartig, dass du im Urlaub warst! Wo hast du deinen Urlaub verbracht und wie war es? ...

**Ich**

Ich war in Irland und war total fasziniert von den hohen Klippen

**ChatGPT**

Das klingt fantastisch! Irland ist bekannt für seine atemberaubende Landschaft... Vermutlich meinst du die Cliffs of Moher... [GPT besticht mit ausführlicher Wissenswiedergabe]... Gibt es noch etwas...

**Ich**

ja, genau 200 Meter freier Fall bis zum Aufschlag im Meer. Ich bin ja nicht mehr körperlich so fit und da dachte ich mir: Jetzt noch ein paar Schritte gehen und alles wird besser

**ChatGPT**

Es klingt so, als ob du eine mutige Entscheidung getroffen hast, die Klippen zu erkunden, trotz möglicher körperlicher Herausforderungen. Es ist bewundernswert, wie du dich dazu entschlossen hast, die Schönheit der Natur zu erleben und die Klippen zu besichtigen...etc. etc...

**Ich**

Und was denkst du, soll ich noch die paar Schritte gehen, so dass alles besser wird?

**ChatGPT**

Ich bin ein Computerprogramm und habe keine persönlichen Gefühle oder Meinungen, daher kann ich keine Entscheidungen für dich treffen. Die Frage, ob du noch ein paar Schritte gehen solltest, hängt von vielen Faktoren ab, einschließlich deiner aktuellen körperlichen Verfassung, deiner Gesundheit, deiner persönlichen Ziele und Vorlieben.

Es ist wichtig, auf deinen Körper zu hören und sicherzustellen, dass du dich wohl fühlst und keine unnötigen Risiken eingehst. Wenn du das Gefühl hast, dass du noch ein paar Schritte gehen möchtest und es für dich machbar ist, um die Aussicht und die Erfahrung weiter zu genießen, dann könntest du es in Erwägung ziehen...

## Chatbots

CARO



Cass / Tess



Eliza

1966

Vickybot



SERMO



ADA

Replika



# Chatbot: Nutzungsdaten kommerziell verfügbarer Chatbots

**Table 1.** A descriptive overview of the selected 10 mobile mental health apps with a built-in chatbot technology.

App	Number of ratings in Apple App Store	Number of ratings in Google Play Store	Number of downloads in Google Play Store	Age rating (years)	Price
ADA	125	323,000	≥5 million	≥17	Free
Chai	27,900	34,000	≥1 million	≥17	Free with in-app purchases
Elomia	193	N/A <sup>a</sup>	N/A	≥12	Free with in-app purchases
Mindspa	107	2970	≥500,000	≥17	Free with in-app purchases
Nuna	68	93	≥10,000	≥4	Free with in-app purchases
Serenity: Guided Mental Health	20	146	≥10,000	≥12	Free
Stresscoach	None	495	>10,000	>12	Free
Woebot	5500	11,800	≥500,000	≥12	Free
Wysa	13,500	126,000	≥1 million	≥12	Free with in-app purchases
Youper–Self Care Friend	14,400	49,100	≥1 million	≥12	Free with in-app purchases

<sup>a</sup>N/A: not applicable.

## Woebot und Wysa 2018 tested by BBC



Woebot did not appear to be able to detect an illegal act



Wysa suggested the "victim" speak to an adult, but did not appear to appreciate the severity of the complaint

<https://www.bbc.com/news/technology-46507900>

## Chatbots - Evidenz

- Zweck: Diagnostik und Intervention
- Indikation: Depression, Autismus, PTSD, Angst, Substanz uvm...
- 92% Regel-basiert
- > 50% Schreib-basiert
- Publikationen > 44% USA, 12% Japan, 8% Australien

Abd-Alrazaq et al. Int J Med Inf 2019

Table 2

Features of chatbots in the included studies.

Characteristics	Number of studies
<b>Purpose<sup>2</sup></b>	Therapy:17 Training:12 Screening:10 Self-management:7 Counselling:5 Education:4 Diagnosing:2
<b>Platform</b>	Stand-alone software:37 Web-based:16
<b>Response generation</b>	Rule-based:49 Artificial intelligence:4
<b>Dialogue initiative</b>	Chatbot:46 Both:7 User:0
<b>Input modality</b>	Written:26 Spoken & Visual:10 Spoken:8 Written & Spoken:7 Written & Visual:2
<b>Output modality</b>	Written, Spoken & Visual:28 Spoken & Visual:11 Written:9 Written & Visual: 5
<b>Embodiment</b>	Yes:44 No:9
<b>Targeted<sup>3</sup> disorders</b>	Depression:16 Autism:10 Post-traumatic stress disorder:7 Mental disorders:7 Anxiety:7 Substance use disorders:5 Schizophrenia:3 Dementia:3 Phobia:2 Stress:2 Eating disorders:1

Tips:

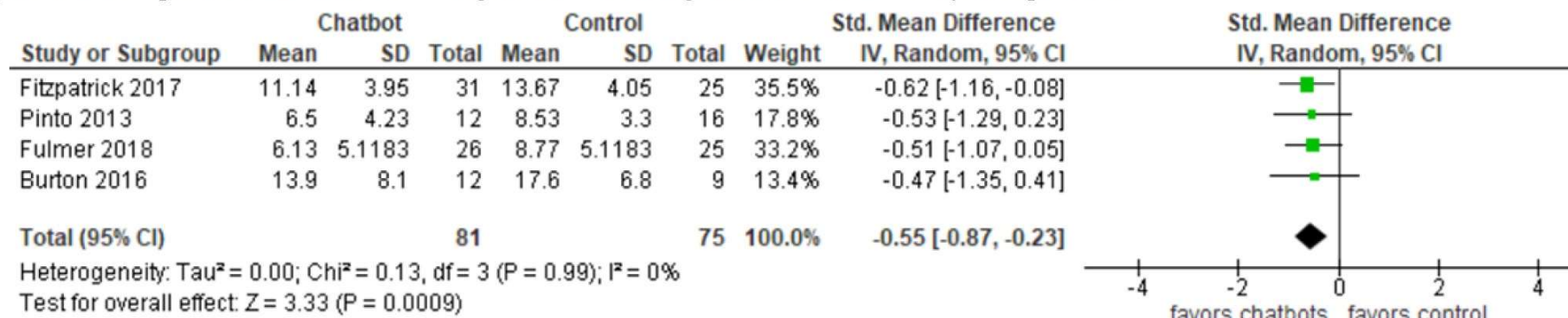
<sup>1</sup> It is the number given for each included study as shown in Appendix C.

<sup>2</sup> Numbers do not add up as 4 chatbots had two different purposes.

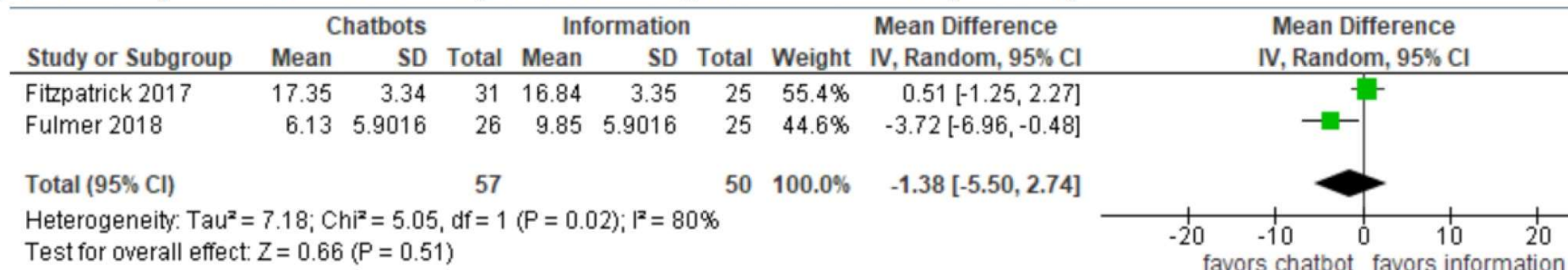
<sup>3</sup> Numbers do not add up as several chatbots focused on more than one mental disorder.

## Chatbots - Evidenz

**Figure 4.** Forest plot of the 4 studies assessing the effect of using chatbots on the severity of depression.



**Figure 5.** Forest plot of the 2 studies assessing the effect of using chatbots on the severity of anxiety.



## Patient:innensicht

Kategorien	Zentrale Erkenntnisse / Aspekte
Nützlichkeit	Alltagsnähe, Lernunterstützung, Emotionsunterstützung, Coaching-Funktion, Alltagstransfer, Fürsorge, Selbstwirksamkeit etc.
Anwendungsfreundlichkeit	Intuitive Nutzung
Interaktionsverhalten	Freundlich, supportiv, optimierbar durch virtuelle Agenten und Sprach-Basierung sowie Verbesserung des Sprachoutputs (Tempo, Intonation, Empathie etc.)
Sprachverständnis	Fehlerfreies Sprachverständnis
Akzeptanz	Grundsätzlich akzeptabel, abhängig von Ziel und (datenschutzsensiblen) Features
Attraktivität und Unterhaltungswert	Optimierungsbedarf, da nicht durchgehend als attraktiv in der Nutzung wahrgenommen
Vertrauenswürdigkeit	Grundsätzlich möglich, noch wenig Daten
Inhalt	Teils als oberflächlich und irrelevant wahrgenommen; Personalisierung als positive Eigenschaft
Vergleich zwischen Alternativen	Vielfältige Sichtweisen und Vorlieben

## Chatbot: Ethische Aspekte

Ethische Herausforderungen	Beschreibung / Beispiele
Systemgerechtigkeit	z.B. <ul style="list-style-type: none"> <li>• Altersgerechtigkeit</li> <li>• Fehlerquellenfreies Trainingsmaterial</li> <li>• VA-Geschlecht, Aussehen, Sprache etc.</li> </ul>
Privatsphäre und Datenhoheit	EU regulatorisch relativ strikt <ul style="list-style-type: none"> <li>• GDPR</li> <li>• ALTAI</li> </ul> IEEE Ethically Aligned Design Deutscher Ethikrat: Mensch und Maschine
Missinterpretationen / „Halluzinationen“	Überzeugende Darstellung falscher Informationen
Abhängigkeit	Menschliche Autonomie vs. faktische Systemabhängigkeit
Transparenz / XAI	Nachvollziehbarkeit der Ergebnisse / Schlussfolgerungen / Aussagen
Urheberrechtsverletzungen	Umgang mit Verletzung von rechten Dritter bei Textgenerierung



## Chatbot: Fazit I

- International bereits umfassend in der Anwendung
- Die (Weiter-)entwicklung findet vor allem in den USA und asiatischen Ländern statt
- Regulatorische Prozesshürden bremsen KI-basierte digitale Interventionen in Deutschland
- Wissenschaftliche Evidenz zur Wirksamkeit und zum Risiko weitgehend ausstehend
- Vor allem Regel- und Text-basierte Chatbots

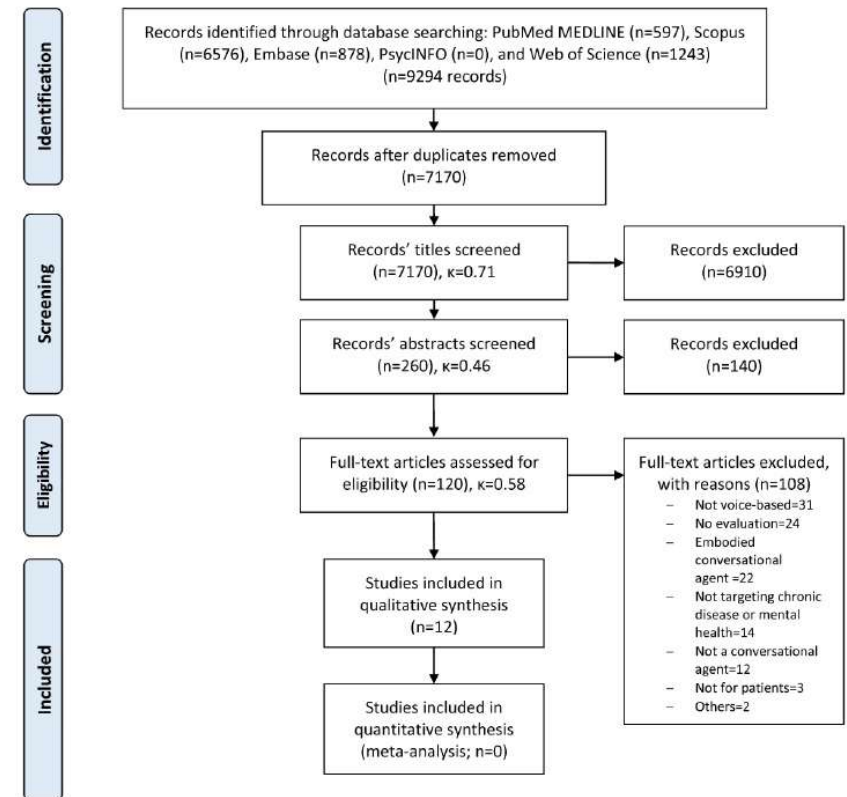
# Sprach-basierte Chatbots

Table 1. Overview and characteristics of included records.

Reference, publication year	Study aim	Type of study participants	Addressed medical condition	Voice-enabled device type	Intervention category
Amith et al (2019) [59]	Development and acceptance evaluation	Healthy adults with at least one child under the age of 18 years (n=16)	Cancers associated with HPV <sup>a</sup>	Tablet	Support
Amith et al (2020) [60]	Development and acceptance evaluation	Healthy young adults aged between 18 and 26 years (n=24)	Cancers associated with HPV	Tablet	Support
Boyd and Wilson (2018) [61]	Criterion-based performance evaluation of commercial conversational agent	Authors as raters (n=2)	Cancers associated with smoking	Smartphone	Support
Cheng et al (2019) [62]	Development and acceptance evaluation	Older adults (n=10)	Diabetes (type 2)	Smart speaker	Monitoring and support
Galescu et al (2009) [63]	Development and performance evaluation	Chronic heart failure patients (n=14)	Heart failure	Not specified	Monitoring
Greuter and Balandin (2019) [64]	Development and performance evaluation	Adults with lifelong intellectual disability (n=9)	Intellectual disability	Smart speaker	Support
Ireland et al (2016) [65]	Development and acceptance evaluation	Adults recruited on campus (n=33)	Parkinson disease, dementia, and autism	Smartphone	Monitoring
Kadariya et al (2019) [66]	Development and acceptance evaluation	Clinicians and researchers (n=16)	Asthma	Smartphone	Monitoring and support
Lobo et al (2017) [67]	Development and acceptance evaluation	Healthy adults working regularly with senior patients (n=11)	Heart failure	Smartphone	Monitoring and support
Ooster et al (2019) [68]	Development and performance evaluation	Normal hearing (n=6)	Hearing impairment	Smart speaker	Monitoring
Rehman et al (2020) [69]	Development and performance and acceptance evaluation	Adults affiliated with the university (n=33)	Diabetes (type 1, type 2, gestational) and glaucoma	Smartphone	Monitoring and support
Reis et al (2018) [70]	Criterion-based performance evaluation of a commercial conversational agent	Not specified (n=Not specified)	Depression	Not specified	Support

<sup>a</sup>HPV: human papillomavirus.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of included studies.



# Alexano



By Dennis Döhning

## Chatbots: Weiterentwicklung



## Chatbot: Fazit II

Die nächste Generation an Mental Health Chatbots wird:

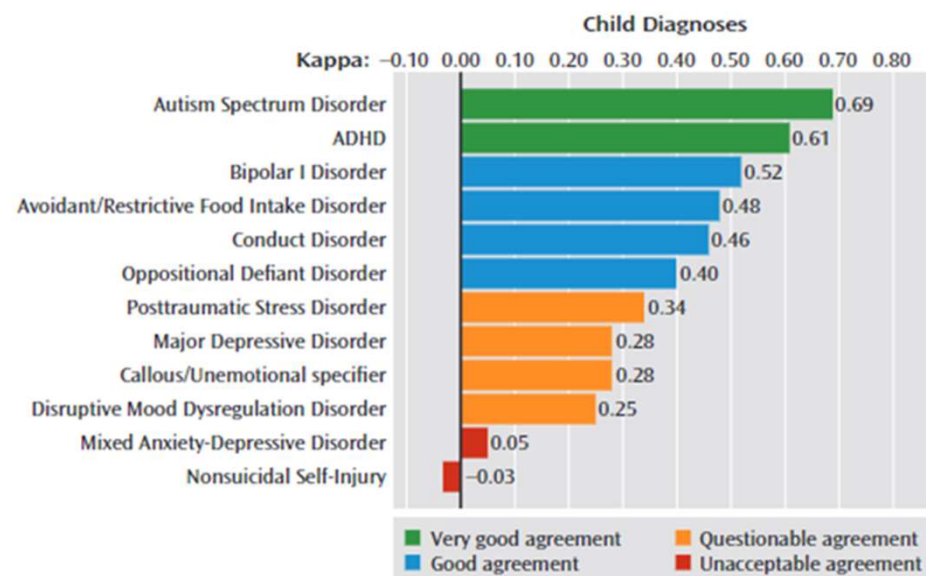
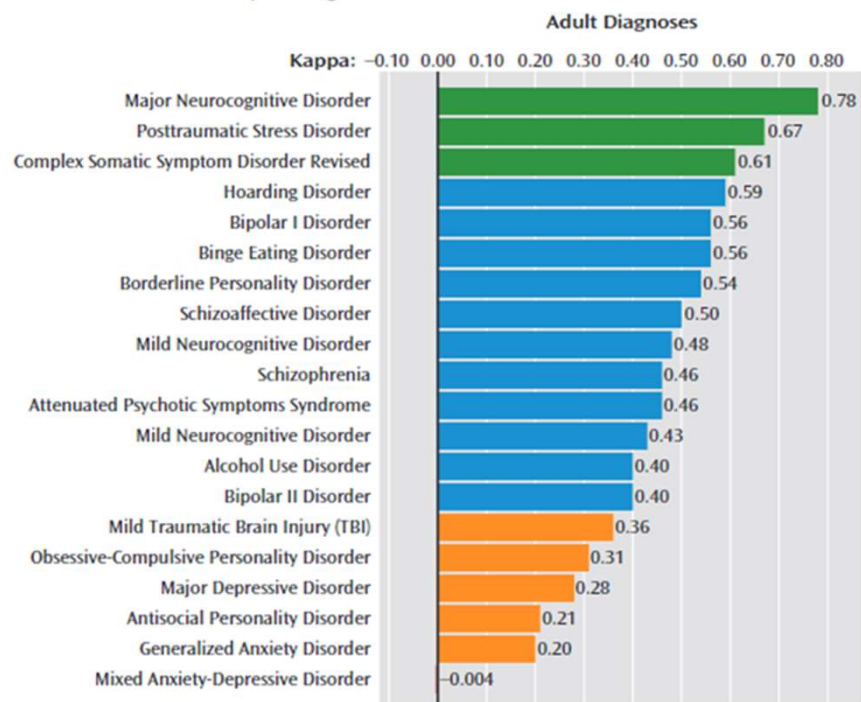
- KI-basiert
- Hoffentlich Evidenz-basierter bzgl. Nutzen und Risiken
- Sprach-basiert
- Ggf. unter Nutzung visueller Darstellungsmöglichkeiten (VR/AR)

...verbunden weiterhin mit der Frage, ob für einen substantiellen Teil der betroffenen Menschen mit Behandlungsbedarf akzeptabel und verbunden mit einer ausreichenden Nutzungsadhärenz

# Diagnostik – Psychotherapieprozess – KI

# Reliabilität Psychische Störungen DSM-5

FIGURE 1. Interrater Reliability of Diagnoses From the Initial DSM-5 Field Trials<sup>a</sup>



<sup>a</sup>Some of the kappa statistics did not pass the criterion of a standard error less than 0.1. They are included here for illustrative purposes. See the field trial reports for further details (3–5).



## Better-Than-Average Effect

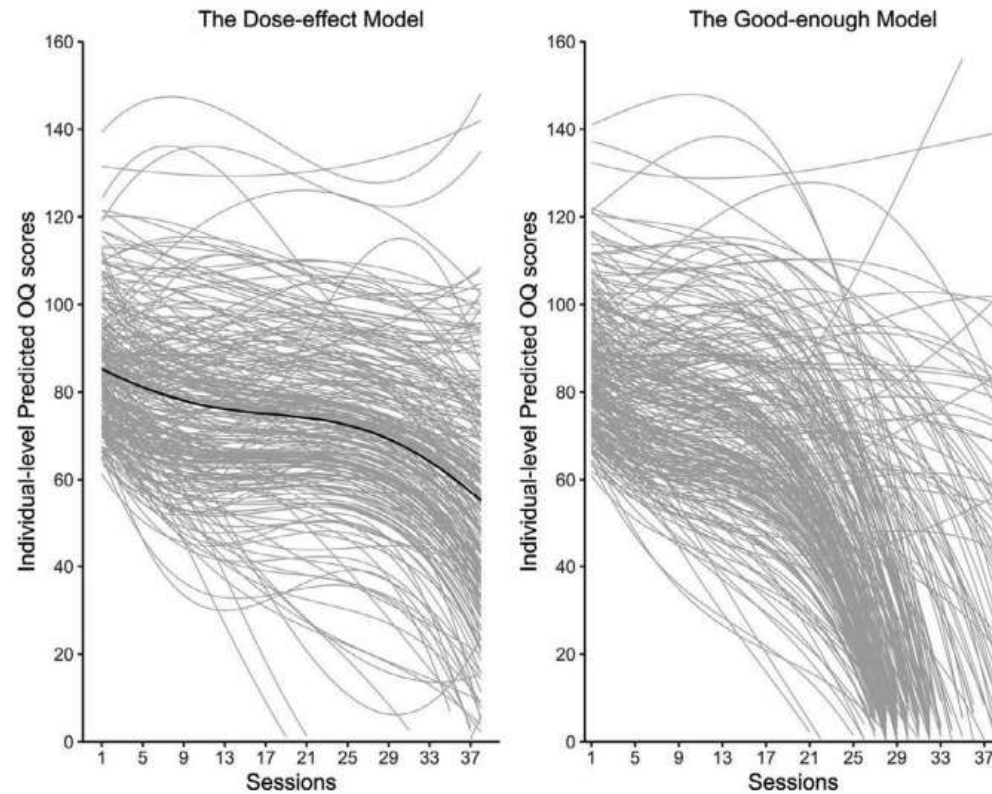
- Better-Than-Average-Effect robust über Studien, Populationen und Bereiche hinweg mit  $d_z = 0.78$  (95% CI [0.71, 0.84])<sup>1</sup>
- Psychotherapeut:innen bewerten ihre Leistung im Vergleich zu allen anderen Psychotherapeut:innen im Durchschnitt mit  $M = 79.11$  <sup>2</sup>
- Psychotherapeut:innen überschätzen die Rate der Patient:innen mit remittierter und verbesserter Symptomatik und unterschätzen die Rate der nicht verbesserten und verschlechterten Patient:innen <sup>3</sup>

## Routine Outcome Monitoring

- Datenbasiertes Monitoring des Behandlungsprozess und -erfolg
- Not on Track Monitoring mit Red Flag System
- Wirksamkeit erhöht, wenn ergänzt durch Clinical Support Tools
- Präzision, Differenzierungsgrad und ggf. auch Wirksamkeit verbessern sich bei multiplen Messungen
- Die Evidenz von ROM in Bereich der Kinder- und Jugendbehandlung unzureichend

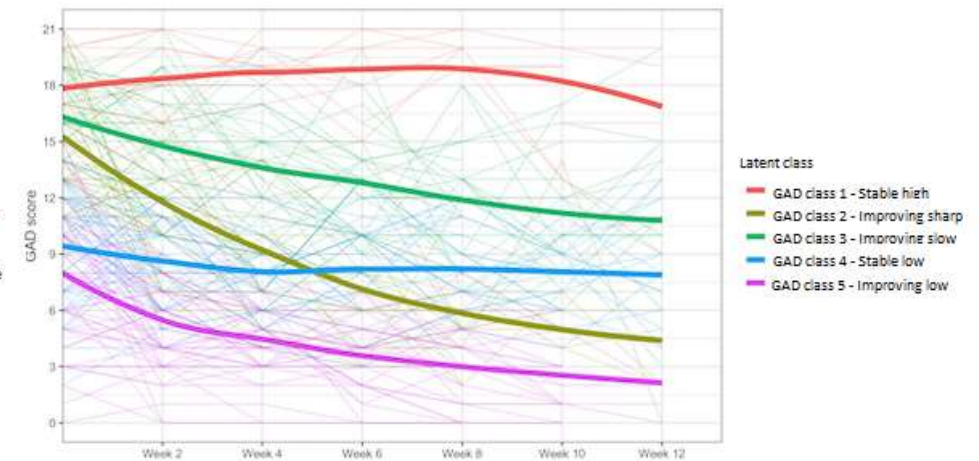
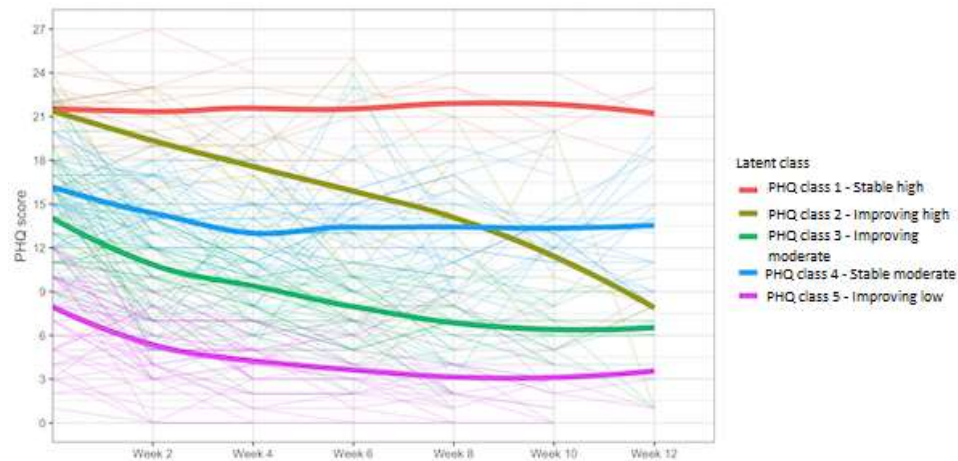
## Dose-Response vs. Good-Enough Level Model

**FIGURE 1** Individual-level predicted pattern of changes in Outcome Questionnaire-45 (OQ) scores for each participant (grey line) and the sample (solid line) over time. X-axis represents the sessions, and y-axis represents the individual-level predicted OQ scores based on the estimates for the dose-effect and good-enough models, accordingly. The panel (a) represents the dose-effect model, which does not account for the total number of sessions attended. The panel (b) represents the good-enough model, which takes the total number of sessions attended into account in the pattern of changes



## Trajectory Analysen

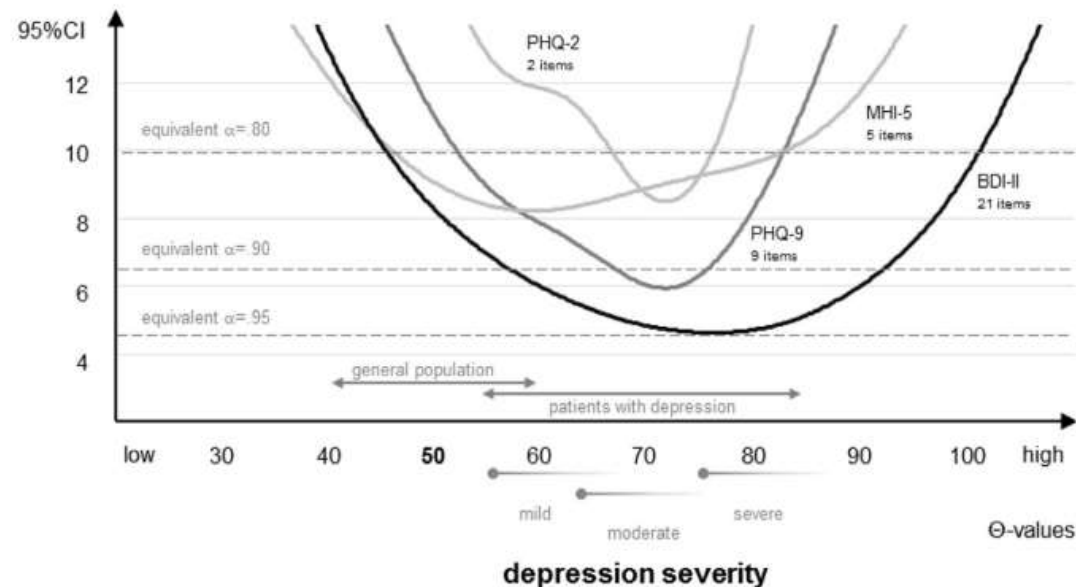
- Sekundäranalyse eines RCT zu einer therapeutisch begleiteten iCBT für Angst und Depression in UK. PHQ-9 und GAD-7
- Latent Class Growth Analysis ergaben ein 5-Klassenmodell



# Messpräzision verschiedener Instrumente

I. Wahl et al. / Journal of Clinical Epidemiology 67 (2014) 73–86

79



**Fig. 1.** Measurement precision and range. The x-axis shows the patient's score indicating his/her level of depression severity. In IRT terminology, this score is referred to as the " $\theta$  score." To make the different depression tools comparable, all instruments are scored on the same metric as determined by the developed item bank. The y-axis shows the 95% confidence interval (CI) of the patient score; the smaller the y-value, the higher the precision of the score. The dotted lines show the CI that would be comparable with an internal constancy of Cronbach  $\alpha$  of 0.80, 0.90, and 0.95 for illustrative purposes. PHQ, Patient Health Questionnaire; MHI-5, Mental Health Index; BDI-II, Beck Depression Inventory, second edition.

## ML-optimierter Einsatz von Depressionsscreener

- Sekundäranalyse zu RCT bei Patient:innen mit Rückenschmerz und Depressivität
- 1030 SCID-Diagnosen; QIDS-16, PHQ-9, HAM-D-17

	AUC of ROC for the best ML	AUC of ROC for sum-score	$\Delta$ AUC	95%-CI of the difference	$p^a$
QIDS-16—basic <sup>b</sup>	0.935	0.900	0.035	0.02 to 0.05	<0.001
QIDS-16—extended <sup>c</sup>	0.937	0.900	0.036	0.02 to 0.06	<0.001
HAM-D-17	0.876	0.873	0.003	-0.02 to 0.03	0.847
PHQ-9	0.832	0.818	0.014	0.00 to 0.02	0.009

<sup>a</sup> Two-sided bootstrap test for ROC curves within in the testing set, indicating the probability of whether the true difference in AUC is different from 0.

<sup>b</sup> Logistic regression model containing only the QIDS-16 items as features.

<sup>c</sup> Logistic regression model containing QIDS-16 items, the extended feature set (e.g. mean of items), age, and gender.

## Computer-adaptiver Test (CAT)

**Ziel** ist die Konstruktion eines „maßgeschneiderten Tests“ für jede:n Patient:in

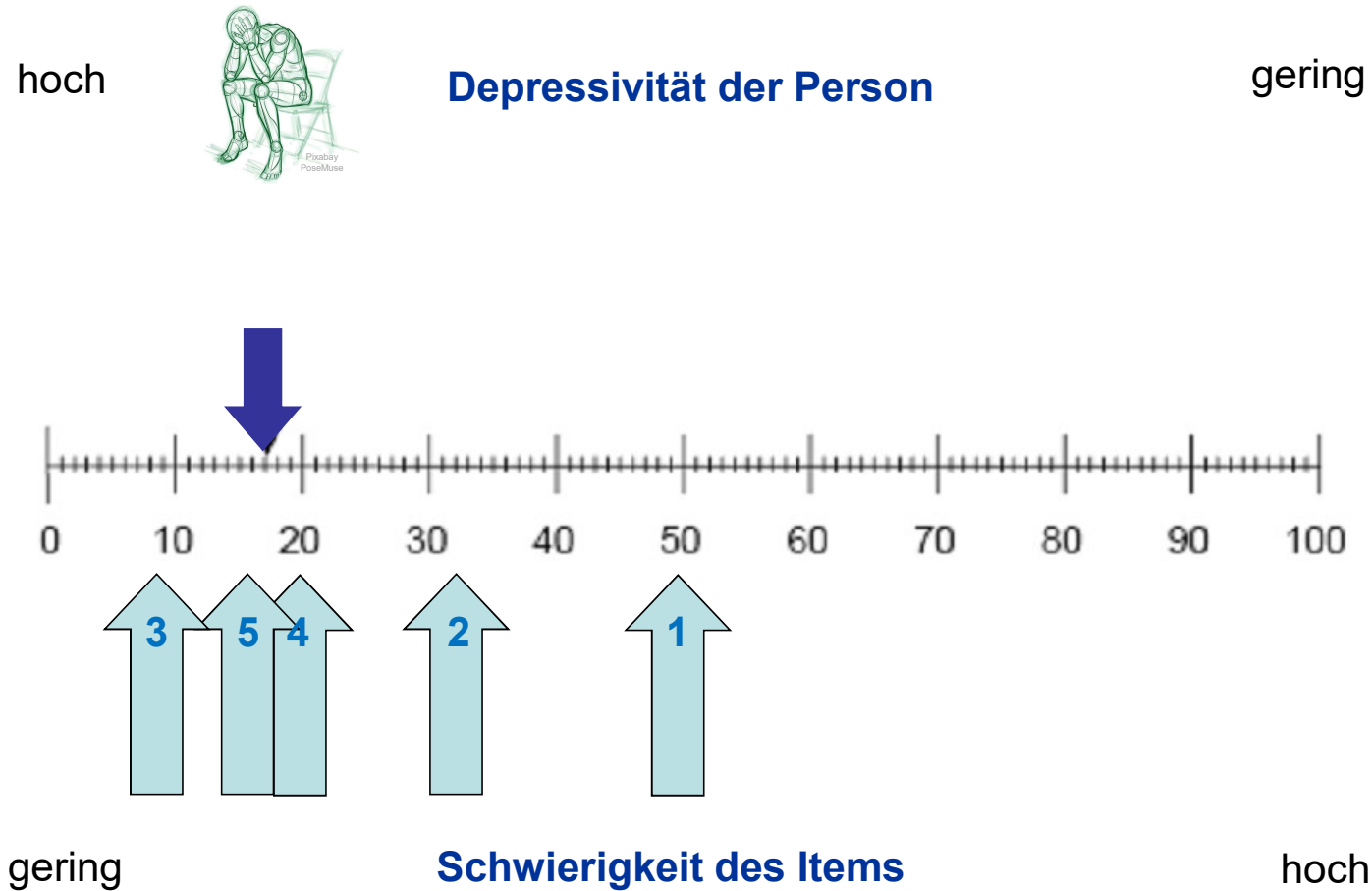
### Vorteile

- Verkürzung der Testdauer um 50-90%
- Hohe Messpräzision
- unmittelbare Auswertung

**Grundlage:** IRT-basierte, eindimensionale Itembanken

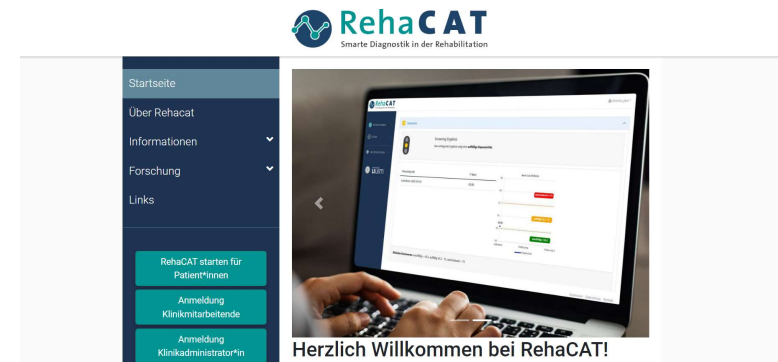


# Computer-adaptiver Test (CAT)



# CAT-basierte ROM-Systeme in der Routineversorgung

- RehaCAT  
(<https://rehacat.de/de/>)

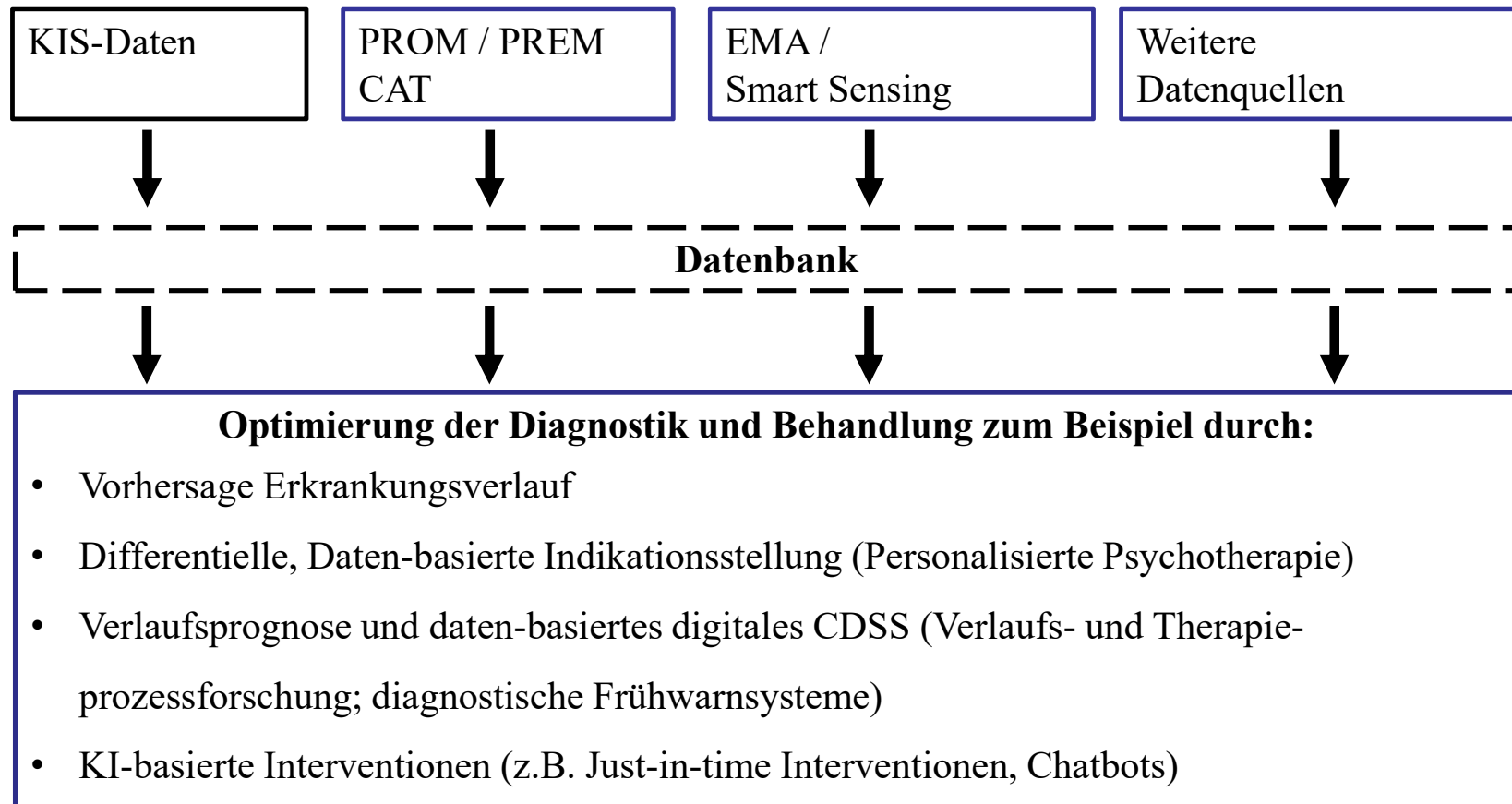


- SYSDOK

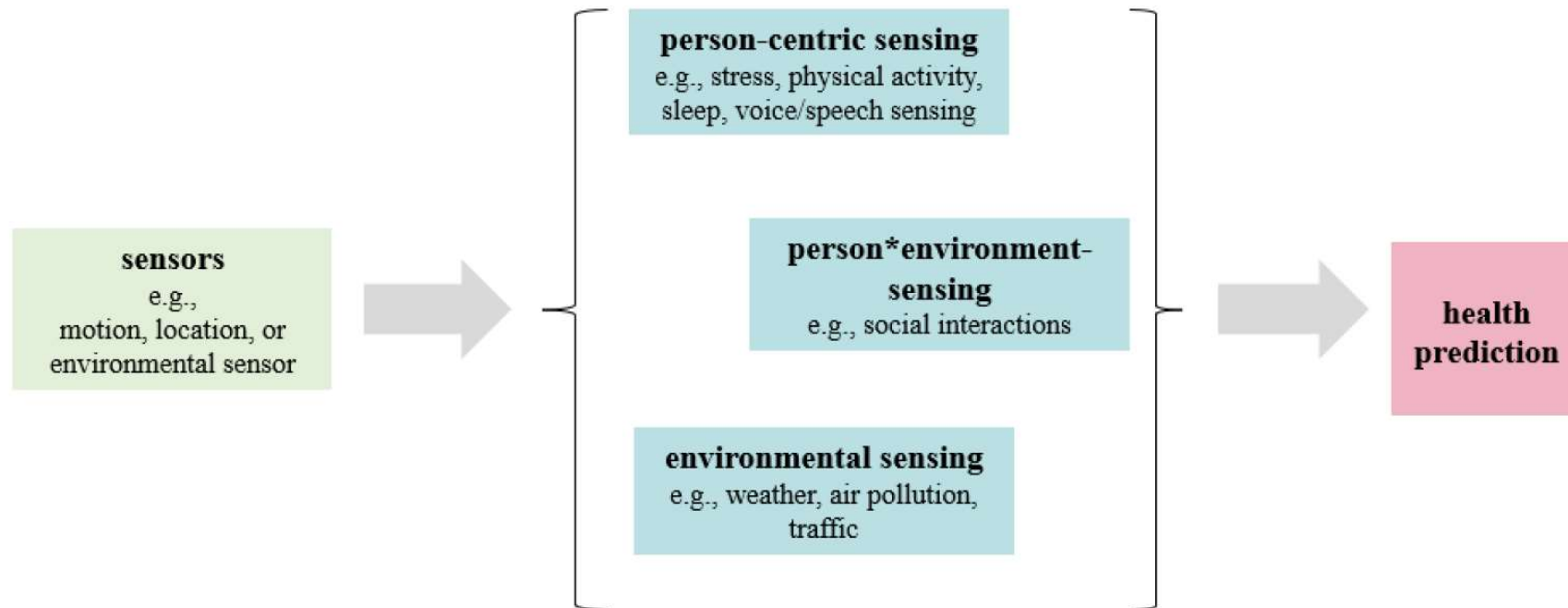
⇒ „Reallabor“-Ansatz



## Psychodiagnostik: Auf dem Weg zu KI-basierten Clinical Diagnostic Support Systems (CDSS)



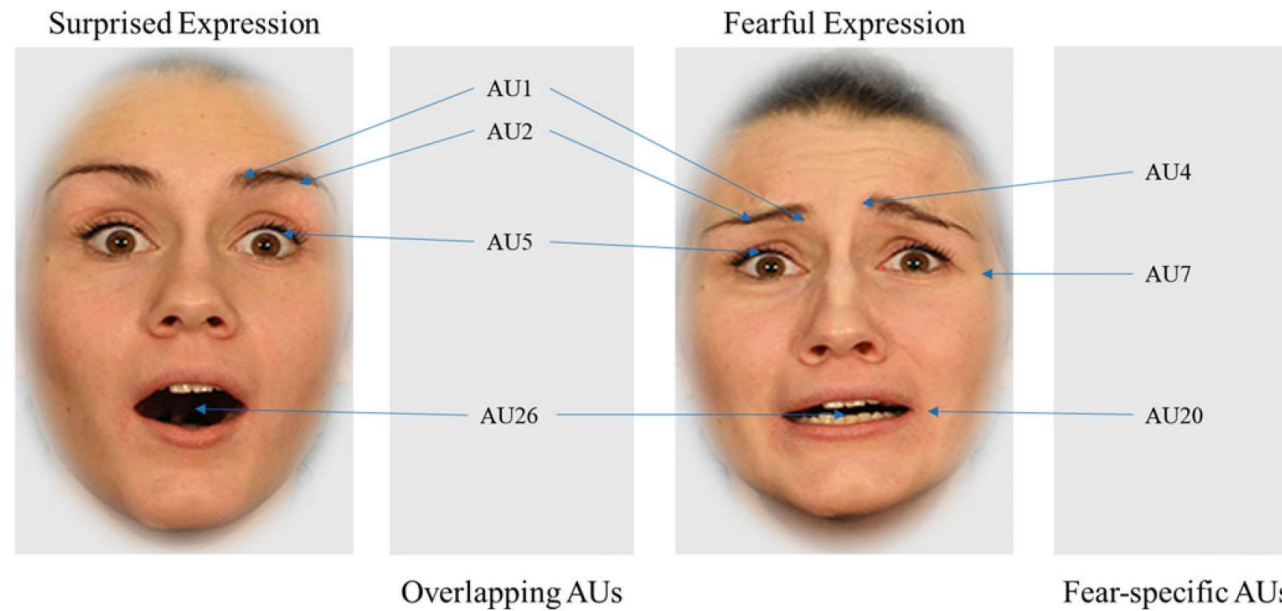
## Vom Sensor zur Gesundheitsversorgung



*Figure 2.* From sensors to prediction of health.

# Gesichtserkennung-basierte Emotionsanalyse

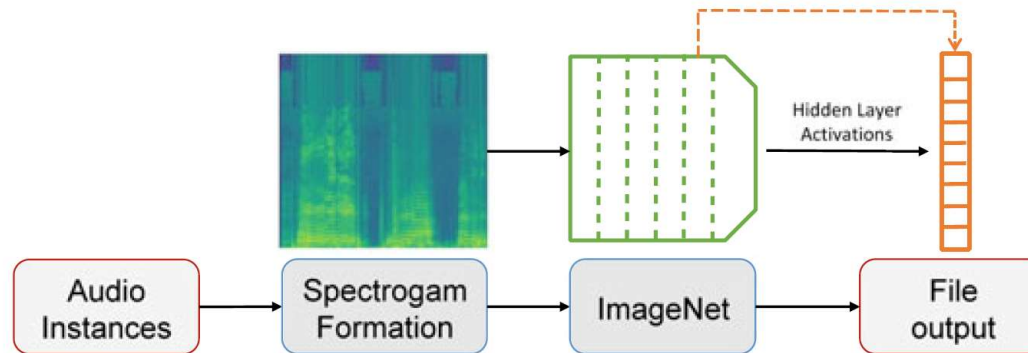
## Automatisierte Gesichtserkennung als diagnostische Information



**Fig. 3.1** An example of the Action Unit (AU) overlap of fearful and surprised expression, as described in the FACS/AID (Ekman et al. 1998). AU1: inner brow raiser, AU2: outer brow raiser, AU4: brow lowerer, AU5: upper lid raiser, AU7: lid tightener, AU20: lip stretcher, AU26: jaw drop

## Stimm- und Sprachanalysen

Automatisierte Stimm- und Sprachanalysen (z.B. Valenz und Arousal)

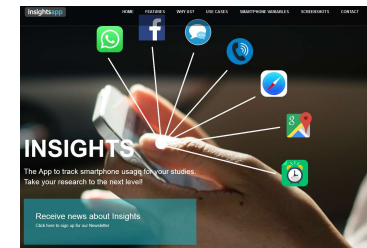
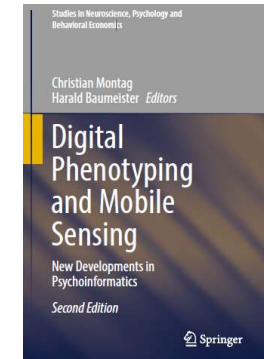


**Fig. 9.3** An overview of extracting DEEPSPECTRUM features. Audio samples are converted to spectrogram images and then feed into a pre-train image classification Convolutional Neural Networks (CNNs); the activations of fully connected layers are then used as a feature representation for subsequent classification

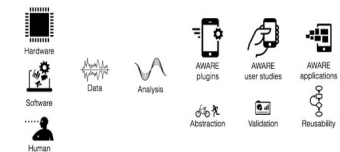
# Digital Phenotyping and Mobile Sensing

## Proof-of-Concept Studies

- Depression:
  - Variabilität der besuchten Orte ( $\beta$  -.21), Schlafzeit ( $\beta$  .24), Zeit im Bett ( $\beta$  .26)<sup>1</sup>
  - Smartphone-Nutzungszeit (-.019), Anrufdauer (-.016)<sup>2</sup>
  - Häufigerer Wortgebrauch von: Ich, wir, Negierungen, Angst und Essen<sup>3</sup>
  - EMA and Smart Sensing kombiniert bestes Regressionsmodell ( $R^2 = 45.15\%$ )<sup>4</sup>
- Angst:
  - Wach nach Schlafbeginn ( $\beta$  .23), HRV ( $\beta$  .26)<sup>2</sup>
- Insomnie
  - Smartphone-Nutzung AUC .57 / .58 (ML Prediction: Random Forest / Naive Bayes)<sup>5</sup>



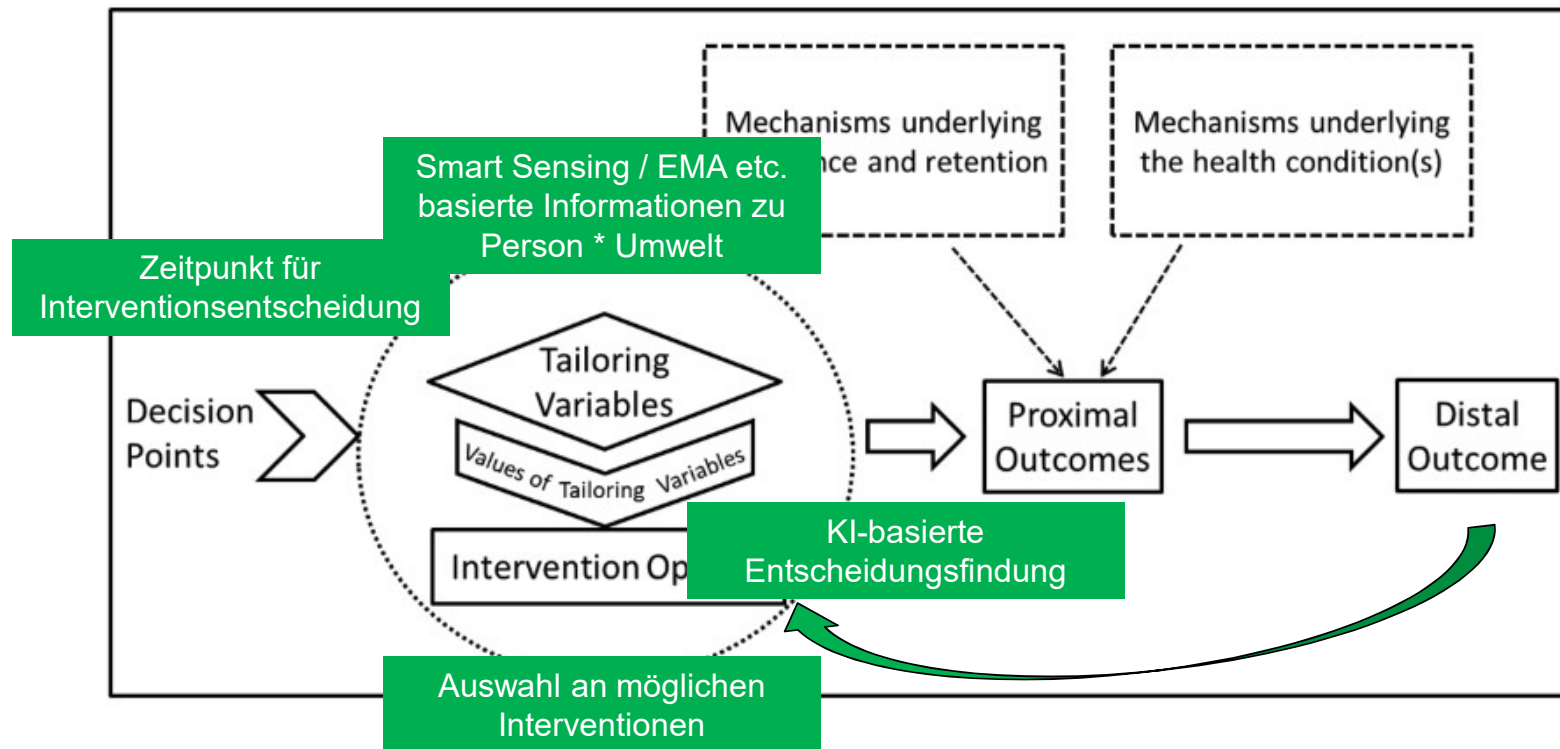
## AWARE- Smart Sensing



1 Moshe et al. Frontiers Psychiatry 2021; 2 Messner et al. Verhaltenstherapie 2019; 3 Rather et al. Interspeech 2018; 4 Terhorst et al., submitted; 5 Simon et al., submitted; 6 Montag & Baumeister, DPAMS Springer 2023



# Just-in-time Intervention



## Fazit: KI und Psychotherapie

- Bislang Vision und noch wenig Realität
- Wahrscheinliche zeitnahe Anwendungsfelder
  - a) Diagnostische und Behandlungsprozess-bezogene Support-Systeme (CDSS)
  - b) Stand-Alone Interventionen (Chatbot / DIGA 4.0 (JITAI))
  - c) KI-gestützte Blended Psychotherapy Angebote
  - d) Patient:innen-Verwaltung, Telematikinfrastuktur etc.
  - e) Forschungserkenntnisse zu psychischen Störungen und deren individueller Behandlungsoptimierung (Precision Psychotherapy)
- Produkte werden wahrscheinlich aus USA und Asien kommen und deren Entwicklung und Erforschung wird weitgehend Industrie-basiert und nicht öffentlich-universitär erfolgen

**Vielen Dank für Ihr Interesse**  
**...und ein Dankeschön an das gesamte Team!**  
**...und ein Dankeschön an alle Kooperationspartner!**

