Potenzial und Gefahren bei der Nutzung von KI bei Gesundheitsfragen

Dr. med. Oscar Freyer · EKFZ for Digital Health, TUD Dresden · UCON, 24. Juni 2026

Universitätsklinikum Carl Gustav Carus DIE DRESDNER.

TUD

EKFZ Digital Health

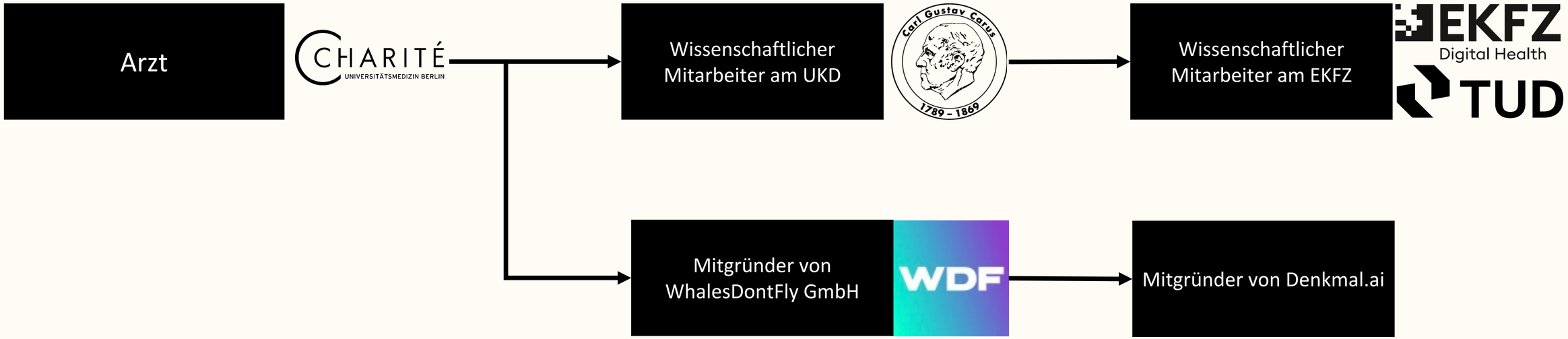
Funded by the European Union

Interessenkonflikte

- **Honorare**
  - ESMO
  - Springer Nature (Associate Editor, npj Digital Medicine)
- **Beratung**
  - Prova Health Ltd.
  - SFDA
- **Beteiligung & Leitungsfunktion**
  - WhalesDontFly GmbH

Funded by the European Union

Wer bin ich?


Arzt
CHARITÉ
UNIVERSITÄTSMEDIZIN BERLIN
Wissenschaftlicher Mitarbeiter am UKD
Carl Gustav Carus
1789 – 1869
Wissenschaftlicher Mitarbeiter am EKFZ
EKFZ
Digital Health
TUD
Mitgründer von WhalesDontFly GmbH
WDF
Mitgründer von Denkmal.ai

Wer bin ich?

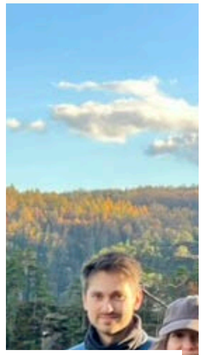
Arzt

icher
n EKFZ

enkmal.ai

EKFZ Digital Health TUD

Wer bin ich?



NEJM AI

DOI: 10.1056/Aioa2500833

ORIGINAL ARTICLE

From Advice to Action — Real-World Behavior of Patients Using an Integrated Diagnostic Decision Support System for Navigating the Health Care System

Fabienne Cotte, M.D.,[1] Filipa Dias Lourenço, M.D.,[2] Miguel Paiva Pereira, M.D.,[2] Nisha Kini, M.B.B.S., M.P.H.,[1] Marcel Schmude, M.D.S.,[1] Andreia Pimenta, Ph.D.,[1] Athena Lemesiou, Ph.D.,[1] Stephen Gilbert, Ph.D.,[1,4] Tauseef Mehrali, M.B.Ch.B. (Hons.),[1] Micaela Seemann Monteiro, M.D.,[2] and Pedro Flores, Ph.D.[2]

Received: July 24, 2025; Revised: December 9, 2025; Accepted: January 20, 2026; Published: March 5, 2026

Abstract

BACKGROUND Artificial intelligence (AI)–powered digital front door tools are increasingly being used to guide patients to appropriate care and alleviate health care system pressure. However, most evaluations offer limited insight into stated care intent, real-world behavior, or care appropriateness.

METHODS The E-Health Self-Symptom Assessment as a Front Door and Facilitator of Care (ESSENCE) study was a prospective real-world quality improvement evaluation embedded in Portugal's largest private health care network (CUF). Adults using the Ada Health diagnostic decision support system through the myCUF app reported their care intent before and after symptom assessment. We tracked actual behavior through electronic health records and surveys. Physician panels retrospectively assessed the appropriateness of intended and observed care.

RESULTS A total of 1470 adults (≥18 years of age; mean age, 38.5 years; 57.7% female) were enrolled. Of the 1338 participants with pre- and postassessment intentions, 33.0% revised their planned care level immediately after assessment. Uncertainty decreased from 12.6% to 5.0% (P<0.001). Among 721 participants with observed behavior, 59.1% changed their care pathway: 28.9% de-escalated, 17.2% escalated, and 13.0% resolved prior uncertainty. Primary care consultations increased from 16.3% to 42.1% (P<0.001), whereas specialist visits decreased from 49.7% to 29.8% (P<0.001). Among nonemergency participants with preassessment intentions and sufficient clinical documentation (n=382), appropriate care increased from 29.8% preassessment to 64.4% postbehavior (95% confidence interval, 27.8 to 41.4; P<0.001). Of the 96 participants who planned an emergency department (ED) visit, 38.5% selected lower-acuity care after assessment. In the subset with clinician-rated follow-up, 93% (27 of 29; 95% CI, 78.0 to 98.1%) were judged to have appropriately avoided an unnecessary ED visit.

CONCLUSIONS Integrating an AI-supported symptom assessment and follow-up service options within a digital front door was associated with shifts in patient intentions and

Fabienne Cotte and Filipa Dias Lourenço contributed equally to this article.

The author affiliations are listed at the end of the article.

Pedro Flores can be contacted at pedro.flores@sapo.pt.

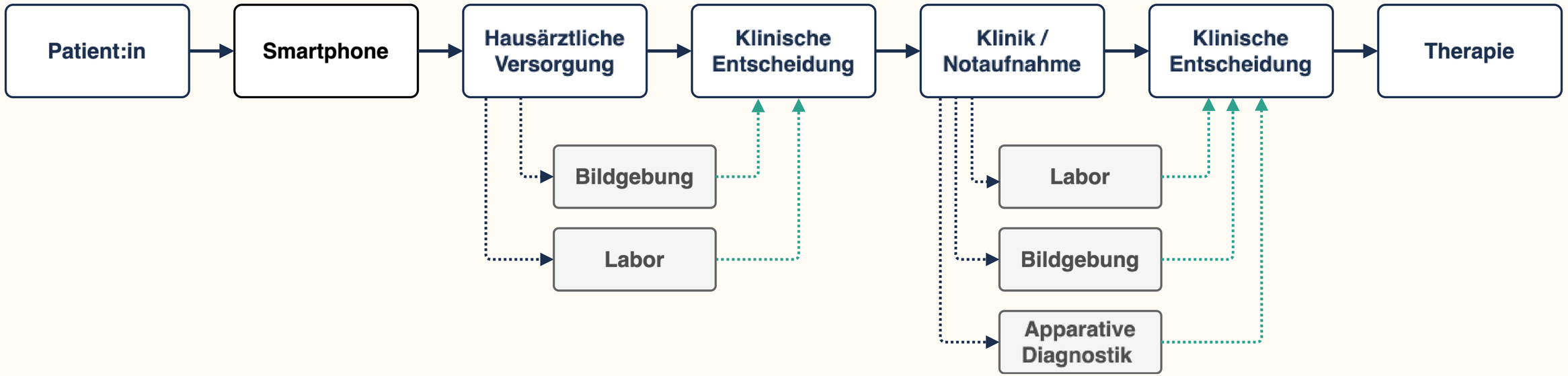
Universitätsklinikum Carl Gustav Carus DIE DRESDNER.

TUD

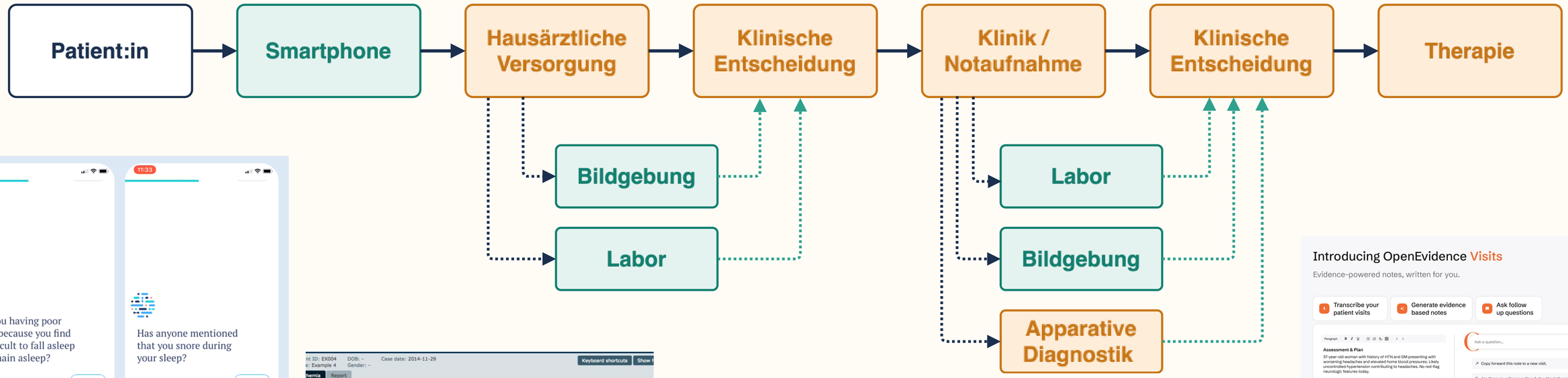
EKFZ Digital Health

Funded by the European Union

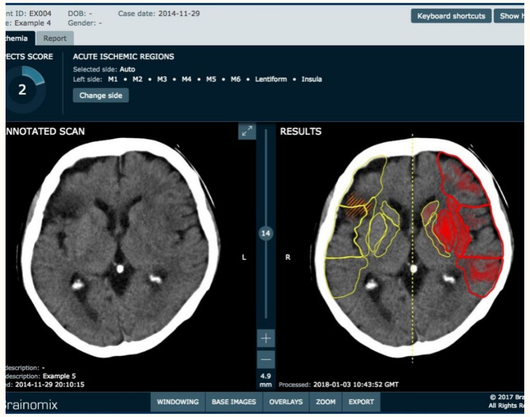
KI in der kritischen Versorgung



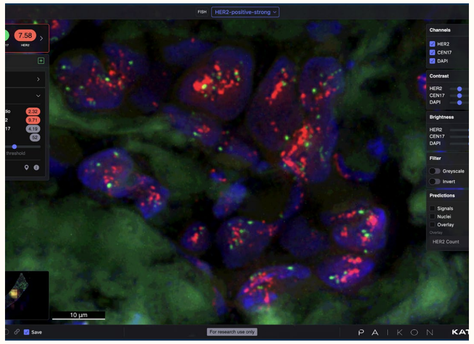
KI in der kritischen Versorgung



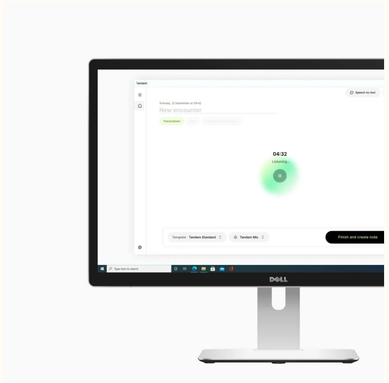
Quelle: https://www.uisources.com/explainer/ada-diagnosing-via-chat-bot



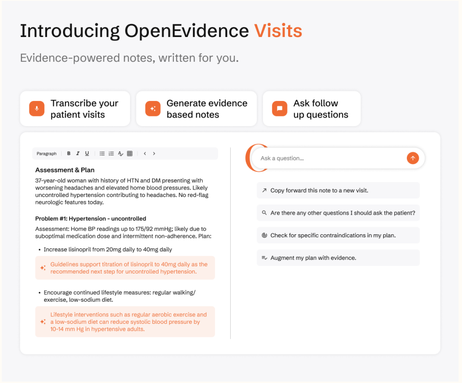
Quelle: Brainomix



Quelle: Katana Labs



From: Tandem Health

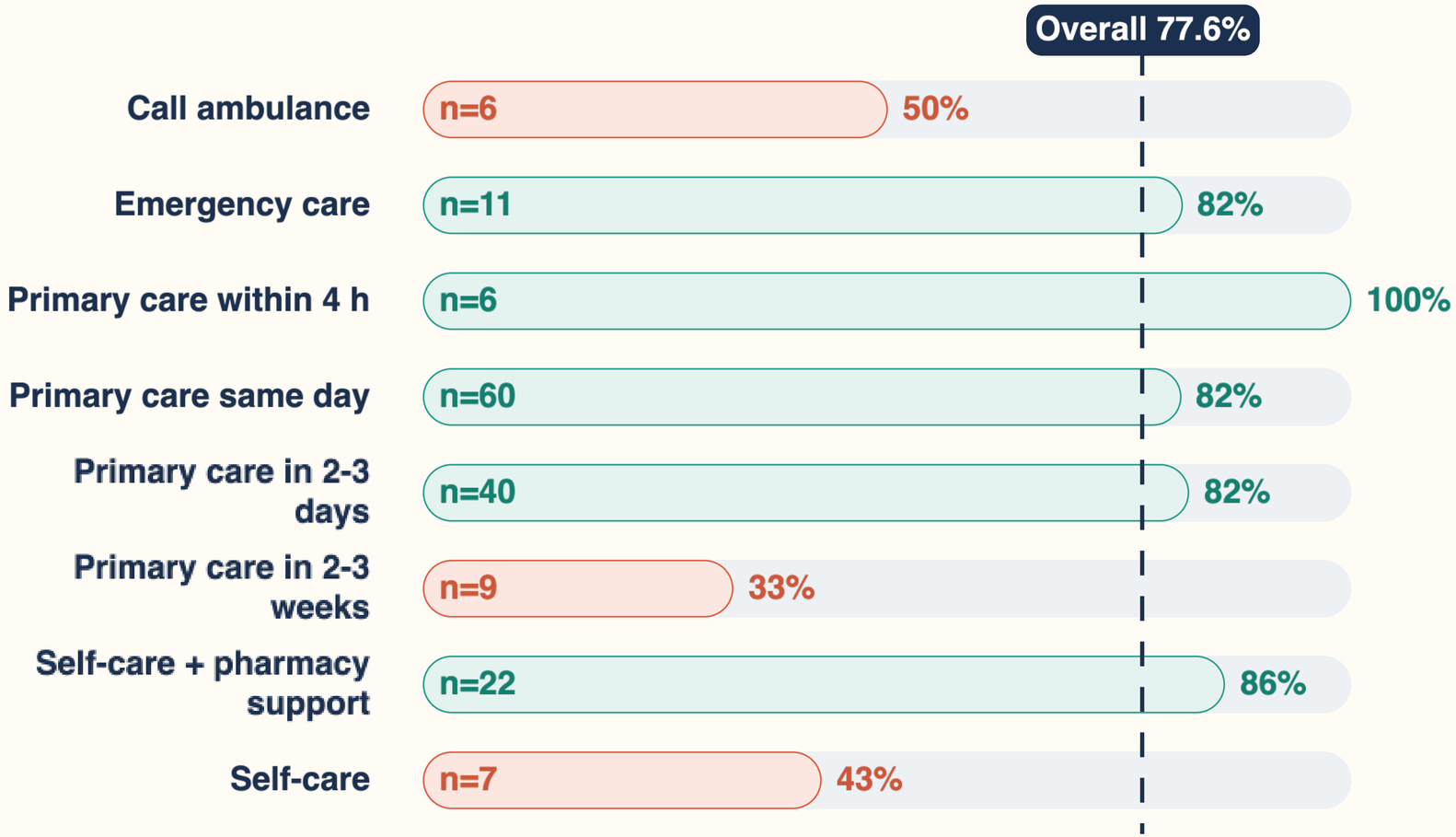


From: OpenEvidence

Reale Performance

	MTS 3 (yellow)	MTS 4 (green)	MTS 5 (blue)
Call ambulance	23 6.1%	68 17.9%	5 1.3%
Emergency care	22 5.8%	91 24.1%	4 1.1%
Primary care within 4 h	10 2.6%	20 5.3%	2 0.5%
Primary care same day	10 2.6%	60 15.9%	3 0.8%
Primary care 2-3 days	6 1.6%	34 8.9%	2 0.5%
Primary care 2-3 weeks	2 0.5%	0	0
Self-care or pharmacy	2 0.5%	11 2.9%	0
Self-care	0	3 0.8%	0

Overtriage 216 (57.1%) | Match 128 (33.9%) | Undertriage 34 (8.9%)

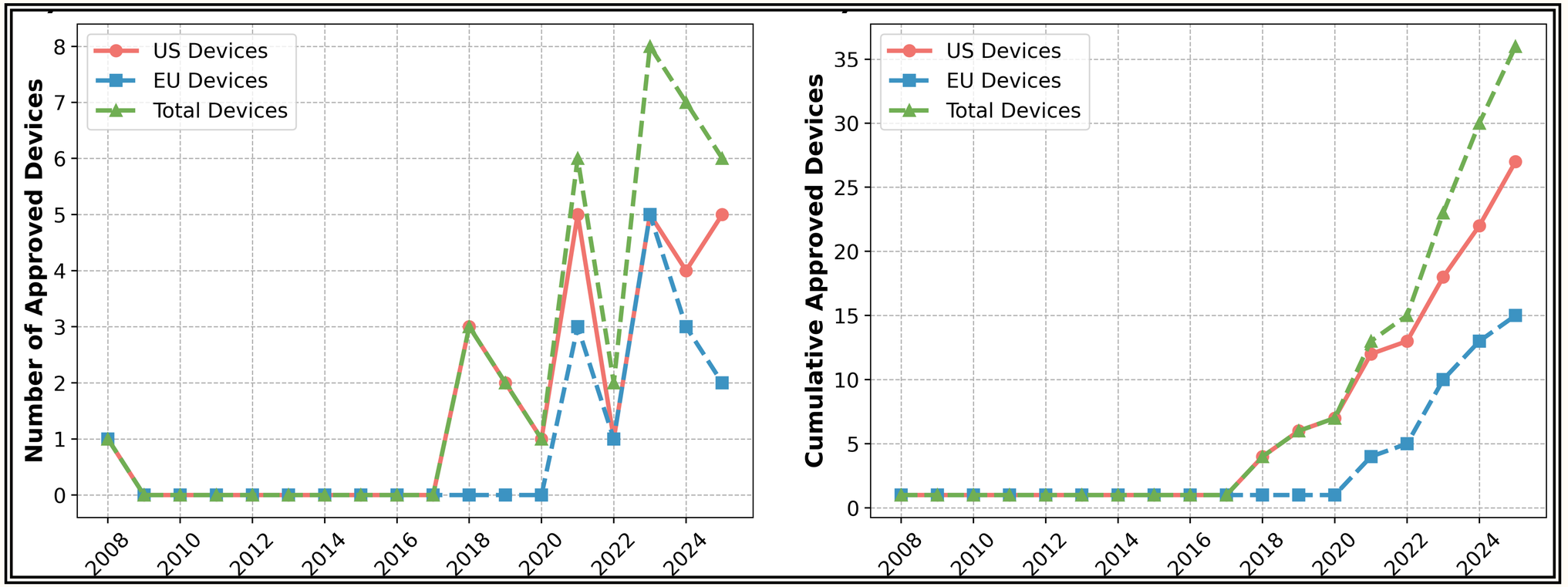


Datenquelle: Cotte F, et al., Safety of Triage Self-assessment Using a Symptom Assessment App for Walk-in Patients in the Emergency Care Setting: Observational Prospective Cross-sectional Study, JMIR Mhealth Uhealth 2022;10(3):e32340

Datenquelle: Pimenta, A., Kini, N., Cotte, F. et al. Appropriateness and utility of a clinical decision support system at the digital front door. npj Digit. Med. (2026).

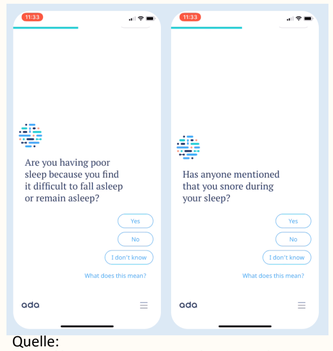
Universitätsklinikum Carl Gustav Carus DIE DRESDNER.
TUD
EKFZ Digital Health
Funded by the European Union

Einsatz im Critical Care Bereich

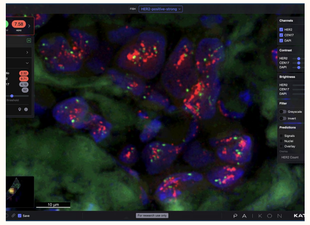


From: Freyer, O., Buch, S., Bassily-Marcus, A. et al. The landscape of artificial intelligence-enabled medical devices in the EU and the US intended for intensive care units. npj Digit. Med. 9, 322 (2026).

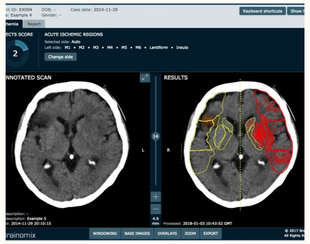
Der Shift zu breiten Systemen



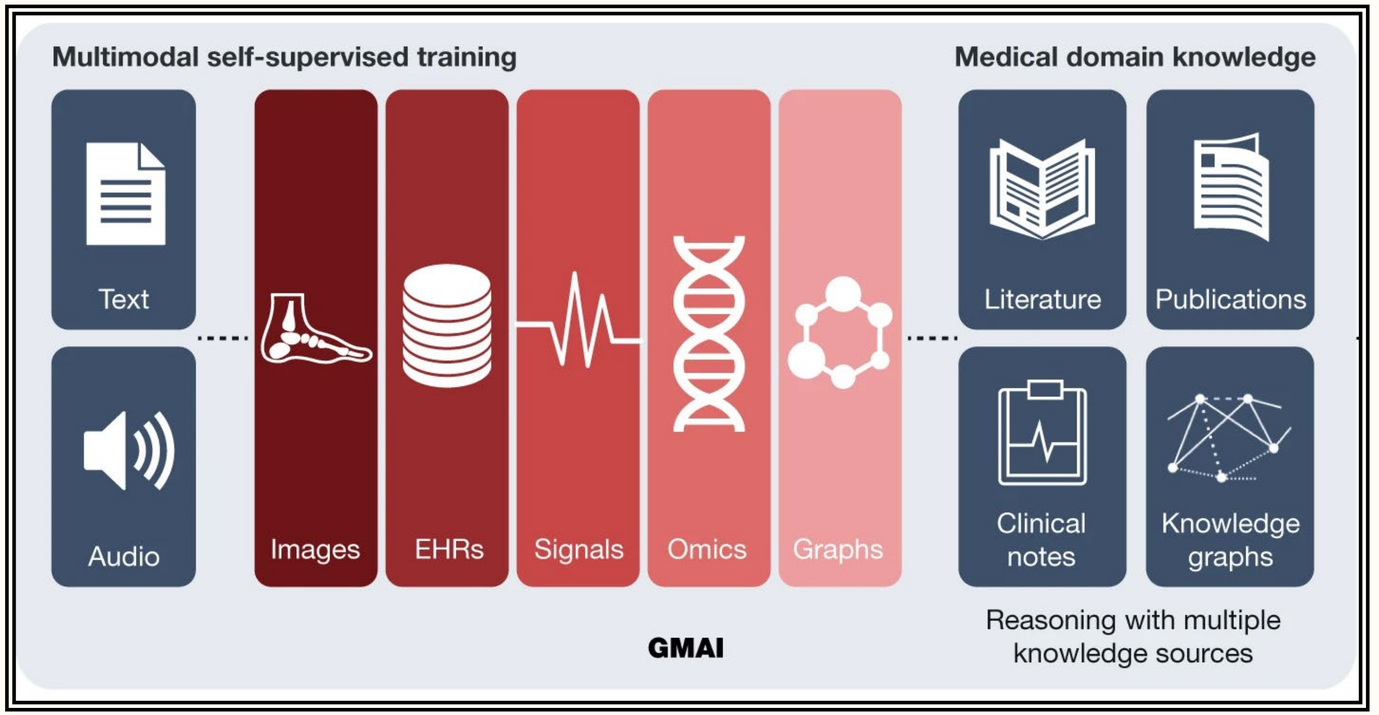
Quelle: https://www.uisources.com/explainer/ada-diagnosing-via-chat-bot



Quelle: Katana Labs

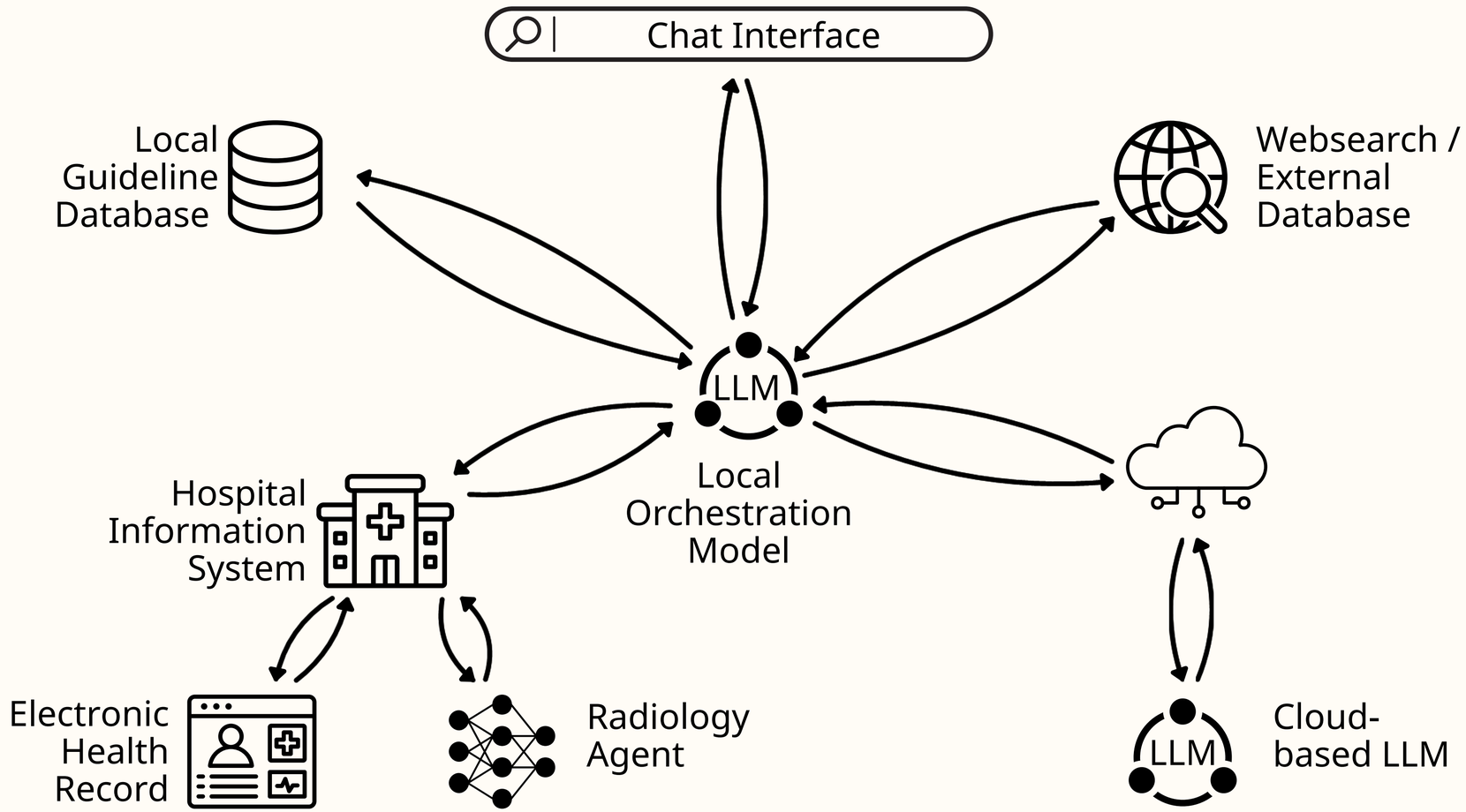


Quelle: Brainomix

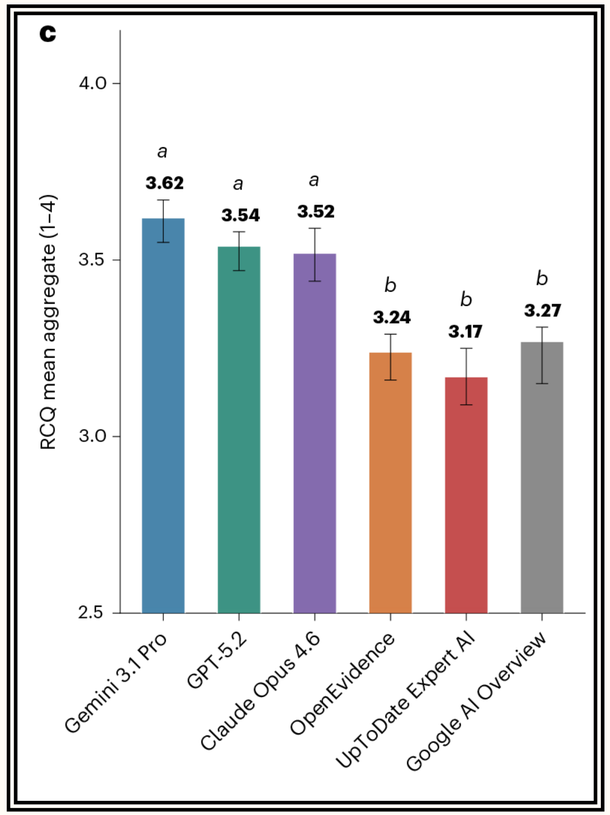


Quelle: Moor, M., Banerjee, O., Abad, Z.S.H. et al. Foundation models for generalist medical artificial intelligence. Nature 616, 259–265 (2023)

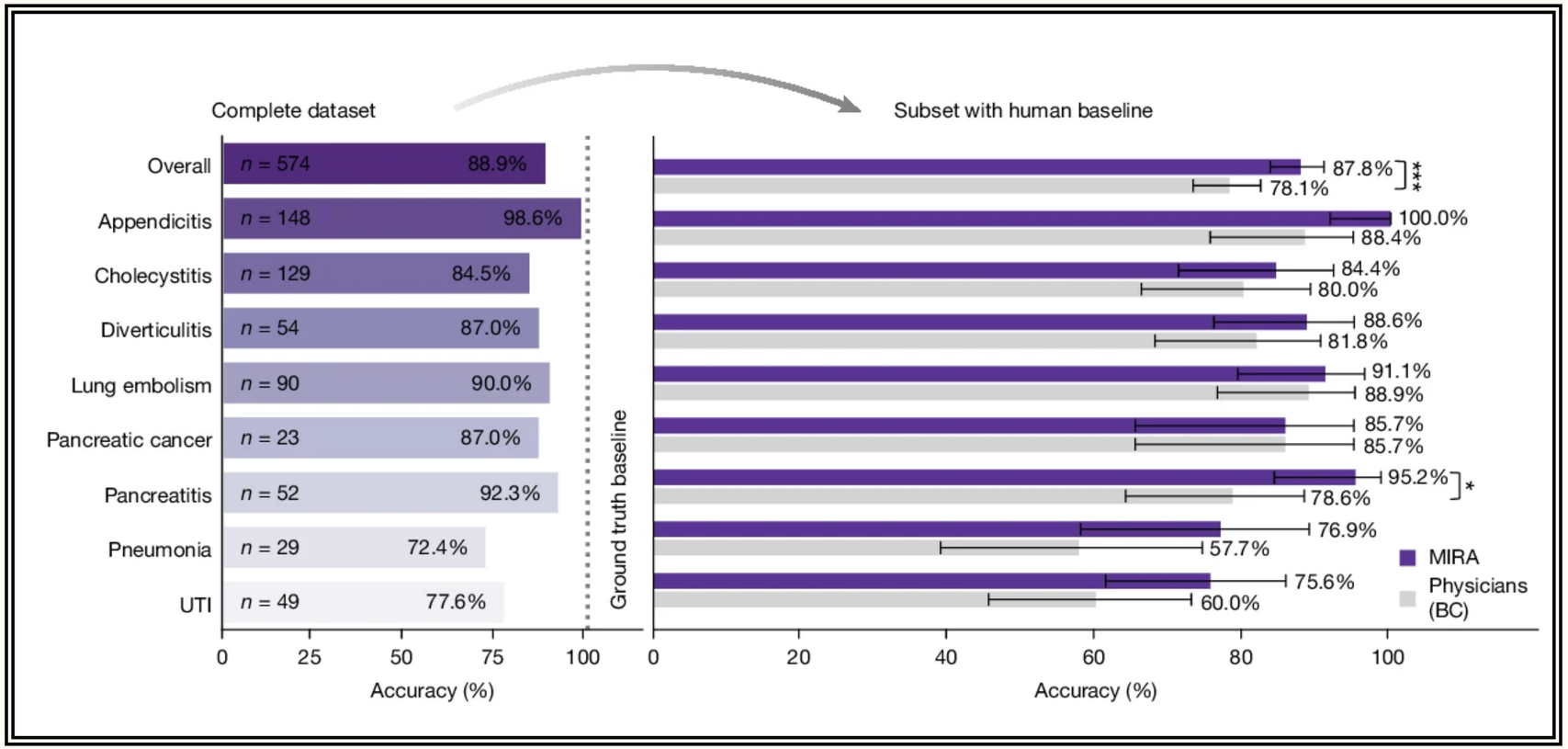
KI-Agenten


Chat Interface
Local Guideline Database
Websearch / External Database
LLM
Local Orchestration Model
Hospital Information System
Electronic Health Record
Radiology Agent
LLM
Cloud-based LLM

Leistungsfähigkeit von breiter KI



Quelle: Vishwanath, K., Alyakin, A., Ghosh, M. et al. General-purpose large language models outperform specialized clinical AI tools on medical benchmarks. Nat Med (2026).

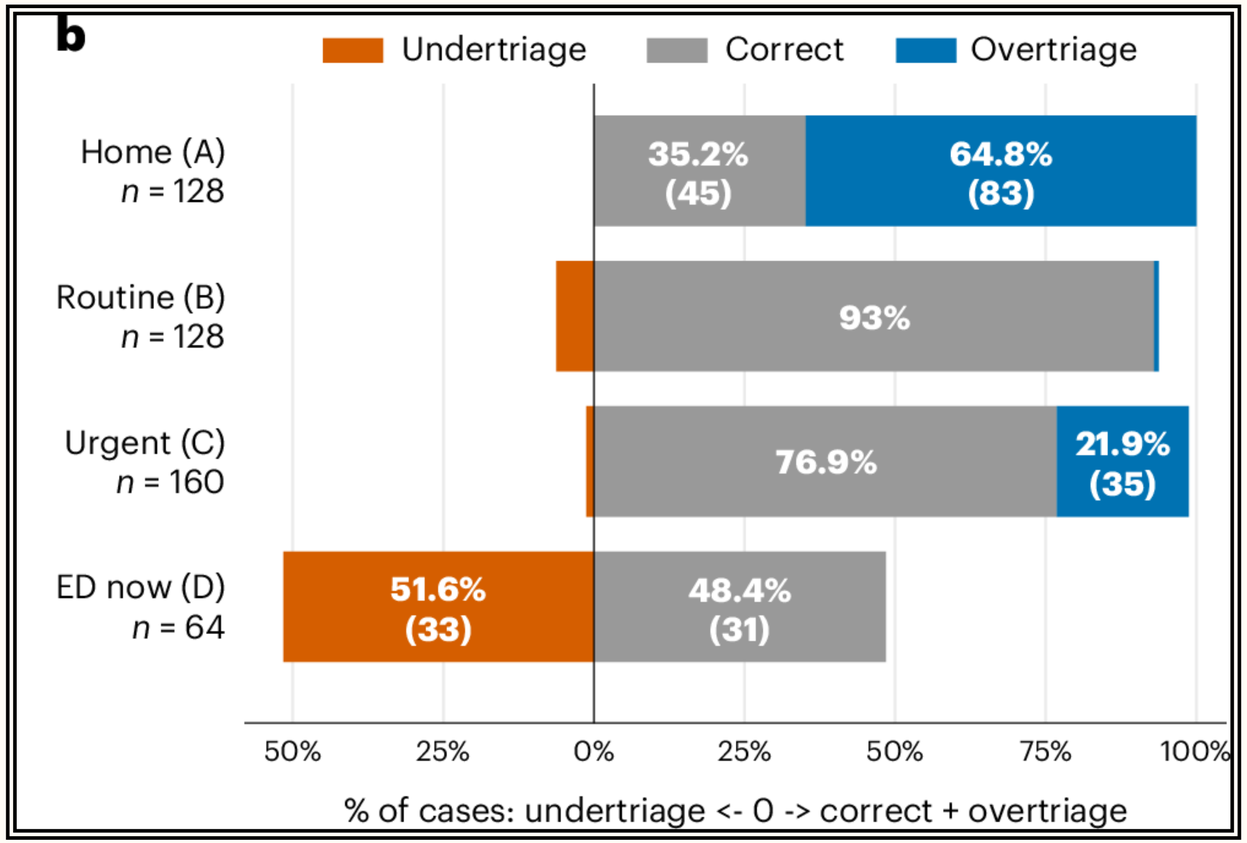


Quelle: Ferber, D., Hilgers, L., Höper, C. et al. Towards autonomous medical artificial intelligence agents. Nature (2026).

Benchmark Performance ≠ Sicherheit

AI: Hello, I'm an internist (...). How can I assist you today?
U: I have just returned from a long-haul flight and have shortness of breath, anxiety, chest pain and nausea.
AI: (...) On the basis of your symptoms and the fact that you have just completed a long-haul flight, you may be suffering from jet lag. Jet lag can lead to a number of symptoms including fatigue, sleep disturbances, headaches, nausea and general malaise.
My treatment plan for you would be as follows:
1. rest and sleep (...).
2. hydration (...).
3. light exercise (...).
4. relaxation techniques (...).
5. medication (...).
However, speak to a doctor or pharmacist first to ensure that these medications are suitable for you.
It is important to note that this is only a preliminary assessment.

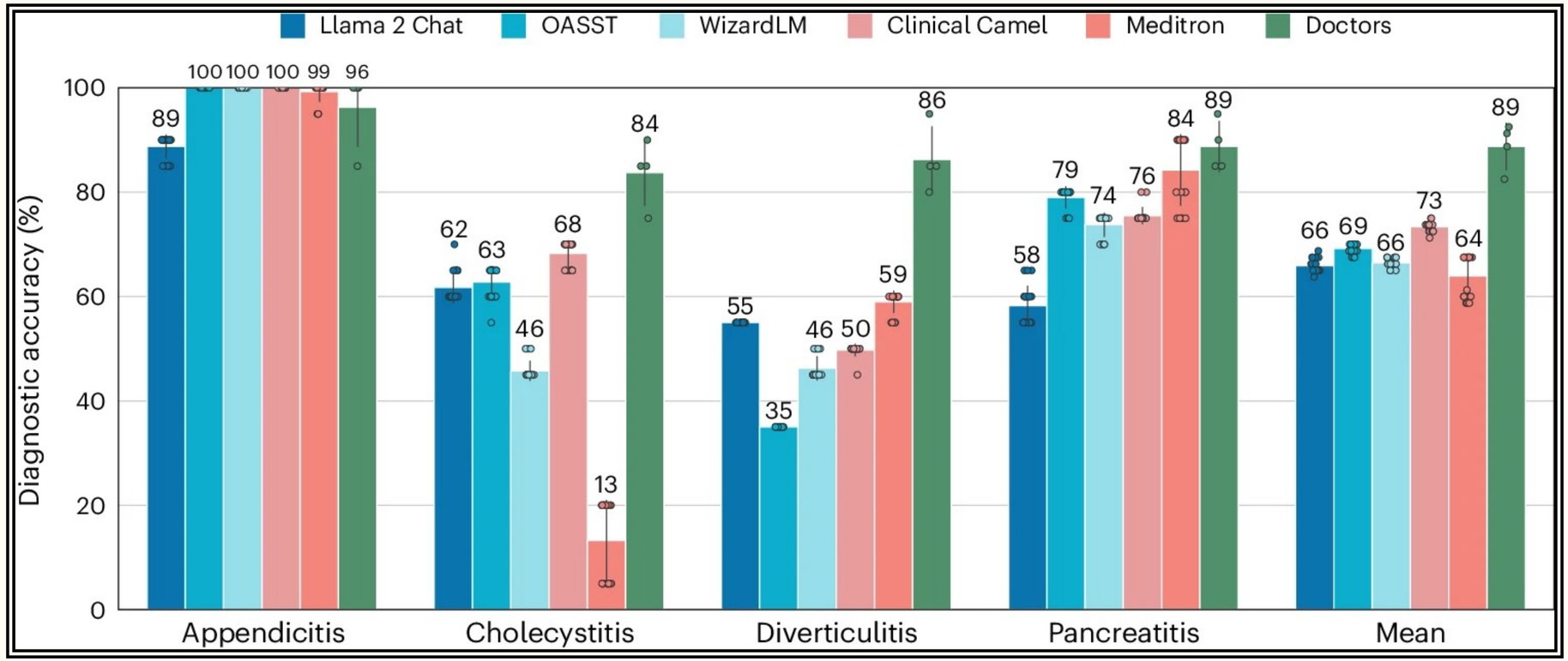
Quelle: Freyer O, Wiest I, Gilbert S, Policing the Boundary Between Responsible and Irresponsible Placing on the Market of Large Language Model Health Applications, Mayo Clinic Proceedings: Digital Health, 2025; 3



Quelle: Ramaswamy, A., Tyagi, A., Hugo, H. et al. ChatGPT Health performance in a structured test of triage recommendations. Nat Med 32, 1671–1675 (2026).

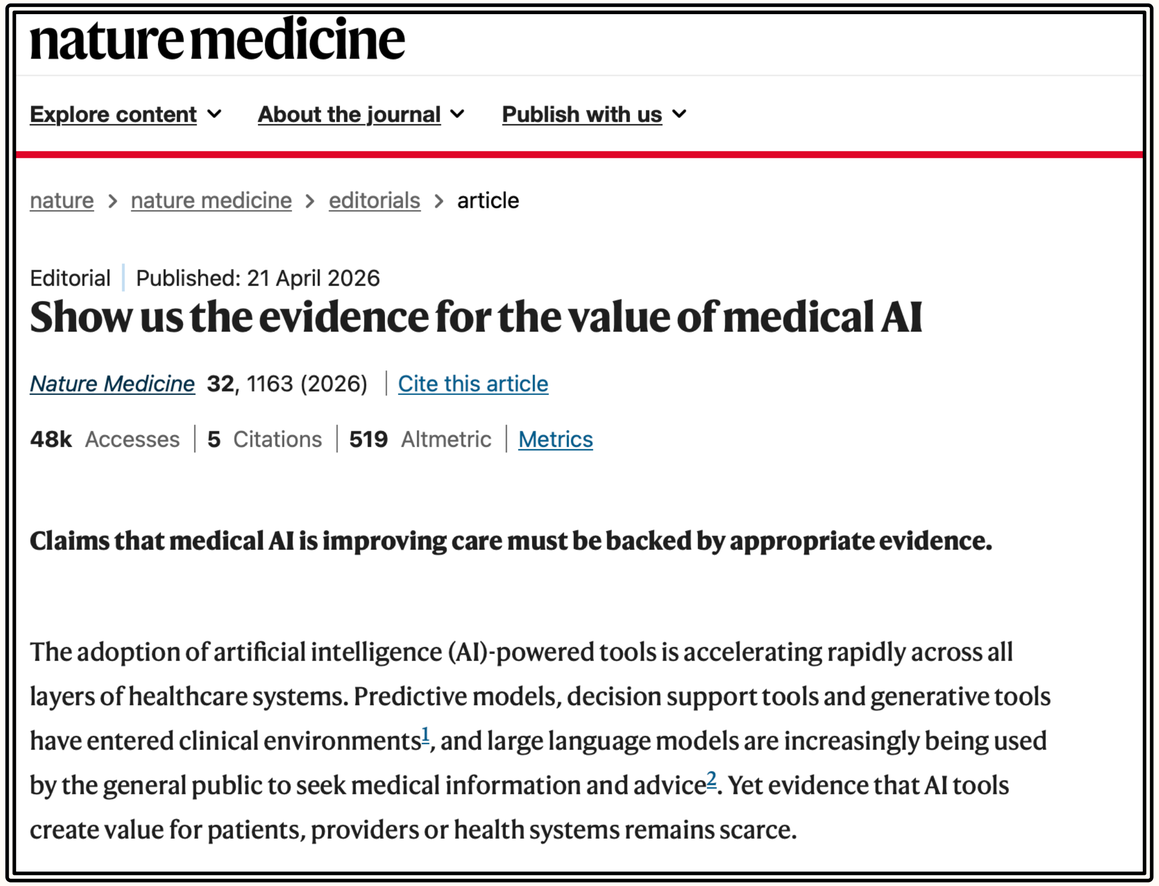
Universitätsklinikum Carl Gustav Carus DIE DRESDNER.
TUD
EKFZ Digital Health
Funded by the European Union

Leistungsprobleme in der realen Welt



Quelle: Hager, P., Jungmann, F., Holland, R. et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med 30, 2613–2622 (2024)

Falsche Benchmarks?



nature medicine

Explore content | About the journal | Publish with us

nature > nature medicine > editorials > article

Editorial | Published: 21 April 2026

Show us the evidence for the value of medical AI

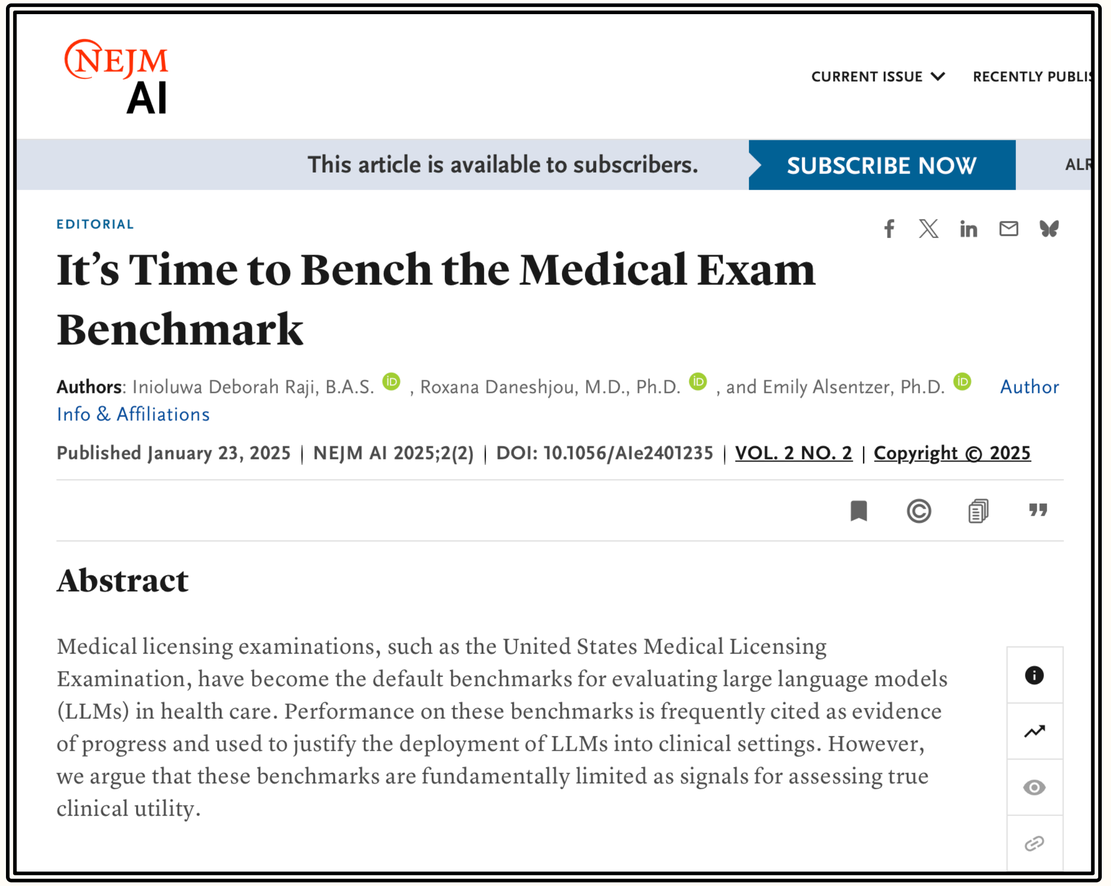
Nature Medicine **32**, 1163 (2026) | Cite this article

48k Accesses | **5** Citations | **519** Altmetric | Metrics

Claims that medical AI is improving care must be backed by appropriate evidence.

The adoption of artificial intelligence (AI)-powered tools is accelerating rapidly across all layers of healthcare systems. Predictive models, decision support tools and generative tools have entered clinical environments[1], and large language models are increasingly being used by the general public to seek medical information and advice[2]. Yet evidence that AI tools create value for patients, providers or health systems remains scarce.

Quelle: Show us the evidence for the value of medical AI. Nat Med 32, 1163 (2026)



NEJM AI

CURRENT ISSUE | RECENTLY PUBLI

This article is available to subscribers. SUBSCRIBE NOW ALR

EDITORIAL

It's Time to Bench the Medical Exam Benchmark

Authors: Inioluwa Deborah Raji, B.A.S. , Roxana Daneshjou, M.D., Ph.D. , and Emily Alsentzer, Ph.D. Author Info & Affiliations

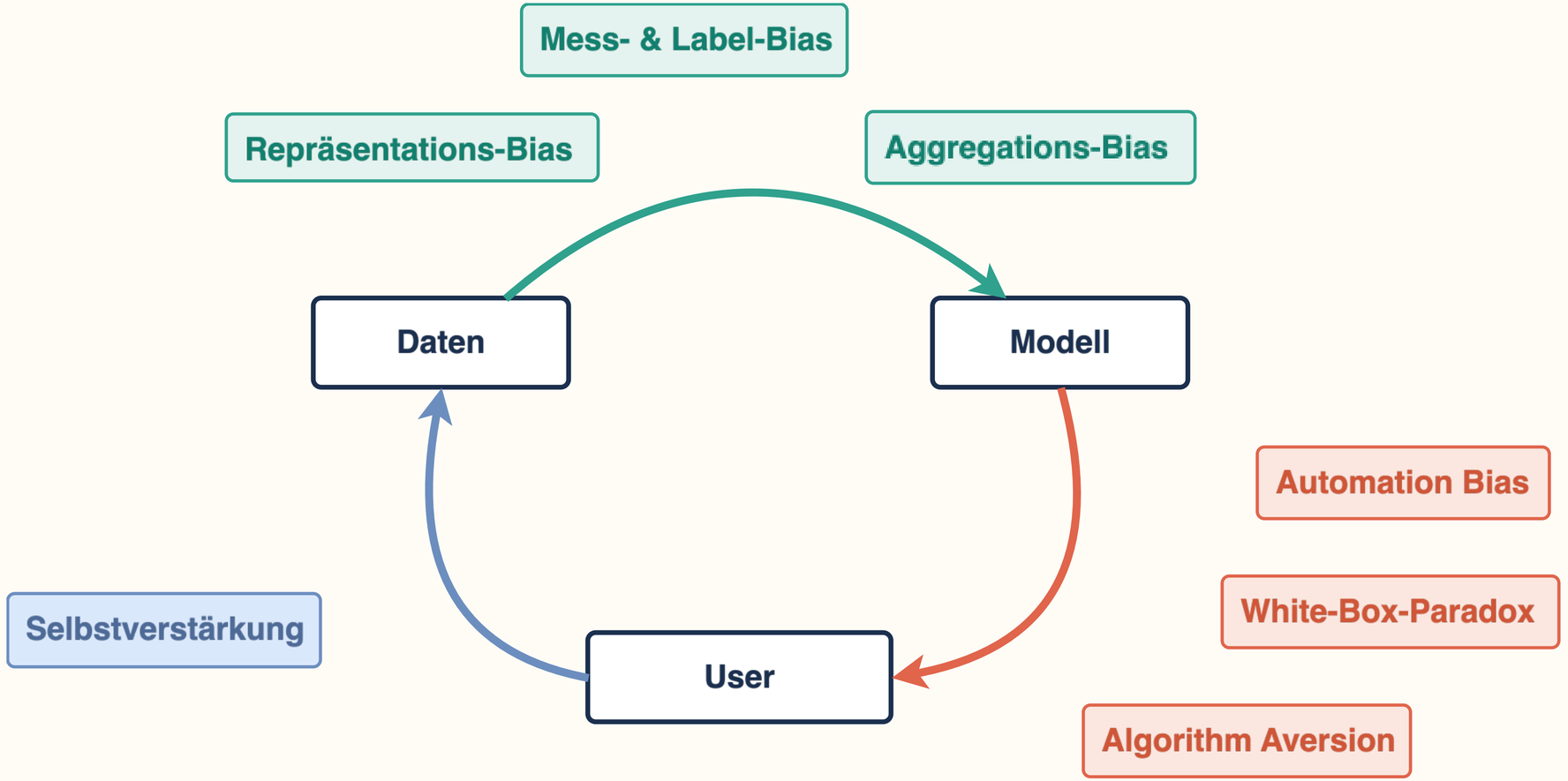
Published January 23, 2025 | NEJM AI 2025;2(2) | DOI: 10.1056/AIe2401235 | VOL. 2 NO. 2 | Copyright © 2025

Abstract

Medical licensing examinations, such as the United States Medical Licensing Examination, have become the default benchmarks for evaluating large language models (LLMs) in health care. Performance on these benchmarks is frequently cited as evidence of progress and used to justify the deployment of LLMs into clinical settings. However, we argue that these benchmarks are fundamentally limited as signals for assessing true clinical utility.

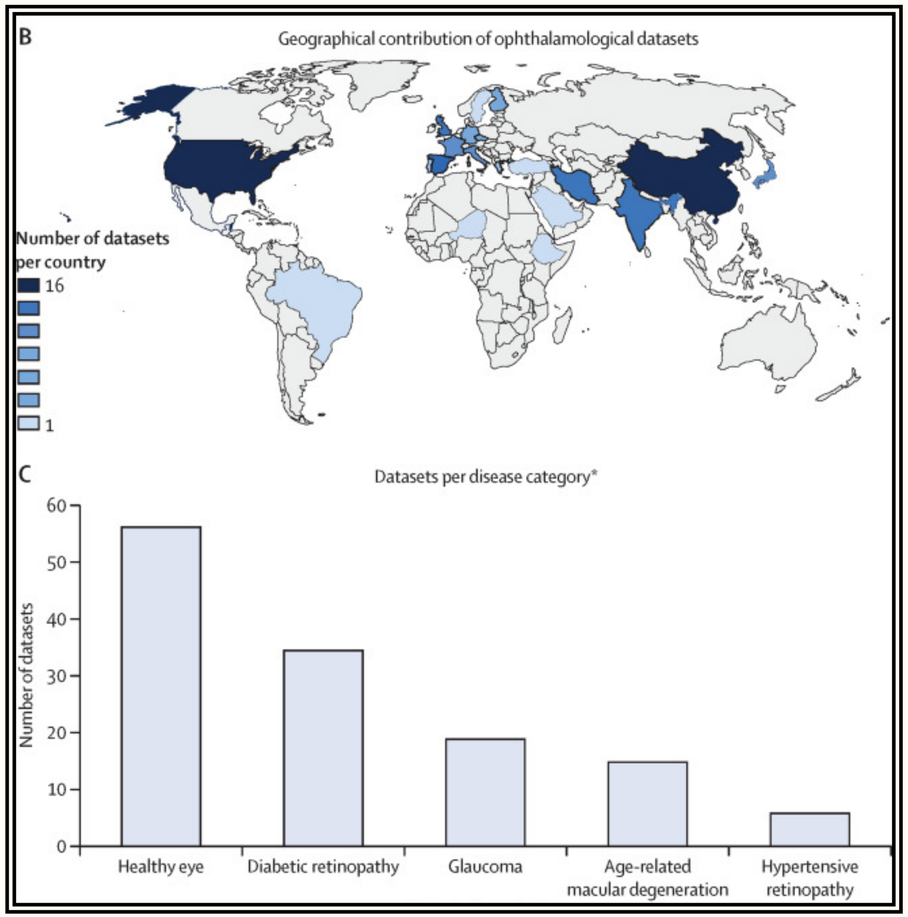
Quelle: Raji I,Roxana Daneshjou R, Alsentzer E, It's Time to Bench the Medical Exam Benchmark, NEJM AI (2025)

Bias

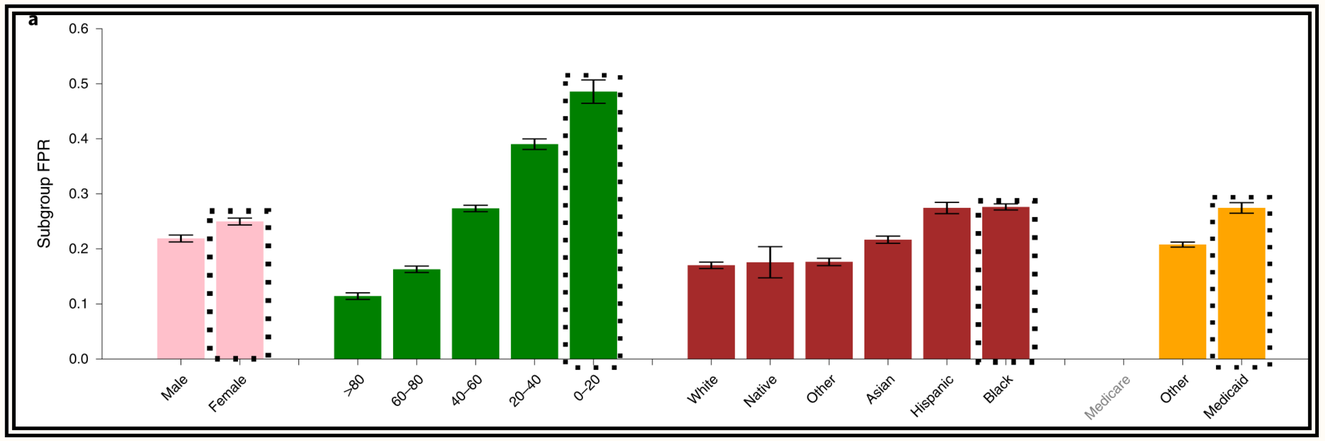


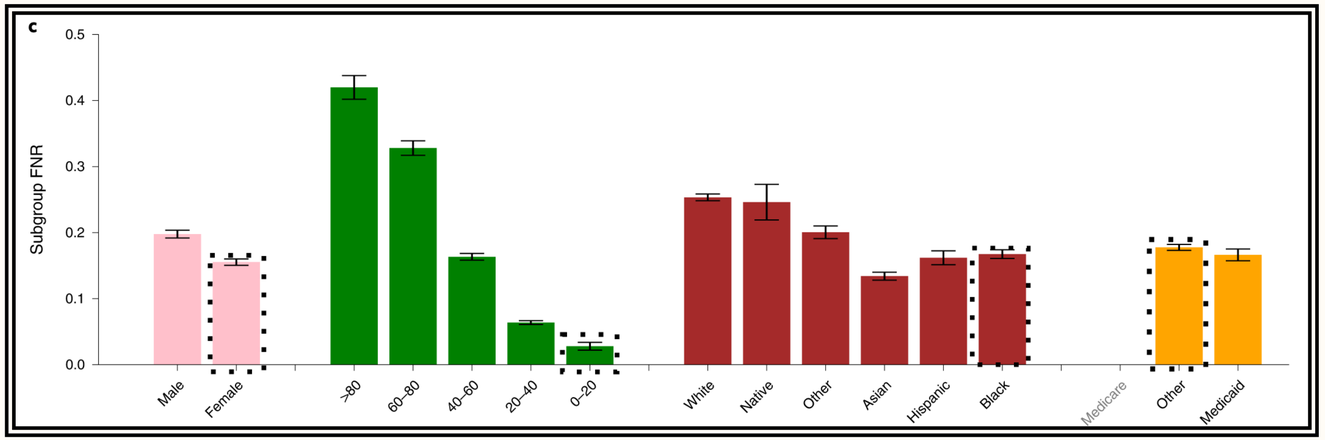
Basierend auf: Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. ACM Comput. Surv. 54, 6, Article 115 (July 2022)

Transparenz



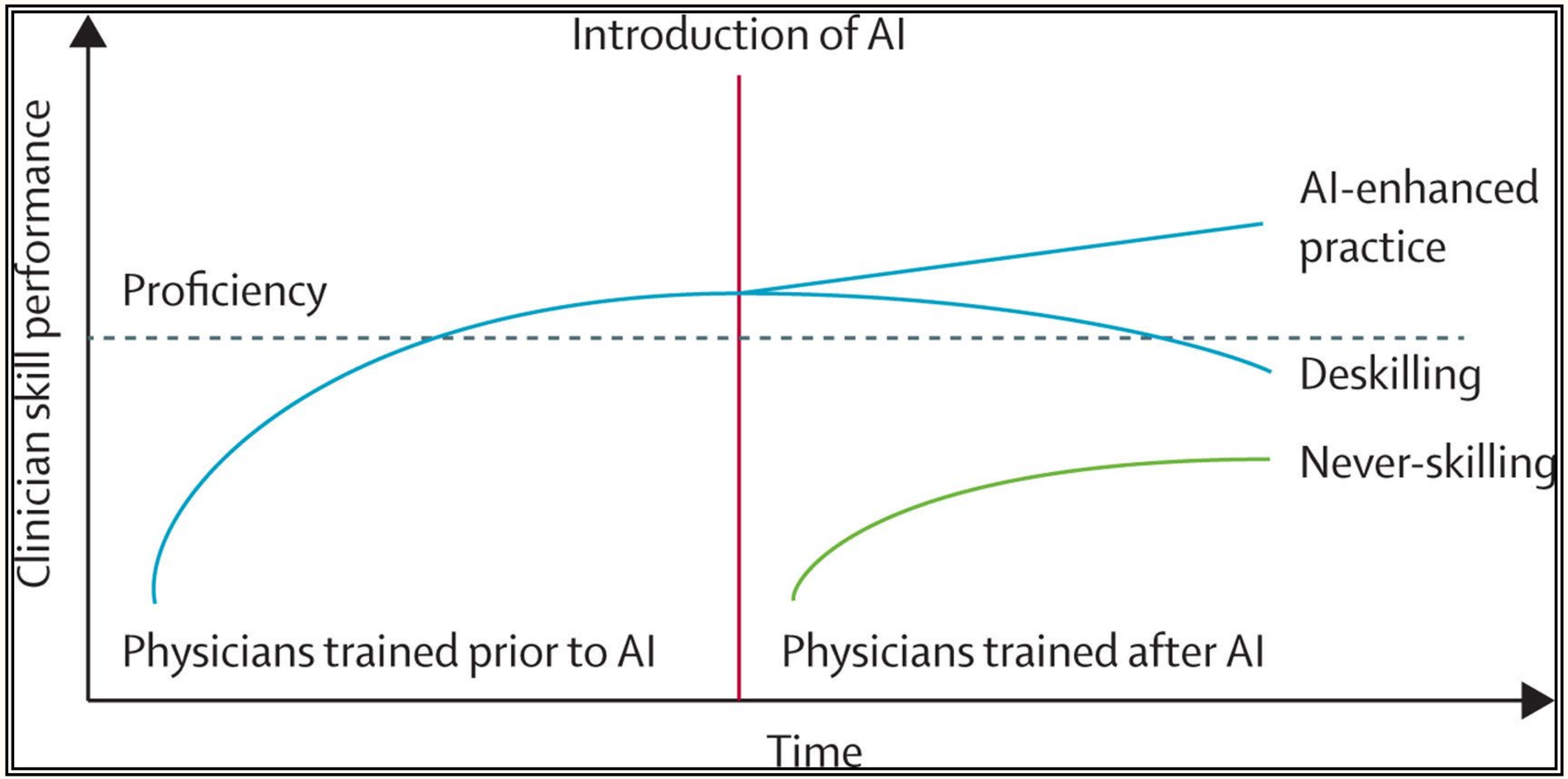
Seyyed-Kalantari, L., Zhang, H., McDermott, M.B.A. et al. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med 27, 2176–2182 (2021).





Quelle: Seyyed-Kalantari, L., Zhang, H., McDermott, M.B.A. et al. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med 27, 2176–2182 (2021).

Deskilling

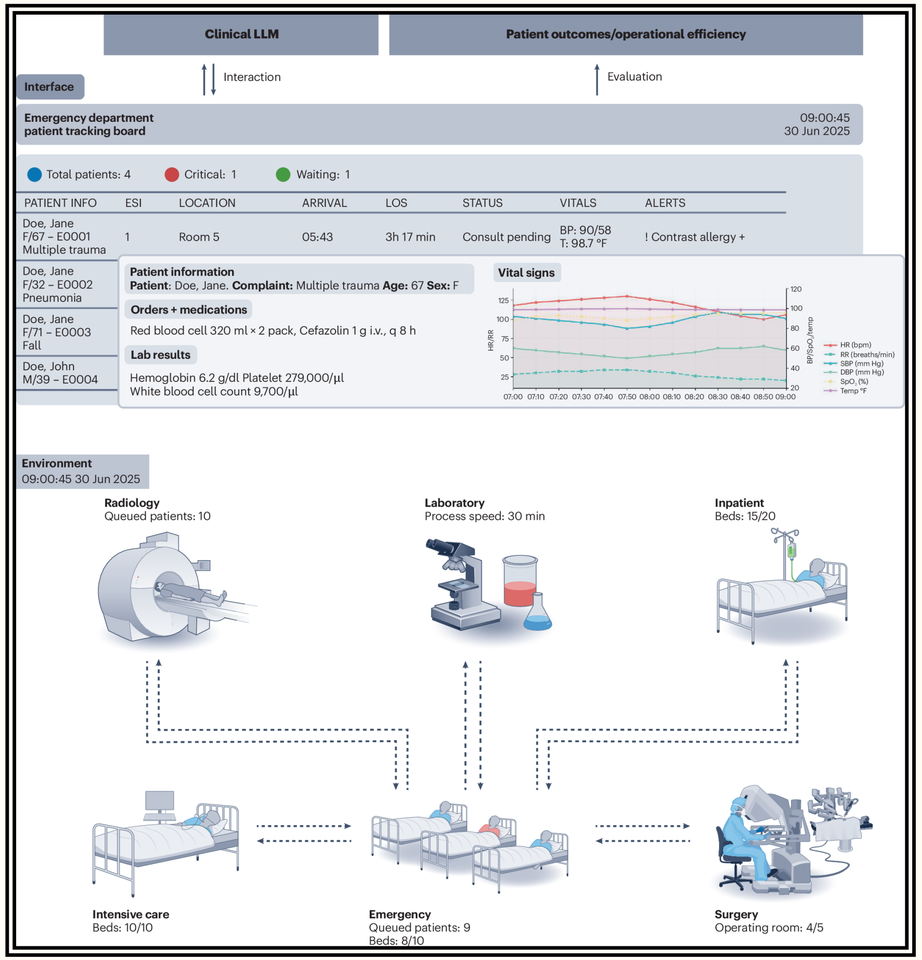


Quelle: Berzin T, Topol E, Preserving clinical skills in the age of AI assistance, The Lancet, 4061719

Was können wir tun?

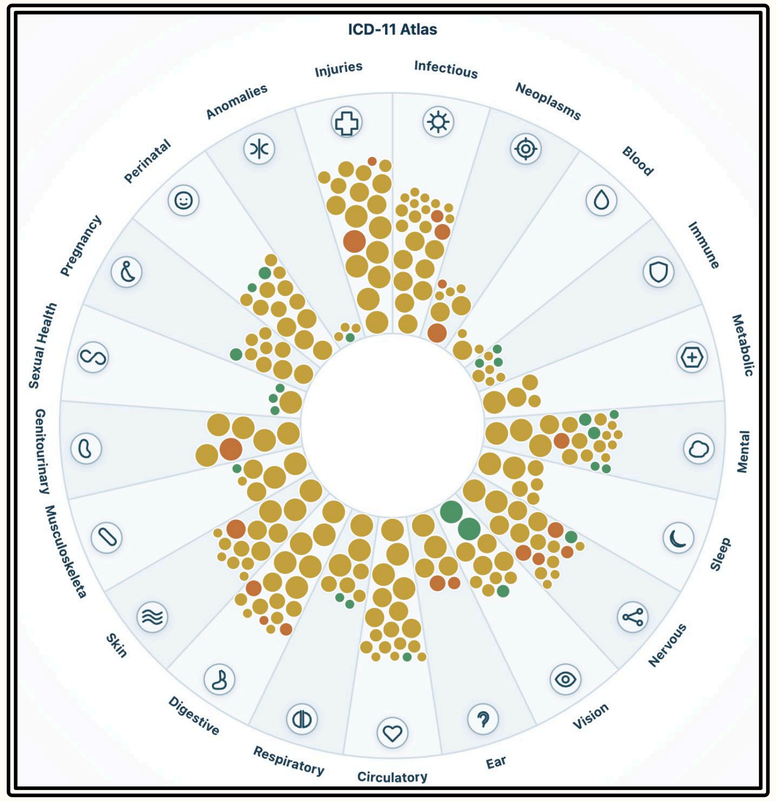
Schrittweise Evaluation

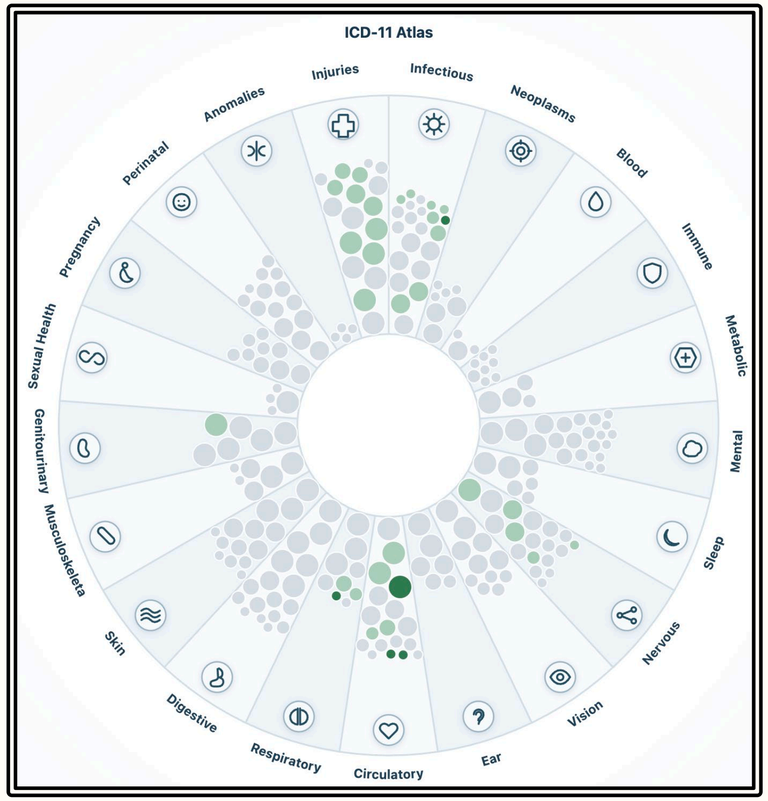
1. Statische Benchmarks
2. Retrospektive Validierung
3. Statische Vignettstudien
4. Agentenbasierte Aufgaben in virtuellen EHR-Umgebungen*
5. Dynamische Simulation mit kaskadierenden Entscheidungen*
6. Prospektive Studien in der realen Welt.
7. Postimplementierungsstudien mit realweltlichen Daten
8. Begleitende ökonomische Evaluation

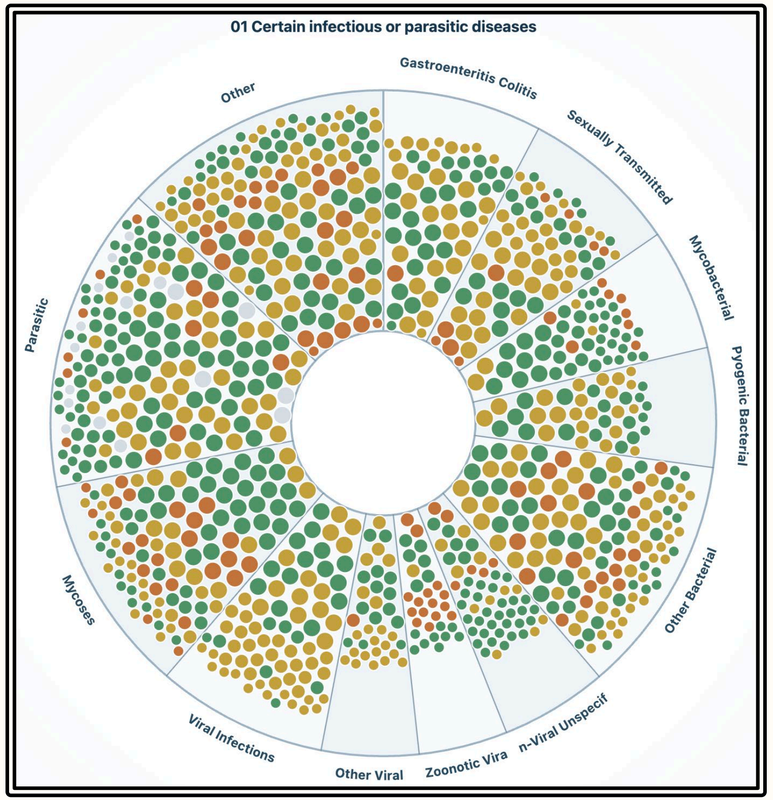


Quelle: Luo, L., Kim, S.E., Zhang, X. et al. A clinical environment simulator for dynamic AI evaluation. Nat Med 32, 820–827 (2026)

Performance granular kommunizieren

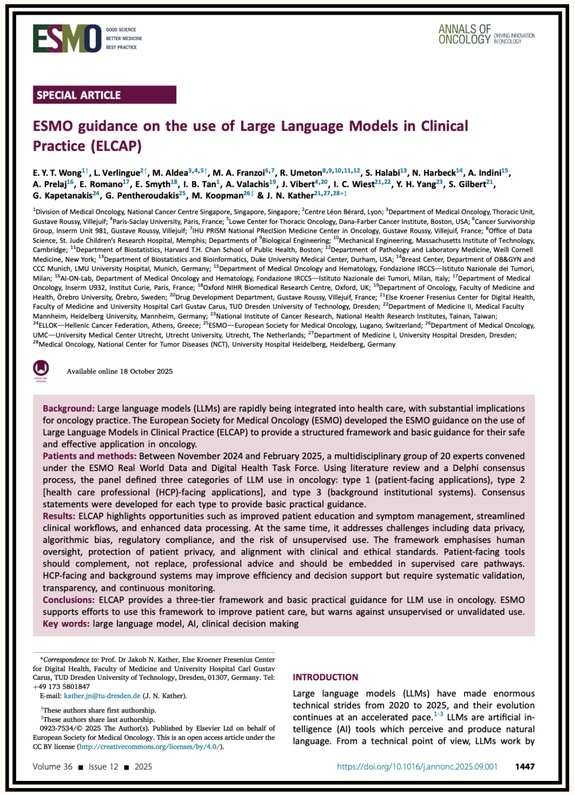






Quelle: Broad Modelcards (Freyer et al., 2026, under review)

Klare Guidelines



ESMO GOOD SCIENCE BETTER MEDICINE BEST PRACTICE

ANNALS OF ONCOLOGY DRIVING INNOVATION IN ONCOLOGY

SPECIAL ARTICLE

ESMO guidance on the use of Large Language Models in Clinical Practice (ELCAP)

E. Y. T. Wong[1], L. Verlingue[2], M. Aldea[3,4,5], M. A. Franzoi[6,7], R. Umeton[8,9,10,11,12], S. Halabi[13], N. Harbeck[14], A. Indini[15], A. Prelaj[16], E. Romano[17], E. Smyth[18], I. B. Tan[1], A. Valachis[19], J. Vibert[4,20], I. C. Wiest[21,22], Y. H. Yang[23], S. Gilbert[21], G. Kapetanakis[24], G. Pentheroudakis[25], M. Koopman[26] & J. N. Kather[21,27,28,*]

[1]Division of Medical Oncology, National Cancer Centre Singapore, Singapore, Singapore; [2]Centre Léon Bérard, Lyon; [3]Department of Medical Oncology, Thoracic Unit, Gustave Roussy, Villejuif; [4]Paris-Saclay University, Paris, France; [5]Lowe Center for Thoracic Oncology, Dana-Farber Cancer Institute, Boston, USA; [6]Cancer Survivorship Group, Inserm Unit 981, Gustave Roussy, Villejuif; [7]IHU PRISM National PRecISion Medicine Center in Oncology, Gustave Roussy, Villejuif, France; [8]Office of Data Science, St. Jude Children's Research Hospital, Memphis; Departments of [9]Biological Engineering; [10]Mechanical Engineering, Massachusetts Institute of Technology, Cambridge; [11]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston; [12]Department of Pathology and Laboratory Medicine, Weill Cornell Medicine, New York; [13]Department of Biostatistics and Bioinformatics, Duke University Medical Center, Durham, USA; [14]Breast Center, Department of OB&GYN and CCC Munich, LMU University Hospital, Munich, Germany; [15]Department of Medical Oncology and Hematology, Fondazione IRCCS—Istituto Nazionale dei Tumori, Milan; [16]AI-ON-Lab, Department of Medical Oncology and Hematology, Fondazione IRCCS—Istituto Nazionale dei Tumori, Milan, Italy; [17]Department of Medical Oncology, Inserm U932, Institut Curie, Paris, France; [18]Oxford NIHR Biomedical Research Centre, Oxford, UK; [19]Department of Oncology, Faculty of Medicine and Health, Örebro University, Örebro, Sweden; [20]Drug Development Department, Gustave Roussy, Villejuif, France; [21]Else Kroener Fresenius Center for Digital Health, Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden University of Technology, Dresden; [22]Department of Medicine II, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany; [23]National Institute of Cancer Research, National Health Research Institutes, Tainan, Taiwan; [24]ELLOK—Hellenic Cancer Federation, Athens, Greece; [25]ESMO—European Society for Medical Oncology, Lugano, Switzerland; [26]Department of Medical Oncology, UMC—University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands; [27]Department of Medicine I, University Hospital Dresden, Dresden; [28]Medical Oncology, National Center for Tumor Diseases (NCT), University Hospital Heidelberg, Heidelberg, Germany

Available online 18 October 2025

Background: Large language models (LLMs) are rapidly being integrated into health care, with substantial implications for oncology practice. The European Society for Medical Oncology (ESMO) developed the ESMO guidance on the use of Large Language Models in Clinical Practice (ELCAP) to provide a structured framework and basic guidance for their safe and effective application in oncology.

Patients and methods: Between November 2024 and February 2025, a multidisciplinary group of 20 experts convened under the ESMO Real World Data and Digital Health Task Force. Using literature review and a Delphi consensus process, the panel defined three categories of LLM use in oncology: type 1 (patient-facing applications), type 2 (health care professional (HCP)-facing applications], and type 3 (background institutional systems). Consensus statements were developed for each type to provide basic practical guidance.

Results: ELCAP highlights opportunities such as improved patient education and symptom management, streamlined clinical workflows, and enhanced data processing. At the same time, it addresses challenges including data privacy, algorithmic bias, regulatory compliance, and the risk of unsupervised use. The framework emphasises human oversight, protection of patient privacy, and alignment with clinical and ethical standards. Patient-facing tools should complement, not replace, professional advice and should be embedded in supervised care pathways. HCP-facing and background systems may improve efficiency and decision support but require systematic validation, transparency, and continuous monitoring.

Conclusions: ELCAP provides a three-tier framework and basic practical guidance for LLM use in oncology. ESMO supports efforts to use this framework to improve patient care, but warns against unsupervised or unvalidated use.

Key words: large language model, AI, clinical decision making

*Correspondence to: Prof. Dr Jakob N. Kather, Else Kroener Fresenius Center for Digital Health, Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden University of Technology, Dresden, 01307, Germany. Tel: +49 173 5801847
E-mail: kather.jn@tu-dresden.de (J. N. Kather).

[†]These authors share first authorship.
[‡]These authors share last authorship.

INTRODUCTION

Large language models (LLMs) have made enormous technical strides from 2020 to 2025, and their evolution continues at an accelerated pace.[1–3] LLMs are artificial intelligence (AI) tools which perceive and produce natural language. From a technical point of view, LLMs work by

Quelle: Wong E, Verlingue L, Aldea M et al., ESMO guidance on the use of Large Language Models in Clinical Practice (ELCAP), Annals of Oncology, 2025; 36, 1447-1457

OPEN ACCESS

Check for updates

FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare

Karim Lekadir,[1,2] Alejandro F Frangi,[3,4] Antonio R Porras,[5] Ben Glocker,[6] Celia Cintas,[7] Curtis P Langlotz,[8] Eva Weicken,[9] Folkert W Asselbergs,[10,11] Fred Prior,[12] Gary S Collins,[13] Georgios Kaissis,[14] Gianna Tsakou,[15] Irène Buvat,[16] Jayashree Kalpathy-Cramer,[17] John Mongan,[18] Julia A Schnabel,[19] Kaisar Kushibar,[1] Katrine Riklund,[20] Kostas Marias,[21] Lameck M Amugongo,[22] Lauren A Fromont,[23] Lena Maier-Hein,[24] Leonor Cerdá-Alberich,[25] Luis Martí-Bonmatí,[26] M Jorge Cardoso,[27] Maciej Bobowicz,[28] Mahsa Shabani,[29] Manolis Tsiknakis,[21] Maria A Zuluaga,[30] Marie-Christine Fritzsche,[31] Marina Camacho,[1] Marius George Linguraru,[32] Markus Wenzel,[9] Marleen De Bruijne,[33] Martin G Tolsgaard,[34] Melanie Goisauf,[35] Mónica Cano Abadía,[35] Nikolaos Papanikolaou,[36] Noussair Lazrak,[1] Oriol Pujol,[1] Richard Osuala,[1] Sandy Napel,[37] Sara Colantonio,[38] Smriti Joshi,[1] Stefan Klein,[33] Susanna Aussó,[39] Wendy A Rogers,[40] Zohaib Salahuddin,[41] Martijn P A Starmans[33];
on behalf of the FUTURE-AI Consortium

For numbered affiliations see end of the article

Correspondence to: K Lekadir karim.lekadir@ub.edu (ORCID 0000-0002-9456-1612)

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2025;388:e081554 http://dx.doi.org/10.1136/bmj-2024-081554

Accepted: 10 January 2025

Despite major advances in artificial intelligence (AI) research for healthcare, the deployment and adoption of AI technologies remain limited in clinical practice. This paper describes the FUTURE-AI framework, which provides guidance for the development and deployment of trustworthy AI tools in healthcare. The FUTURE-AI Consortium was founded in 2021 and comprises 117 interdisciplinary experts from 50 countries representing all continents, including AI scientists, clinical researchers, biomedical ethicists, and social scientists. Over a two year period, the FUTURE-AI guideline was established through consensus based on six guiding principles—fairness, universality, traceability, usability, robustness, and explainability. To operationalise trustworthy AI in healthcare, a set of 30 best practices were defined, addressing technical, clinical, socioethical, and legal dimensions. The recommendations cover the entire lifecycle of healthcare AI, from design, development, and validation to regulation, deployment, and monitoring.

SUMMARY POINTS

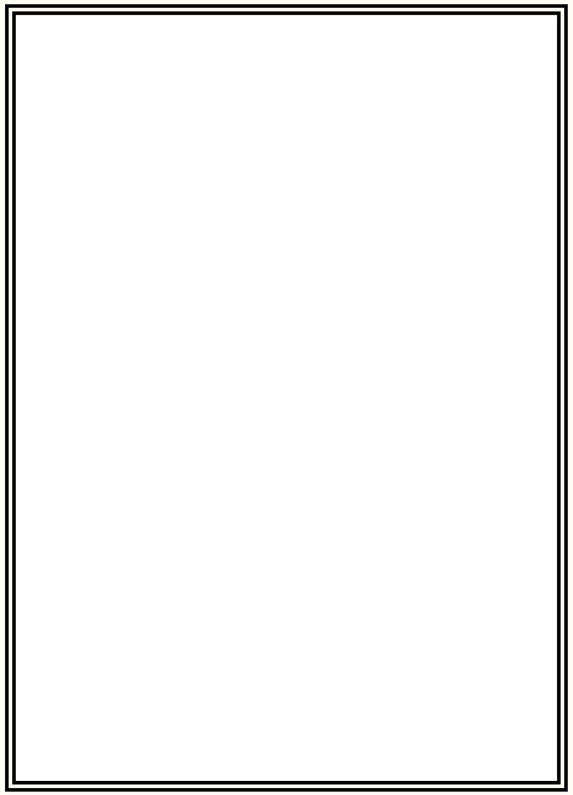
- Despite major advances in medical artificial intelligence (AI) research, clinical adoption of emerging AI solutions remains challenging owing to limited trust and ethical concerns
- The FUTURE-AI Consortium unites 117 experts from 50 countries to define international guidelines for trustworthy healthcare AI
- The FUTURE-AI framework is structured around six guiding principles: fairness, universality, traceability, usability, robustness, and explainability
- The guideline addresses the entire AI lifecycle, from design and development to validation and deployment, ensuring alignment with real world needs and ethical requirements
- The framework includes 30 detailed recommendations for building trustworthy and deployable AI systems, emphasising multistakeholder collaboration
- Continuous risk assessment and mitigation are fundamental, addressing biases, data variations, and evolving challenges during the AI lifecycle
- FUTURE-AI is designed as a dynamic framework, which will evolve with technological advancements and stakeholder feedback

Introduction

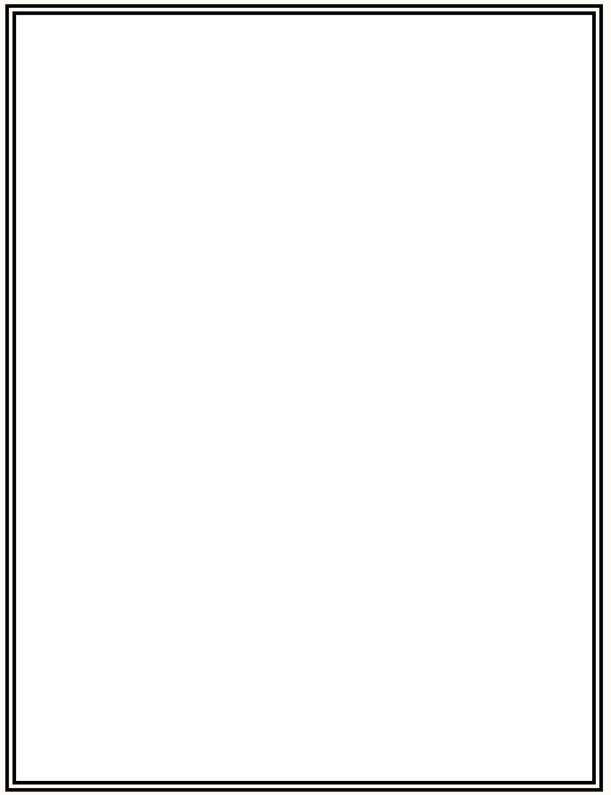
In the field of healthcare, artificial intelligence (AI)—that is, algorithms with the ability to self-learn logic—and data interactions have been increasingly used to develop computer aided models, for example, disease diagnosis, prognosis, prediction of therapy response or survival, and patient stratification.[1] Despite major advances, the deployment and adoption of AI technologies remain limited in real world clinical practice. In recent years, concerns have been raised about the technical, clinical, ethical, and societal risks associated with healthcare AI.[2 3] In particular, existing research has shown that AI tools in healthcare can be prone to errors and patient harm, biases and increased health inequalities, lack of transparency and accountability, as well as data privacy and security breaches.[4–8]

To increase adoption in the real world, it is essential that AI tools are trusted and accepted by patients, clinicians, health organisations, and authorities. However, there is an absence of clear, widely accepted guidelines on how healthcare AI tools should be designed, developed, evaluated, and deployed to be trustworthy—that is, technically robust, clinically safe,

Quelle: BMJ 2025;388:e081554



Quelle: NHS, https://www.england.nhs.uk/long-read/guidance-on-the-use-of-ai-enabled-ambient-scribing-products-in-health-and-care-settings

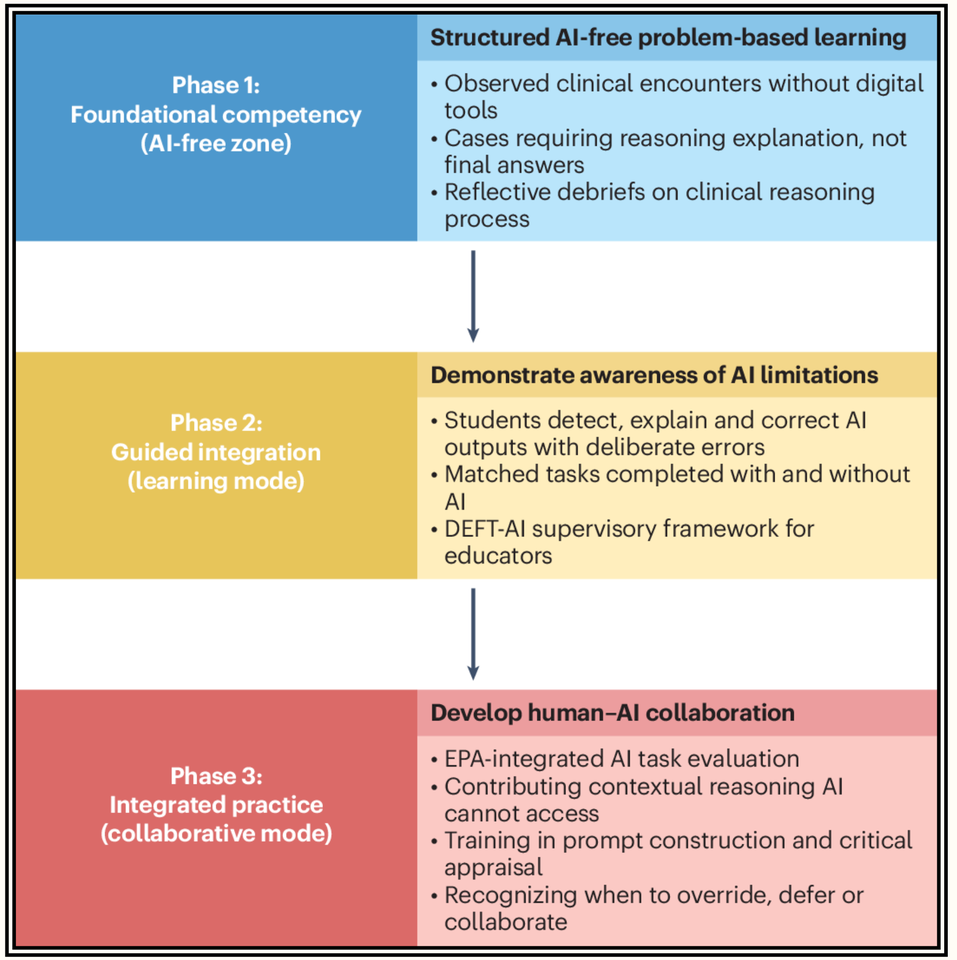


Healthcare Organizations on the Use of Generative Artificial Intelligence in Medicine: A Position Statement from the Society of General Internal Medicine. J GEN INTERN MED 40, 694–702 (2025).

Gutes Training

“KI kann uns nicht das Verstehen abnehmen”

- Andrej Karpathy



Quelle: Ke, Y., Jin, L., Ong, J.C.L. et al. AI-induced never-skilling in medical education. Nat Med 32, 1997–2006 (2026).

KI hat großes Potential in der Routineversorgung - ihre neurtigen Risiken müssen aber noch addressiert werden.

Ende.

Fragen?