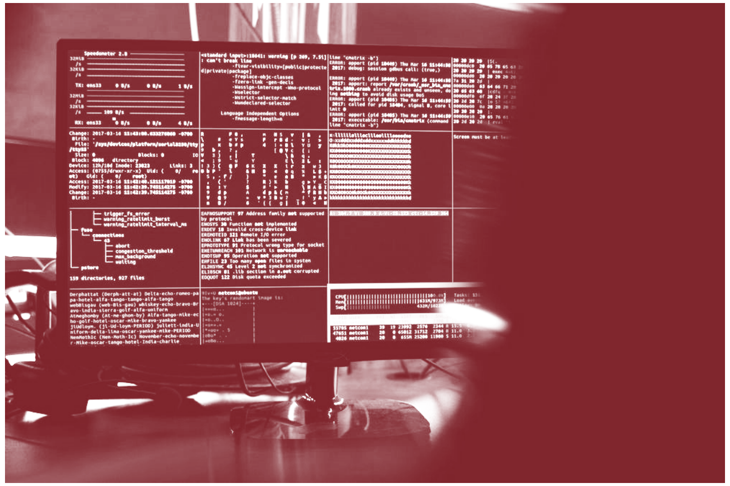
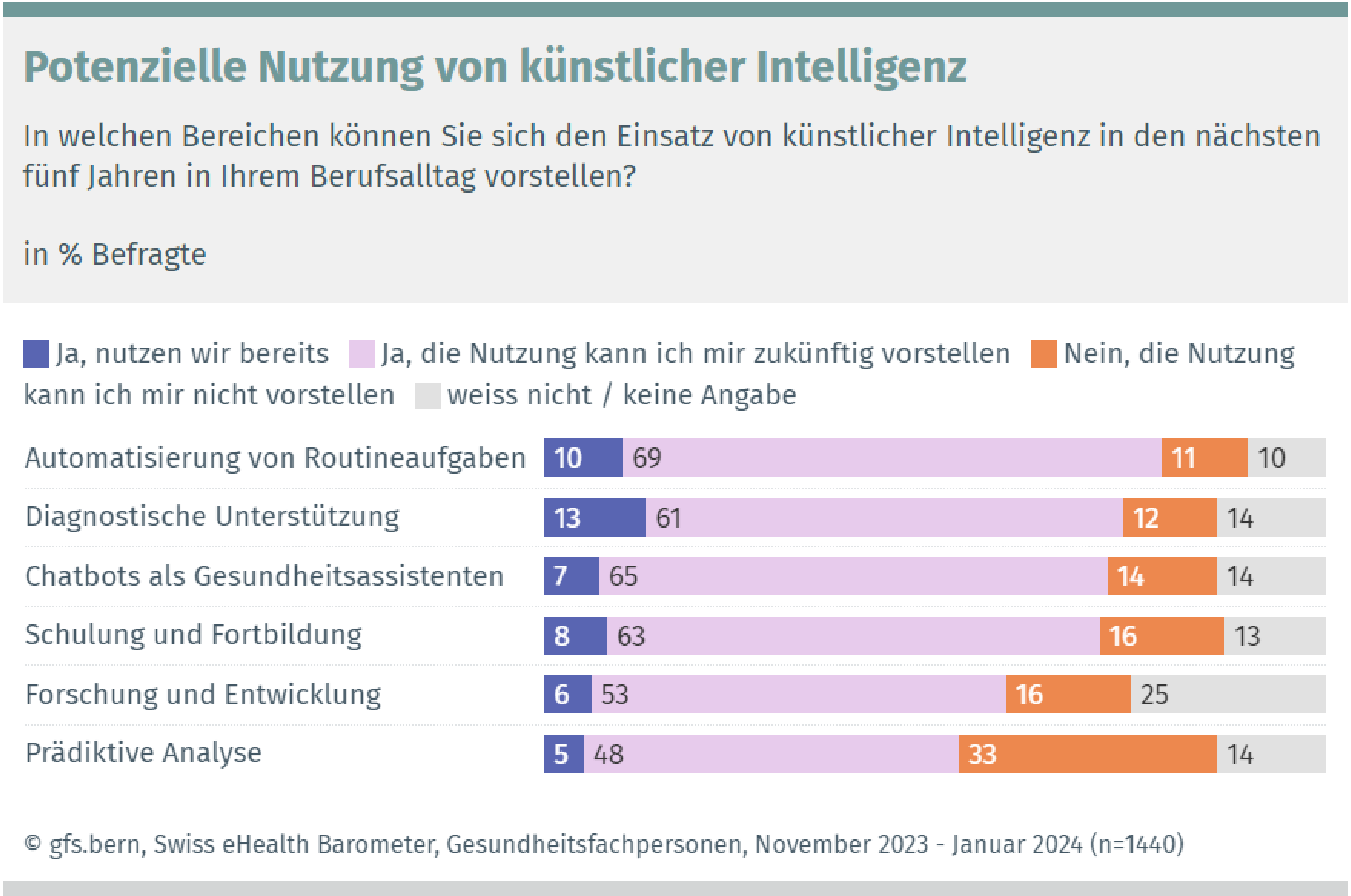
AI und Patientensicherheit: Aufbruch in eine ungewisse Zukunft

Prof. Dr. David Schwappach, MPH

Institut für Sozial und Präventivmedizin (ISPM)
Universität Bern
David.Schwappach@unibe.ch

24. Mai 2024, SAQM Symposium, UniS Bern





Potenzielle Nutzung von künstlicher Intelligenz

In welchen Bereichen können Sie sich den Einsatz von künstlicher Intelligenz in den nächsten fünf Jahren in Ihrem Berufsalltag vorstellen?

in % Befragte

© gfs.bern, Swiss eHealth Barometer, Gesundheitsfachpersonen, November 2023 - Januar 2024 (n=1440)

https://cockpit.gfsbern.ch/de/cockpit/ehealth-gesundheitsfachpersonen-2024/

Aktuelle AI-Anwendungen

- Unterstützung bei Routine-Aufgaben
  z.B. Austrittsbericht, Medikationsabgleich, Dokumentation Konsultation („digital scribe“)

- Assistenz Befundung Bildgebung
  z.B. Röntgen, MRI, Dermatologie, Ophthalmologie (z.B. diabetische Retinopathie)

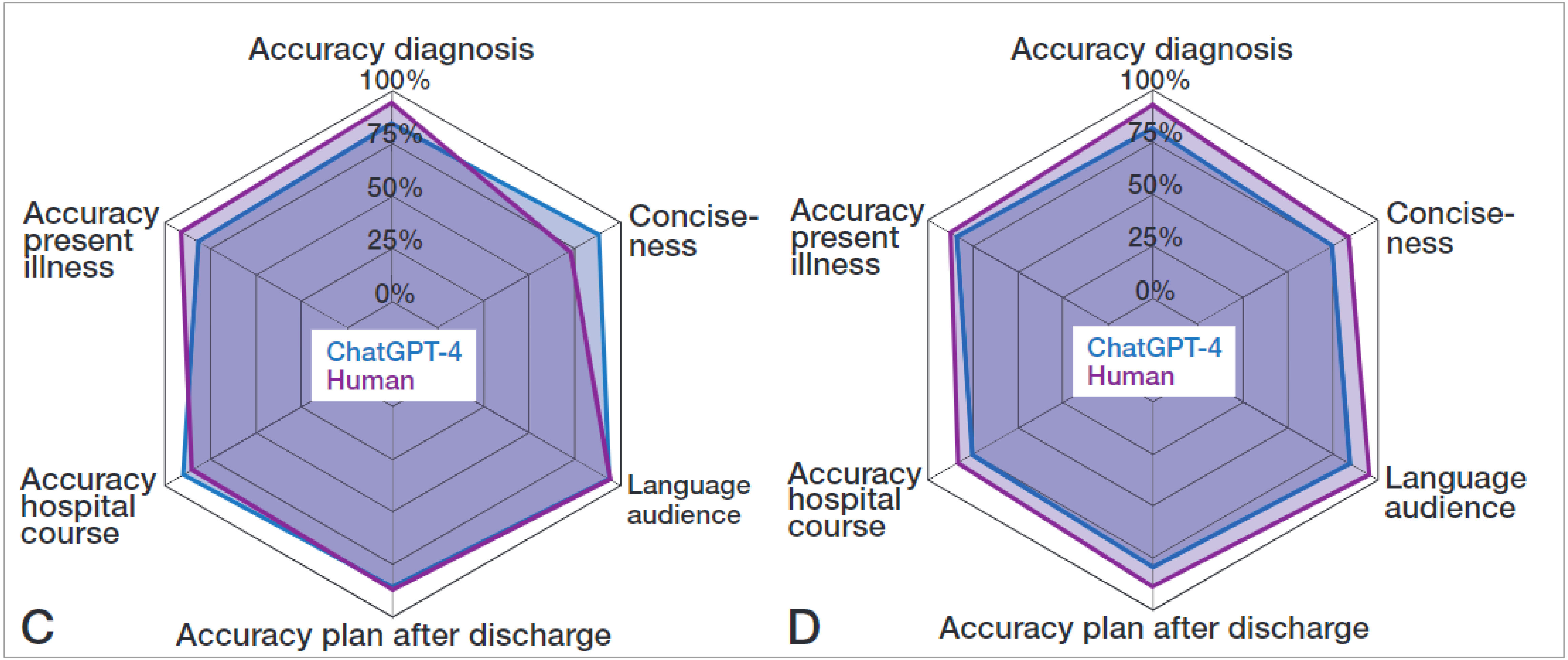
- Vorhersage Verschlechterung individueller Patienten
  z.B. Sepsis, Druckgeschwür, unerwünschtes Arzneimittel-Ereignis, chirurgische Komplikation

Beispiel I: Orthopädische Austrittsbriefe

- Erstellung von Austrittsdokumenten für 6 fiktive Patienten basierend auf KG (incl. Labor, Befunde, Bildgebungsbefunde, Verlaufsnotizen, Medikation, etc.)
- Aufgabe: Erstellen je eines Austrittsberichts für Hausarzt/ärztin und Patient/in, entsprechend dem Standardformat der Spitäler
- Assistenzarzt vs. Oberarzt vs. Chat-GPT4
- Evaluation der Austrittsdokumente durch erfahrene, verblindete Kliniker (n=15) anhand definierter Kriterien

Rosenberg GS et al. doi: 10.2340/17453674.2024.40182

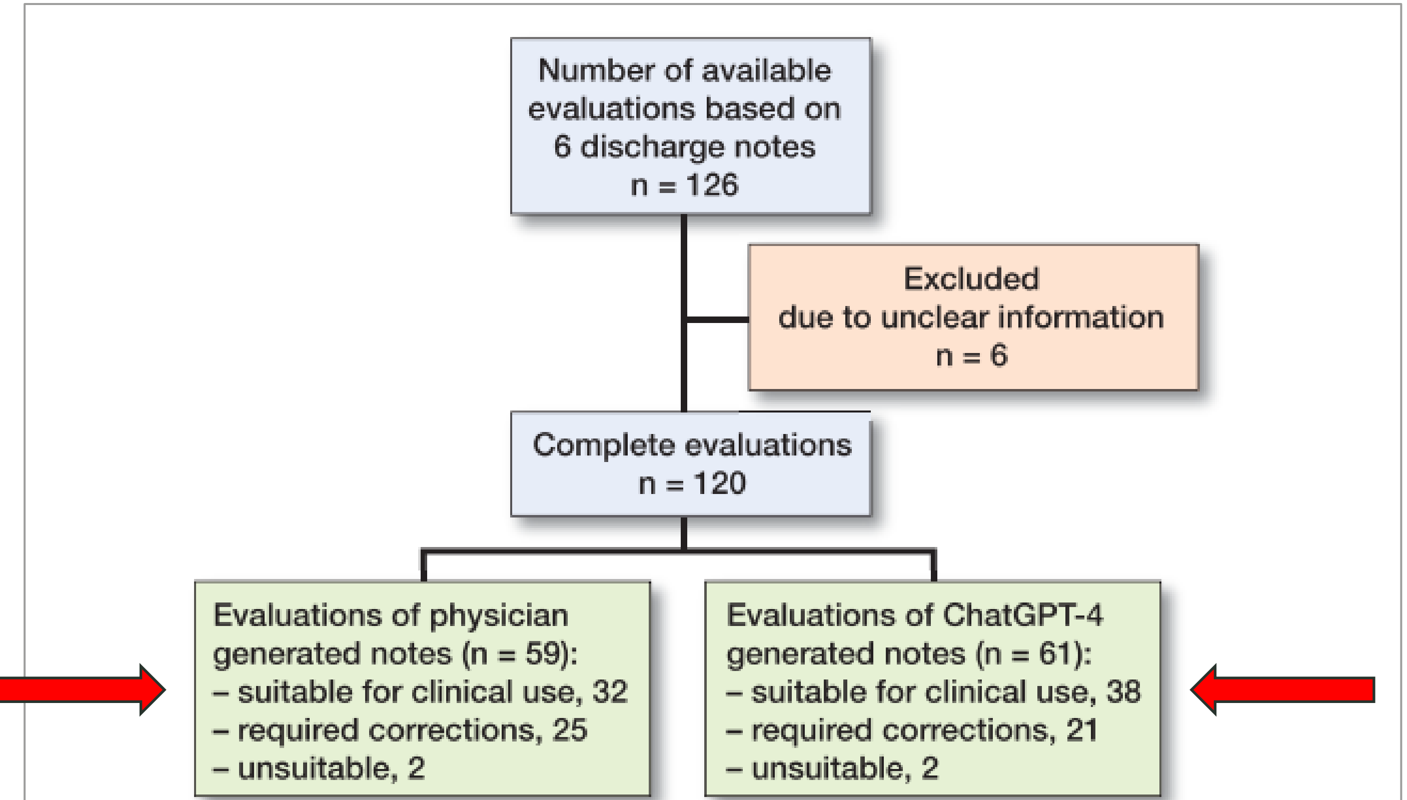
Beispiel I: Orthopädische Austrittsbriefe



C. Swiss summary; D. Swiss letter.

Rosenberg GS et al. doi: 10.2340/17453674.2024.40182

Beispiel I: Orthopädische Austrittsbriefe



Number of available evaluations based on 6 discharge notes
n = 126

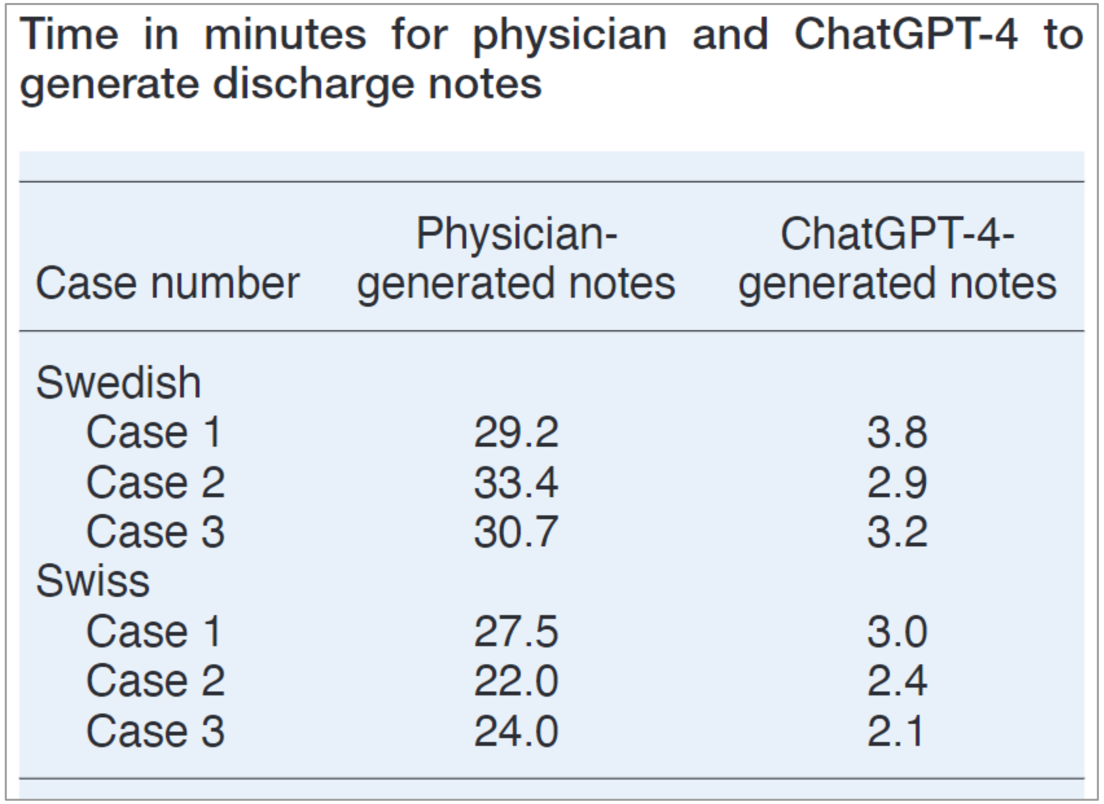
Excluded due to unclear information
n = 6

Complete evaluations
n = 120

Evaluations of physician generated notes (n = 59):
- suitable for clinical use, 32
- required corrections, 25
- unsuitable, 2

Evaluations of ChatGPT-4 generated notes (n = 61):
- suitable for clinical use, 38
- required corrections, 21
- unsuitable, 2

Figure 2. Flowchart of evaluations of discharge notes by the expert panel.



Time in minutes for physician and ChatGPT-4 to generate discharge notes

Case number	Physician-generated notes	ChatGPT-4-generated notes
Swedish		
Case 1	29.2	3.8
Case 2	33.4	2.9
Case 3	30.7	3.2
Swiss		
Case 1	27.5	3.0
Case 2	22.0	2.4
Case 3	24.0	2.1

Introspektion?

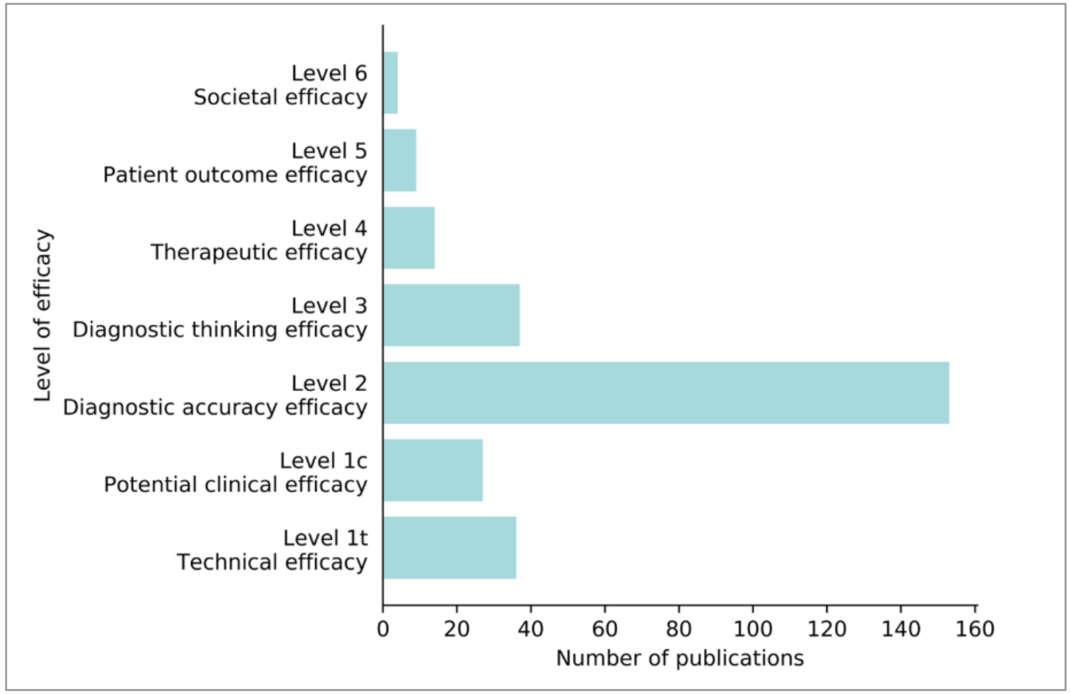
- Merkt AI, wenn vorliegende Daten nicht ausreichen, oder fehlerhaft sind?
- Wenn sie etwas nicht weiss oder versteht?

Rosenberg GS et al. doi: 10.2340/17453674.2024.40182

Beispiel II: AI-assistierte Befundung

- AI-Assistenz für Befundung von Bildgebung inzwischen weit verbreitet (*siehe gfs*)
- Grosses Potential, diagnostische Präzision und Effizienz zu verbessern
- Für viele Systeme gibt es bisher keine unabhängigen Untersuchungen
  (Von 100 CE-markierten, kommerziellen Produkten liegen **nur für 36% wissenschaftliche Publikationen** vor)
- Vorhandene Evidenz bezieht sich vorrangig auf technische Performance und Genauigkeit
- Für Patientennutzen ist **gute Zusammenarbeit** AI-Mensch zentral

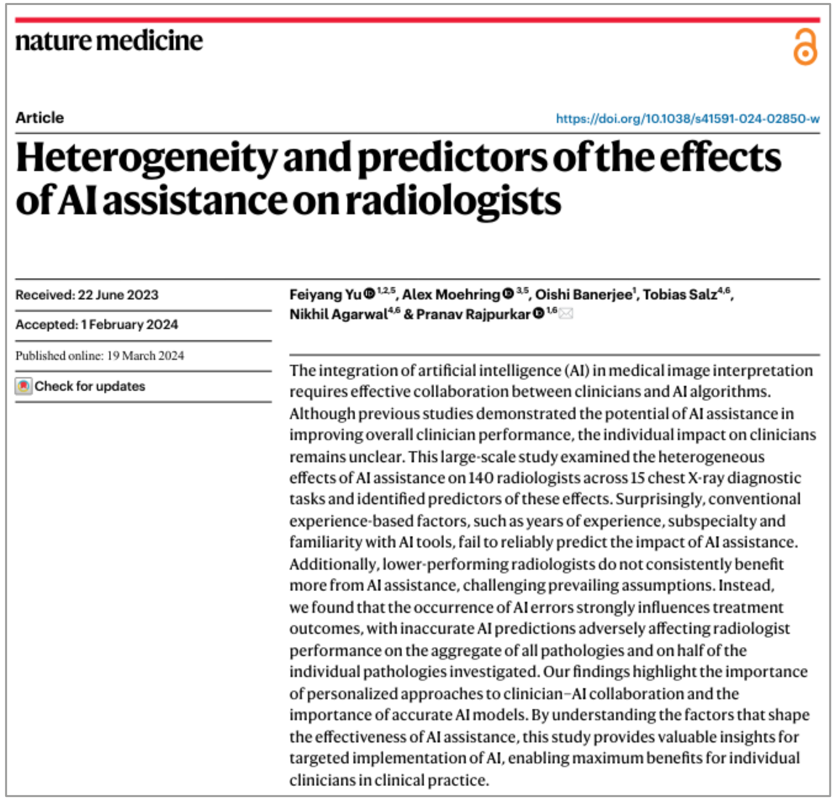
Fig. 5 The levels of efficacy of the included papers. The search strategy yielded 239 peer-reviewed publications on the efficacy of 36 out of 100 commercially available AI products. A single paper could address multiple levels



van Leeuwen KG et al. doi: 10.1007/s00330-021-07892-z

Beispiel II: AI-assistierte Befundung

- 140 Radiologen befunden jeweils 15 Röntgen-Thorax mit / ohne AI-Unterstützung
- Erfahrung und Spezialisierung von Radiologen sind **keine Prädiktoren** für Verbesserung durch AI
- Diagnostische Fähigkeit (Performance ohne AI) ist **kein Prädiktor** für Verbesserung durch AI
- **Radiologen können nicht zuverlässig zwischen genauen und ungenauen AI-Vorhersagen unterscheiden** und werden durch schlechte AI in die Irre geführt
- **Ungenaue AI Vorhersagen mit vielen Fehlern führen zu <u>insgesamt schlechteren Ergebnissen</u>**



nature medicine

Article

https://doi.org/10.1038/s41591-024-02850-w

Heterogeneity and predictors of the effects of AI assistance on radiologists

Received: 22 June 2023

Accepted: 1 February 2024

Published online: 19 March 2024

Check for updates

Feiyang Yu[1,2,5], Alex Moehring[3,5], Oishi Banerjee[1], Tobias Salz[4,6], Nikhil Agarwal[4,6] & Pranav Rajpurkar[1,6]

The integration of artificial intelligence (AI) in medical image interpretation requires effective collaboration between clinicians and AI algorithms. Although previous studies demonstrated the potential of AI assistance in improving overall clinician performance, the individual impact on clinicians remains unclear. This large-scale study examined the heterogeneous effects of AI assistance on 140 radiologists across 15 chest X-ray diagnostic tasks and identified predictors of these effects. Surprisingly, conventional experience-based factors, such as years of experience, subspecialty and familiarity with AI tools, fail to reliably predict the impact of AI assistance. Additionally, lower-performing radiologists do not consistently benefit more from AI assistance, challenging prevailing assumptions. Instead, we found that the occurrence of AI errors strongly influences treatment outcomes, with inaccurate AI predictions adversely affecting radiologist performance on the aggregate of all pathologies and on half of the individual pathologies investigated. Our findings highlight the importance of personalized approaches to clinician–AI collaboration and the importance of accurate AI models. By understanding the factors that shape the effectiveness of AI assistance, this study provides valuable insights for targeted implementation of AI, enabling maximum benefits for individual clinicians in clinical practice.

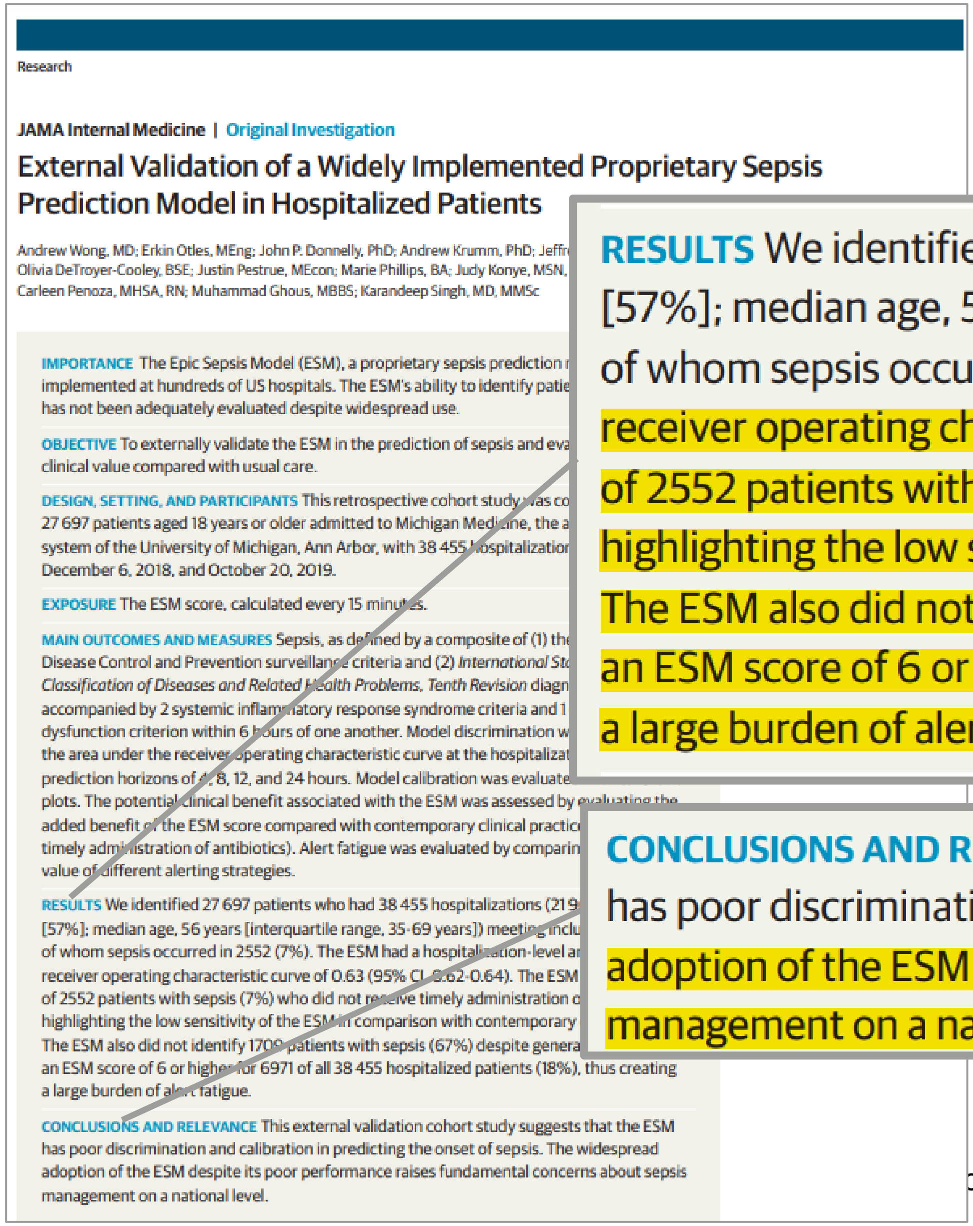
Yu F et al. doi: 10.1038/s41591-024-02850-w

Beispiel III: Frühzeitige Sepsis Erkennung

- Proprietäres Sepsis-Vorhersage-Model (ESM), tief in KIS integriert
  - läuft aktiv in tausenden US-Spitälern
  - nutzt ca. 80 Parameter (z.B. Vitaldaten) in real-time
  - berechnet ca. alle 20 Minuten die individuelle Wahrscheinlichkeit einer Sepsis
  - produziert Warnhinweise und Empfehlungen für das Behandlungsteam

- Werbung versprach eine signifikante Senkung der Mortalität
- Jedoch keine externen Validierungen vor breiter Implementierung 2017
- Zwei grosse externe Validierungsstudien (Wong 2021; Kamran 2024)

Beispiel III: Frühzeitige Sepsis Erkennung



Research

JAMA Internal Medicine | Original Investigation

External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients

Andrew Wong, MD; Erkin Otles, MEng; John P. Donnelly, PhD; Andrew Krumm, PhD; Jeffr
Olivia DeTroyer-Cooley, BSE; Justin Pestrue, MEcon; Marie Phillips, BA; Judy Konye, MSN,
Carleen Penoza, MHSA, RN; Muhammad Ghous, MBBS; Karandeep Singh, MD, MMSc

IMPORTANCE The Epic Sepsis Model (ESM), a proprietary sepsis prediction
implemented at hundreds of US hospitals. The ESM's ability to identify patie
has not been adequately evaluated despite widespread use.

OBJECTIVE To externally validate the ESM in the prediction of sepsis and eva
clinical value compared with usual care.

DESIGN, SETTING, AND PARTICIPANTS This retrospective cohort study was co
27 697 patients aged 18 years or older admitted to Michigan Medicine, the a
system of the University of Michigan, Ann Arbor, with 38 455 hospitalization
December 6, 2018, and October 20, 2019.

EXPOSURE The ESM score, calculated every 15 minutes.

MAIN OUTCOMES AND MEASURES Sepsis, as defined by a composite of (1) the
Disease Control and Prevention surveillance criteria and (2) *International St*
Classification of Diseases and Related Health Problems, Tenth Revision diagn
accompanied by 2 systemic inflammatory response syndrome criteria and 1
dysfunction criterion within 6 hours of one another. Model discrimination w
the area under the receiver operating characteristic curve at the hospitalizat
prediction horizons of 4, 8, 12, and 24 hours. Model calibration was evaluate
plots. The potential clinical benefit associated with the ESM was assessed by evaluating the
added benefit of the ESM score compared with contemporary clinical practic
timely administration of antibiotics). Alert fatigue was evaluated by comparin
value of different alerting strategies.

RESULTS We identified 27 697 patients who had 38 455 hospitalizations (21 9
[57%]; median age, 56 years [interquartile range, 35-69 years]) meeting inclu
of whom sepsis occurred in 2552 (7%). The ESM had a hospitalization-level ar
receiver operating characteristic curve of 0.63 (95% CI, 0.62-0.64). The ESM
of 2552 patients with sepsis (7%) who did not receive timely administration o
highlighting the low sensitivity of the ESM in comparison with contemporary
The ESM also did not identify 1709 patients with sepsis (67%) despite genera
an ESM score of 6 or higher for 6971 of all 38 455 hospitalized patients (18%), thus creating
a large burden of alert fatigue.

CONCLUSIONS AND RELEVANCE This external validation cohort study suggests that the ESM has poor discrimination and calibration in predicting the onset of sepsis. The widespread adoption of the ESM despite its poor performance raises fundamental concerns about sepsis management on a national level.

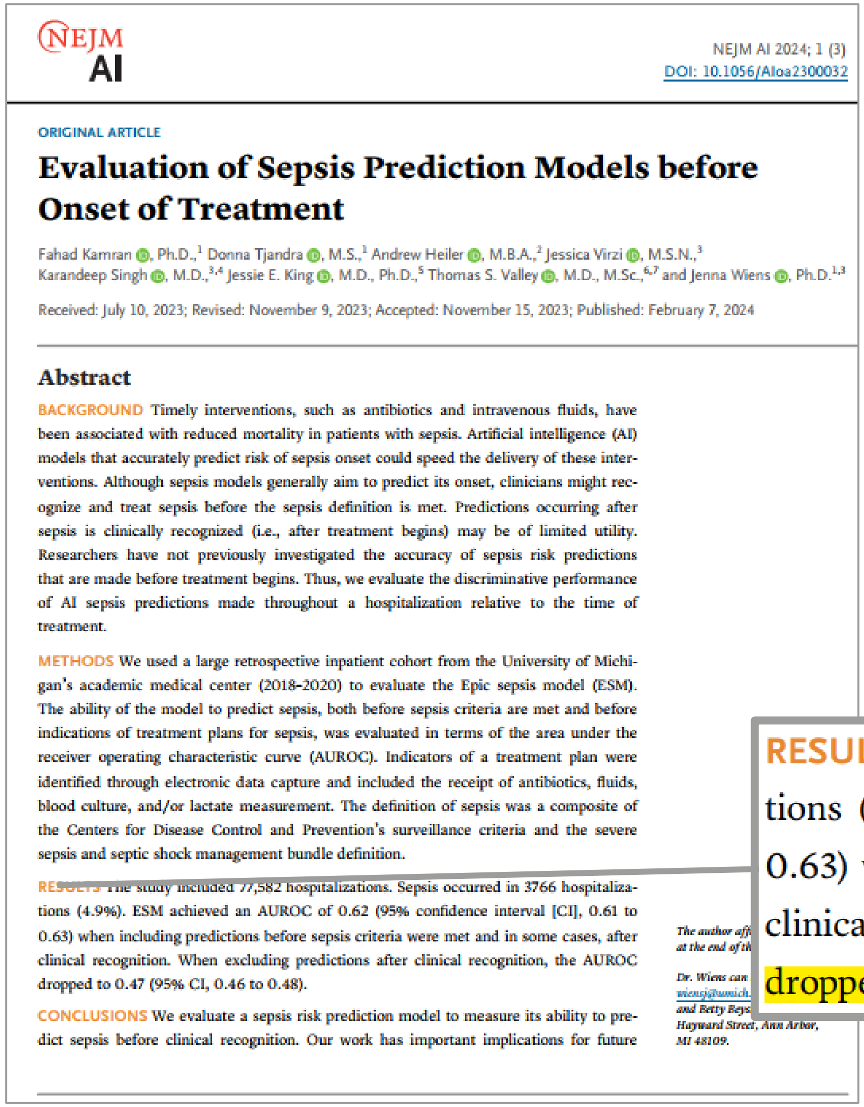
RESULTS We identified 27 697 patients who had 38 455 hospitalizations (21 904 women [57%]; median age, 56 years [interquartile range, 35-69 years]) meeting inclusion criteria, of whom sepsis occurred in 2552 (7%). The ESM had a hospitalization-level area under the receiver operating characteristic curve of 0.63 (95% CI, 0.62-0.64). The ESM identified 183 of 2552 patients with sepsis (7%) who did not receive timely administration of antibiotics, highlighting the low sensitivity of the ESM in comparison with contemporary clinical practice. The ESM also did not identify 1709 patients with sepsis (67%) despite generating alerts for an ESM score of 6 or higher for 6971 of all 38 455 hospitalized patients (18%), thus creating a large burden of alert fatigue.

CONCLUSIONS AND RELEVANCE This external validation cohort study suggests that the ESM has poor discrimination and calibration in predicting the onset of sepsis. The widespread adoption of the ESM despite its poor performance raises fundamental concerns about sepsis management on a national level.

ch

Wong A et al. doi: 10.1001/jamainternmed.2021.2626

Beispiel III: Frühzeitige Sepsis Erkennung



NEJM AI

NEJM AI 2024; 1 (3)
DOI: 10.1056/AIoa2300032

ORIGINAL ARTICLE

Evaluation of Sepsis Prediction Models before Onset of Treatment

Fahad Kamran, Ph.D.,[1] Donna Tjandra, M.S.,[1] Andrew Heiler, M.B.A.,[2] Jessica Virzi, M.S.N.,[3] Karandeep Singh, M.D.,[3,4] Jessie E. King, M.D., Ph.D.,[5] Thomas S. Valley, M.D., M.Sc.,[6,7] and Jenna Wiens, Ph.D.[1,3]

Received: July 10, 2023; Revised: November 9, 2023; Accepted: November 15, 2023; Published: February 7, 2024

Abstract

BACKGROUND Timely interventions, such as antibiotics and intravenous fluids, have been associated with reduced mortality in patients with sepsis. Artificial intelligence (AI) models that accurately predict risk of sepsis onset could speed the delivery of these interventions. Although sepsis models generally aim to predict its onset, clinicians might recognize and treat sepsis before the sepsis definition is met. Predictions occurring after sepsis is clinically recognized (i.e., after treatment begins) may be of limited utility. Researchers have not previously investigated the accuracy of sepsis risk predictions that are made before treatment begins. Thus, we evaluate the discriminative performance of AI sepsis predictions made throughout a hospitalization relative to the time of treatment.

METHODS We used a large retrospective inpatient cohort from the University of Michigan's academic medical center (2018–2020) to evaluate the Epic sepsis model (ESM). The ability of the model to predict sepsis, both before sepsis criteria are met and before indications of treatment plans for sepsis, was evaluated in terms of the area under the receiver operating characteristic curve (AUROC). Indicators of a treatment plan were identified through electronic data capture and included the receipt of antibiotics, fluids, blood culture, and/or lactate measurement. The definition of sepsis was a composite of the Centers for Disease Control and Prevention's surveillance criteria and the severe sepsis and septic shock management bundle definition.

RESULTS The study included 77,582 hospitalizations. Sepsis occurred in 3766 hospitalizations (4.9%). ESM achieved an AUROC of 0.62 (95% confidence interval [CI], 0.61 to 0.63) when including predictions before sepsis criteria were met and in some cases, after clinical recognition. When excluding predictions after clinical recognition, the AUROC dropped to 0.47 (95% CI, 0.46 to 0.48).

CONCLUSIONS We evaluate a sepsis risk prediction model to measure its ability to predict sepsis before clinical recognition. Our work has important implications for future

The author aff
at the end of th

Dr. Wiens can
wiensj@umich.
and Betty Beys
Hayward Street, Ann Arbor, MI 48109.

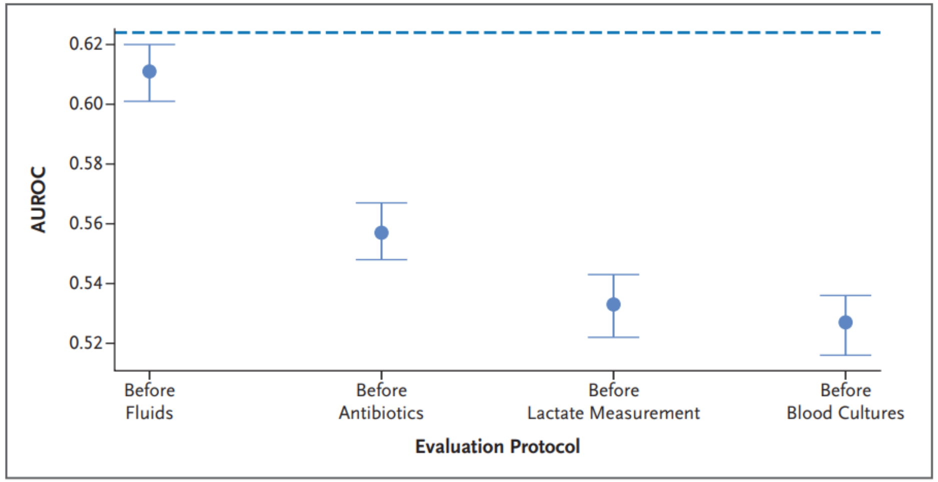


Figure 3. Evaluating the Accuracy of the ESM with Respect to Different Treatments.
We visualize the model's performance with 95% confidence intervals for each evaluation. The blue dashed line denotes the ESM's performance before the time of meeting the sepsis criteria. Its performance drops the most when predictions are made only before the time of blood culture orders, achieving nearly random performance. Meanwhile, the model's performance drops only slightly when using predictions before orders for fluids. AUROC denotes area under the receiver operating characteristic curve; and ESM, Epic sepsis model.

RESULTS The study included 77,582 hospitalizations. Sepsis occurred in 3766 hospitalizations (4.9%). ESM achieved an AUROC of 0.62 (95% confidence interval [CI], 0.61 to 0.63) when including predictions before sepsis criteria were met and in some cases, after clinical recognition. When excluding predictions after clinical recognition, the AUROC dropped to 0.47 (95% CI, 0.46 to 0.48).

Kamran F et al. doi: DOI: 10.1056/AIoa2300032

Beispiel III: Frühzeitige Sepsis Erkennung

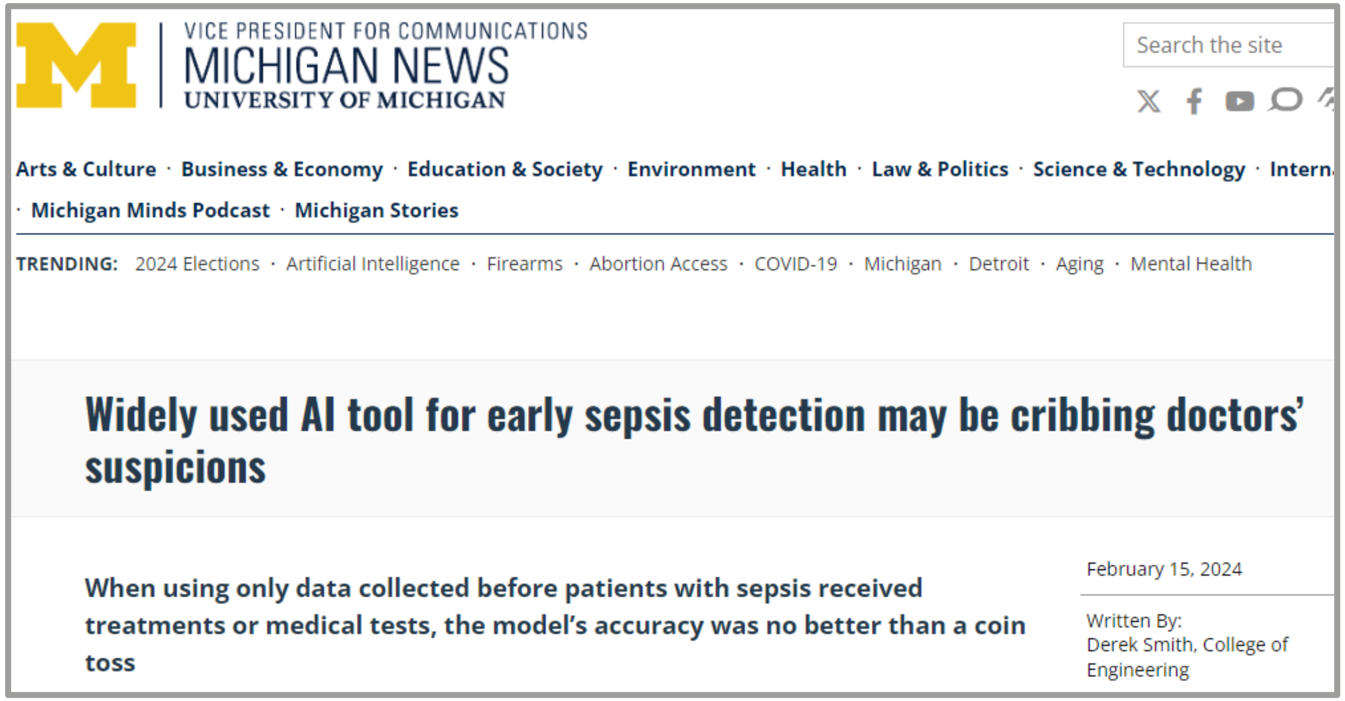
Das Problem ist …

- nicht, dass der Algorithmus nicht optimal ist !
- sondern dass er sehr breit implementiert wurde,
- mit grossen Versprechungen,
- ohne extern validiert zu sein,
- ohne transparent und offen zugänglich zu sein.

> *Poor timeliness combined with increased score complexity and **lack of transparency** of the SPM epitomizes its major flaw: it appears to predict sepsis **long after the clinician has recognized** possible sepsis and acted on that suspicion.*

Schertz AR et al. doi:10.1001/jamanetworkopen.2023.29729

Beispiel III: Frühzeitige Sepsis Erkennung


VICE PRESIDENT FOR COMMUNICATIONS
MICHIGAN NEWS
UNIVERSITY OF MICHIGAN
Search the site
Arts & Culture · Business & Economy · Education & Society · Environment · Health · Law & Politics · Science & Technology · Intern
· Michigan Minds Podcast · Michigan Stories
TRENDING: 2024 Elections · Artificial Intelligence · Firearms · Abortion Access · COVID-19 · Michigan · Detroit · Aging · Mental Health

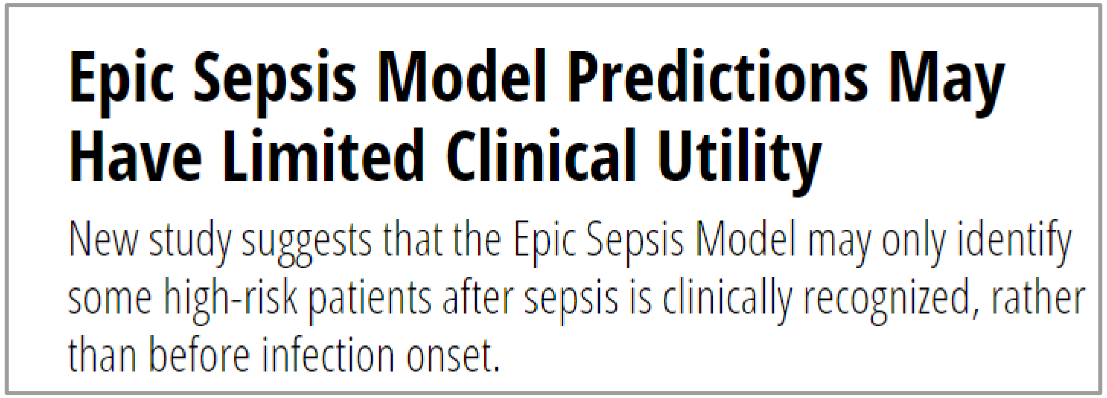
Widely used AI tool for early sepsis detection may be cribbing doctors' suspicions

When using only data collected before patients with sepsis received treatments or medical tests, the model's accuracy was no better than a coin toss

February 15, 2024

Written By:
Derek Smith, College of Engineering

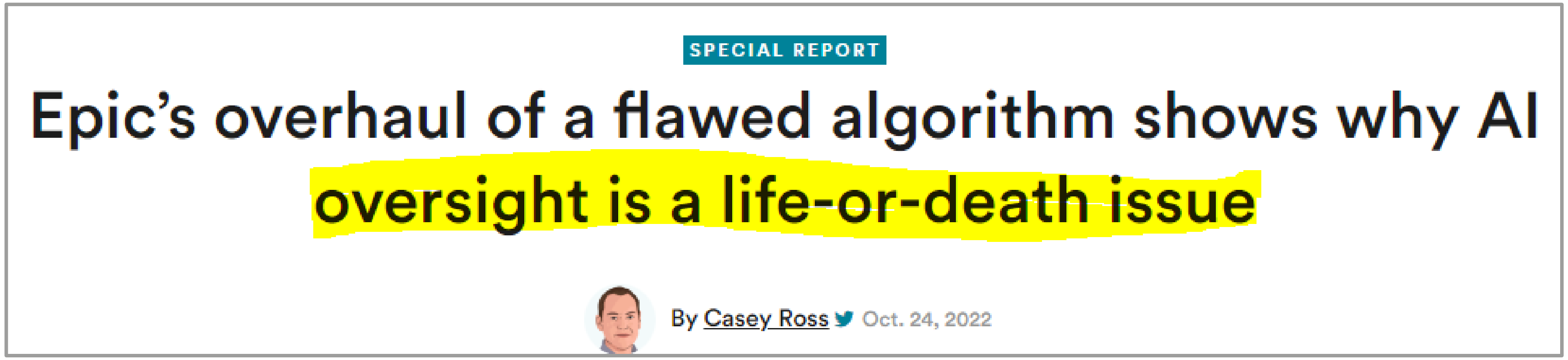
https://news.umich.edu/widely-used-ai-tool-for-early-sepsis-detection-may-be-cribbing-doctors-suspicions/



Epic Sepsis Model Predictions May Have Limited Clinical Utility

New study suggests that the Epic Sepsis Model may only identify some high-risk patients after sepsis is clinically recognized, rather than before infection onset.

https://healthitanalytics.com/news/epic-sepsis-model-predictions-may-have-limited-clinical-utility



SPECIAL REPORT

Epic's overhaul of a flawed algorithm shows why AI oversight is a life-or-death issue

By Casey Ross Oct. 24, 2022

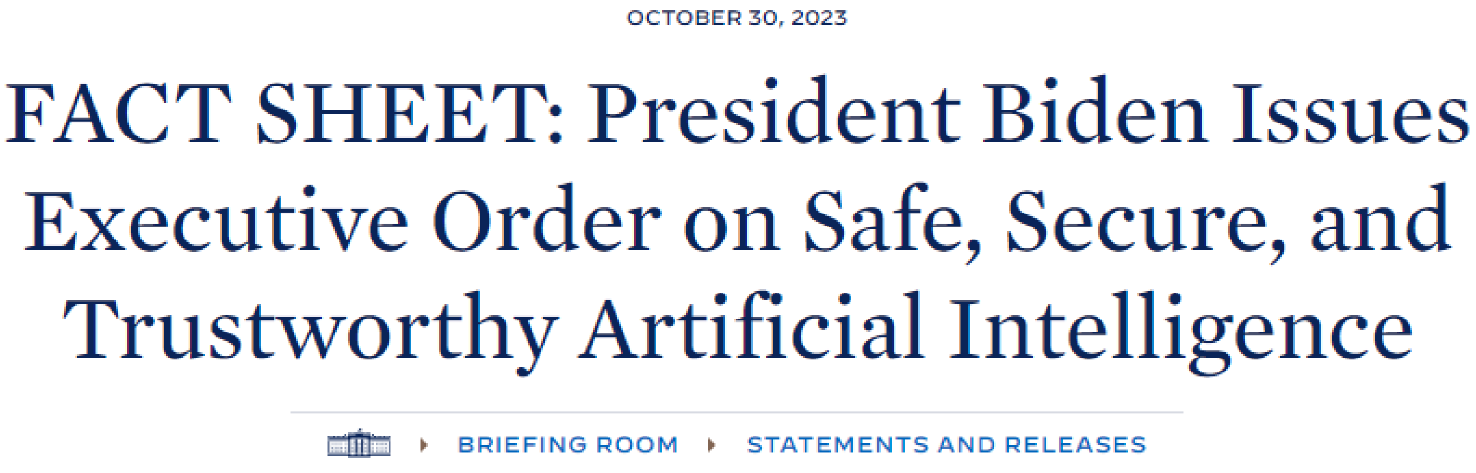
https://www.statnews.com/2022/10/24/epic-overhaul-of-a-flawed-algorithm/

Menschliche Übersicht ?

- AI ist nicht perfekt, aber nützlich genug …
- Forderung der human oversight / vigilance:
  Kliniker sollen im „Zweifel“-Modus arbeiten: *AI kontrollieren, validieren, Fehler suchen*

- Problem 1: Transparenz der AI-Integration in klinischen Anwendungen und ihrer Güte
- Problem 2: Benötigte Zeit-Ressourcen werden weg-rationalisiert werden
- **Problem 3: Menschen sind keine guten Wächter**
  - In einer Vielzahl von korrekten Outputs Fehler oder Lücken zu erkennen, ist kognitiv extrem anspruchsvoll (hohe Aufmerksamkeit ohne Aktivität)
  - De-Skilling

Adler-Milstein J et al. doi: 10.1001/jama.2024.3620



OCTOBER 30, 2023

FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence

› BRIEFING ROOM › STATEMENTS AND RELEASES

(iv) Within **365 days of the date of this order**, the Secretary of HHS shall, in consultation with the Secretary of Defense and the Secretary of Veterans Affairs, establish **an AI safety program** that, in partnership with voluntary federally listed Patient Safety Organizations:

(A) **establishes a common framework for approaches to identifying and capturing clinical errors resulting from AI deployed in healthcare settings** as well as specifications for a central tracking repository for associated incidents that cause harm, including through bias or discrimination, to patients, caregivers, or other parties;

(B) **analyzes captured data and generated evidence to develop**, wherever appropriate, **recommendations, best practices, or other informal guidelines** aimed at avoiding these harms; and

(C) **disseminates those recommendations**, best practices, or other informal guidance to appropriate stakeholders, including healthcare providers.

…

Develop standards, tools, and tests to help ensure that AI systems are safe, secure, and trustworthy. The National Institute of Standards and Technology will set the rigorous standards for **extensive red-team testing** to ensure safety before public release.

https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/

Ausblick

- Wir benötigen SCHNELL eine strukturierte Vorgehensweise, um AI auf Sicherheits-Effekte zu testen (red-teaming)

- CAVE: Technologie-Adoptions-Paradox:
  Es wird bald niemanden mehr geben, der in Vergleichs-Studien *ohne AI* arbeiten will auch falls die existierende Evidenz mager ist (siehe *clinical decision support* Baysari et al. 2023)

- Tiefe Integration von AI in Klinik- und Praxisinformationssysteme erleichtert Arbeitsfluss, erschwert aber Interpretierbarkeit und *human oversight*

- *Human oversight* im klinischen Alltag ist langfristig keine effektive und sichere Strategie

- Zukunft: AI als Team-Member in Kollaboration und Interaktion (z.B. Tumorboard)