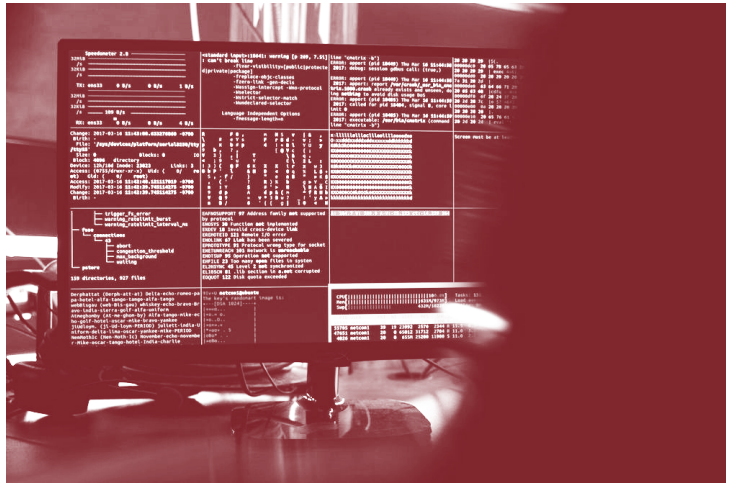
IA et sécurité des patients : vers un avenir incertain

Prof. David Schwappach, MPH

Institut de médecine sociale et préventive (ISPM)
Université de Berne
David.Schwappach@unibe.ch

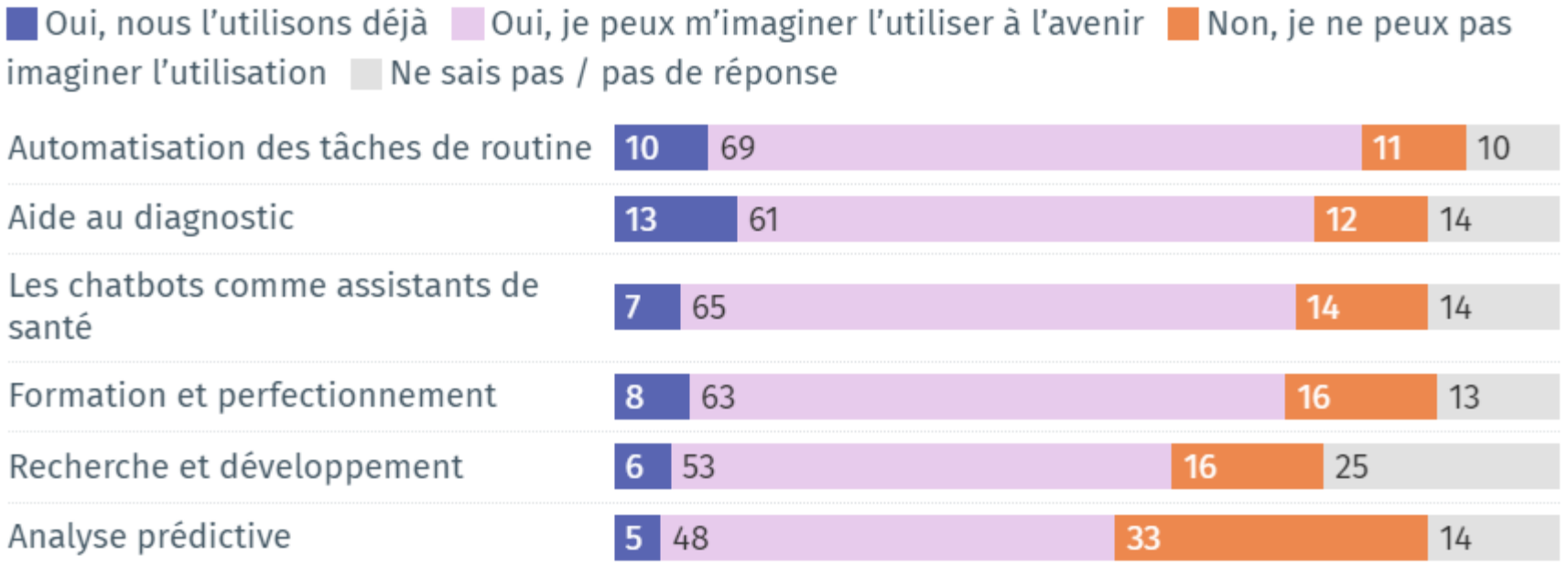
24 mai 2024, Symposium de l'ASQM, UniS Berne



Utilisation potentielle de l'intelligence artificielle

Dans quels domaines pouvez-vous imaginer l'utilisation de l'intelligence artificielle dans votre quotidien professionnel au cours des cinq prochaines années?

en % des répondants



© gfs.bern, Le Baromètre cybersanté, professionels de santé, novembre 2023 - janvier 2024 (n=1440)

https://cockpit.gfsbern.ch/fr/cockpit/ehealth-professionnels-de-sante-2024/

Applications actuelles de l'IA

- Soutien dans les tâches de routine
  p. ex. rapport de sortie, vérification de la médication, documentation de la consultation (« digital scribe »)

- Aide à l'interprétation des résultats d'imagerie
  p. ex. radiologie, IRM, dermatologie, ophtalmologie (p. ex. rétinopathie diabétique)

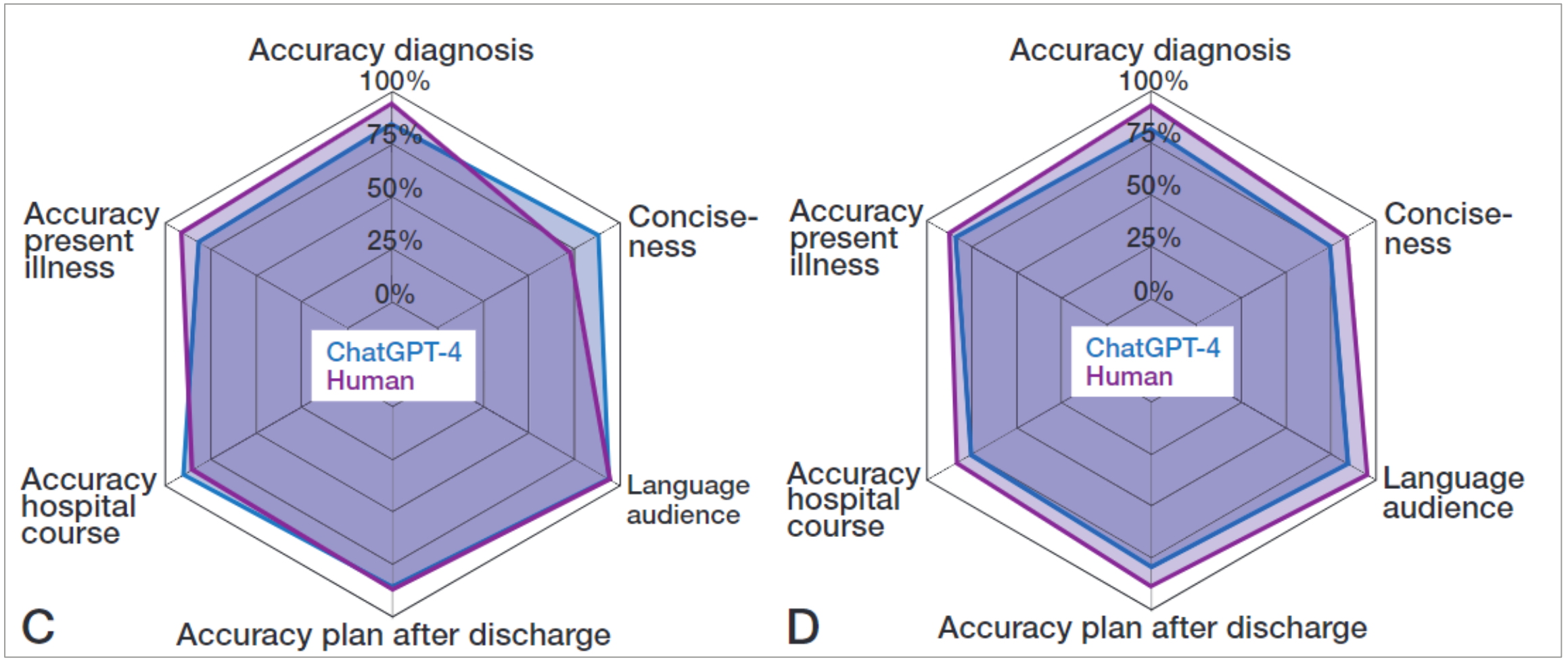
- Détection précoce de la détérioration de l'état de santé de patients
  p. ex. septicémie, escarres, événement médicamenteux indésirable, complication chirurgicale

Exemple I : lettres de sortie en orthopédie

- Rédaction de documents de sortie pour 6 patients fictifs sur la base du dossier médical (y compris résultats de laboratoire, d'analyses et d'imagerie, notes médicales, médication, etc.)
- Tâche : rédaction d'un rapport de sortie à l'intention du médecin de famille et d'un autre rapport à l'intention du patient, conformément au format standard des hôpitaux
- Médecin-assistant vs chef de clinique vs Chat-GPT4
- Évaluation des documents par des cliniciens expérimentés et en aveugle (n=15) sur la base de critères définis

Rosenberg GS et al. doi: 10.2340/17453674.2024.40182

Exemple I : lettres de sortie en orthopédie



C. Swiss summary; D. Swiss letter.

Rosenberg GS et al. doi: 10.2340/17453674.2024.40182

Exemple I : lettres de sortie en orthopédie

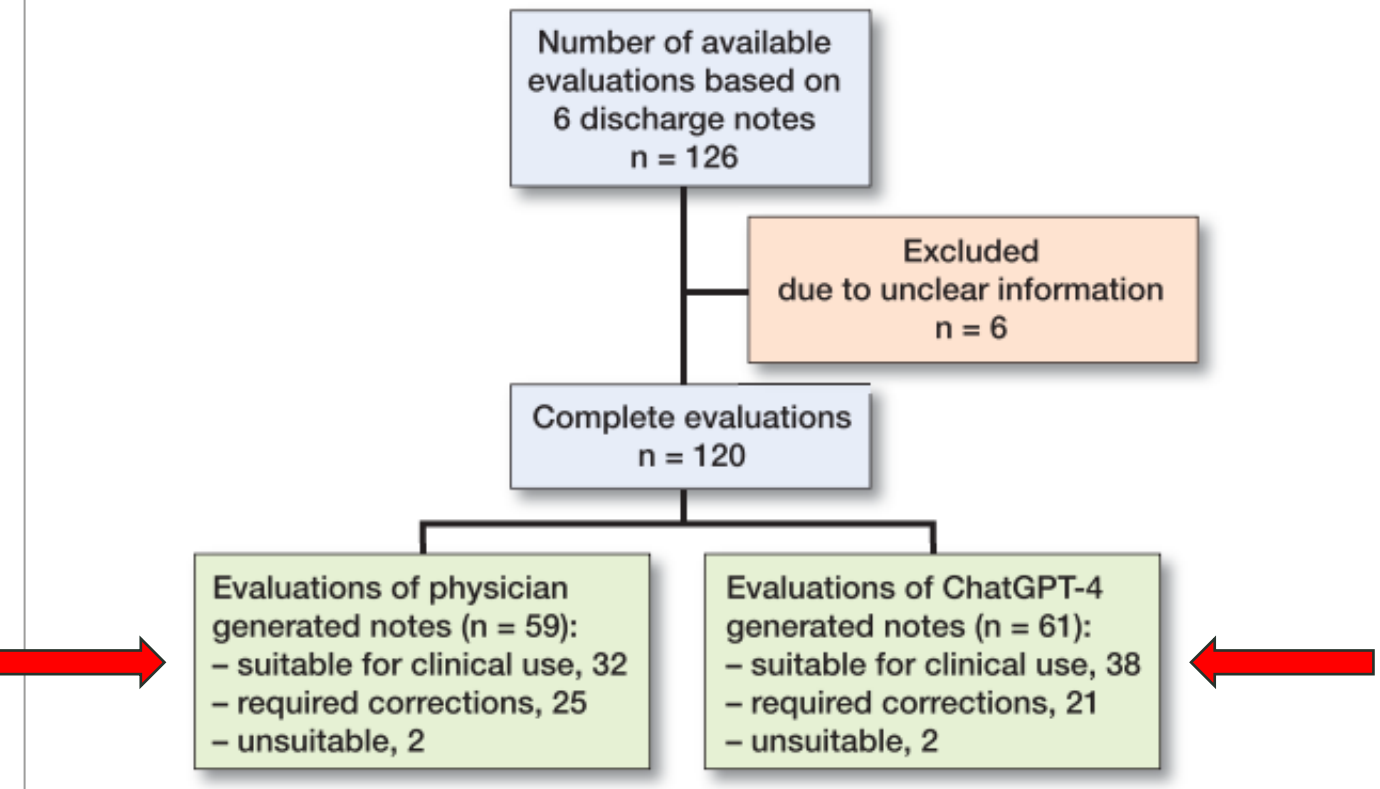


Figure 2. Flowchart of evaluations of discharge notes by the expert panel.

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

Introspection ?

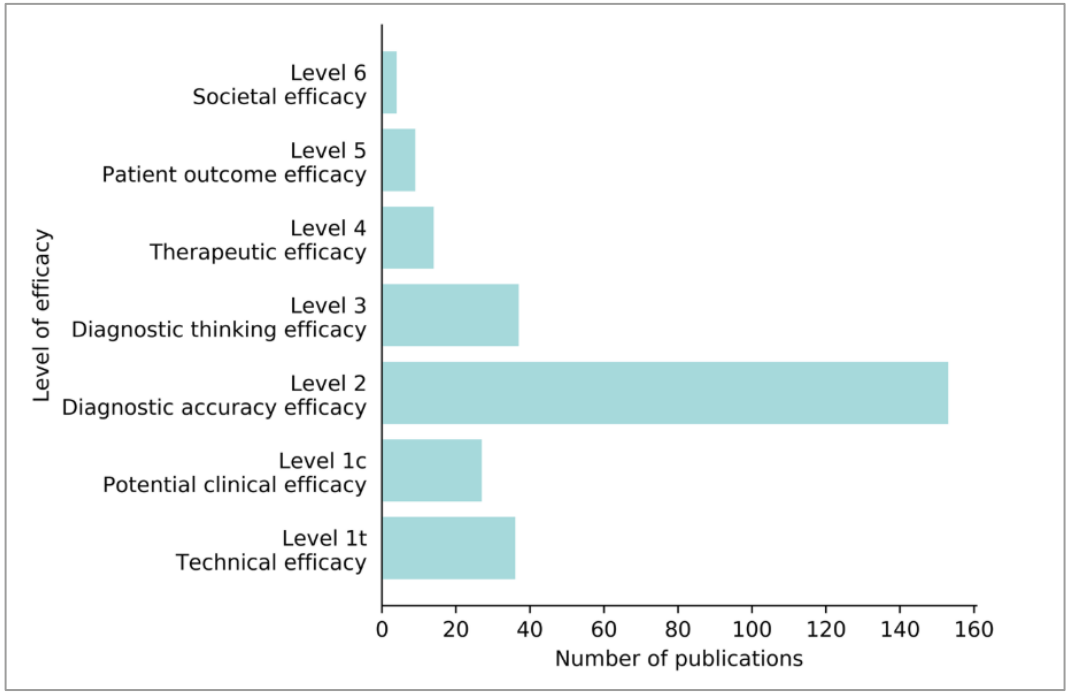
- L'IA remarque-t-elle si les données disponibles ne sont pas suffisantes ou sont erronées ?
- Si elle ne sait pas ou ne comprend pas quelque chose ?

Rosenberg GS et al. doi: 10.2340/17453674.2024.40182

Exemple II : aide au diagnostic d'imagerie

- L'IA est désormais largement répandue en radiologie *(cf. gfs)*
- Grand potentiel pour améliorer la précision et l'efficacité du diagnostic
- Pour de nombreux systèmes, il n'existe à ce jour aucune étude indépendante
  (sur 100 produits commerciaux marqués CE, il n'existe **des publications scientifiques que pour 36 %** d'entre eux)
- Les preuves existantes concernent principalement la performance et la précision techniques
- Une **bonne collaboration** entre l'IA et l'humain est essentielle pour le bénéfice du patient

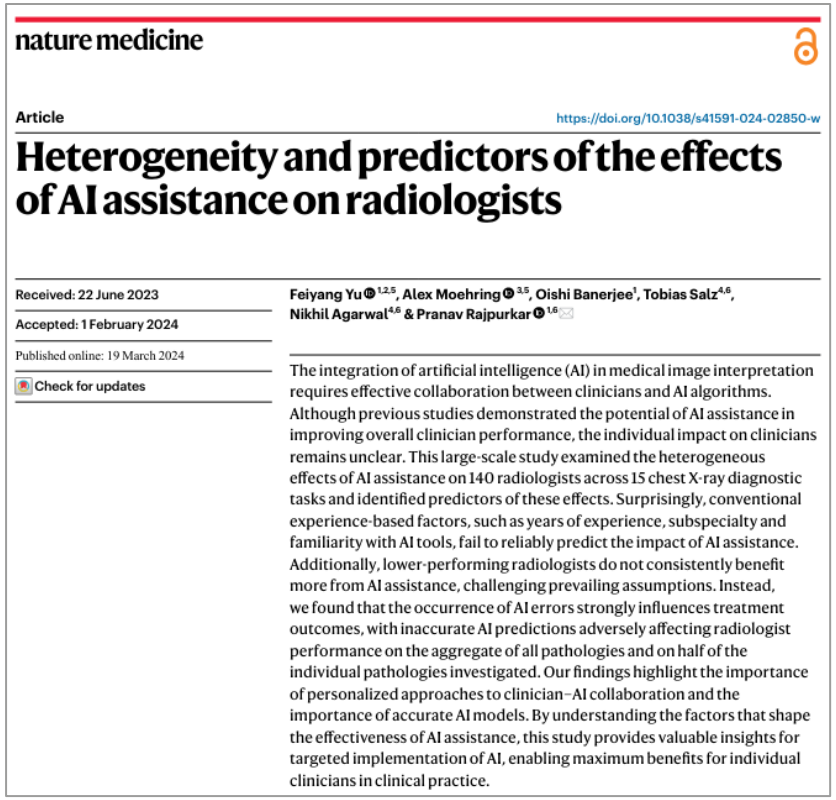
Fig. 5 The levels of efficacy of the included papers. The search strategy yielded 239 peer-reviewed publications on the efficacy of 36 out of 100 commercially available AI products. A single paper could address multiple levels



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

Exemple II : aide au diagnostic d'imagerie

- 140 radiologues interprètent chacun 15 clichés du thorax avec / sans assistance par l'IA
- L'expérience et la spécialisation des radiologues ne sont **pas des prédicteurs** d'amélioration par l'IA
- La capacité diagnostique (performance sans IA) n'est **pas un prédicteur** d'amélioration par l'IA
- **Les radiologues ne peuvent pas faire de distinction fiable entre les prédictions IA précises et imprécises** et sont induits en erreur par une IA de mauvaise qualité
- **Des prédictions IA imprécises avec de nombreuses erreurs entraînent <u>des résultats globalement moins bons</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.

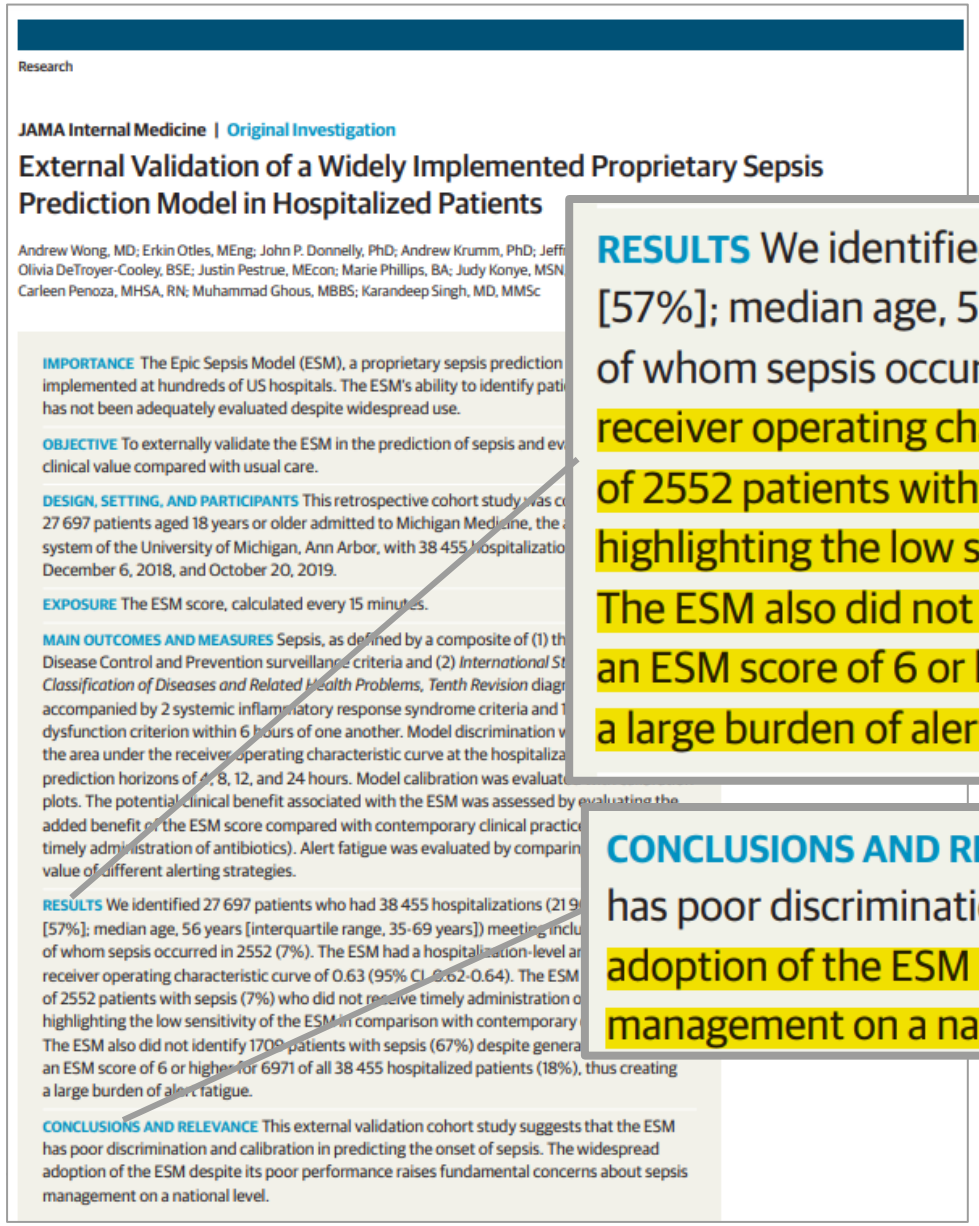
Yu F et al. doi: 10.1038/s41591-024-02850-w

Exemple III : détection précoce du sepsis

- Un modèle propriétaire de prédiction du sepsis (ESM), profondément intégré dans les systèmes informatiques hospitaliers,
  - est actuellement utilisé dans des milliers d'hôpitaux américains
  - recourt à environ 80 paramètres (p. ex. données vitales) en temps réel
  - calcule toutes les 20 minutes environ la probabilité individuelle de sepsis
  - génère des avertissements et des recommandations pour l'équipe soignante

- La publicité promettait une réduction significative de la mortalité
- Cependant, aucune validation externe avant une large implémentation en 2017
- Deux grandes études de validation externes (Wong 2021 ; Kamran 2024)

Exemple III : détection précoce du sepsis



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; Jeffi
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 pati
has not been adequately evaluated despite widespread use.

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

DESIGN, SETTING, AND PARTICIPANTS This retrospective cohort study was c
27 697 patients aged 18 years or older admitted to Michigan Medicine, the
system of the University of Michigan, Ann Arbor, with 38 455 hospitalizatio
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) th
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
dysfunction criterion within 6 hours of one another. Model discrimination
the area under the receiver operating characteristic curve at the hospitaliza
prediction horizons of 4, 8, 12, and 24 hours. Model calibration was evaluat
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.

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

Exemple III : détection précoce du sepsis

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

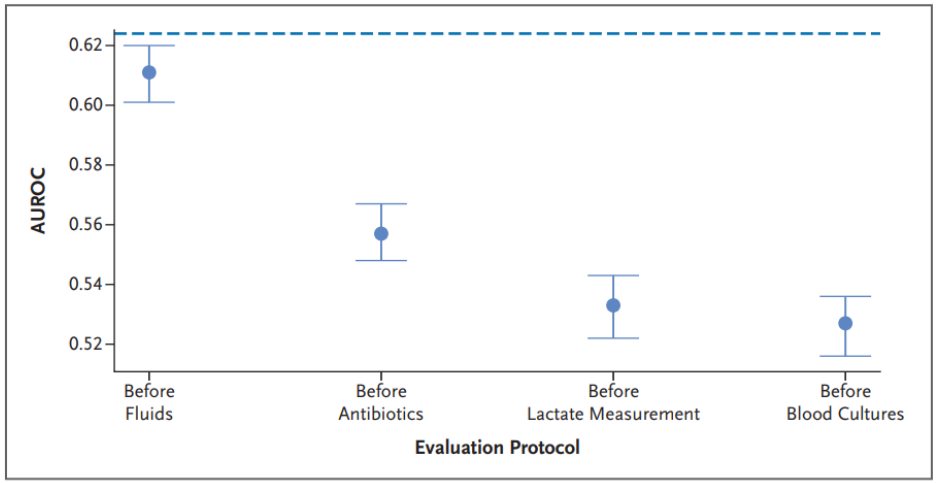


Figure 3. Evaluating the Accuracy of the ESM with Respect to Different Treatments.
We visualize the model's performance with 95% confidence intervals for each evaluation. The blue dashed line denotes the ESM's performance before the time of meeting the sepsis criteria. Its performance drops the most 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

Exemple III : détection précoce du sepsis

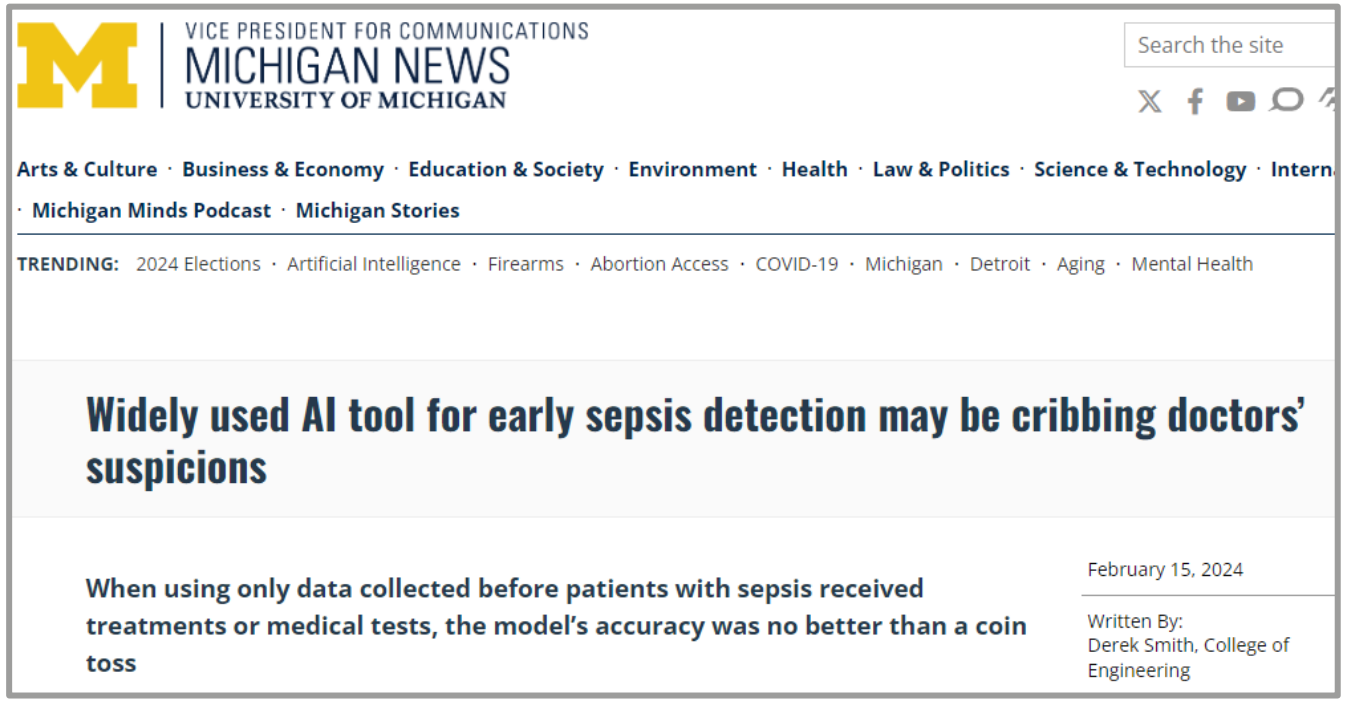
Le problème...

- n'est pas que l'algorithme ne soit pas optimal !
- mais qu'il ait été implémenté à très large échelle,
- à grand renfort de promesses,
- sans validation externe,
- sans être transparent ni librement accessible.

*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

Exemple III : détection précoce du sepsis



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

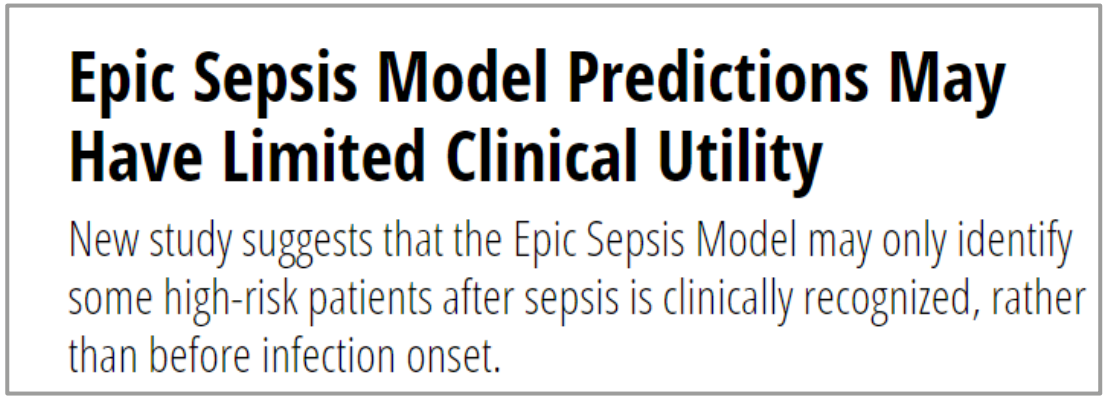
Widely used AI tool for early sepsis detection may be cribbing doctors' suspicions

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/

Surveillance humaine ?

- L'IA n'est pas parfaite, mais suffisamment utile...
- Requiert une vigilance / surveillance humaine :
  les cliniciens doivent travailler en mode « doute » : *contrôler l'IA, la valider, rechercher les erreurs*

- Problème 1 : transparence de l'intégration de l'IA dans les applications cliniques et de sa qualité
- Problème 2 : disparition des ressources en temps nécessaires (rationalisation)
- **Problème 3 : l'humain n'est pas un bon gardien**
  - Identifier des erreurs ou des lacunes parmi une multitude de données correctes est extrêmement exigeant sur le plan cognitif (attention élevée sans activité)
  - 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/

Perspectives

- Nous avons besoin RAPIDEMENT d'une approche structurée pour tester les effets de l'IA sur la sécurité (red-teaming)
- ATTENTION : paradoxe de l'adoption de la technologie :
  il n'y aura bientôt plus personne qui voudra travailler dans des études comparatives *sans IA* même si les preuves existantes sont maigres (cf. *clinical decision support* Baysari et al. 2023)
- L'intégration profonde de l'IA dans les systèmes informatiques des cliniques et des cabinets médicaux facilite le flux de travail, mais complique l'interprétabilité et la surveillance humaine
- La surveillance humaine dans le quotidien clinique n'est pas une stratégie efficace et sûre à long terme
- L'avenir : l'IA en tant que membre de l'équipe dans la collaboration et l'interaction (p. ex. tumorboard)