RADIATION ONCOLOGY QUALITY & SAFETY GRAND ROUNDS

AI in the Clinic:

What Could Go Wrong, and How We Can Catch It

Andrew Y. K. Foong, Ph.D.

November 21st, 2025

MAYO CLINIC

Radiation Oncology
AI & Data Analytics
AIDA

Objectives

- Examine real-world AI failures to see how things can go wrong in practice.
- Understand similarities and differences between AI safety and safety in other disciplines.
- Identify how *automation bias* and *data drift* undermine AI performance over time.
- Explore strategies for *cognitive forcing* and *continuous monitoring* to mitigate these problems

Today's talk

1. Case studies
   - *Computer aided diagnosis of mammography*
   - *The Epic sepsis model*
   - *Colonoscopy deskilling*
2. Lessons for safety in other fields
   - *Traditional software*
   - *Humans*
   - *Drugs*
3. Mitigation strategies
   - *Tackling drift through continuous monitoring*
   - *Tackling automation bias through cognitive forcing*

Computer aided diagnosis

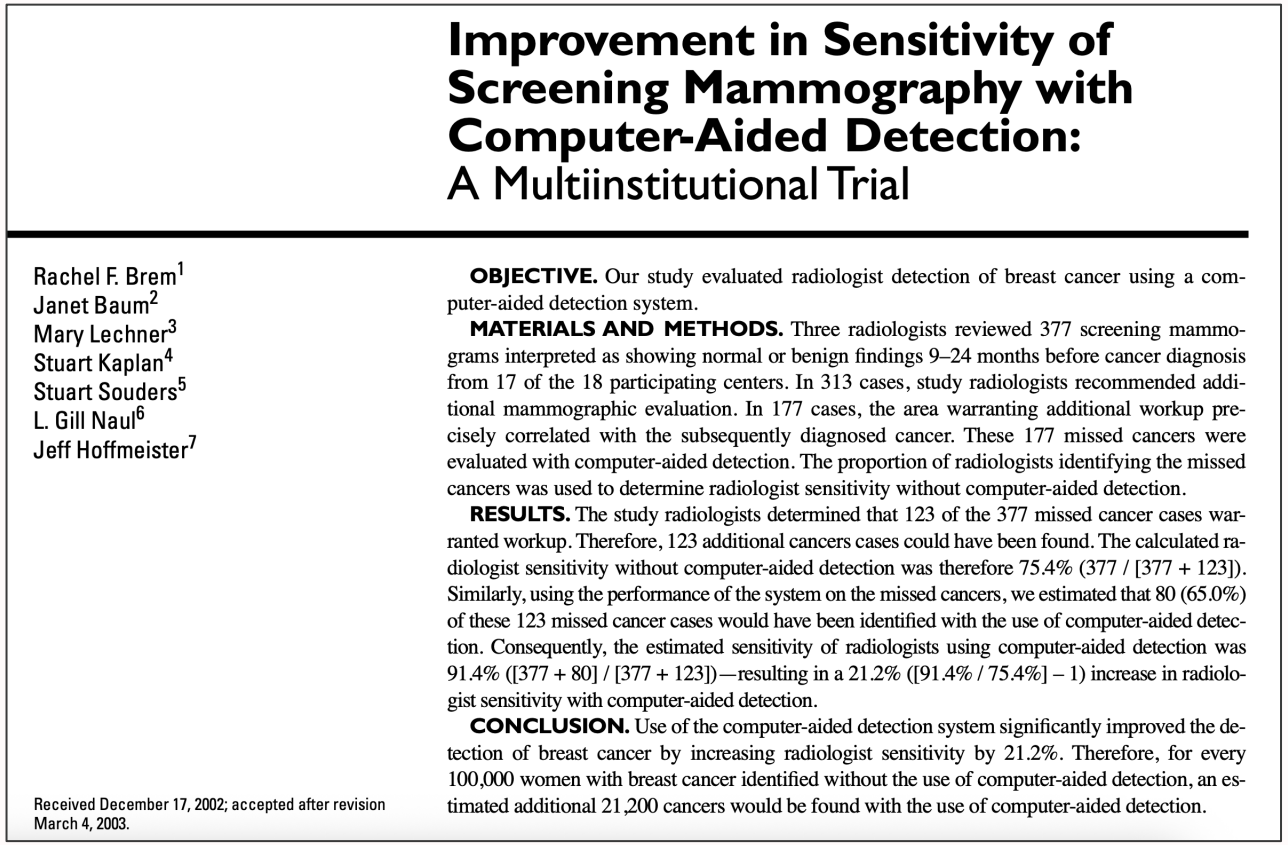
Case study 1

Mammogram reading

- Since the 1990s, studies had shown that accuracy increases when there's more than one reader.
- AI systems were introduced to act as a virtual second reader.
- System was called *SecondLook*, introduced in the late 1990s
  - *Note*: this was a very early system, and much less accurate than modern AI tools for reading mammograms

Mammogram reading

Early papers showed promising results:



Improvement in Sensitivity of Screening Mammography with Computer-Aided Detection:
A Multiinstitutional Trial

Rachel F. Brem[1]
Janet Baum[2]
Mary Lechner[3]
Stuart Kaplan[4]
Stuart Souders[5]
L. Gill Naul[6]
Jeff Hoffmeister[7]

OBJECTIVE. Our study evaluated radiologist detection of breast cancer using a computer-aided detection system.

MATERIALS AND METHODS. Three radiologists reviewed 377 screening mammograms interpreted as showing normal or benign findings 9–24 months before cancer diagnosis from 17 of the 18 participating centers. In 313 cases, study radiologists recommended additional mammographic evaluation. In 177 cases, the area warranting additional workup precisely correlated with the subsequently diagnosed cancer. These 177 missed cancers were evaluated with computer-aided detection. The proportion of radiologists identifying the missed cancers was used to determine radiologist sensitivity without computer-aided detection.

RESULTS. The study radiologists determined that 123 of the 377 missed cancer cases warranted workup. Therefore, 123 additional cancers cases could have been found. The calculated radiologist sensitivity without computer-aided detection was therefore 75.4% (377 / [377 + 123]). Similarly, using the performance of the system on the missed cancers, we estimated that 80 (65.0%) of these 123 missed cancer cases would have been identified with the use of computer-aided detection. Consequently, the estimated sensitivity of radiologists using computer-aided detection was 91.4% ([377 + 80] / [377 + 123])—resulting in a 21.2% ([91.4% / 75.4%] – 1) increase in radiologist sensitivity with computer-aided detection.

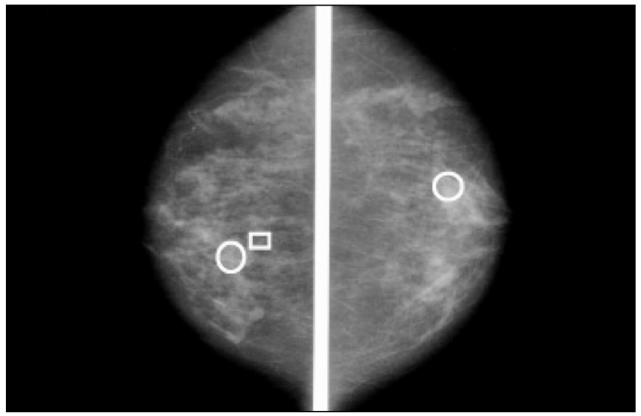
CONCLUSION. Use of the computer-aided detection system significantly improved the detection of breast cancer by increasing radiologist sensitivity by 21.2%. Therefore, for every 100,000 women with breast cancer identified without the use of computer-aided detection, an estimated additional 21,200 cancers would be found with the use of computer-aided detection.

Received December 17, 2002; accepted after revision March 4, 2003.

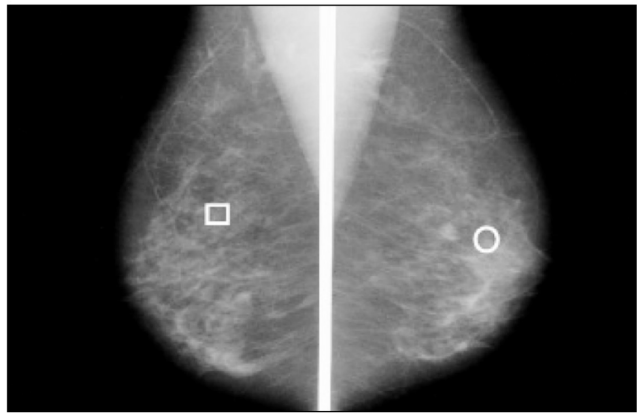
Brem, Rachel F., Janet Baum, Mary Lechner, Stuart Kaplan, Stuart Souders, L. Gill Naul, and Jeff Hoffmeister. “Improvement in sensitivity of screening mammography with computer-aided detection: a multiinstitutional trial.” *American Journal of Roentgenology* 181, no. 3 (2003): 687-693.

Mammogram reading

Early papers showed promising results:



A



B

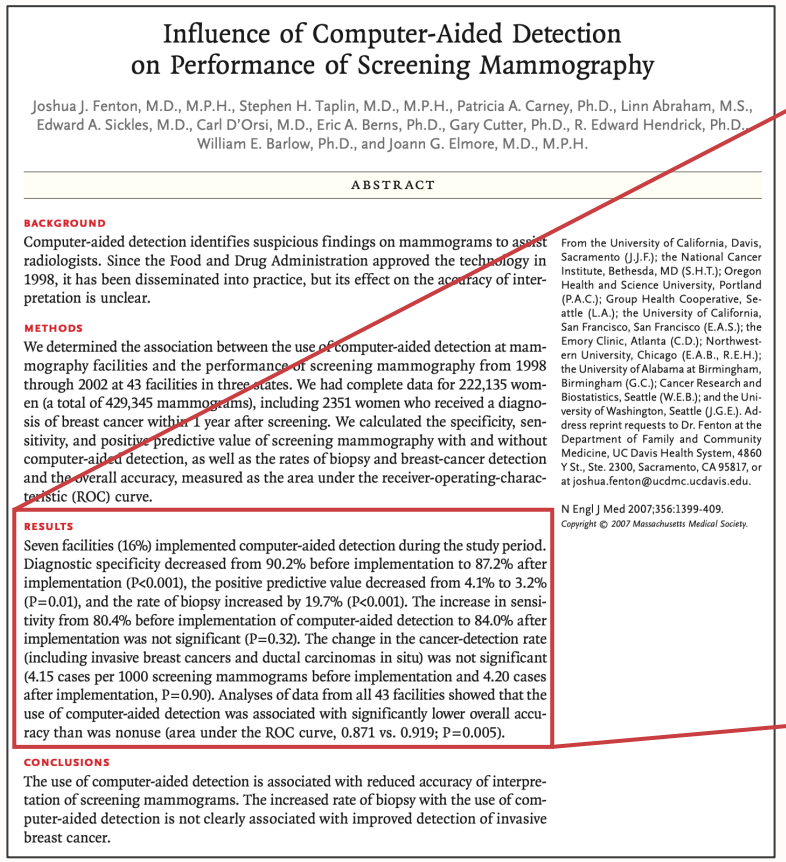
Fig. 2.—Printouts of digitized mammograms in 56-year-old woman who underwent screening mammography for breast cancer. Printouts were generated by computer-aided detection software (Second Look [version 3.4], CADx Systems, Beavercreek, OH).
A and **B,** Printouts of craniocaudal (**A**) and mediolateral oblique (**B**) views highlight areas with suspicious findings. Ellipses mark potential masses (nonspiculated masses, spiculated masses, architectural distortions, and asymmetric densities), and rectangles mark potential microcalcifications.

CONCLUSION. Use of the computer-aided detection system significantly improved the detection of breast cancer by increasing radiologist sensitivity by 21.2%. Therefore, for every 100,000 women with breast cancer identified without the use of computer-aided detection, an estimated additional 21,200 cancers would be found with the use of computer-aided detection.

Brem, Rachel F., Janet Baum, Mary Lechner, Stuart Kaplan, Stuart Souders, L. Gill Naul, and Jeff Hoffmeister. “Improvement in sensitivity of screening mammography with computer-aided detection: a multiinstitutional trial.” *American Journal of Roentgenology* 181, no. 3 (2003): 687-693.

Mammogram reading

But an *NEJM* follow-up study in 2007 raised significant concerns:



Influence of Computer-Aided Detection on Performance of Screening Mammography

Joshua J. Fenton, M.D., M.P.H., Stephen H. Taplin, M.D., M.P.H., Patricia A. Carney, Ph.D., Linn Abraham, M.S., Edward A. Sickles, M.D., Carl D'Orsi, M.D., Eric A. Berns, Ph.D., Gary Cutter, Ph.D., R. Edward Hendrick, Ph.D., William E. Barlow, Ph.D., and Joann G. Elmore, M.D., M.P.H.

ABSTRACT

BACKGROUND

Computer-aided detection identifies suspicious findings on mammograms to assist radiologists. Since the Food and Drug Administration approved the technology in 1998, it has been disseminated into practice, but its effect on the accuracy of interpretation is unclear.

METHODS

We determined the association between the use of computer-aided detection at mammography facilities and the performance of screening mammography from 1998 through 2002 at 43 facilities in three states. We had complete data for 222,135 women (a total of 429,345 mammograms), including 2351 women who received a diagnosis of breast cancer within 1 year after screening. We calculated the specificity, sensitivity, and positive predictive value of screening mammography with and without computer-aided detection, as well as the rates of biopsy and breast-cancer detection and the overall accuracy, measured as the area under the receiver-operating-characteristic (ROC) curve.

RESULTS

Seven facilities (16%) implemented computer-aided detection during the study period. Diagnostic specificity decreased from 90.2% before implementation to 87.2% after implementation (P<0.001), the positive predictive value decreased from 4.1% to 3.2% (P=0.01), and the rate of biopsy increased by 19.7% (P<0.001). The increase in sensitivity from 80.4% before implementation of computer-aided detection to 84.0% after implementation was not significant (P=0.32). The change in the cancer-detection rate (including invasive breast cancers and ductal carcinomas in situ) was not significant (4.15 cases per 1000 screening mammograms before implementation and 4.20 cases after implementation, P=0.90). Analyses of data from all 43 facilities showed that the use of computer-aided detection was associated with significantly lower overall accuracy than was nonuse (area under the ROC curve, 0.871 vs. 0.919; P=0.005).

CONCLUSIONS

The use of computer-aided detection is associated with reduced accuracy of interpretation of screening mammograms. The increased rate of biopsy with the use of computer-aided detection is not clearly associated with improved detection of invasive breast cancer.

From the University of California, Davis, Sacramento (J.J.F.); the National Cancer Institute, Bethesda, MD (S.H.T.); Oregon Health and Science University, Portland (P.A.C.); Group Health Cooperative, Seattle (L.A.); the University of California, San Francisco, San Francisco (E.A.S.); the Emory Clinic, Atlanta (C.D.); Northwestern University, Chicago (E.A.B., R.E.H.); the University of Alabama at Birmingham, Birmingham (G.C.); Cancer Research and Biostatistics, Seattle (W.E.B.); and the University of Washington, Seattle (J.G.E.). Address reprint requests to Dr. Fenton at the Department of Family and Community Medicine, UC Davis Health System, 4860 Y St., Ste. 2300, Sacramento, CA 95817, or at joshua.fenton@ucdmc.ucdavis.edu.

N Engl J Med 2007;356:1399-409.

RESULTS

Seven facilities (16%) implemented computer-aided detection during the study period. Diagnostic specificity decreased from 90.2% before implementation to 87.2% after implementation (P<0.001), the positive predictive value decreased from 4.1% to 3.2% (P=0.01), and the rate of biopsy increased by 19.7% (P<0.001). The increase in sensitivity from 80.4% before implementation of computer-aided detection to 84.0% after implementation was not significant (P=0.32). The change in the cancer-detection rate (including invasive breast cancers and ductal carcinomas in situ) was not significant (4.15 cases per 1000 screening mammograms before implementation and 4.20 cases after implementation, P=0.90). Analyses of data from all 43 facilities showed that the use of computer-aided detection was associated with significantly lower overall accuracy than was nonuse (area under the ROC curve, 0.871 vs. 0.919; P=0.005).

Fenton, Joshua J., Stephen H. Taplin, Patricia A. Carney, Linn Abraham, Edward A. Sickles, Carl D'Orsi, Eric A. Berns et al. “Influence of computer-aided detection on performance of screening mammography.” *New England Journal of Medicine* 356, no. 14 (2007): 1399-1409.

What went wrong?

It's likely that many factors contributed, but two stand out:

1. AUTOMATION BIAS
   - Humans over-rely on the AI over time
   - Less vigilant in practice (when under time pressure) than in controlled research studies
   - Performance metrics are only part of the story—the real effect of AI is only revealed when it is integrated into clinical workflows
2. DATA DRIFT
   - Sometimes called *model drift* or *model aging*—although the model doesn't change at all
   - Happens when patient populations, clinical practice, or devices evolve
   - The model *can't respond to these changes* unless it is retrained

The Epic Sepsis Model

Case study 2

Sepsis detection

- Early detection and treatment associated with significant mortality benefit
- Many models developed, but one was by far the most adopted: the Epic Sepsis Model.
- *Trained on*:
  - 405,000 patient encounters
  - 3 health systems
- Proprietary, no independent validations carried out before it was rolled out.

Positive signs?

Epic's director of nursing discusses new research on its sepsis early warning model

September 14, 2021

A new independent study in the Journal of Critical Care Medicine found that Epic's sepsis early warning system led to faster antibiotic administration and better patient outcomes without an increase in harmful clinical interventions, like antibiotic or IV fluid overdose.

The model, used by hospitals nationwide, detects the first risk factors of infection in patients, allowing clinicians to enact early treatment measures and save lives.

https://www.epic.com/epic/post/epics-director-of-nursing-discusses-new-research-on-its-sepsis-early-warning-model
Accessed 21 November 2025

Improving Timeliness of Antibiotic Administration Using a Provider and Pharmacist Facing Sepsis Early Warning System in the Emergency Department Setting: A Randomized Controlled Quality Improvement Initiative*

Yasir Tarabichi, MD, MSCR[1–3]
Aurelia Cheng, MD[3,4]
David Bar-Shain, MD[2,3]
Brian M. McCrate, PharmD, BCPS, BCCCP[5]
Lewis H. Reese, PharmD, BCPS[5]
Charles Emerman, MD[3,4]
Jonathan Siff, MD, MBA[2–4]
Christine Wang, BS[3]
David C. Kaelber, MD, PhD, MPH[3,6,7]
Brook Watts, MD, MS[3,8]
Michelle T. Hecker, MD[3,9]

OBJECTIVES: Results of pre-post intervention studies of sepsis early warning systems have been mixed, and randomized clinical trials showing efficacy in the emergency department setting are lacking. Additionally, early warning systems can be resource-intensive and may cause unintended consequences such as antibiotic or IV fluid overuse. We assessed the impact of a pharmacist and provider facing sepsis early warning systems on timeliness of antibiotic administration and sepsis-related clinical outcomes in our setting.

DESIGN: A randomized, controlled quality improvement initiative.

SETTING: The main emergency department of an academic, safety-net healthcare system from August to December 2019.

PATIENTS: Adults presenting to the emergency department.

INTERVENTION: Patients were randomized to standard sepsis care or standard care augmented by the display of a sepsis early warning system–triggered flag in the electronic health record combined with electronic health record–based emergency department pharmacist notification.

MEASUREMENTS AND MAIN RESULTS: The primary process measure was time to antibiotic administration from arrival. A total of 598 patients were included in the study over a 5-month period (285 in the intervention group and 313 in the standard care group). Time to antibiotic administration from emergency department arrival was shorter in the augmented care group than that in the standard care group (median, 2.3 hr [interquartile range, 1.4–4.7 hr] vs 3.0 hr [interquartile range, 1.6–5.5 hr]; $p = 0.039$). The hierarchical composite clinical outcome measure of days alive and out of hospital at 28 days was greater in the augmented care group than that in the standard care group (median, 24.1 vs 22.5 d; $p = 0.011$). Rates of fluid resuscitation and antibiotic utilization did not differ.

CONCLUSIONS: In this single-center randomized quality improvement initiative, the display of an electronic health record–based sepsis early warning system–triggered flag combined with electronic health record–based pharmacist notification was associated with shorter time to antibiotic administration without an increase in undesirable or potentially harmful clinical interventions.

Tarabichi, Yasir, Aurelia Cheng, David Bar-Shain, Brian M. McCrate, Lewis H. Reese, Charles Emerman, Jonathan Siff et al. "Improving timeliness of antibiotic administration using a provider and pharmacist facing sepsis early warning system in the emergency department setting: a randomized controlled quality improvement initiative." *Critical care medicine* 50, no. 3 (2022): 418-427.

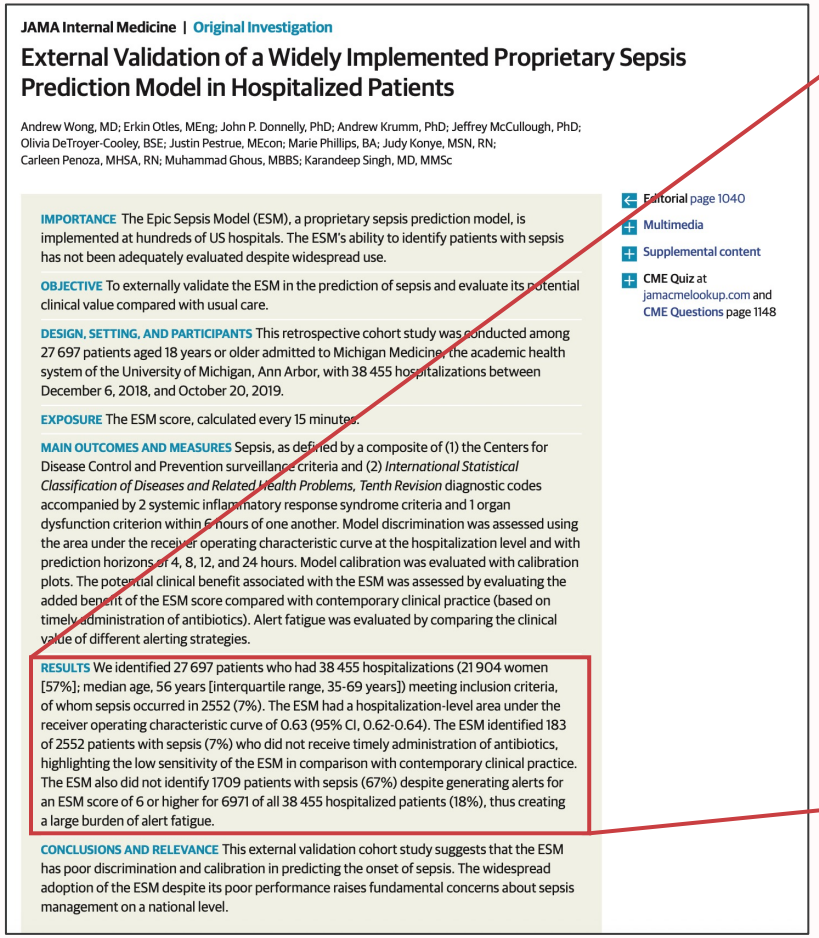
Initial positive signs

MEASUREMENTS AND MAIN RESULTS: The primary process measure was time to antibiotic administration from arrival. A total of 598 patients were included in the study over a 5-month period (285 in the intervention group and 313 in the standard care group). Time to antibiotic administration from emergency department arrival was shorter in the augmented care group than that in the standard care group (median, 2.3 hr [interquartile range, 1.4–4.7 hr] vs 3.0 hr [interquartile range, 1.6–5.5 hr]; $p = 0.039$). The hierarchical composite clinical outcome measure of days alive and out of hospital at 28 days was greater in the augmented care group than that in the standard care group (median, 24.1 vs 22.5 d; $p = 0.011$). Rates of fluid resuscitation and antibiotic utilization did not differ.

CONCLUSIONS: In this single-center randomized quality improvement initiative, the display of an electronic health record–based sepsis early warning system–triggered flag combined with electronic health record–based pharmacist notification was associated with shorter time to antibiotic administration without an increase in undesirable or potentially harmful clinical interventions.

Tarabichi, Yasir, Aurelia Cheng, David Bar-Shain, Brian M. McCrate, Lewis H. Reese, Charles Emerman, Jonathan Siff et al. "Improving timeliness of antibiotic administration using a provider and pharmacist facing sepsis early warning system in the emergency department setting: a randomized controlled quality improvement initiative." *Critical care medicine* 50, no. 3 (2022): 418-427.

Further validation reveals problems



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; Jeffrey McCullough, PhD; Olivia DeTroyer-Cooley, BSE; Justin Pestrue, MEcon; Marie Phillips, BA; Judy Konye, MSN, RN; Carleen Penoza, MHSA, RN; Muhammad Ghous, MBBS; Karandeep Singh, MD, MMSc

Editorial page 1040
Multimedia
Supplemental content
CME Quiz at jamacmelookup.com and CME Questions page 1148

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

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

DESIGN, SETTING, AND PARTICIPANTS This retrospective cohort study was conducted among 27 697 patients aged 18 years or older admitted to Michigan Medicine, the academic health system of the University of Michigan, Ann Arbor, with 38 455 hospitalizations between 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 Centers for Disease Control and Prevention surveillance criteria and (2) *International Statistical Classification of Diseases and Related Health Problems, Tenth Revision* diagnostic codes accompanied by 2 systemic inflammatory response syndrome criteria and 1 organ dysfunction criterion within 6 hours of one another. Model discrimination was assessed using the area under the receiver operating characteristic curve at the hospitalization level and with prediction horizons of 4, 8, 12, and 24 hours. Model calibration was evaluated with calibration plots. The potential clinical benefit associated with the ESM was assessed by evaluating the added benefit of the ESM score compared with contemporary clinical practice (based on timely administration of antibiotics). Alert fatigue was evaluated by comparing the clinical value of different alerting strategies.

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.

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, Andrew, Erkin Otles, John P. Donnelly, Andrew Krumm, Jeffrey McCullough, Olivia DeTroyer-Cooley, Justin Pestrue et al. "External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients." *JAMA internal medicine* 181, no. 8 (2021): 1065-1070.

What went wrong?

- Rolled out with external studies that verified the *main endpoint*: how accurate is the model when used to predict sepsis for an individual patient?
- Wong et. al. show that the model AUC is higher *per alert*.
  - Epic reports 0.76–0.83
  - Wong et. al. find 0.72–0.76 *per alert*.
  - *But this is not what matters in practice!*
- In practice, the *first warning* for a patient matters. Warnings beyond cause alert fatigue
- Per-alert accuracy can be high even for a false-positive case
- Failure to analyze the *workflow* within which the AI model would be deployed

Colonoscopy deskilling

Case study 3

AI use can lead to deskilling

Findings

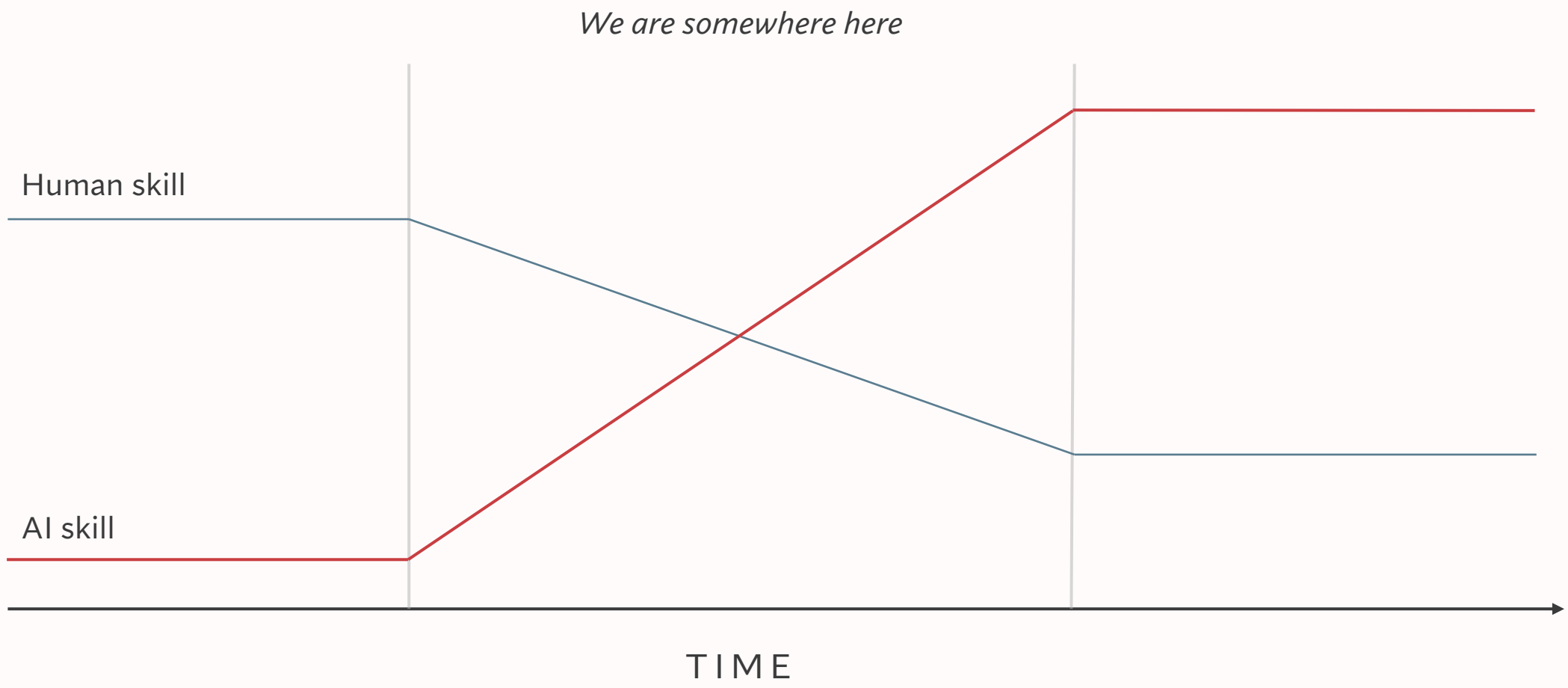
Between Sept 8, 2021, and March 9, 2022, 1443 patients underwent non-AI assisted colonoscopy before (n=795) and after (n=648) the introduction of AI (median age 61 years [IQR 45–70], 847 [58·7%] female, 596 [41·3%] male). The ADR of standard colonoscopy decreased significantly from 28·4% (226 of 795) before to 22·4% (145 of 648) after exposure to AI, corresponding with an absolute difference of –6·0% (95% CI –10·5 to –1·6; p=0·0089). In multivariable logistic regression analysis, exposure to AI (odds ratio 0·69 [95% CI 0·53–0·89]), male versus female patient sex (1·78 [1·38–2·30]), and patient age ≥60 years versus <60 years (3·60 [2·74–4·72]) were the independent factors significantly associated with ADR.

Interpretation

Continuous exposure to AI might reduce the ADR of standard non-AI assisted colonoscopy, suggesting a negative effect on endoscopist behaviour.

Budzyń, Krzysztof, Marcin Romańczyk, Diana Kitala, Paweł Kołodziej, Marek Bugajski, Hans O. Adami, Johannes Blom et al. "Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: a multicentre, observational study." *The Lancet Gastroenterology & Hepatology* 10, no. 10 (2025): 896-903.

The unhappy middle



Today's talk

1. Case studies
   - *Computer aided diagnosis of mammography*
   - *The Epic sepsis model*
   - *Colonoscopy deskilling*
2. Lessons for safety in other fields
   - *Traditional software*
   - *Humans*
   - *Drugs*
3. Mitigation strategies
   - *Tackling drift through continuous monitoring*
   - *Tackling automation bias through cognitive forcing*

AI vs Traditional Software

The black box of traditional software

- Modern medicine is *impossible* without software
  - Electronic health records
  - Treatment planning software
  - Image analysis
  - Clinical decision support systems
- These systems are (largely) black boxes to their users:
  - The modern software stack is *extremely* complex, and built on thousands of lines of computer code
  - Dedicated software engineers understand all the nuances, but not typical users.

How do we keep software safe?

- The use of computers in healthcare only really began in the 1960s.
- Now, software regularly processes *life and death decisions.*
- Since then, we've learned how to deploy this black box safely and reliably.
  - *Unit testing*
    - Engineers can define expected behavior for every case, and run automated and regular tests to verify it
  - *Continuous monitoring*
    - Logs, alerts and scripts detect failures in real time
  - *Human-in-the-loop*
    - Clinicians, physicists, double check confirm every high-impact step
    - We design safe *workflows.* We don't assume any one piece of software is 100% infallible

Lessons for AI safety

- Does this translate to AI?
  - *Unit testing* ✗
    - Engineers cannot define expected behavior for every case
    - If we knew exactly what the output should be for every input, we wouldn't need AI!
    - AI fails silently—poor predictions rather than computer crashes
  - *Continuous monitoring* ✓
    - AI use and metrics should be monitored automatically—more on this later
  - *Human-in-the-loop* ✓
    - Humans must confirm every high-impact step
    - BUT: checking AI output is often harder than for traditional software
      - AI can make subtle predictions and automate much more difficult tasks
      - Checking those predictions is a much greater cognitive workload
      - Risk of AUTOMATION BIAS—more on this later

AI vs Humans

The black box of human beings

- We don't traditionally think this way, but it's useful to view *human beings* as black boxes!
- *Humans are opaque systems*: we can't see the *exact* inner workings of another human's perception/bias/intuition
- Humans *can* report explanations about themselves, but:
  - Explanations are not always 100% accurate
  - Cannot explain themselves in exhaustive detail like traditional software

How do humans deliver safe healthcare?

- Humans are not *inherently* safe
- But medicine has spent centuries building layers of training, redundancy, and feedback that make human care predictable and reliable.
  - *Training and certification*
    - Standardized education
    - A continuous process—like today's talk!
  - *Team redundancy*
    - Multiple clinicians review high-risk steps
  - *Learning from failure*
    - Near-miss analysis
    - Institutional memory

Lessons for AI safety

- Does this translate to AI?
- In many ways, AI systems are *more similar to humans* than they are to traditional software:
  - Their output is *stochastic*: the *same input* doesn't always generate the *same output.*
  - It's impossible to describe with *infinite* precision *why and AI system made a specific decision.*
- POSSIBLE MENTAL MODEL:
  - Treat AI like a very knowledgeable trainee in the workflow.
  - How would you verify and guarantee the quality of their output?

Lessons for AI safety

- Do human safety principles translate to AI?
  - *Training and certification* ✓
    - AI systems are tested in specific environments
      - E.g., "detects cancer with an AUC of 0.85"
      - Can think of this as "certification"—whether we trust an AI's use depends on how well it's been certified.
    - A continuous process
      - Unlike humans, the code behind AI doesn't change. *But*:
        - Data shift can cause performance to degrade over time. Humans naturally adapt, but AI doesn't.
        - Generative AI depends on services from companies (OpenAI/Google/etc.) that get updated.
  - *Team redundancy* ✓
    - Clinicians review high-risk steps—human-in-the-loop
  - *Learning from failure* ✓
    - Will analyze some AI failure-cases later

AI vs Drugs

The black box of clinical drugs

- The mechanism of clinical drugs is often opaque:
  - Complex biochemical cascades
  - Receptor binding
  - Off-target effects
  - Metabolism
- They work and proven safe *long* before we fully understand *how* they work.

How do we keep drugs safe?

- *Clinical trials pipeline*
  - Randomized control trials
  - Standard pipeline: preclinical → Phase I–III → post-market surveillance
- *Regulatory oversight*
  - FDA approval
- *Human safeguards*
  - Prescribing limits
  - Double checks

Lessons for AI safety

- Do clinical drug safety principles translate to AI?
  - *Clinical trials pipeline* ?
    - AI model validation is akin to trials
    - However: often, we develop AI models in-house, and trial work is not outsourced to a large pharmaceutical company.
      - More of the burden of validation falls to the institution
  - *Regulatory oversight* ?
    - Some AI systems are considered *clinical decision support* and don't require extensive validation
      - More of the burden of validation falls to the clinician-user
  - *Human safeguards* ✓
    - Human-in-the-loop, as always
    - Automation bias, again

Today's talk

1. Case studies
   - *Computer aided diagnosis of mammography*
   - *The Epic sepsis model*
   - *Colonoscopy deskilling*
2. Lessons for safety in other fields
   - *Traditional software*
   - *Humans*
   - *Drugs*
3. Mitigation strategies
   - *Tackling drift through continuous monitoring*
   - *Tackling automation bias through cognitive forcing*

Continuous Monitoring

Tackling drift

The data drift problem

Performance deterioration of deep learning models after clinical deployment: a case study with auto-segmentation for definitive prostate cancer radiotherapy

Biling Wang[1,3,5], **Michael Dohopolski**[1,2,5], **Ti Bai**[1,2], **Junjie Wu**[1,2], **Raquibul Hannan**[1,2], **Neil Desai**[1,2], **Aurelie Garant**[1,2], **Daniel Yang**[1,2], **Dan Nguyen**[1,2], **Mu-Han Lin**[1,2], **Robert Timmerman**[1,2], **Xinlei Wang**[3,4,*] and **Steve B Jiang**[1,2,*]

[1] Medical Artificial Intelligence and Automation Laboratory, University of Texas Southwestern Medical Center, Dallas, TX, United States of America
[2] Department of Radiation Oncology, University of Texas Southwestern Medical Center, Dallas, TX, United States of America
[3] Department of Statistics and Data Science, Southern Methodist University, Dallas, TX, United States of America
[4] Department of Mathematics, University of Texas at Arlington, Dallas, TX, United States of America
[5] Co-first authors.
[*] Authors to whom any correspondence should be addressed.

E-mail: xinlei.wang@uta.edu and steve.jiang@utsouthwestern.edu

Keywords: deep learning, segmentation, model performance deterioration, radiotherapy

Supplementary material for this article is available online

Abstract

Our study aims to explore the long-term performance patterns for deep learning (DL) models deployed in clinic and to investigate their efficacy in relation to evolving clinical practices. We conducted a retrospective study simulating the clinical implementation of our DL model involving 1328 prostate cancer patients treated between January 2006 and August 2022. We trained and validated a U-Net-based auto-segmentation model on data obtained from 2006 to 2011 and tested on data from 2012 to 2022, simulating the model's clinical deployment starting in 2012. We visualized the trends of the model performance using exponentially weighted moving average (EMA) curves. Additionally, we performed Wilcoxon Rank Sum Test and multiple linear regression to investigate Dice similarity coefficient (DSC) variations across distinct periods and the impact of clinical factors, respectively. Initially, from 2012 to 2014, the model showed high performance in segmenting the prostate, rectum, and bladder. Post-2015, a notable decline in EMA DSC was observed for the prostate and rectum, while bladder contours remained stable. Key factors impacting the prostate contour quality included physician contouring styles, using various hydrogel spacers, CT scan slice thickness, MRI-guided contouring, and intravenous (IV) contrast ($p < 0.0001$, $p < 0.0001$, $p = 0.0085$, $p = 0.0012$, $p < 0.0001$, respectively). Rectum contour quality was notably influenced by factors such as slice thickness, physician contouring styles, and the use of various hydrogel spacers. The quality of the bladder contour was primarily affected by IV contrast. The deployed DL model exhibited a substantial decline in performance over time, aligning with the evolving clinical settings.

Figure 2. Trends in auto-generated contour quality. (a), (b), and (c) present the exponential weighted moving average (EMA) of dice similarity coefficient (EMA DSC) over time post-simulated model deployment for: (a) prostate EMA DSC, (b) rectum EMA DSC, and (c). bladder EMA DSC. EMA DSC for the auto-generated prostate and rectum contours declined, but those for the bladder contours remained stable.

Wang, Biling, Michael Dohopolski, Ti Bai, Junjie Wu, Raquibul Hannan, Neil Desai, Aurelie Garant et al. "Performance deterioration of deep learning models after clinical deployment: a case study with auto-segmentation for definitive prostate cancer radiotherapy." *Machine Learning: Science and Technology* 5, no. 2 (2024): 025077.

The data drift problem

- Key factors impacting contour quality:
  - Physician contouring styles
  - Use of hydrogel spacers
  - CT scan slice thickness
  - MRI-guided contouring
  - IV contrast
- Not all of these factors can be foreseen
- The question is not *if* to retrain the models in the future, but *when*:
  - Training too frequently is costly and time-consuming
  - Each new model must be rigorously validated
  - Collecting new training data can be costly too

Continuous monitoring

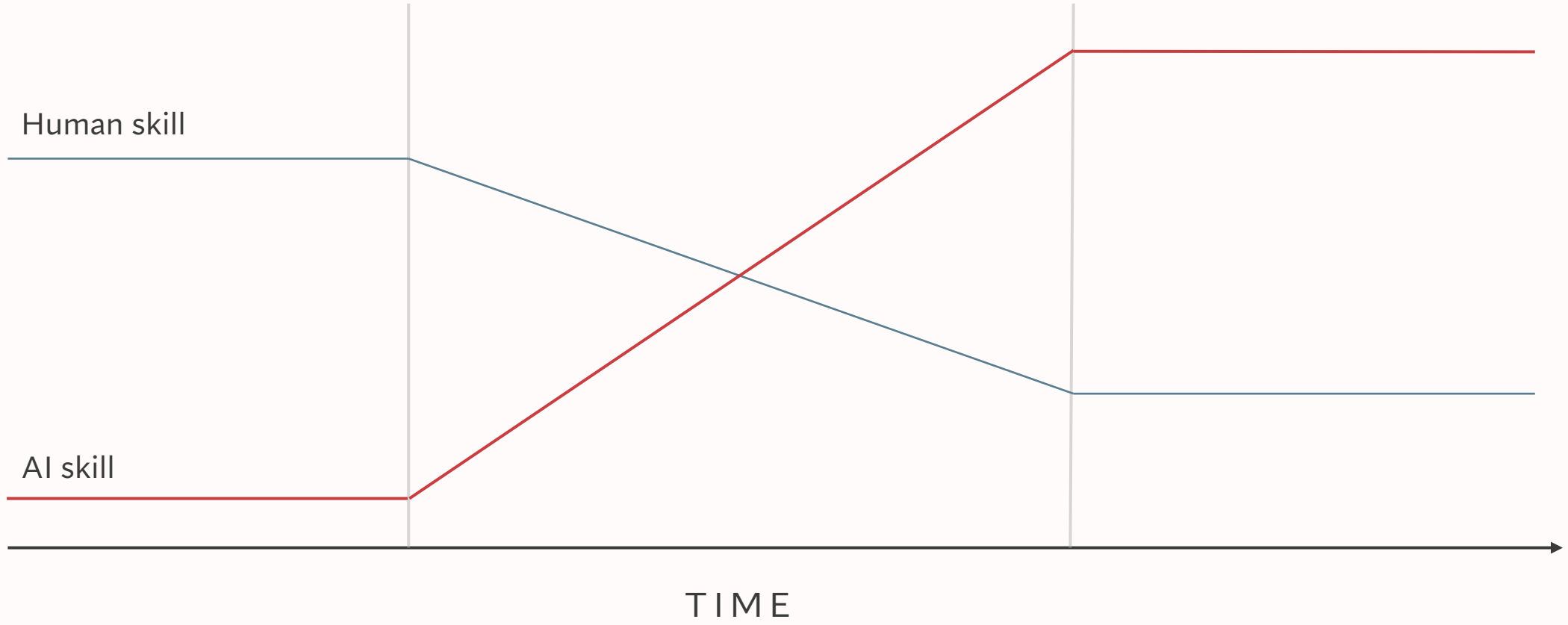
- To know *when* to retrain, we must know *how performance evolves over time.*
- *Example*:
  - For auto-contouring, monitor the Dice similarity coefficient between the AI-generated contour and the final approved contour (*work done by Dr. Satomi Shiraishi*)
  - Stratify by organ, gender, BMI etc.
  - Define acceptable minimum Dice score.
  - If average Dice scores are below acceptable level, trigger the retraining process.
- Note—defining what metrics to monitor does not only (or even mainly) require AI expertise. It requires understanding of:
  - Medical and scientific knowledge
  - Clinical workflows
  - Human interaction

Cognitive Forcing

Tackling automation bias

The unhappy middle

How can we live in the unhappy middle while ensuring we get the best *of humans + AI, not the* worst?



Cognitive forcing

- Need to ensure that human review of AI is genuine independent verification—combat automation bias
- Techniques to address this are known as *cognitive forcing*.
- *Examples*
  - Required confirmation step
  - Choosing between multiple options
  - Mandatory justification
  - Displaying uncertainty
- KEY CHALLENGE: how to encourage thought without annoying users.
  - Effective cognitive forcing strategies likely to be different for each AI application
  - Workflow understanding is critical

Recap

- AI safety is *not* about:
  - Perfectly accurate models
  - Perfectly interpretable/understandable models
- It's about *building robust workflows around imperfect models.*
- AI risks have analogies in other disciplines:
  - Complex like software
  - Unpredictable like humans
  - Opaque like drugs
- Core risks and mitigations:
  - Data-drift—continuous monitoring
  - Automation bias—cognitive forcing

Question & Answer