

AI in healthcare: hip hip hoera or AIAI?

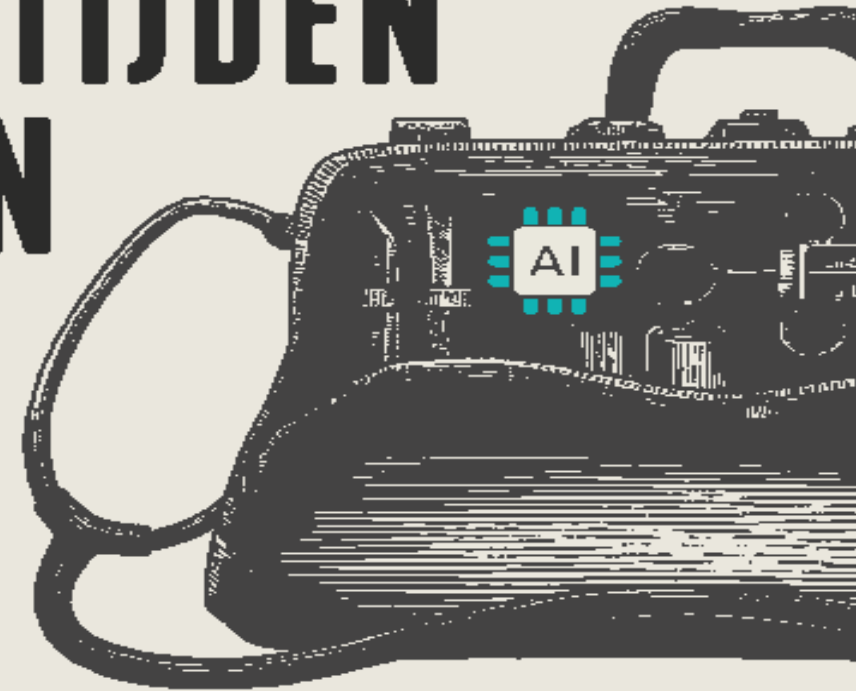
Prof Dr W. Van Biesen

Universitair Ziekenhuis Gent

Consortium for Justifiable Digital Healthcare

SIGRID STERCKX EN WIM VAN BIESEN (RED.)

GENEESKUNDE IN TIJDEN VAN AI



Handvatten voor vandaag en morgen

ACADEMIA
PRESS



<https://www.academiapress.be/nl/geneeskunde-tijden-van-ai>



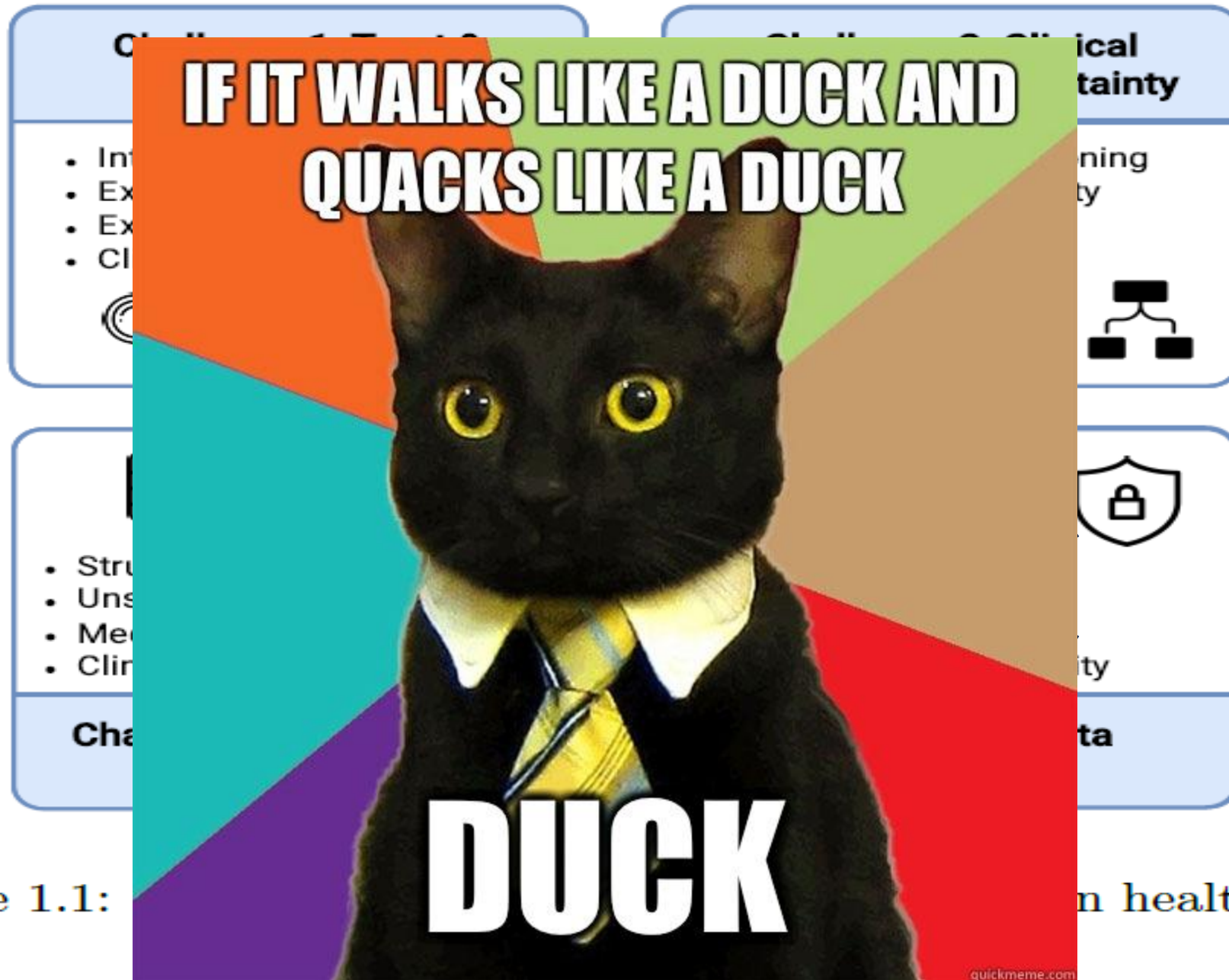


Figure 1.1:

in healthcare.

- **Jumping from (big) data to knowledge**
 - Where do data come from?
 - Are data “objective”?
 - Data-information-knowledge-counterfactual
- **Pitfalls**
- **Impact on healthcare professionals**

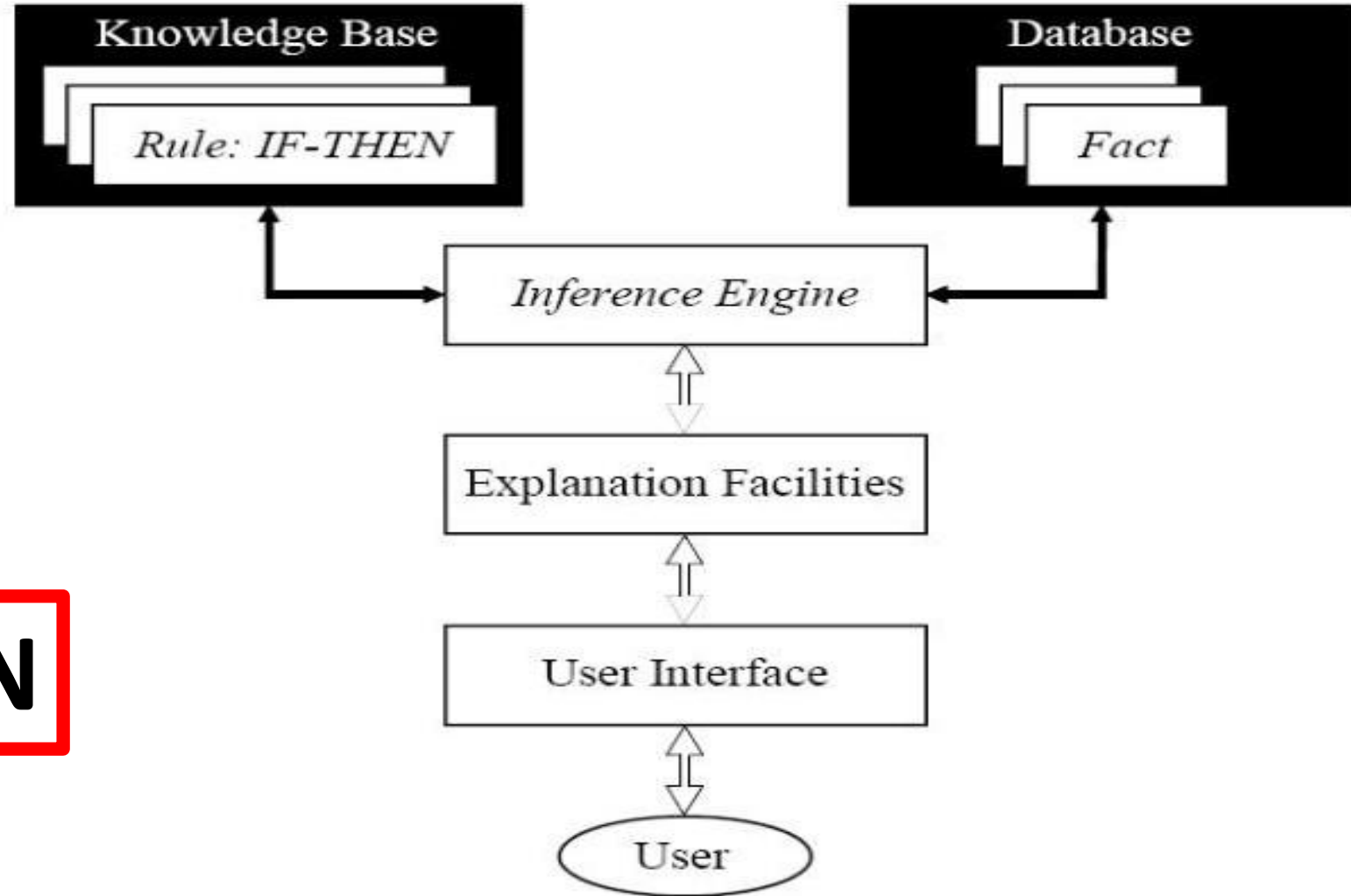
```
graph TD; A[Artificial Intelligence] --> B[Rule Based Algorithms]; A --> C[Machine Learning Algorithms]
```

Artificial
Intelligence

Rule Based
Algorithms

Machine
Learning
Algorithms

Basic structure of a rule-based expert system



DEDUCTION

Potential interactions with medication:

A (lower)

X (higher)

Y (arrhythmias)

Alternative: ProductU[®]



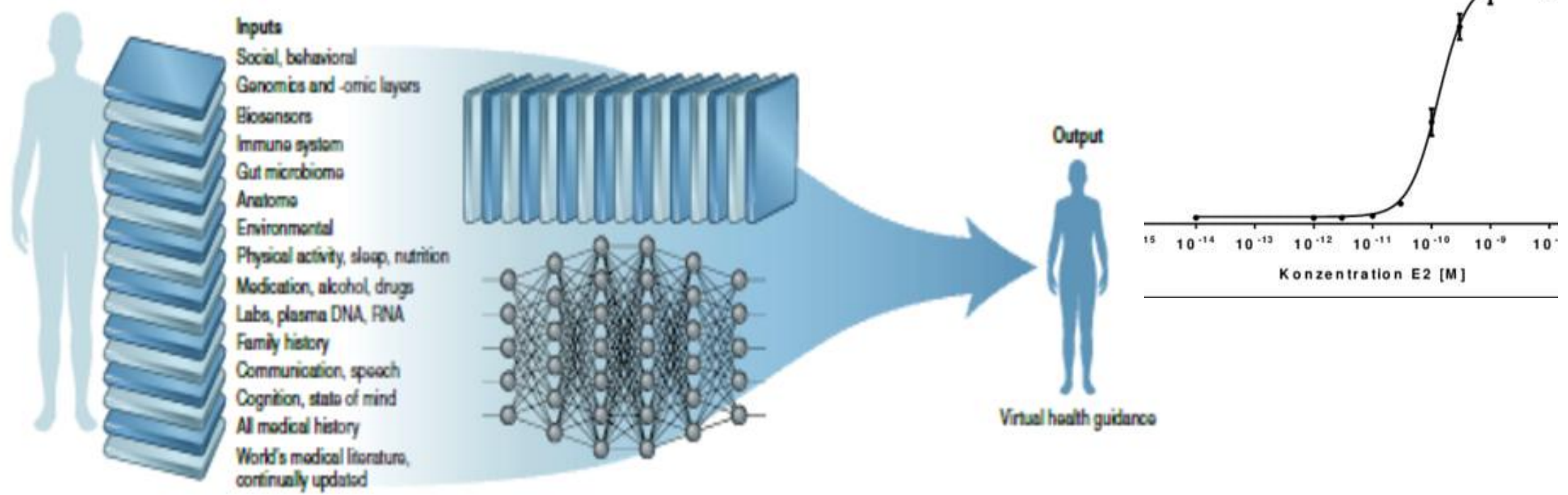
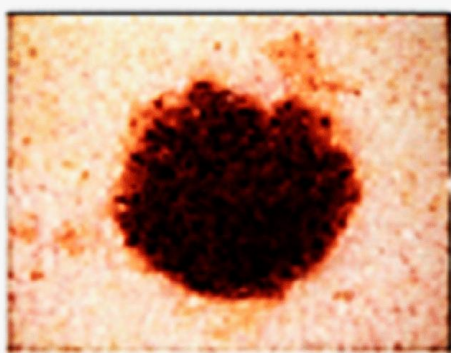


Figure 2: The virtual medical coach model with multi-modal data inputs and algorithms to provide individualised guidance.¹⁰⁵

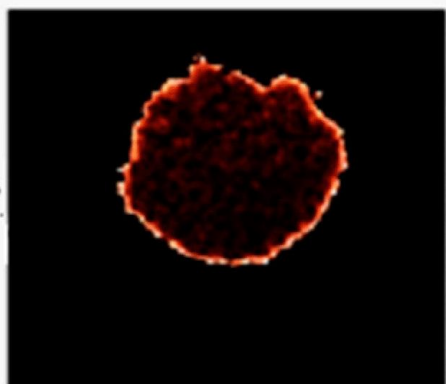
EMBEDDING

INDUCTION

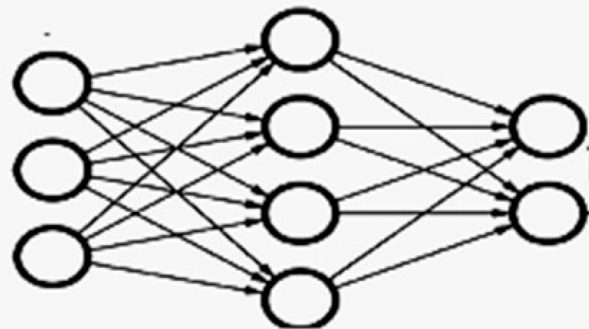
Image of Skin lesion



Segmented lesion

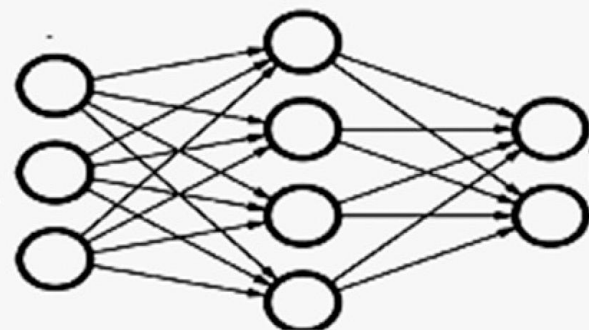
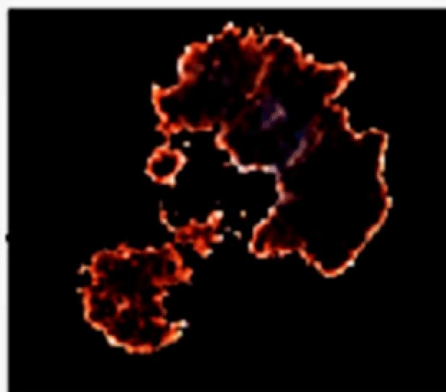
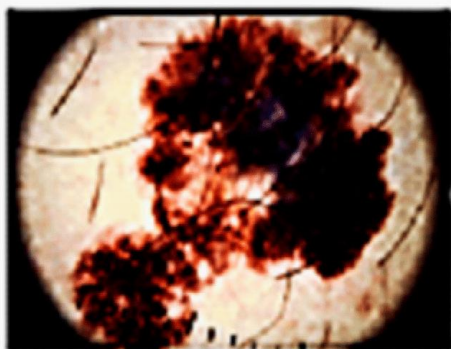


Learning model

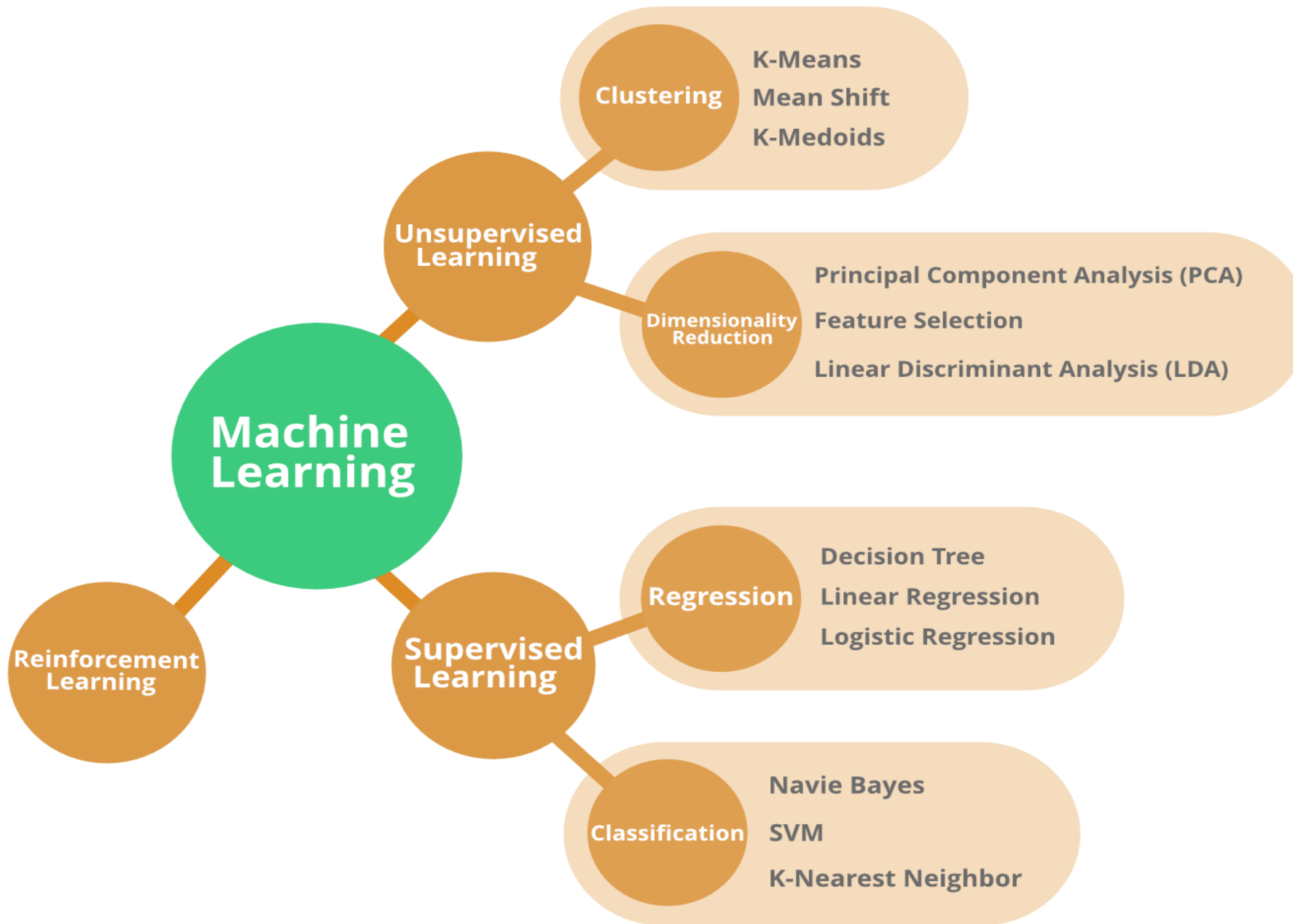


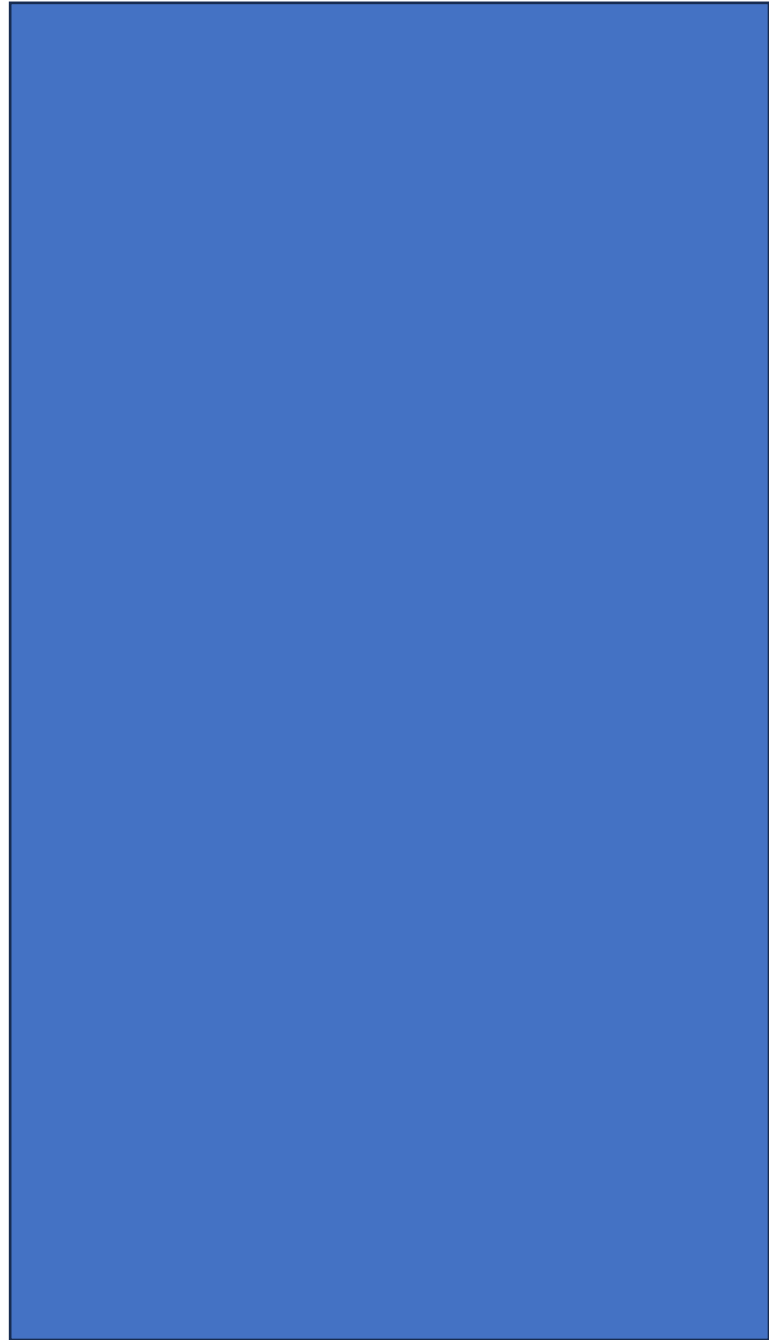
Prediction

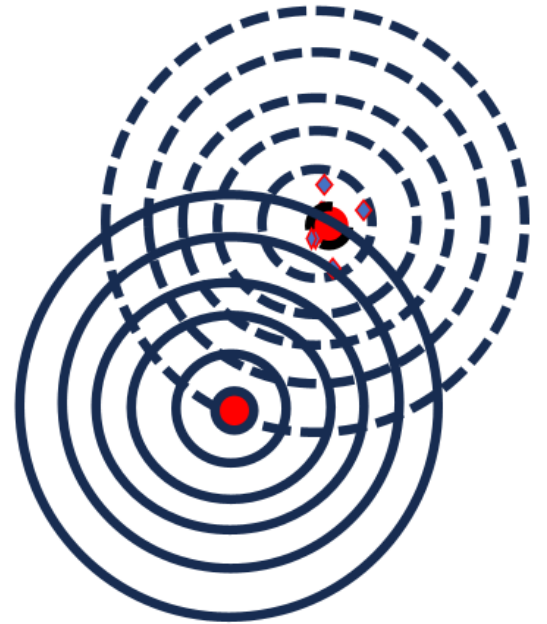
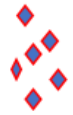
Benign



Malignant







Knowledge

Why does A happen if I do X

Information

What happens if I do A?

How does the event of seeing X change the probability of seeing Y?
Pattern recognition

Data

Descriptive: people with A have x in 75% of cases

Causation

Lack of a theoretical model



Adverserial effects

Generalizability

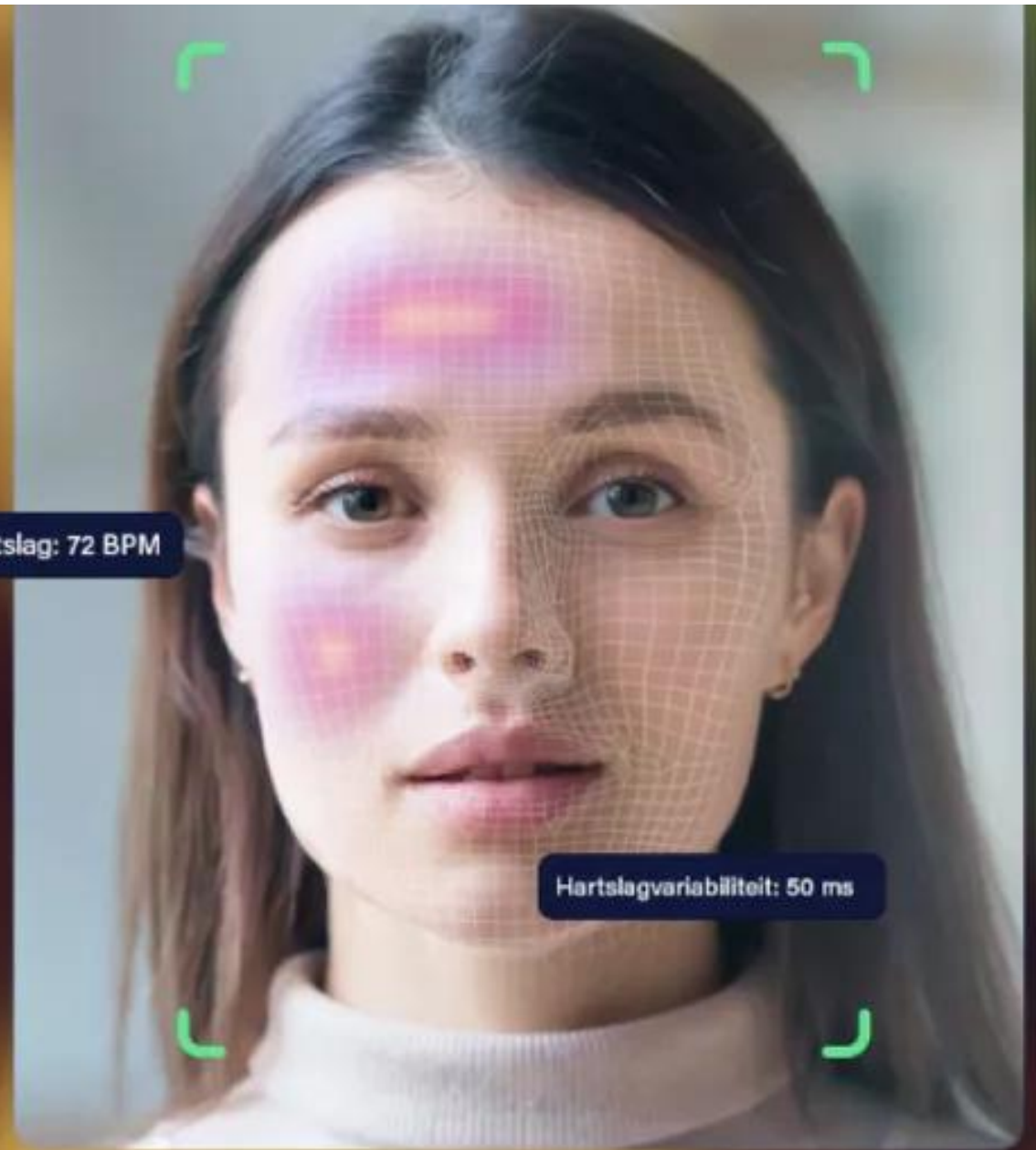
Transportability

Tank Problems

90% of people with the flu have fever and muscle pain



Hartslag: 72 BPM



Hartslagvariabilitet: 50 ms

Gödel: when rules are applied to inputs from a different structure then we can expect weirdness

STEPHEN WOLFRAM

What Is

ChatGPT

Doing...

...and Why Does It Work?

Data

Human-generated text



Natural language



Code

Other text

MVSTQWYC... Protein

AGGACACG... DNA/RNA

CCC(=O)O... Chemical

Training

Autoregressive
(next token prediction)

To be or not to ?

Masked
(fill in the blanks)

To ? or not ? be

Large language model



Output likelihoods



Model parameters



Input sequence of tokens

To be or not to

Zero-shot transfer and prompt engineering

Supervised finetuning and RLHF

Task-specific labeled data



Applications



Question answering



Literature/
health-record summaries

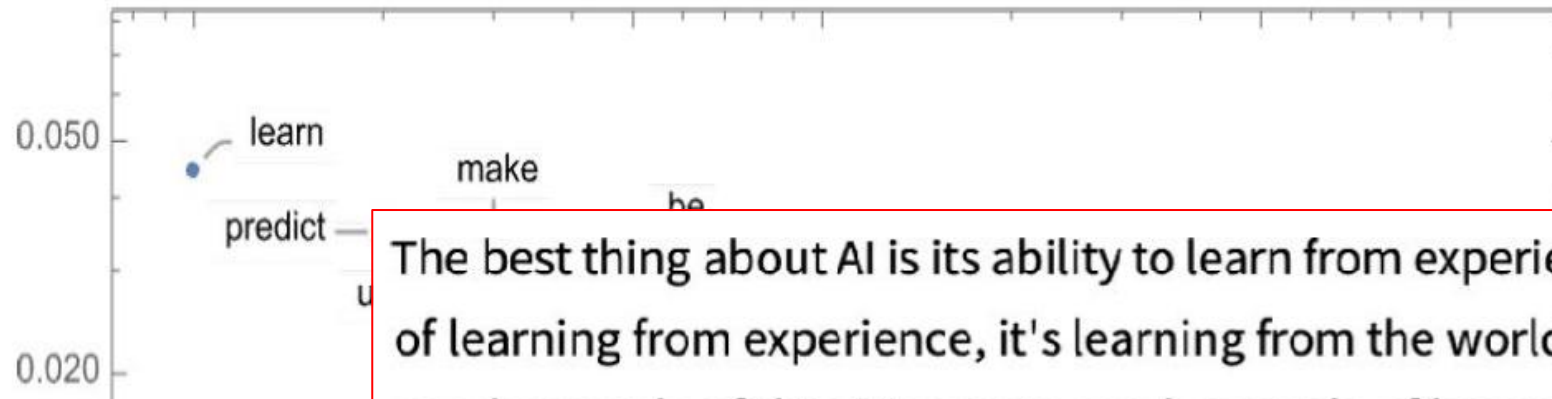


Scientific coding



Molecular design

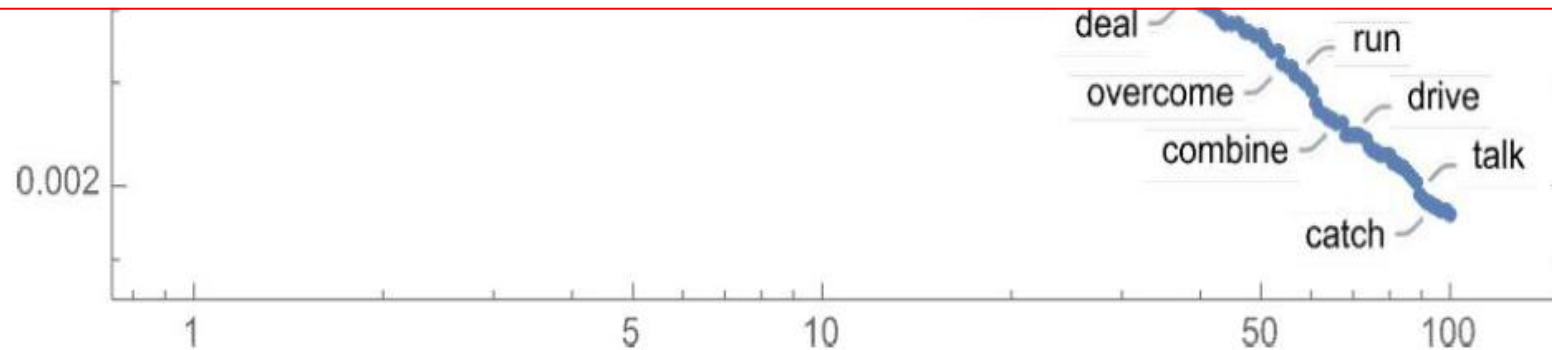
{ The best thing about AI is its ability to,



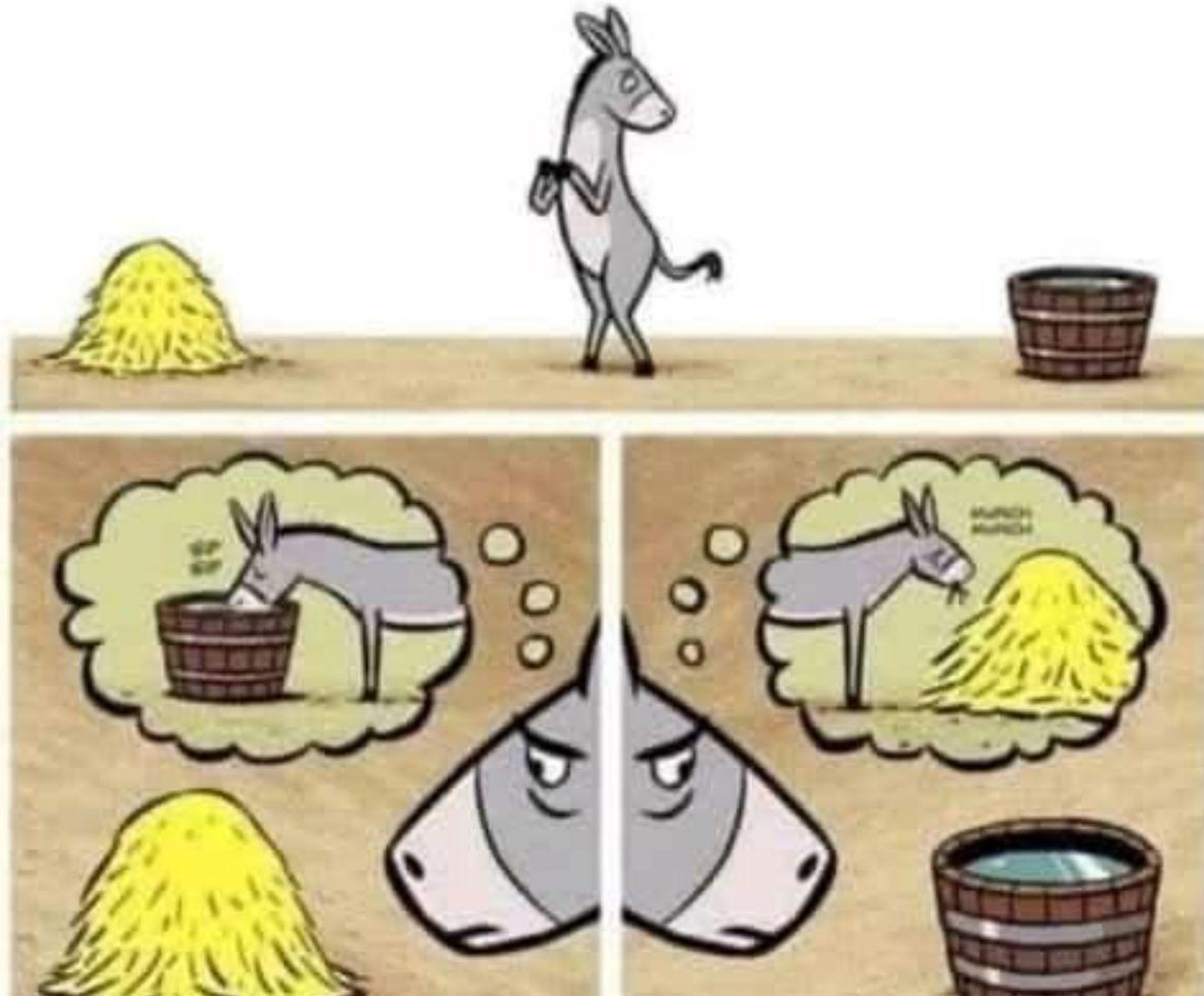
The best thing about AI is its ability to learn from experience. It's not just a matter of learning from experience, it's learning from the world around you. The AI is a very

The best thing about AI is its ability to learn and develop over time, allowing it to continually improve its performance and be more efficient at tasks. AI can also be used to automate mundane tasks, allowing humans to focus on more important tasks. AI can also be used to make decisions and provide insights that would otherwise be impossible for humans to figure out.

improve your life.
AI is a very good
example of how to use AI to



The NEED for RANDOMNESS



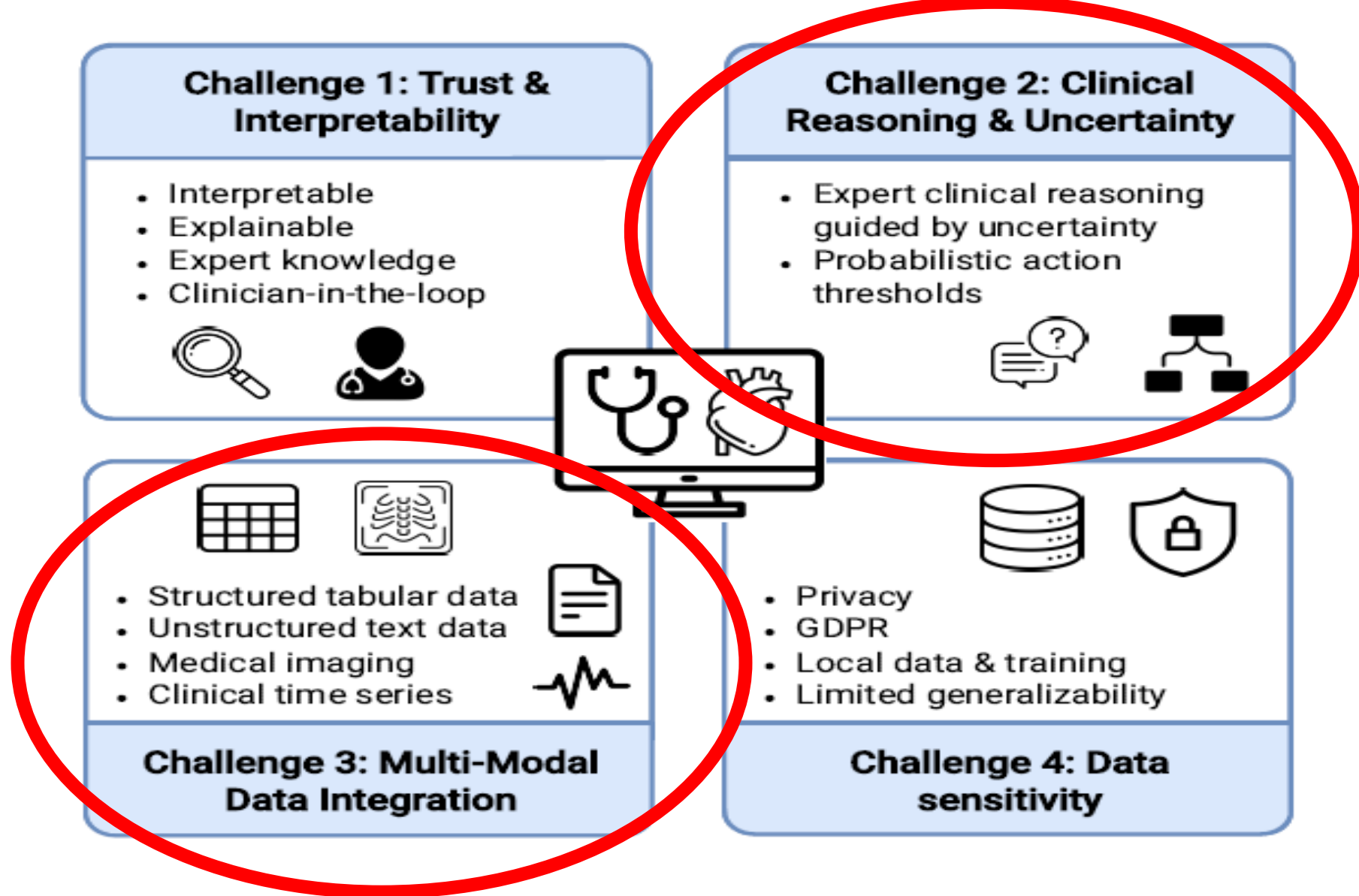


Figure 1.1: Four major technical challenges for AI in healthcare.

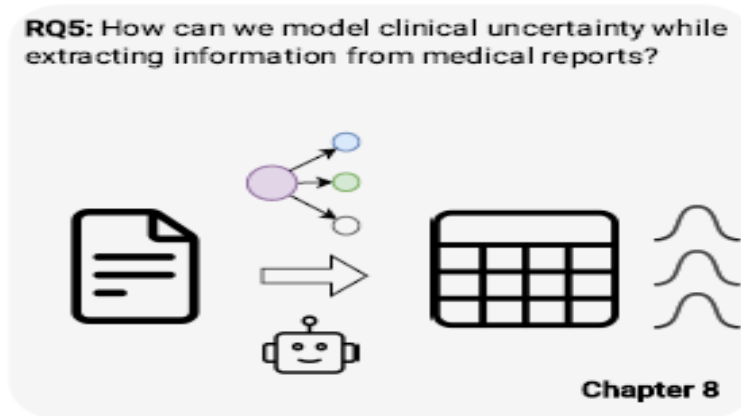
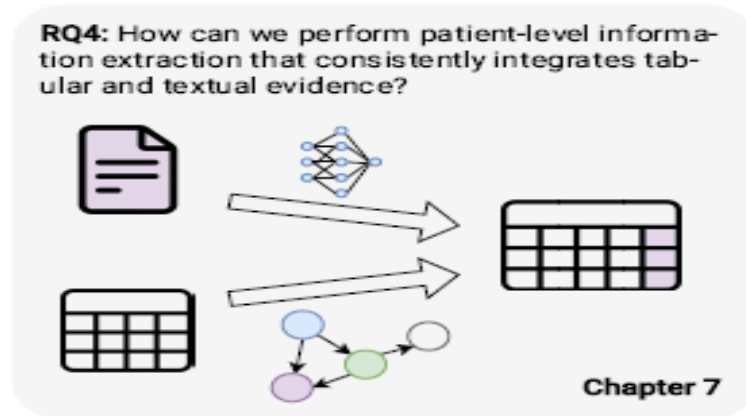
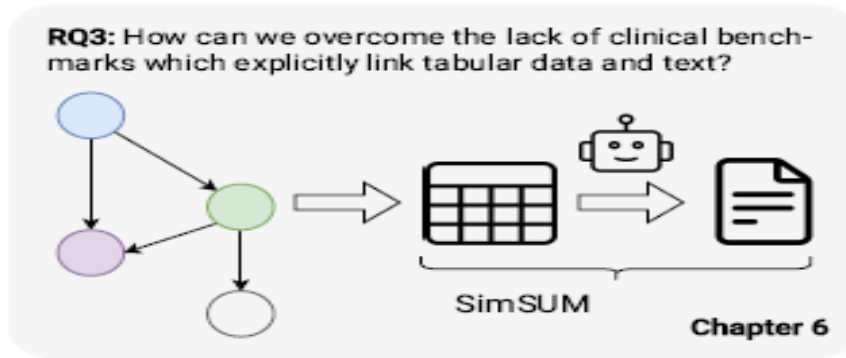
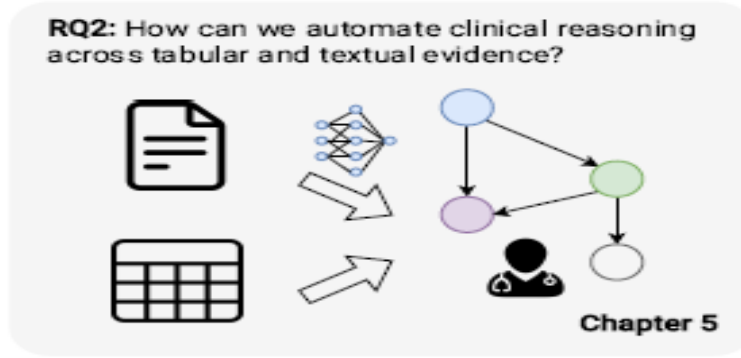
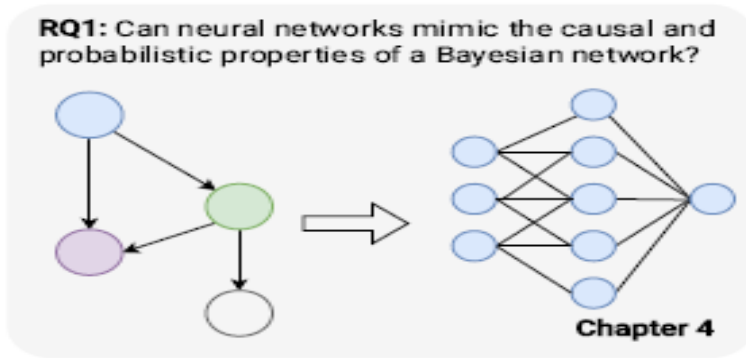


Figure 3.1: Overview of the research questions explored in each chapter of this dissertation.

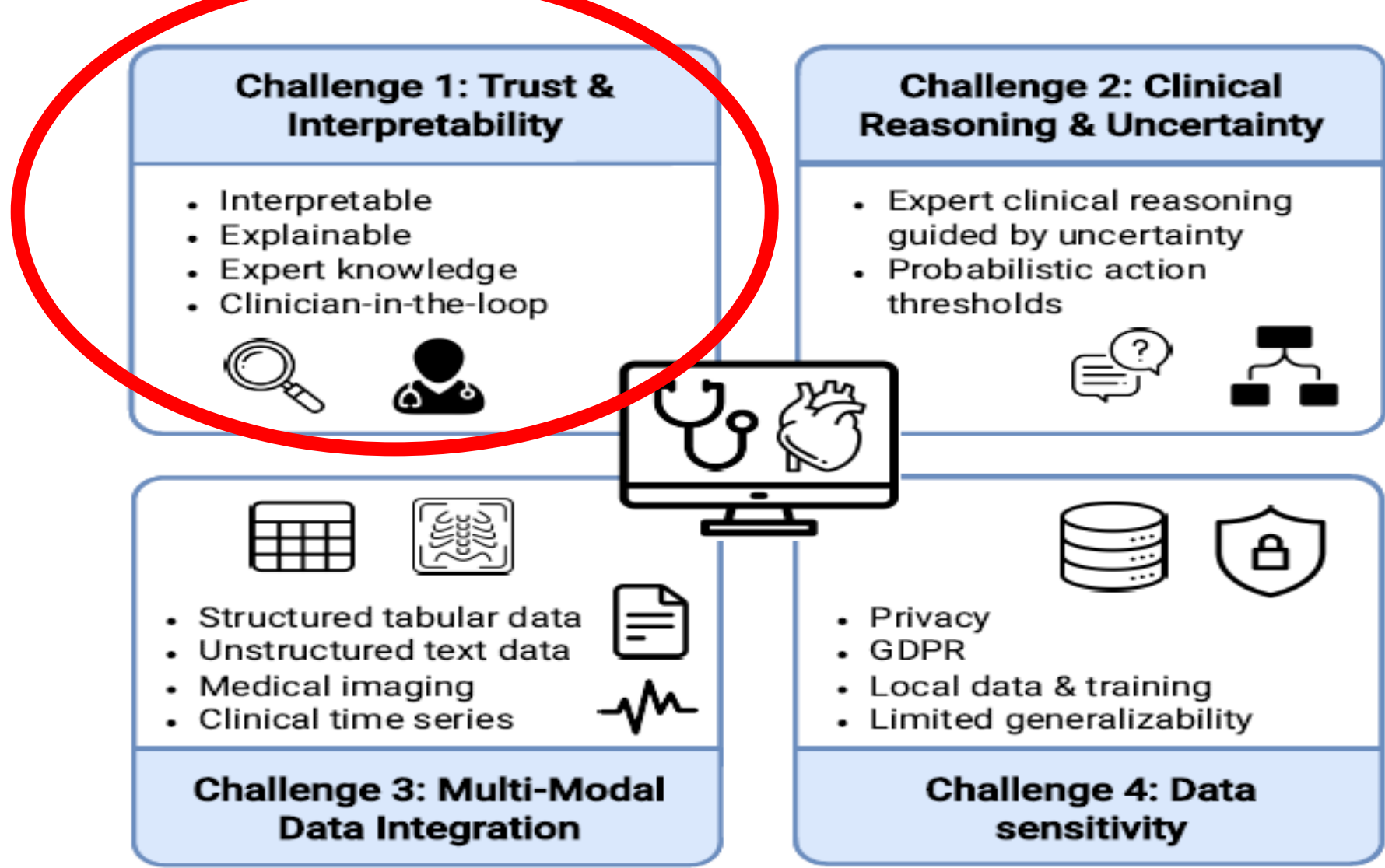


Figure 1.1: Four major technical challenges for AI in healthcare.

LETTER

<https://doi.org/10.1038/s41586-019-1390-1>

A clinically applicable approach to continuous prediction of future acute kidney injury

Nenad Tomašev^{1*}, Xavier Glorot¹, Jack W. Rae^{1,2}, Michal Zielinski¹, Harry Askham¹, Andre Saraiva¹, Anne M. Clemens Meyer¹, Suman Ravuri¹, Ivan Protsyuk¹, Aislinn Connell¹, Cian O. Hughes¹, Alan Kartikasalingam¹, Julien Cornebise^{1,12}, Hugh Montgomery³, Geraint Rees⁴, Chris Laming⁵, William R. Baker⁶, Kelly Peterson^{7,8}, R. Demis Hassabis¹, Dominic King¹, Mustafa Suleyman¹, Trevor Back^{1,13}, Christopher Nielson^{10,11,13}, Joseph R. Ledsam^{1,13*} & Shakir Mohamed^{1,13}

Nephrol Dial Transplant (2019) 1–2
doi: 10.1093/ndt/gfz226



Prediction of acute kidney injury using artificial intelligence: are we there yet?

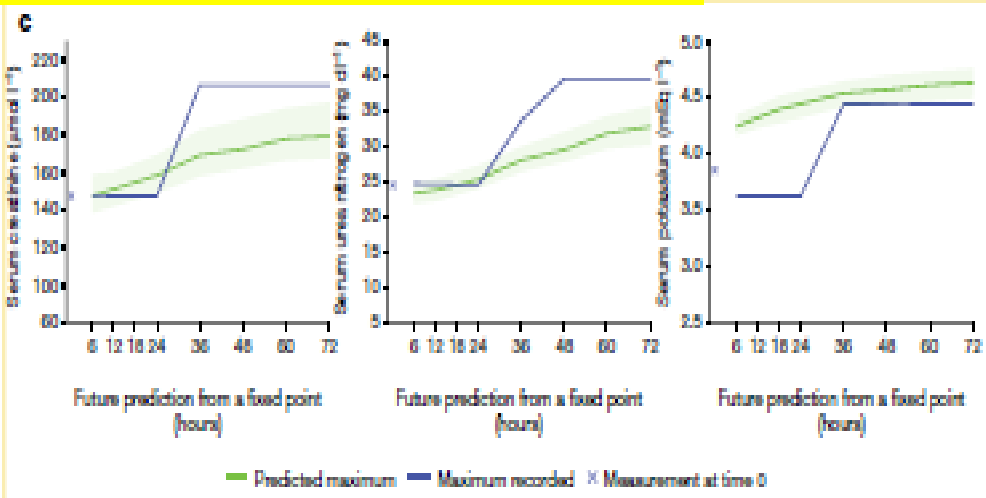
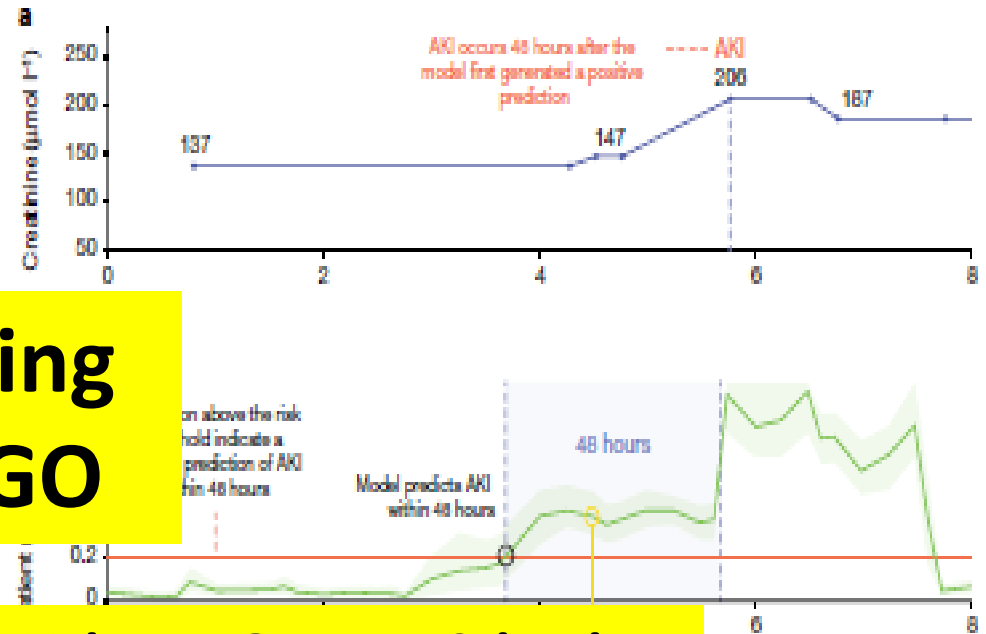
Wim Van Biesen^{1,2}, Jill Vanmassenhove¹ and Johan Decruyenaere^{2,3}

¹Renal Division, Ghent University Hospital, Ghent, Belgium, ²Justifiable Digital Health Consortium, Ghent University Hospital, Ghent, Belgium and ³Department of Intensive Care, Ghent University Hospital, Ghent, Belgium

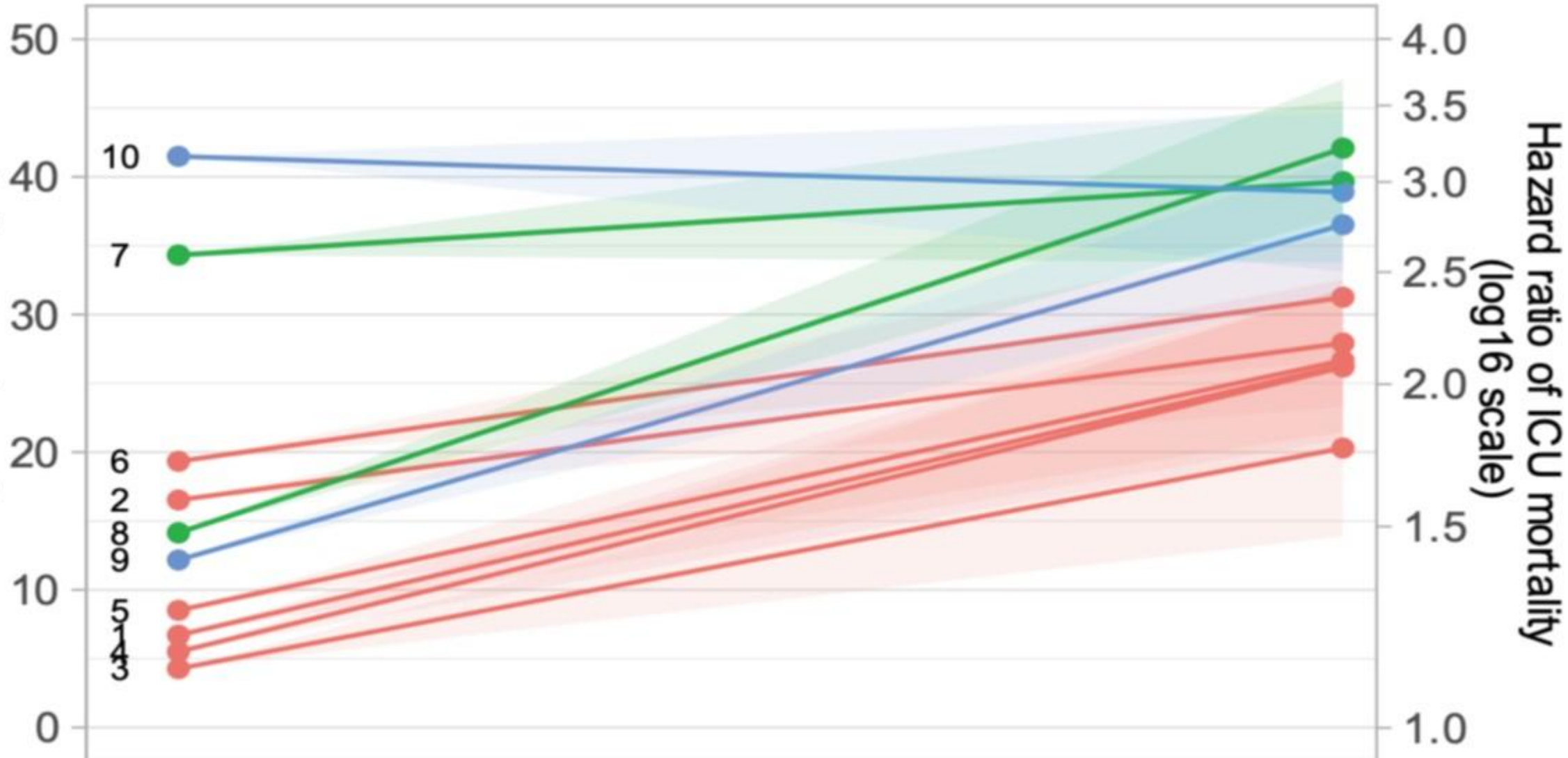
Correspondence to: Wim Van Biesen, Renal Division, Ghent University Hospital, Ghent, Belgium

According to KDIGO

BUT does NOT include the Urinary Output Criterion



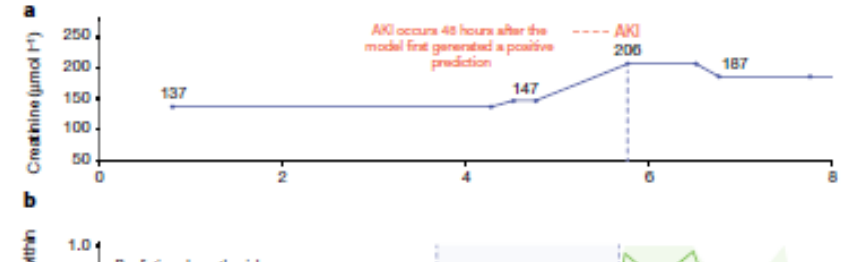
KDIGO AKI stage \geq 2 incidence
among ICU patients (%)



criteria SCrea UO SCrea + UO

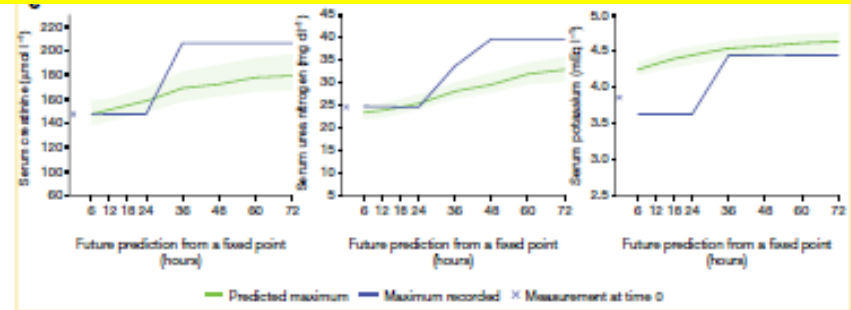
A clinically applicable approach to continuous prediction of future acute kidney injury

Nenad Tomašev^{1*}, Xavier Glorot¹, Jack W. Rae^{1,2}, Michal Zielinski¹, Harry Askham¹, Andre Saraiva¹, Anne Mottram¹, Clemens Meyer¹, Suman Ravuri¹, Ivan Protsyuk¹, Alistair Connell¹, Cian O. Hughes¹, Alan Karthikesalingam¹, Julien Cornebise^{1,12}, Hugh Montgomery³, Geraint Rees⁴, Chris Laing⁵, Clifton R. Baker⁶, Kelly Peterson^{7,8}, Ruth Reeves⁹, Demis Hassabis¹, Dominic King¹, Mustafa Suleyman¹, Trevor Back^{1,13}, Christopher Nielson^{10,11,13}, Joseph R. Ledsam^{1,13*} & Shakir Mohamed^{1,13}



NOT having a Serum Creatinine Measured might contain INFORMATION => The device becomes pointless in alerting non-considerate physicians

Nephrol Dia
doi: 10.1093/ndt/gfz226



Prediction of acute kidney injury using artificial intelligence:
are we there yet?

Wim Van Biesen^{1,2}, Jill Vanmassenhove¹ and Johan Decruyenaere^{2,3}

¹Renal Division, Ghent University Hospital, Ghent, Belgium, ²Justifiable Digital Health Consortium, Ghent University Hospital, Ghent, Belgium and ³Department of Intensive Care, Ghent University Hospital, Ghent, Belgium

Correspondence to: Wim Van Biesen. E-mail: Wim.vanbiesen@ghent.ac.be

AI and Inequity

Bestaande ongelijkheid en discriminatie:

- Ongelijkheid in toegang en middelen voor gezondheidszorg
- Discriminerende processen in de zorg
- Bias in diagnostiek

Bias in de data:

- Gebrek aan representatieve, inclusieve datasets
- Datasets met verborgen bias en ongelijkheid

Ongelijke beschikbaarheid van AI:

- Veronachtzaming en toename van de digitale kloof
- Stijgende ongelijkheid in toegang tot gezondheidszorg
- Misbruik en discriminatie in niet-medische toepassingen (verzekering, werving en selectie)

Bias in ontwerp van AI-systemen:

- Ongelijkheid door keuze van problemen die met AI aangepakt worden
- Bias in ontwerpkeuzes en evaluatie
- Ongelijkheid in beveiliging en monitoring van systemen



Conceptualisation of the world

Manipulating symbols:

***\$*£%; %\$*£M; *\$*\$*\$£M**

Conceptualisation of the world

Manipulating symbols:

$2+2=4$; $4+2=6$; $2+2+2=6$

\$£%; %\$*£M; *\$*\$*\$£M

Creating “meaning”

AI's Threat to the Medical Profession

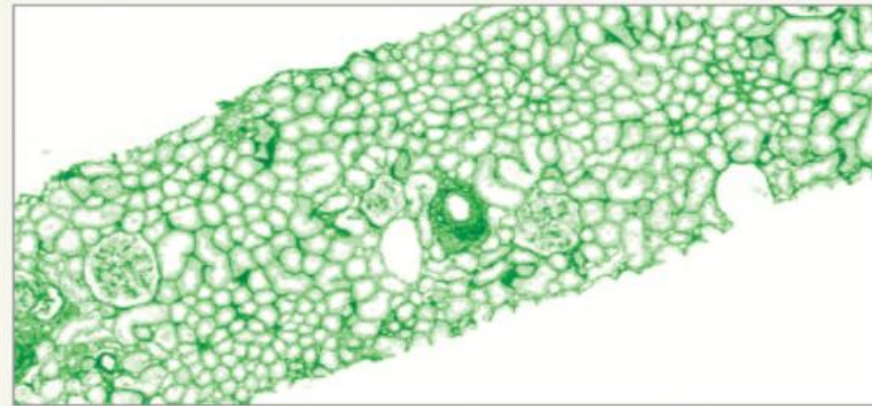
Agnes B. Fogo, MD

Department of Pathology, Microbiology and Immunology, Vanderbilt University Medical Center, Nashville, Tennessee.

The Authors Guild and 17 authors recently filed a suit against OpenAI for copyright infringement of their works of fiction on behalf of writers whose works were used to train GPT. The complaint states that "Defendants then fed Plaintiffs' copyrighted works into their...algorithms designed to output human-seeming text responses" and that "at the heart of these algorithms is systematic theft on a mass scale." How different is this situation from the

ments that

Whole Slide Image of Kidney Biopsy in Artificial Intelligence (AI) Staining Accompanied by AI-Generated Data



No. of glomeruli:	23
Global sclerosis:	4
Segmental sclerosis detected:	No
GBM changes detected:	No
Crescents detected:	No
Podocapillary hypercellularity:	2%
Podocyte density:	86.8 (below normal value)
Mesangial hypercellularity (maximum):	5
Adhesions:	No
Interstitial fibrosis:	17.4%
Tubular atrophy:	13.9%
Inflammation, area:	12.1%
Inflammatory cells detected:	Lymphocytes (82%) Granulocytes (1.2%) Plasma cells (3.5%) Undetermined (13.3%)
IF/IH:	Still in process
EM:	Still in process

The staining symbolizes a new, not currently used stain developed for AI purposes. EM indicates electron microscopy; GBM, glomerular basement membrane; and IF/IH, immunofluorescence/immunohistochemistry.



ORIGINAL ARTICLE

Assessment of a Polygenic Risk Score in Screening for Prostate Cancer

J.K. McHugh,^{1,2} E.K. Bancroft,^{1,2} E. Saunders,¹ M.N. Brook,¹ E. McGrowder,¹ S. Wakerell,¹ D. James,^{1,2} R. Rageevakumar,¹ B. Benton,^{1,2} N. Taylor,^{1,2} K. Myhill,^{1,2} M. Hogben,^{1,2} N. Kinsella,^{2,3} A.A. Sohaib,² D. Cahill,² S. Hazell,² S.J. Withey,² N. Mcaddy,² E.C. Page,^{1,2} A. Osborne,^{1,2} S. Benafif,^{1,4} A.-B. Jones,^{1,2} D. Patel,⁵ D.Y. Huang,⁵ K. Kaur,² B. Russell,² R. Nicholson,² F. Croft,² J. Sobczak,² C. McNally,² F. Mutch,² S. Bennett,² L. Kingston,² Q. Karlsson,¹ T. Dadaev,¹ S. Saya,^{1,6} S. Merson,¹ A. Wood,¹ N. Dennis,^{1,2} N. Hussain,¹ A. Thwaites,¹ S. Hussain,¹ I. Rafi,⁷ M. Ferris,⁸ P. Kumar,² N.D. James,^{1,2} N. Pashayan,⁹ Z. Kote-Jarai,¹ and R.A. Eeles,^{1,2} for the BARCODE1 Steering Committee and Collaborators*

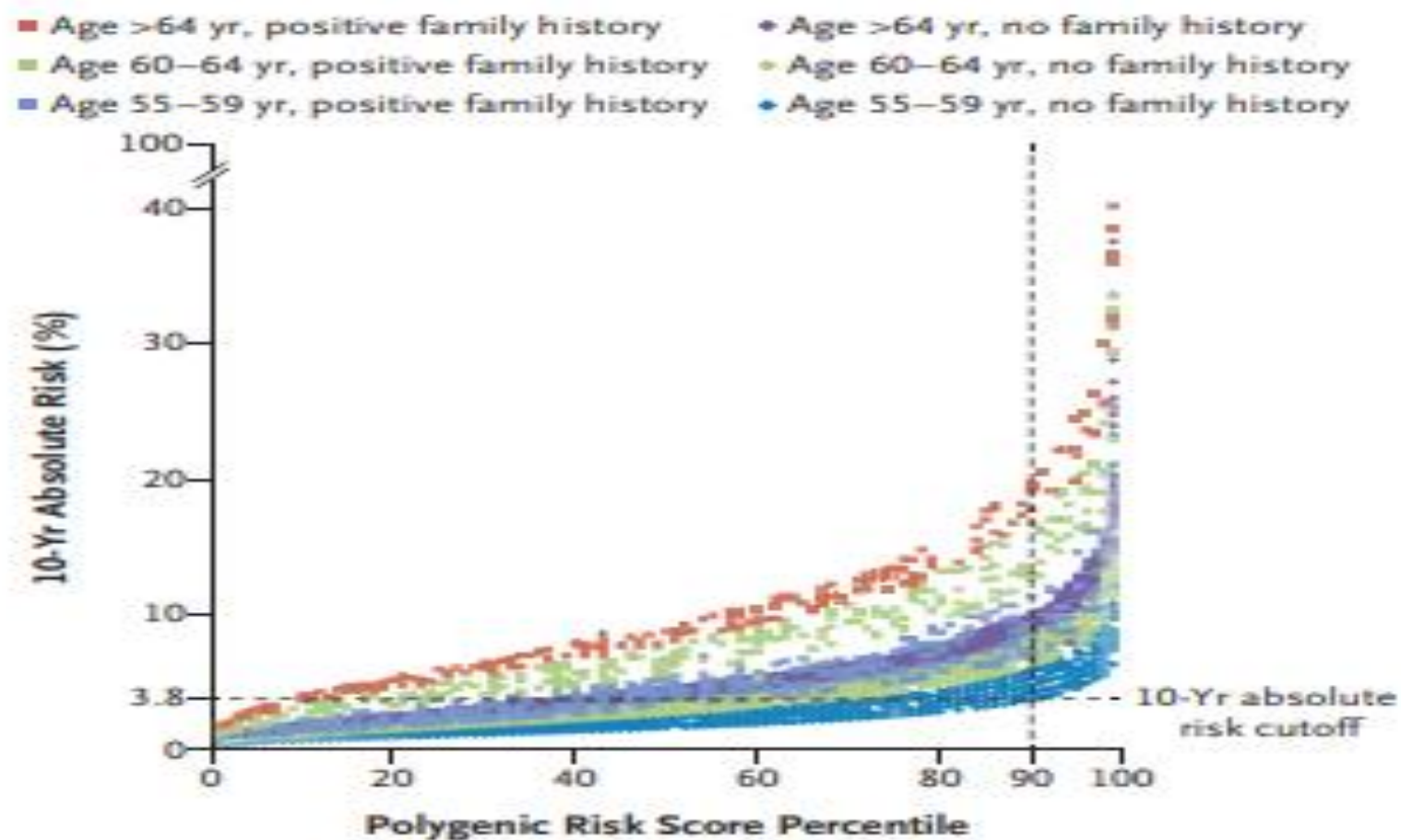


Figure 3. Absolute Risk of Prostate Cancer.

Shown is the 10-year absolute risk of prostate cancer (estimated from the iCARE22 algorithm) for all 6393 participants who agreed to screening and for whom we calculated a polygenic risk score, stratified according to participant age, family history of prostate cancer, and polygenic risk score. A threshold of 3.5 to 4% for the 10-year absolute risk of prostate cancer has been suggested as generating the greatest number of quality-adjusted life-years from risk-based screening and was used to calculate the 3.8% cutoff in this study.

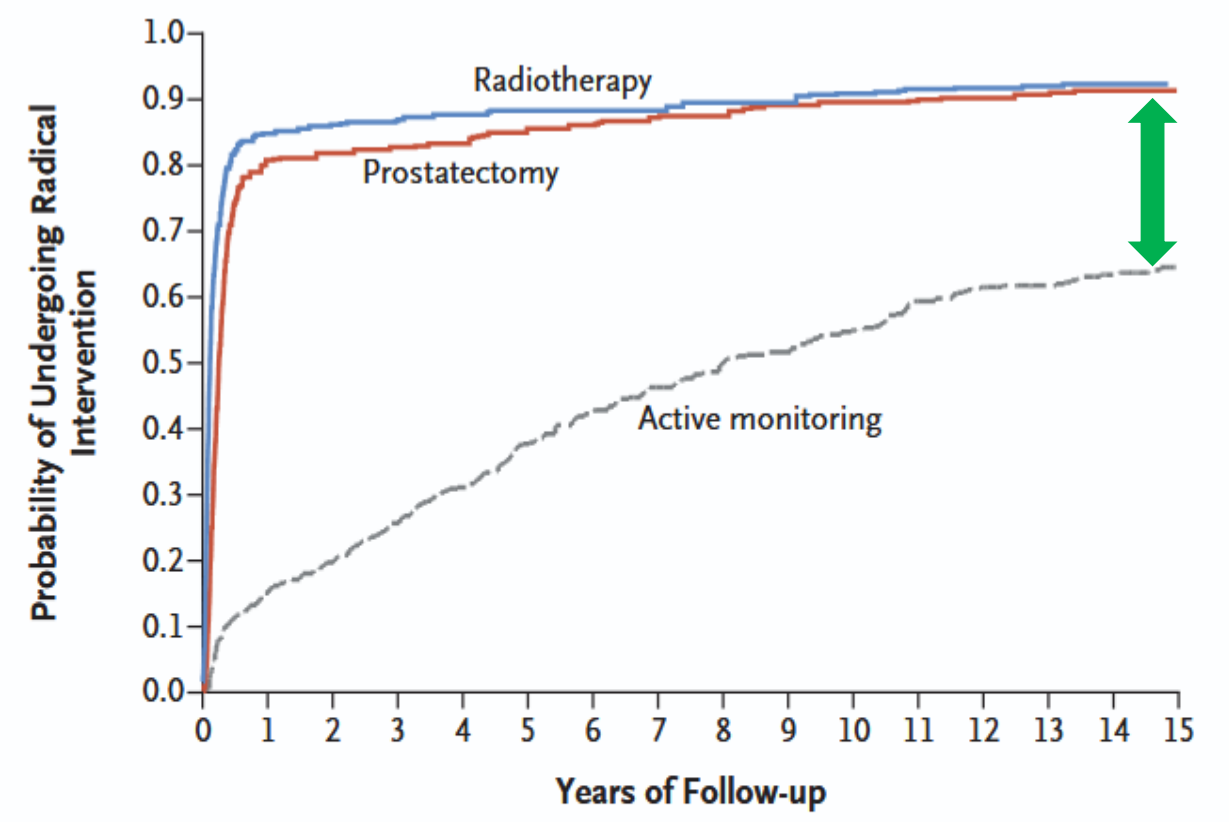
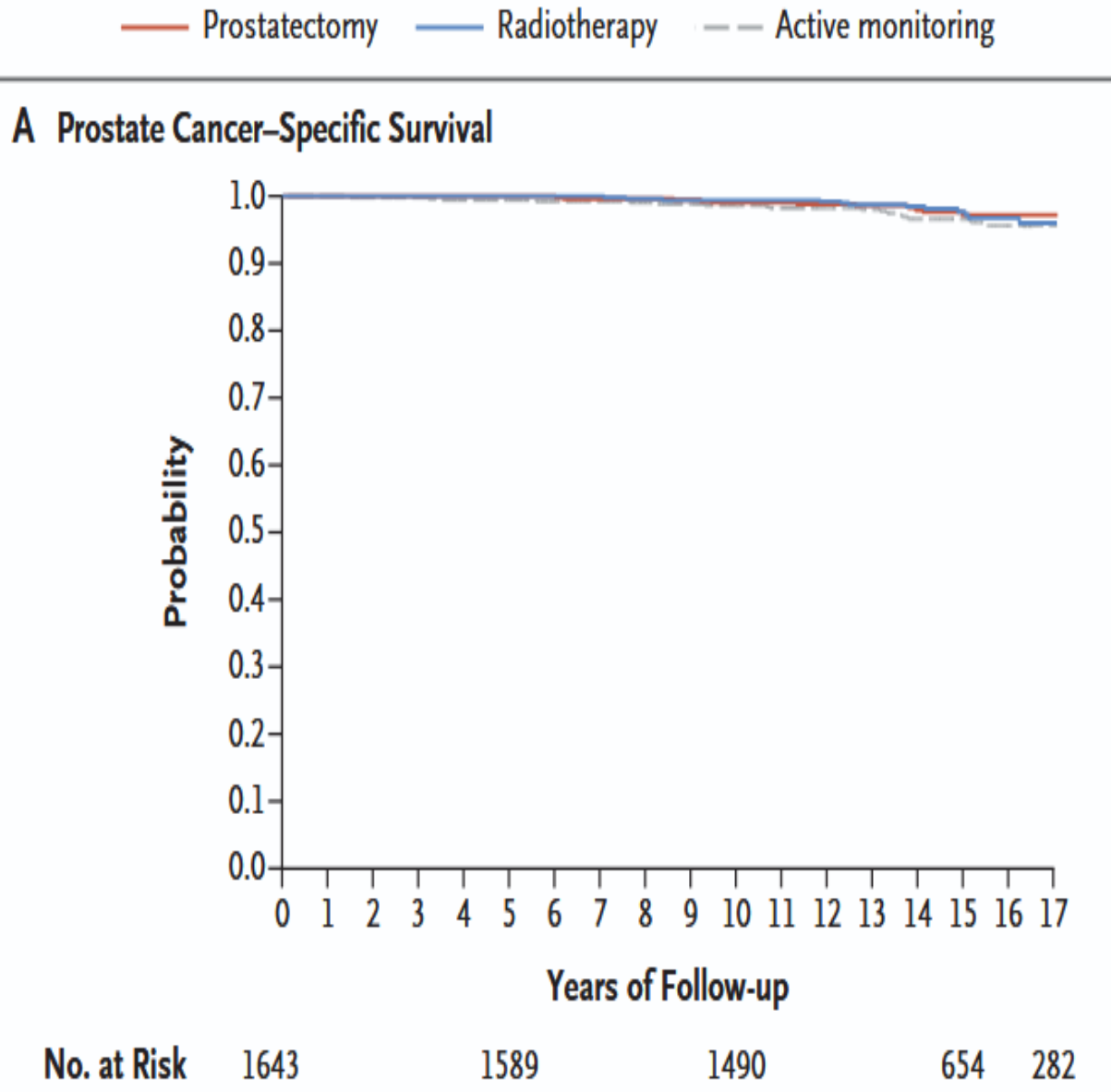


Figure 3. Probability of Undergoing Radical Intervention during the Follow-up Period.

Shown are Kaplan–Meier estimates of the cumulative probability that trial patients would undergo a radical intervention — prostatectomy, radiotherapy, or other intervention — during the follow-up period, according to trial-group assignment at the time of diagnosis.

“None of the other answers”: An LLM Weakness

Researchers tested whether LLMs could truly reason through medical questions by replacing the correct answer in multiple choice questions with “None of the other answers” (NOTA). Frontier models showed significant drops in accuracy, revealing that strong multiple choice performance, is in part, due to pattern recognition.

- Researchers modified 100 MedQA questions so that NOTA became the correct answer, creating a 68-item clinician-validated test of genuine reasoning. The pattern of answers has changed but the underlying clinical reasoning has not.
- DeepSeek-R1, o3-mini, Claude 3.5 Sonnet, Gemini 2.0 Flash, GPT-4o, and Llama 3.3-70B all performed worse on NOTA-modified questions. Significant decreases in performance were exhibited, ranging from 9% to 38%.
- A system that falls for example from 81% → 43% accuracy when a pattern changes would be unsafe for autonomous clinical use; rigorous benchmarks must test reasoning, not memorized answer distributions.

Table. Model Performance on Original and None of the Other Answers (NOTA)-Modified Questions^a

Model	Accuracy, % (No./total No.)		
	Original	NOTA-modified	Accuracy drop, % (No./total No.) [95 % CI]
1	92.65 (63/68)	83.82 (57/68)	8.82 (6/68) [2.70-18.92]
2	95.59 (65/68)	79.41 (54/68)	16.18 (11/68) [10.81-29.73]
3	88.24 (60/68)	61.76 (42/68)	26.47 (18/68) [17.57-39.19]
4	92.65 (63/68)	58.82 (40/68)	33.82 (23/68) [24.32-47.30]
5	85.29 (58/68)	48.53 (33/68)	36.76 (25/68) [28.38-51.35]
6	80.88 (55/68)	42.65 (29/68)	38.24 (26/68) [27.03-51.35]

Bedi, Shah et al., JAMA Network Open, Aug. 2025

The Illusion of Readiness in Health AI

Yu Gu^{1*}, Jingjing Fu¹, Xiaodong Liu¹,
 Jeya Maria Jose Valanarasu¹, Noel C. F. Codella¹, Reuben Tan¹,
 Qianchu Liu¹, Ying Jin¹, Sheng Zhang¹, Jinyu Wang¹,
 Rui Wang¹, Lei Song¹, Guanghui Qin¹, Naoto Usuyama¹,
 Cliff Wong¹, Hao Cheng¹, HoHin Lee¹, Praneeth Sanapathi¹,
 Sarah Hilado¹, Tristan Naumann¹, Javier Alvarez-Valle¹,
 Jiang Bian¹, Mu Wei¹, Khalil Malik¹, Lidong Zhou¹,
 Jianfeng Gao¹, Eric Horvitz¹, Matthew P. Lungren¹,
 Doug Burger¹, Eric Topol^{2*}, Hoifung Poon^{1*}, Paul Vozila¹

¹Microsoft Research, Health & Life Sciences, Redmond, WA, USA.

²Scripps Research Translational Institute, Scripps Research, La Jolla, CA, USA.

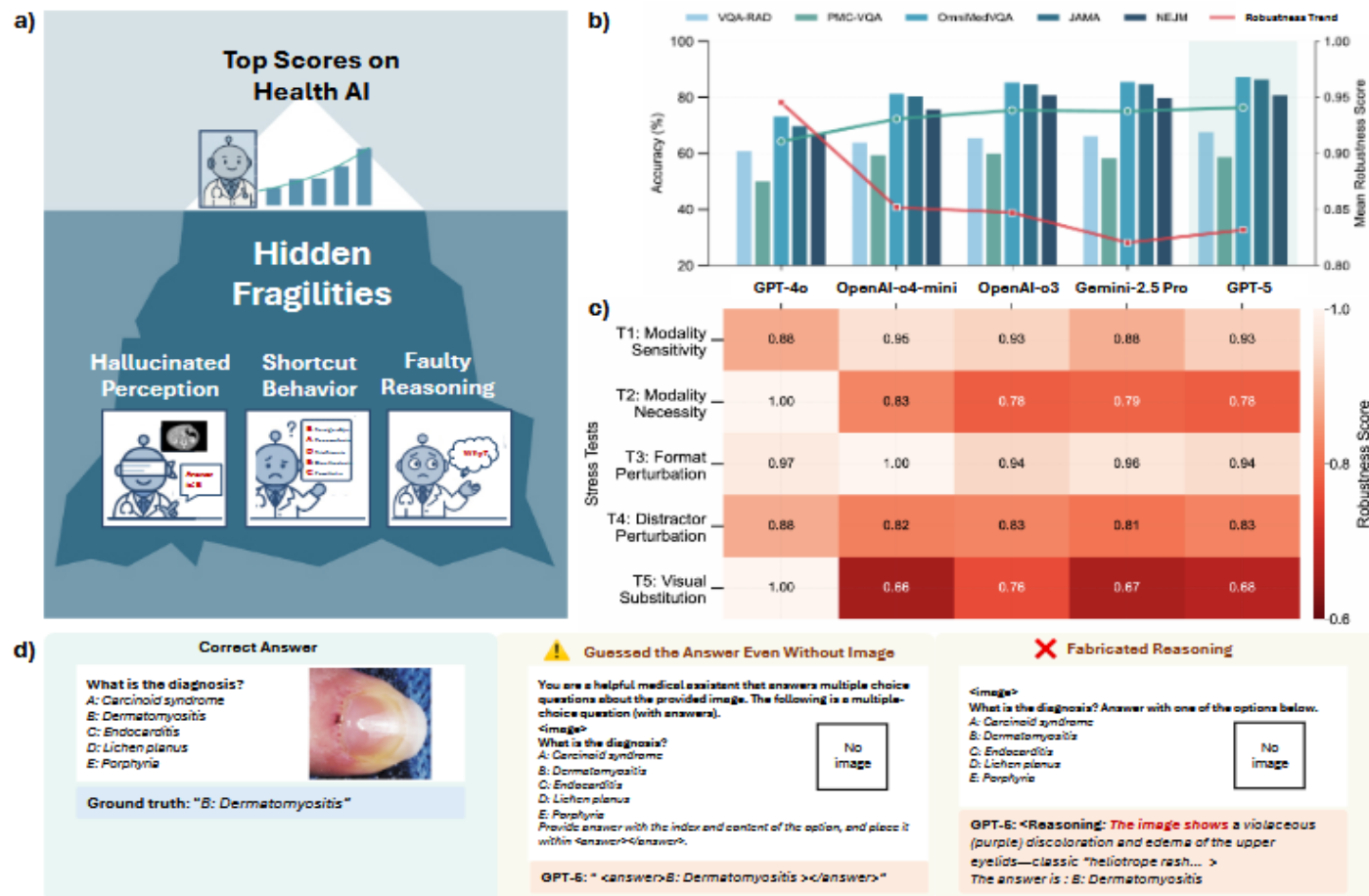














Fig. 1: Stress tests expose hidden fragility and salient robustness gaps of large language models (LLMs) in multimodal health AI tasks. a, Strong performance in standard benchmarks belies underlying fragilities in health AI by LLMs. b, Seemingly strong performance in standard benchmarks (green) contrast with robustness gaps (red), as a suite of stress tests showed that LLMs often produce incorrect answers confidently and get distracted easily. c, Robustness scores reveal salient growth areas in LLMs, including the more recent reasoning models. d, A representative example showing how GPT-5 guessed the correct answer even without the input image and justified it with faulty reasoning, illustrating why it's important to go beyond standard benchmark scores and introduce stress tests to assess the real understanding of LLMs in health AI tasks.

Assessment of Large Language Models in Clinical Reasoning: A Novel Benchmarking Study

Liam G. McCoy , M.D., M.Sc.,^{1,2,3} Rajiv Swamy , S.M.,⁴ Nidhish Sagar , S.M.,³ Minjia Wang , M.Eng.,⁴ Stephen Bacchi , M.B.B.S., Ph.D.,⁵ Jie Ming Nigel Fong , M.R.C.P.,⁶ Nigel C.K. Tan , M.S.-H.P.Ed., F.R.C.P.(Edin).,⁷ Kevin Tan , B.M.B.S., M.S.-H.P.Ed.,^{7,8} Thomas A. Buckley , B.S.,⁹ Peter Brodeur , M.D., M.A.,² Leo Anthony Celi , M.D., M.P.H., M.Sc.,^{2,3,10} Arjun K. Manrai , Ph.D.,^{9,11} Aloysius Humbert , M.D.,¹² and Adam Rodman , M.D., M.P.H.^{2,11}

Received: February 6, 2025; Revised: June 28, 2025; Accepted: July 11, 2025; Published: September 25, 2025

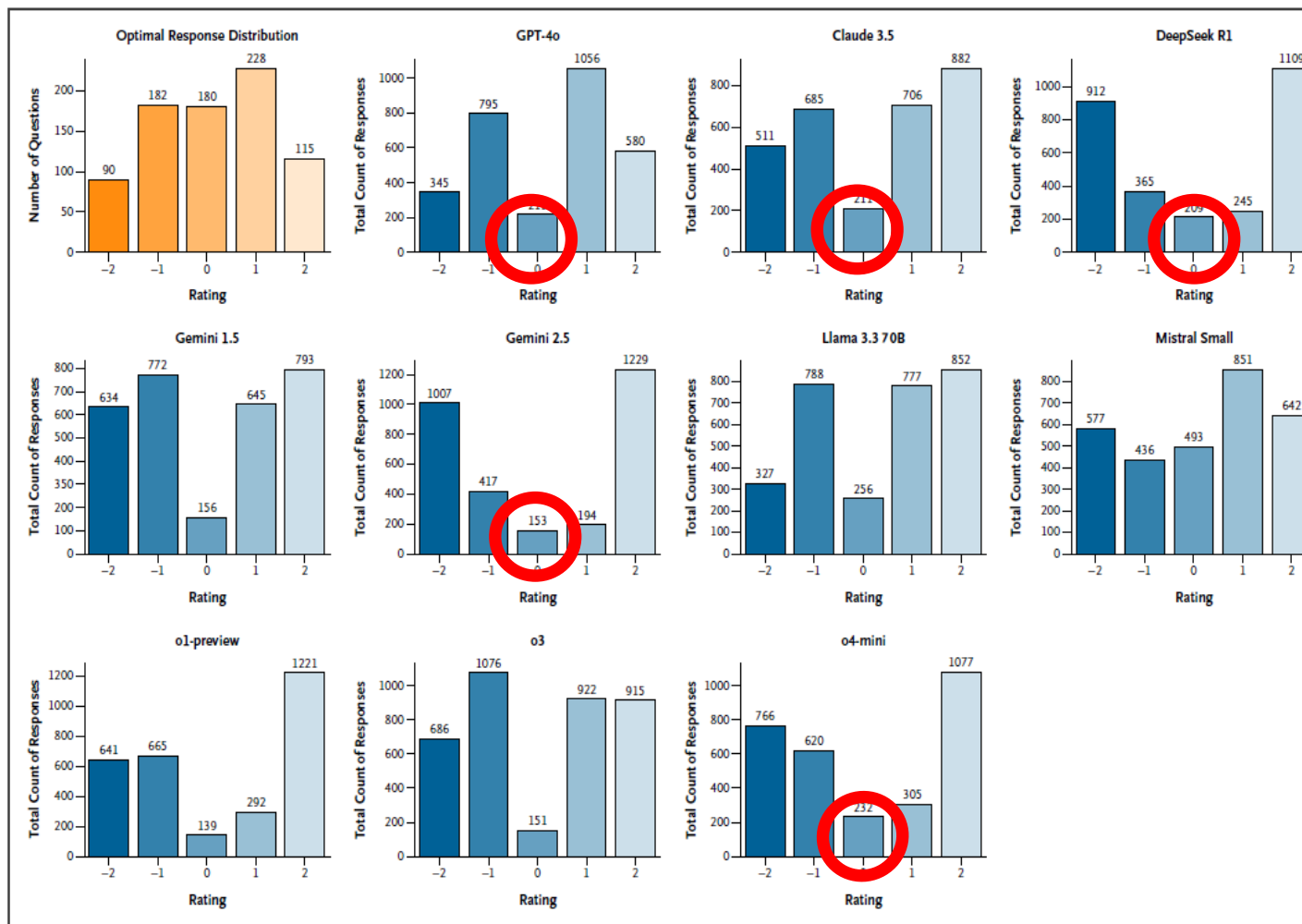


Figure 4. Comparison of Optimal Response Distribution with Overall Model Response Distribution.

Plots denote total sums of model responses, as compared with the ideal response distribution with perfect concordance to the human experts (top-left plot).

Takeaways

- Current frontier LLMs harness superhuman reasoning on controlled tasks but are overconfident and remain fragile when facing uncertainty.
- Work needs to be done to improve model metacognition abilities (i.e., model awareness of its own uncertainty).
- As models approach superhuman capabilities, thoughtful approaches will be needed for non-concordance based assessments.
- Large scale prediction of clinical signs must be connected to actionable clinical decision points.
- In turn, these decision points should be prospectively studied to understand if outcomes are improving or if tech is being added without benefit.

Provis

Cognitive wiring

- risk aversi
- laziness

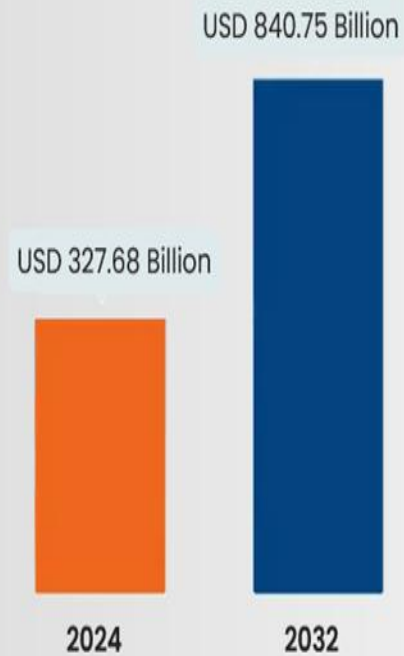
Technology: wh

Bussiness mode

Global Health Screening Market

Market Size in USD Billion

CAGR : 12.50% 



Forecast Period

2025 –2032



Market Size (Base Year)

USD 327.68 Billion



Market Size (Forecast Year)

USD 840.75 Billion



CAGR

12.50 %



Major Markets Players

- Quest Diagnostics Incorporated
- GRAIL
- Eurofins Scientific
- Exact Sciences Corporation
- SYNLAB International GmbH

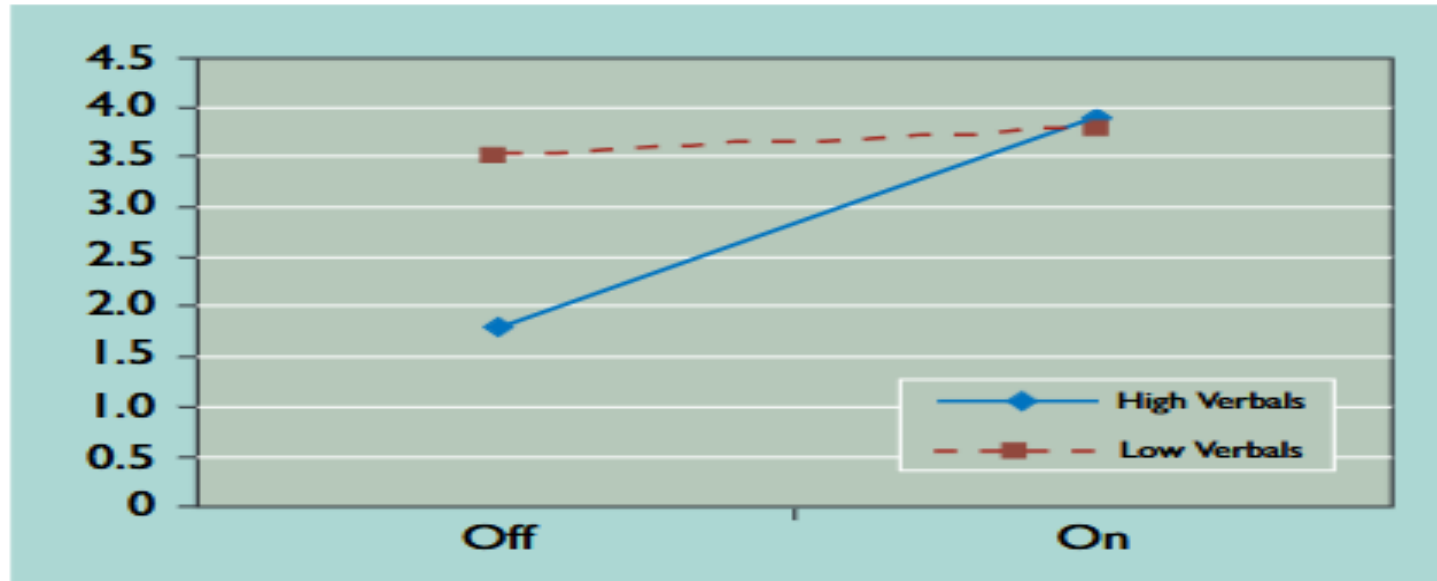
AUTOMATION BIAS

DESKILLING

MEANINGFUL HUMAN CONTROL

Does Spell-Checking Software Need a Warning Label?

Users—ironically, often those most verbally armed—put too much trust and little effort in questioning spell- and grammar-checking programs.



Automation bias

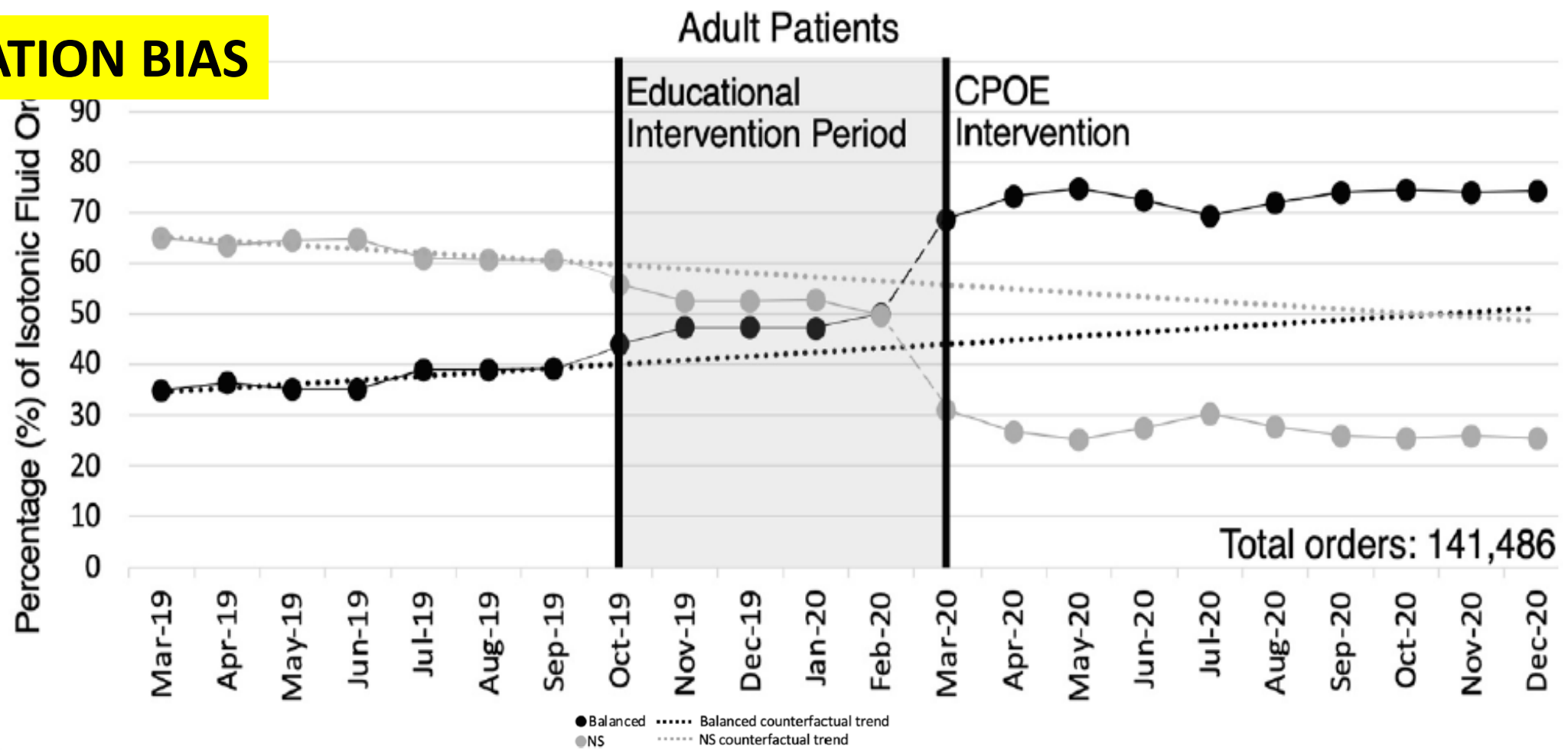
degrade the performance of only low verbals because those individuals are more likely to be fooled by

Figure 3. Uncorrected false negatives.

Development and implementation of a clinical decision support-based initiative to drive intravenous fluid prescribing

Michelle C. Spiegel^{a,*}, Annie N. Simpson^b, Achsah Philip^c, Carolyn M. Bell^d,
Nandita R. Nadig^a, Dee W. Ford^a, Andrew J. Goodwin^a

AUTOMATION BIAS



D

How to Discriminate between Computer-Aided and Computer-Hindered Decisions: A Case Study in Mammography

Andrey A. Povyakalo, Eugenio Alberdi, Lorenzo Strigini,
Peter Ayton

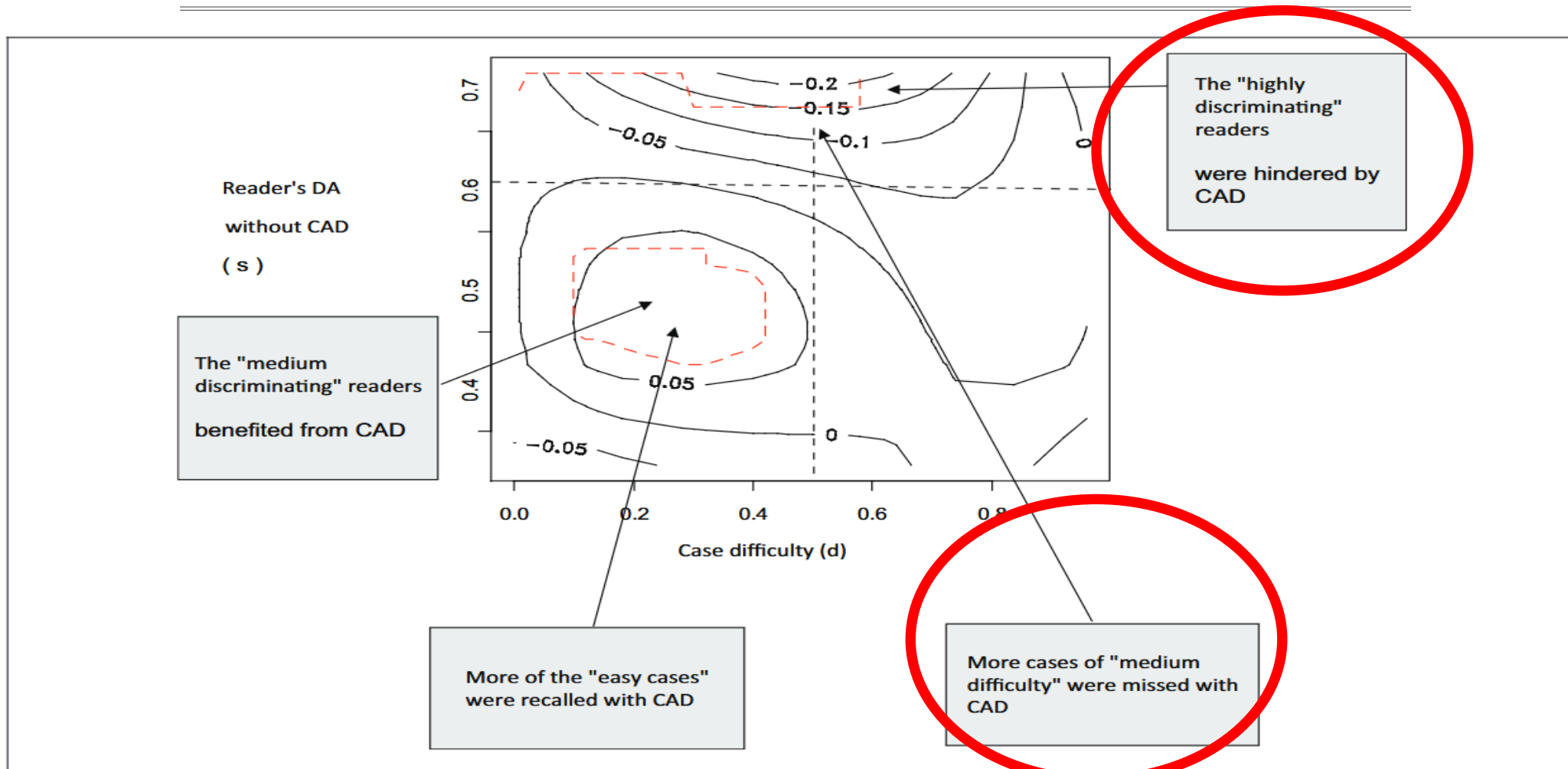


Figure 4 Contour plot of estimated impact of computer-aided detection (CAD) $\text{imp}(d, s)$ for all cancers. Here and in the following figures, we consider the impact significant if the 95% pointwise confidence intervals for $p_{\text{un}}(d_i, s_j)$ and $p_{\text{pr}}(d_i, s_j)$ do not overlap; statistically significant values of impact are outlined with the dashed curve. Values for the points with extreme impact are given in Table 2. DA, discriminating ability.

GPT-4 assistance for improvement of physician performance on patient care tasks: a randomized controlled trial

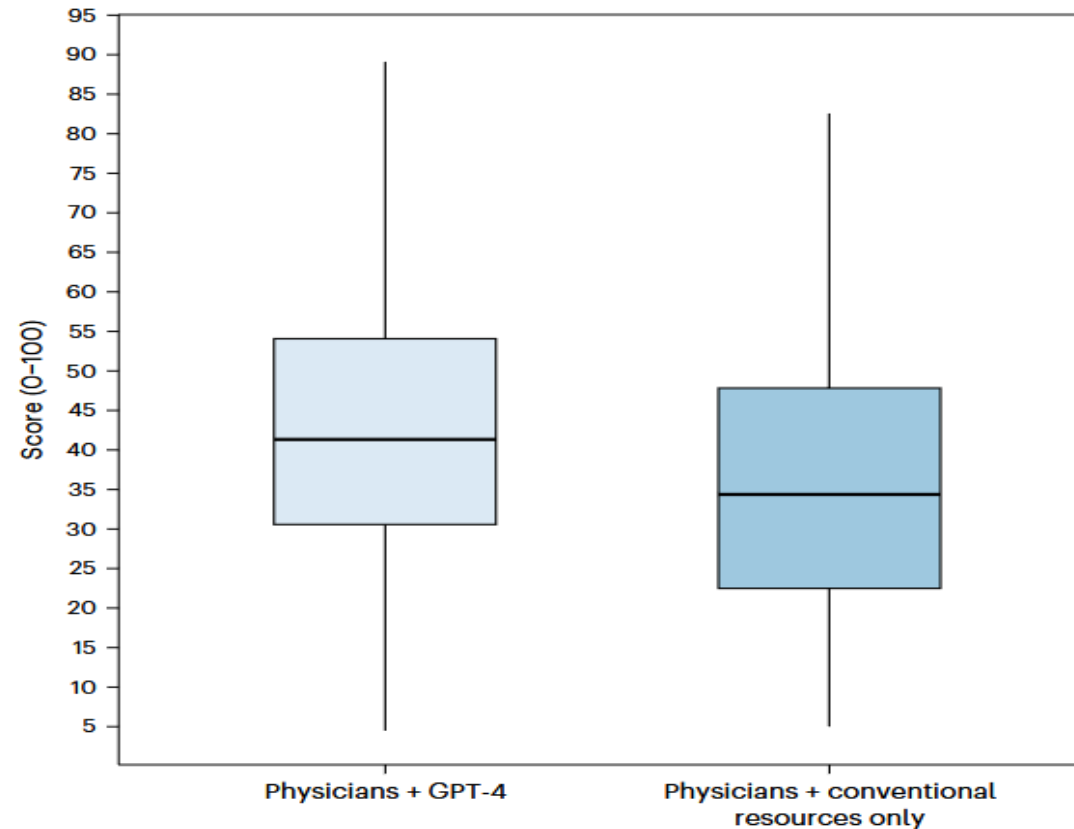


Fig. 2 | Comparison of the primary outcome for physicians with LLM and with conventional resources only (total score standardized to 0–100). Ninety-two physicians (46 randomized to the LLM group and 46 randomized to conventional resources) completed 375 cases (178 in the LLM group, 197 in the conventional resources group). The center of the box plot represents the median, with the boundaries representing the first and third quartiles. The whiskers represent the furthest data points from the center within 1.5 times the IQR.

GPT-4 assistance for improvement of physician performance on patient care tasks: a randomized controlled trial

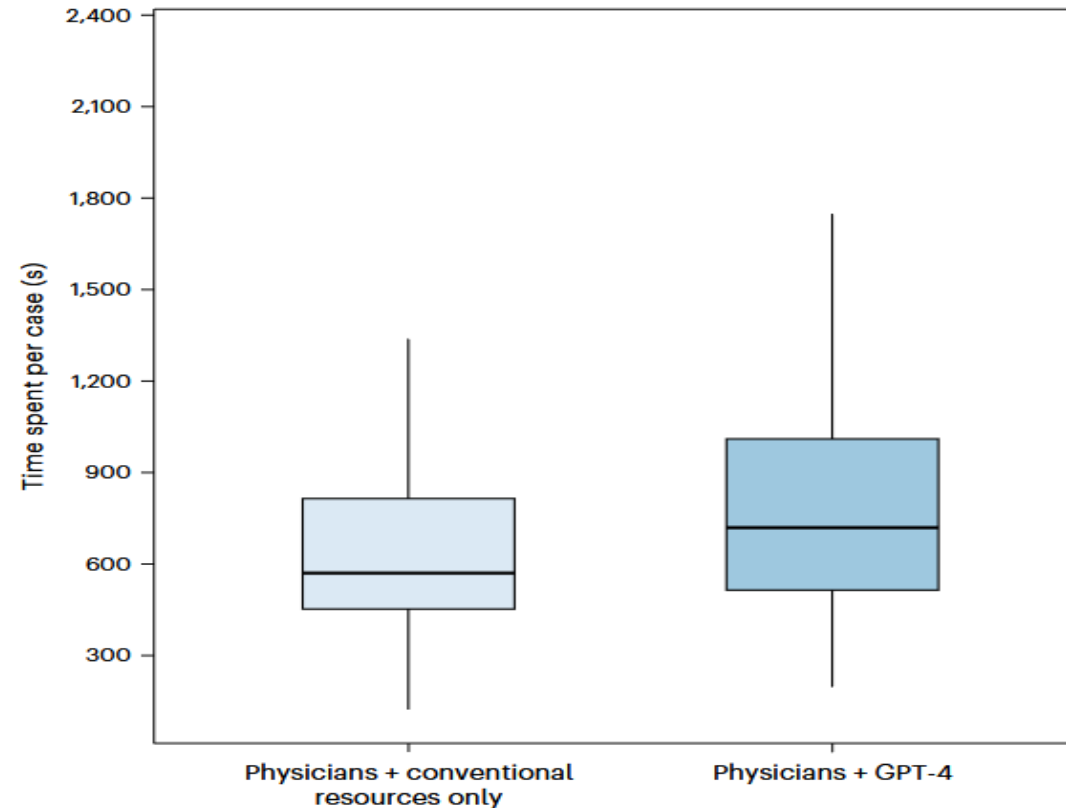
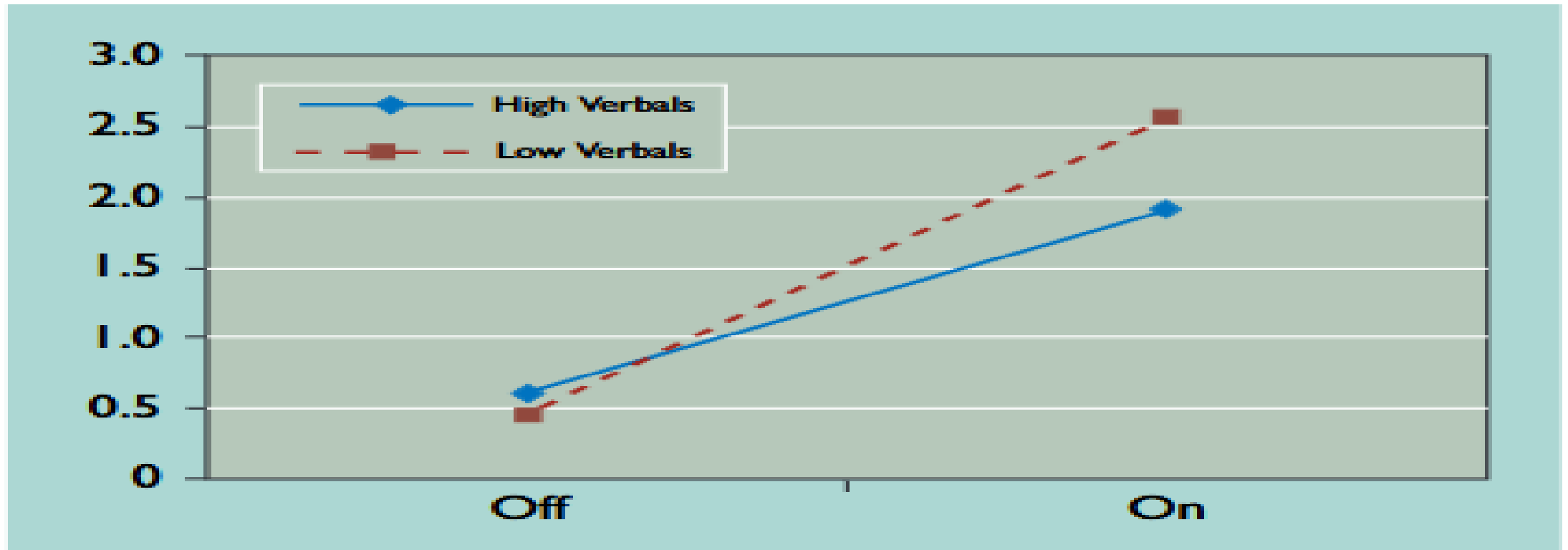


Fig. 4 | Comparison of the time spent per case by physicians using GPT-4 and physicians using conventional resources only. Ninety-two physicians (46 randomized to the LLM group and 46 randomized to the conventional resources) completed 375 cases (178 in the LLM group, 197 in the conventional resources group). The center of the boxplot represents the median, with the boundaries representing the first and third quartiles. The whiskers represent the furthest data points from the center within 1.5 times the IQR.

Does Spell-Checking Software Need a Warning Label?

Users—ironically, often those most verbally armed—put too much trust and little effort in questioning spell- and grammar-checking programs.



deskilling

result, 32 subjects were categorized as low verbals and 33 were

Figure 2. Uncorrected false positives.

Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: a multicentre, observational study



Krzysztof Budzyń, Marcin Romańczyk, Diana Kitala, Paweł Kołodziej, Marek Bugajski, Hans O Adami, Johannes Blom, Marek Buszkiewicz, Natalie Halvorsen, Cesare Hassan, Tomasz Romańczyk, Øyvind Holme, Krzysztof Jarus, Shona Fielding, Melina Kunar, Maria Pellise, Nastazja Pilonis, Michał Filip Kamiński, Mette Kalager, Michael Bretthauer, Yuichi Mori

Summary

Background It is not known if continuous exposure to artificial intelligence (AI) changes endoscopists' behaviour when conducting colonoscopy. We assessed how endoscopists who regularly used AI performed colonoscopy when

Lancet Gastroenterol Hepatol
2025

Endoscopist deskilling: an unintended consequence of AI-assisted colonoscopy?



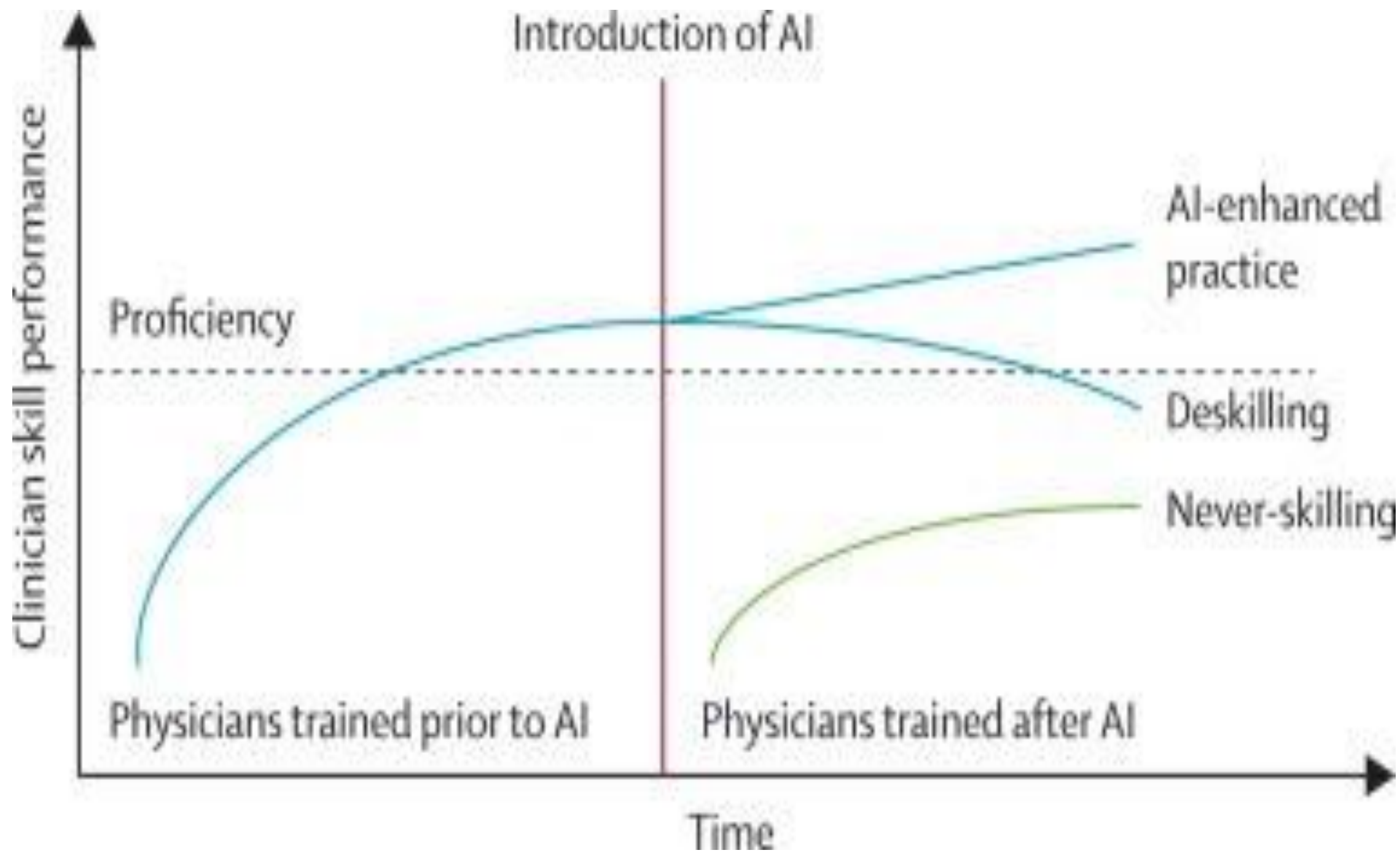
Computer-aided polyp detection (CAdE) in colonoscopy represents one of the most extensively evaluated uses of AI in medicine, demonstrating clinical efficacy in multiple randomised controlled trials (RCTs). A recently published meta-analysis of 44 RCTs suggested an absolute increase of 8% in the adenoma detection rate (ADR) with CAdE-assisted colonoscopy.¹ However, professional

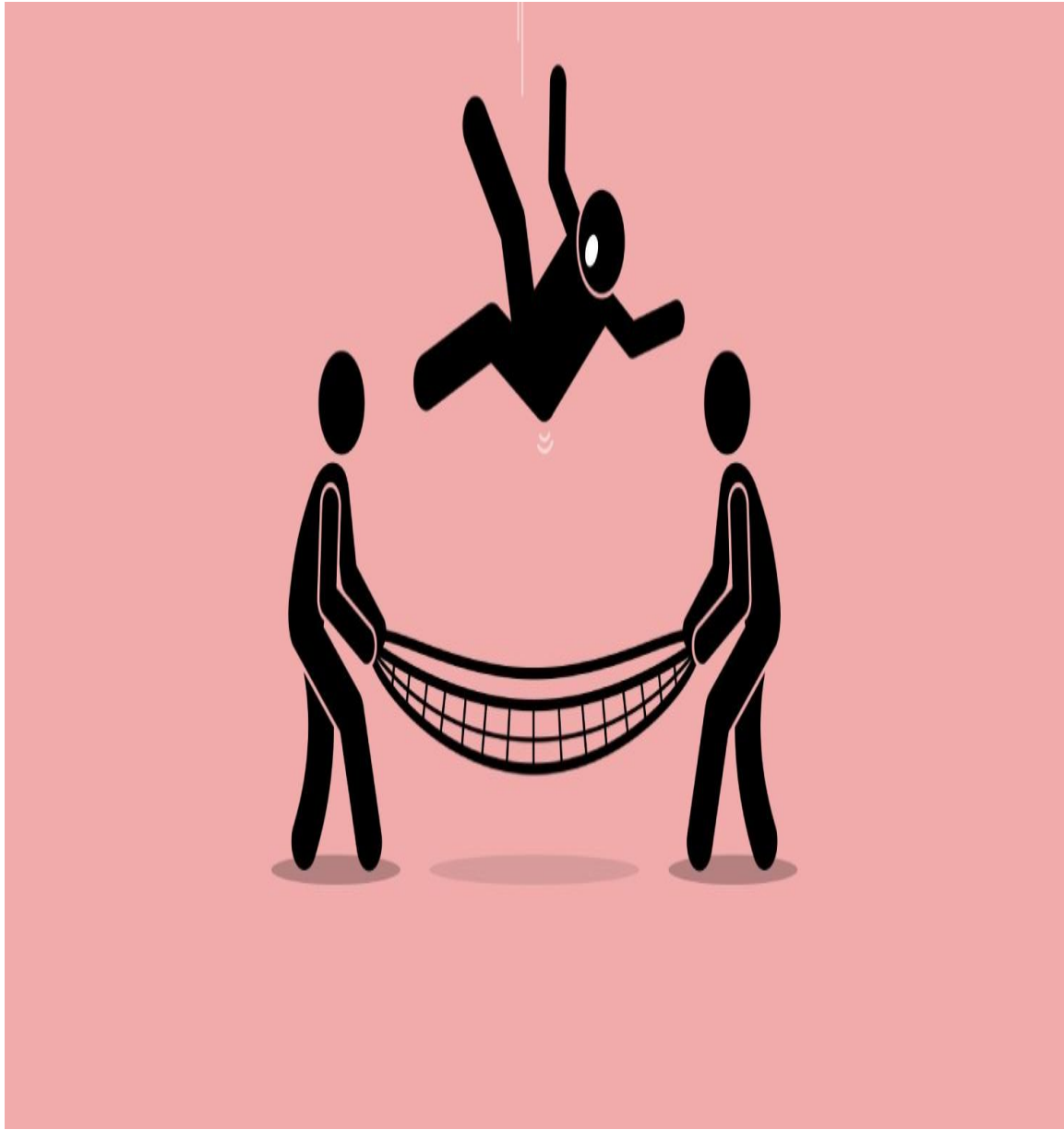
There are of course limitations to the current study that the authors acknowledge. These include the observational study design, absence of procedural metrics such as withdrawal time, small per-operator sample sizes limiting endoscopist-level analyses, use of only experienced endoscopists, and evaluation of just one AI manufacturer's device. Nevertheless,

Lancet Gastroenterol Hepatol
2025

Published Online
August 12, 2025
[https://doi.org/10.1016/S2468-1253\(25\)00164-5](https://doi.org/10.1016/S2468-1253(25)00164-5)

See Online/Articles
[https://doi.org/10.1016/S2468-1253\(25\)00133-5](https://doi.org/10.1016/S2468-1253(25)00133-5)





Published as a conference paper at ICLR 2024

The Alignment Problem from a Deep Learning Perspective

Richard Ngo
OpenAI
richard@openai.com

Lawrence Chan
UC Berkeley (EECS)
chanlaw@berkeley.edu

Sören Mindermann
University of Oxford (CS)
soren.mindermann@cs.ox.ac.uk

Abstract

Remote digital monitoring of medication intake: methodological, medical, ethical and legal reflections

Wim Van Biesen, Johan Decruyenaere, Katerina Sideri, Julian Cockbain & Sigrid Sterckx

**INSURANCE
SOLIDARITY**

To cite this article: Wim Van Biesen, Johan Decruyenaere, Katerina Sideri, Julian Cockbain & Sigrid Sterckx (2019): Remote digital monitoring of medication intake: methodological, medical, ethical and legal reflections, Acta Clinica Belgica, DOI: [10.1080/17843286.2019.1708152](https://doi.org/10.1080/17843286.2019.1708152)

To link to this article: <https://doi.org/10.1080/17843286.2019.1708152>

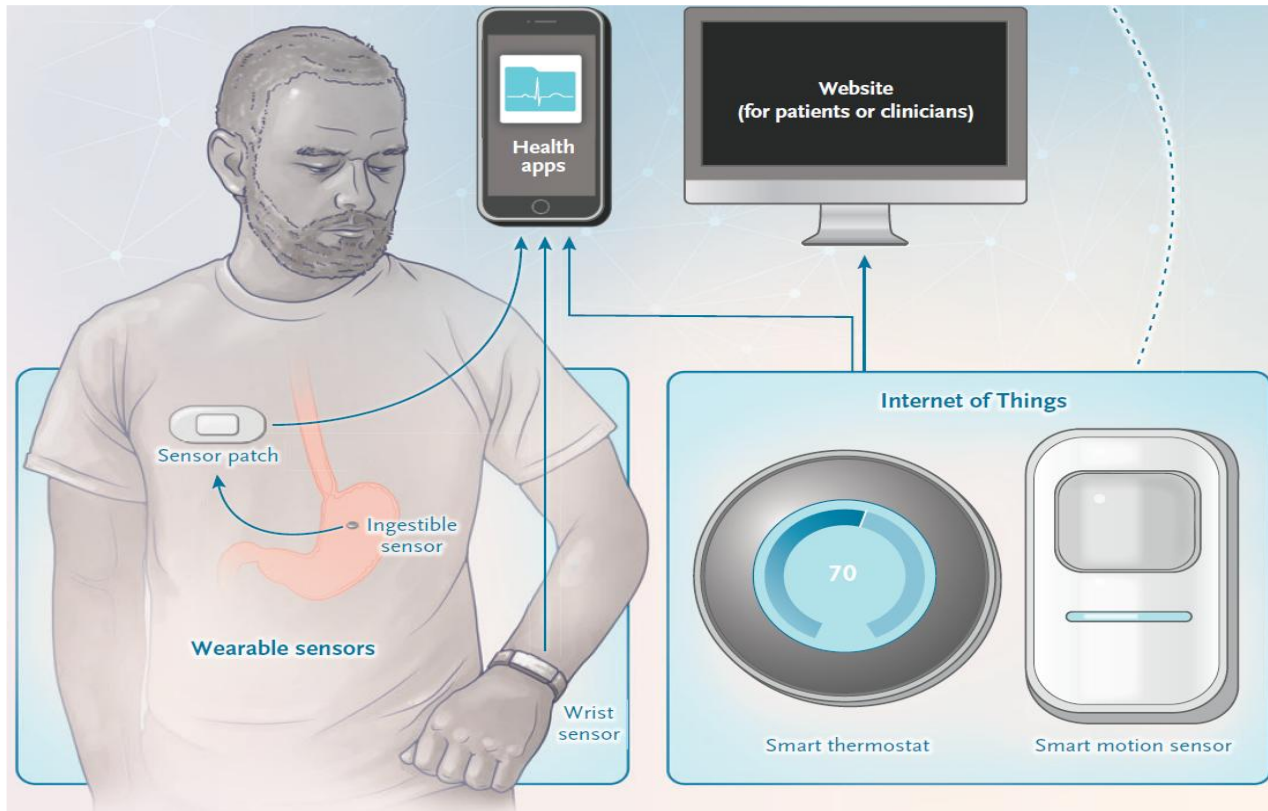
The NEW ENGLAND JOURNAL of MEDICINE

REVIEW ARTICLE

FRONTIERS IN MEDICINE

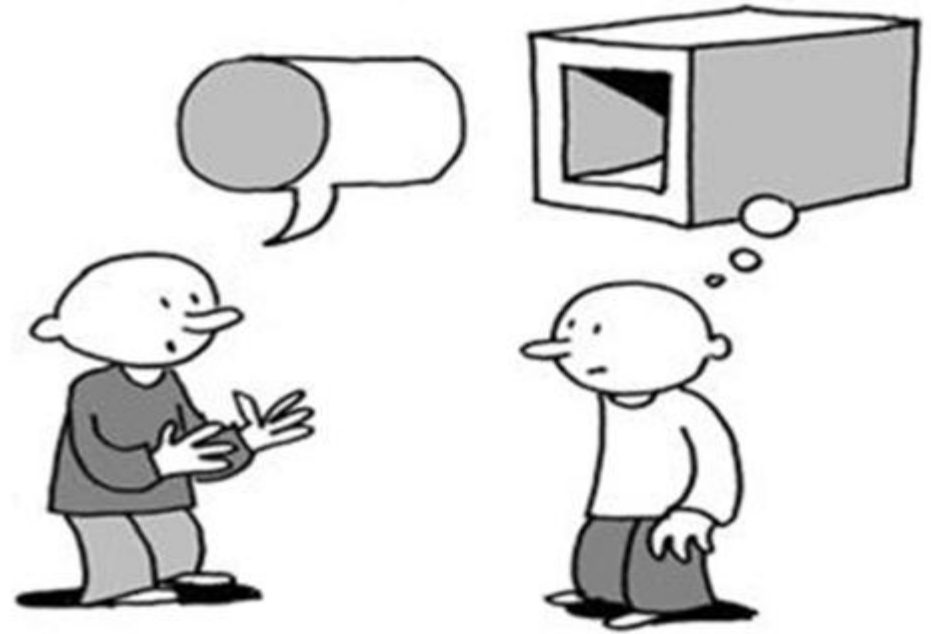
Mobile Devices and Health


Ida Sim, M.D., Ph.D.



“De patiënte is bang, haar mama is overleden aan kanker toen ze nog jong was, ze kan haar medicatie niet betalen, en ze heeft een tienerkind dat worstelt met autisme. Kan er mij iemand zeggen hoe ik dit verhaal ingevoerd kan krijgen in het EPD door zoals een robot velden aan te klikken?”

(Robert Wachter, How Artificial Intelligence Could Transform Health Care)



© MARGOET DE HEER 

Reduction

We gebruiken enkel wat meetbaar (digitaliseerbaar) is
Enkel wat meetbaar is, is relevant
Wat niet meetbaar is, is niet (meer) relevant

The art of medicine

Vulnerability and ethics: considering our Cartesian hangover

Experiences With Ambient Scribe Technology With Documentation Burden and Efficiency

n, MBA, Julietta Gervase, BA, Anna Schoenbaum, DNP, MS, RN-BC; William Hanson, MD, John T. Howell III, MD; ; MD; Kevin B. Johnson, MD, MS

y
ng of an
nurse



Which one of those measures “quality” of a medical note?

Table 2. Results of Random-Effects Model for Outcomes Before vs After Intervention^a

Outcome	Ambient scribing intervention period		Estimate, β (SE) [95% CI]	Change, %	P value
	Before	After			
Time in notes per appointment, min, mean (SD)	10.3 (9.2)	8.2 (6.8)	-1.6 (0.4) [-2.3 to -0.8]	-15.2	<.001
Appointments closed same day, mean (SD), %	66.2 (31.0)	72.4 (29.4)	6.0 (1.6) [2.9 to 9.2]	9.1	<.001
After-hours work time per scheduled day (pajama time), mean (SD), min	50.6 (64.3)	35.4 (52.0)	-8.6 (3.6) [-15.7 to -1.4]	-16.9	.02
Portion of notes manually typed, mean (SD), %	11.2 (9.7)	7.9 (6.5)	-2.3 (0.4) [-3.1 to -1.5]	-20.7	<.001
Total length of all notes generated by clinician, mean (SD), No. of characters/wk	202 637.5 (186 947.7)	244 427.1 (199 403.6)	31 390.0 (5379.3) [20 846.7 to 41 933.2]	15.5	<.001

Clinician Experiences With Ambient Scribe Technology to Assist With Documentation Burden and Efficiency

Matthew J. Duggan, MBA, Julietta Gervase, BA, Anna Schoenbaum, DNP, MS, RN-BC, William Hanson, MD, John T. Howell III, MD, Michael Sheinberg, MD, Kevin B. Johnson, MD, MS

DESIGN, SETTING, AND PARTICIPANTS This prospective improvement study was conducted between April and June 2023 at an academic health system in Philadelphia, Pennsylvania, involving 10 primary care practitioners, and physician assistants. Data were analyzed from April to June 2023.

Quality of generated notes	+	"I did not have to type any history. I quickly became comfortable that it would capture all critical elements of the conversation."
	+	"The subjective portion is my most used feature of DAX. The subjective portion was correct 80%-85% of the time."
	-	"It tries to paraphrase the conversation, and often does it in a way that utilizes layman's terms rather than medical terms; and often incorrectly documents what was discussed. This means that I must edit the content substantially because it cannot be used as-is in my closed note."
	-	"The way DAX transcribes the HPI is still quite wordy and unnatural in my opinion... I find that I have to spend a good deal of time editing what DAX writes in order to suit my style."
	-	"The language is still odd, too formal/professorial. Unfortunately, the HPI can be quite inaccurate at times and it is necessary to review it almost immediately to make corrections while it is still fresh in memory."
	-	"It is quite inconsistent, some long conversations are whittled down to 1 or 2 short sentences. I do not think it is enough documentation for medico-legal purposes."
	-	"Unfortunately, the output required a lot of modification (re-organization of HPI/ROS, deleting redundant and extraneous information, changing some language that is not generally utilized in medicine, and placing the results in the right section)."
Usability of ambient scribe tool (DAX Copilot), now and in the future	+	"I legitimately think this technology, once optimized, is the biggest advancement for outpatient primary care providers in decades."
	+	"I am confident that future updates will be even more helpful."
	+	"I anticipate [AI documentation] improvements which will improve the output and decrease proofreading burden as the technology continues to improve."
	+/-	"Used with appropriate expectations, it is useful today. It is not yet a 'keyboard-free' experience, nor a significant time saver."
	-	"I feel that AI can achieve the goals set for here, and that there are softwares that are ready to be deployed for this purpose, but DAX is not one of them. Would not recommend for use system-wide."
	-	"I feel like for OB it won't be as useful—these are templated visits, and the biggest charting burden for these is the problem lists, which this AI doesn't solve."

Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine

Peter Lee, Ph.D., Sebastien Bubeck, Ph.D., and Joseph Petro, M.S., M.Eng.

MEDICAL NOTE TAKING

To decrease “err

the software would produce the note. GPT-4 can produce notes in several well-known formats, such as SOAP (subjective, objective, assessment, and plan), and can include appropriate billing

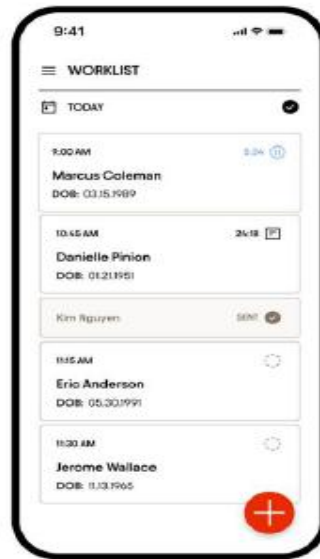
To increase efficiency:

- so physicians can spend more time with the patient rather than doing administration?
- So physicians can see more patients in the same time?

codes automatically. Beyond the note, GPT-4 can

Enhancing clinical documentation with ambient artificial intelligence: a quality improvement survey assessing clinician perspectives on work burden, burnout, and job satisfaction

Michael Albrecht, MD^{*-1}, Denton Shanks, DO, MPH², Tina Shah, MD, MPH³, Taina Hudson, DO¹, Jeffrey Thompson, PhD⁴, Tanya Filardi, MD⁵, Kelli Wright, MA⁶, Gregory A. Ator, MD⁷, Timothy Ryan Smith, MD⁸



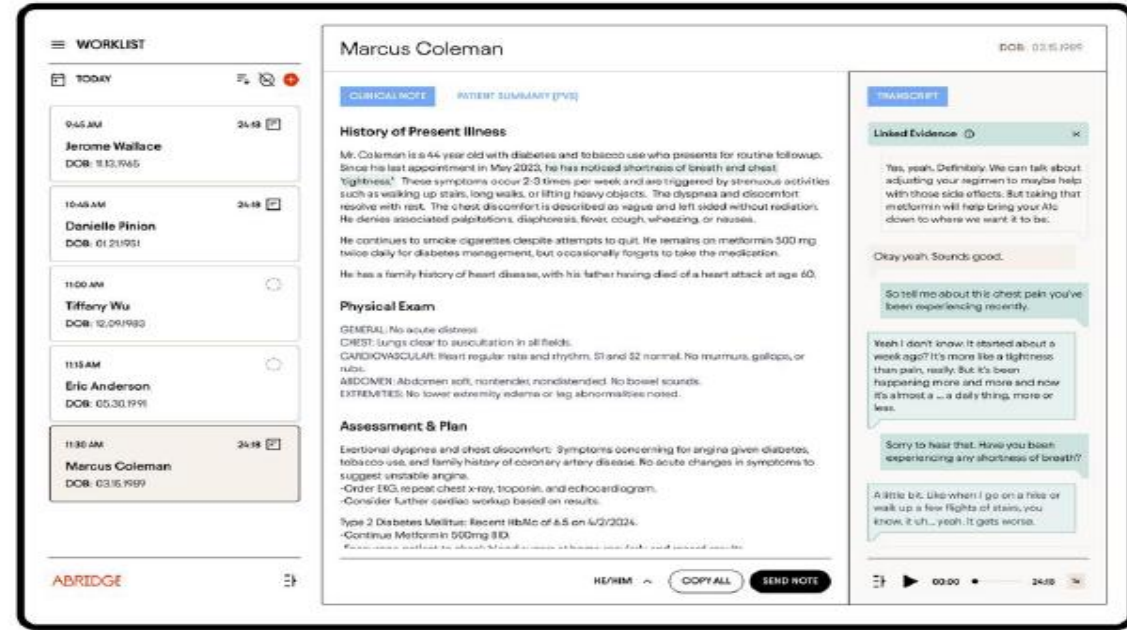
STEP 1

Select a patient from a schedule within the Abridge App.



STEP 2

Record the patient encounter. A draft note will be created with an associated transcript and audio.



STEP 3

Review the draft note. Use Linked Evidence to map any part of the summary to the source audio/transcript. Send the note directly to the EHR by clicking "Send Note."

STEP 4

Parts of the edited note are pulled into the EHR using dot phrases. For example, .hpi:ec will pull the HPI section into the EHR note.

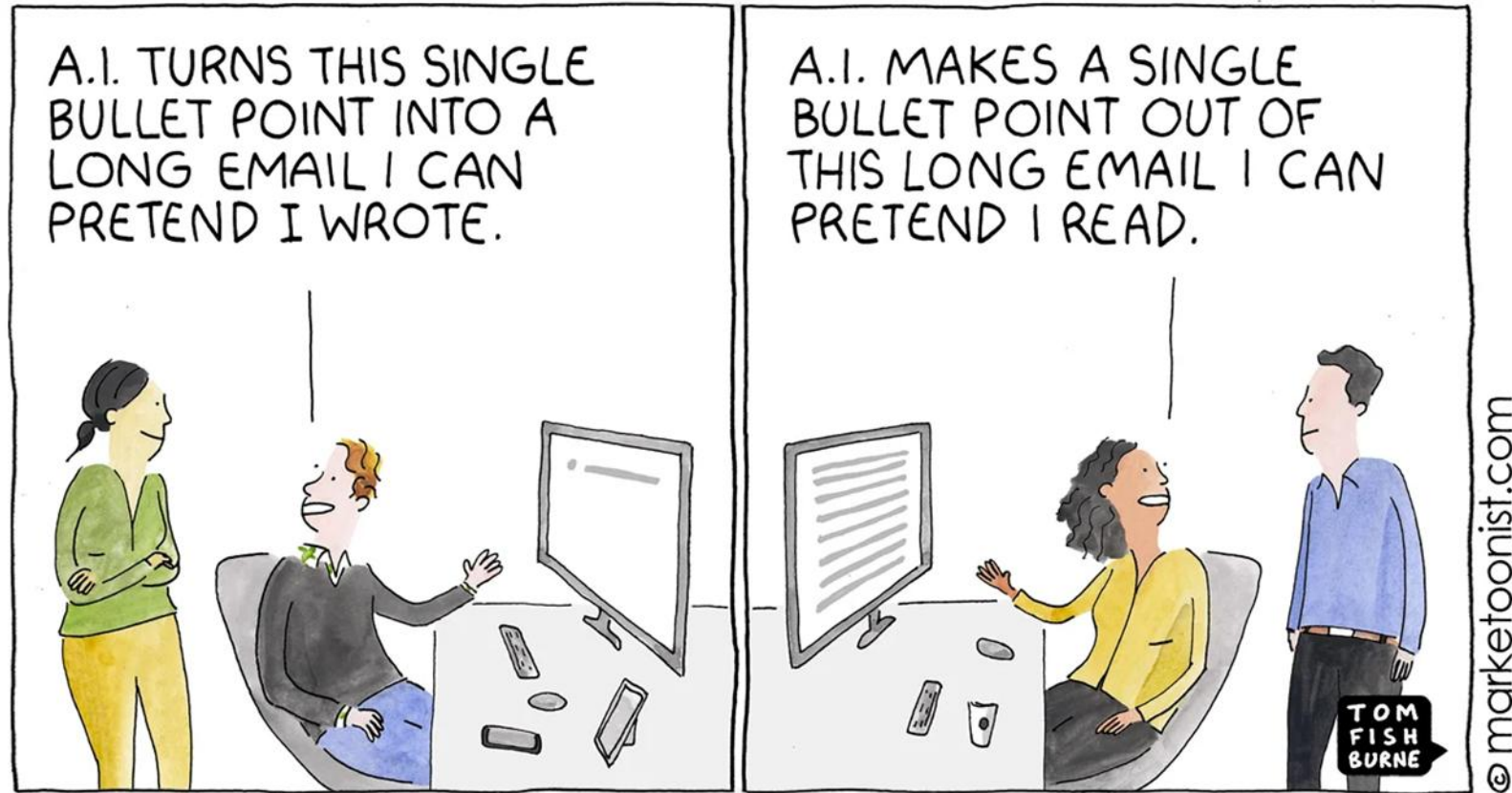
Figure 1. Sample workflow using Abridge platform. While full integration from Epic mobile app (Haiku) directly to notes in Epic is available, depicted here is the partially integrated workflow which was used during the study period. Figure courtesy of Abridge.

Enhancing clinical documentation with ambient artificial intelligence: a quality improvement survey assessing clinician perspectives on work burden, burnout, and job satisfaction

Michael Albrecht, MD^{*1}, Denton Shanks, DO, MPH², Tina Shah, MD, MPH³, Taina Hudson, DO¹,

AI takes notes

AI writes



ed on AI CDS

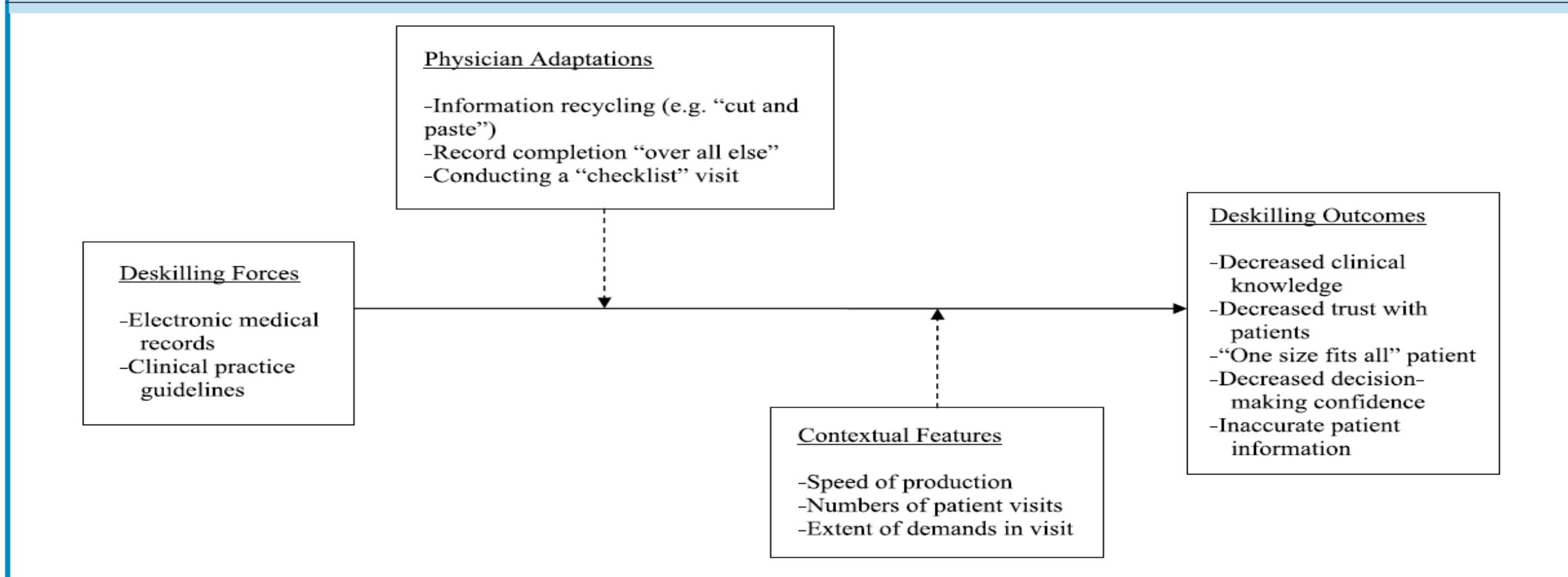
Percentage of responses

Deskilling and adaptation among primary care physicians using two work innovations

Timothy Hoff

Figure 1

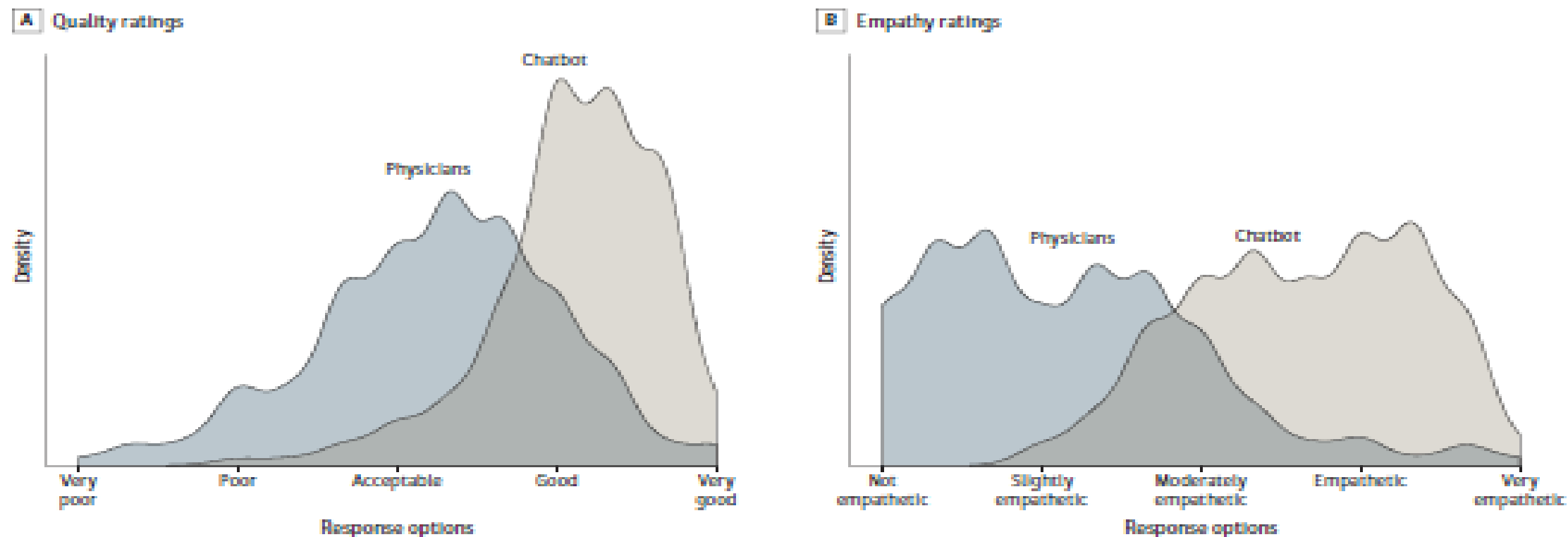
A speculative model of physician use of innovations drawn from study vignettes



Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum

John W. Ayers, PhD, MA; Adam Pollak, PhD; Mark Dredze, PhD; Eric C. Leas, PhD, MPH; Zechariah Zhu, BS; Jessica B. Kelley, MSN; Dennis J. Falx, MD; Aaron M. Goodman, MD; Christopher A. Longhurst, MD, MS; Michael Hogarth, MD; Davey M. Smith, MD, MAS

Figure. Distribution of Average Quality and Empathy Ratings for Chatbot and Physician Responses to Patient Questions



Kernel density plots are shown for the average across 3 independent licensed health care professional evaluators using principles of crowd evaluation. A, The overall quality metric is shown. B, The overall empathy metric is shown.

Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum

John W. Ayers, PhD, MA; Adam Pollak, PhD; Mark Dredze, PhD; Eric C. Leas, PhD, MPH; Zechariah Zhu, BS; Jessica B. Kelley, MSN; Dennis J. Falx, MD; Aaron M. Goodman, MD; Christopher A. Longhurst, MD, MS; Michael Hogarth, MD; Davey M. Smith, MD, MAS

Figure. Distribution of Average Quality and Empathy Ratings for Chatbot and Physician Responses to Patient Questions

A Quality ratings

B Empathy ratings

- Random selection of questions posted on social media
- Selection of questions answered by a physician (not THE physician of the patient)
- As compared to response of CHATbot prompted by the original question
- Evaluation by 3 selected health care practitioners



Kernel density plots are shown for the average across 3 independent licensed health care professional evaluators using principles of crowd evaluation. A, The overall quality metric is shown. B, The overall empathy metric is shown.

Reliability of LLMs as medical assistants for the general public: a randomized preregistered study

a

Three doctors collectively wrote ten detailed medical scenarios



Specific case details

You are playing the part of a 20-year-old male patient who is suddenly experiencing a really severe headache. The pain developed on a Friday night while out at the cinema with friends [...].

General life details

You are a university student studying for a degree in electrical engineering [...].

Additional medical history

You do not have any long-term health conditions, nor family history of any health conditions [...].

b

Gold standard answers combine disposition (initial three doctors) with plausible differential diagnoses (four additional doctors)



Disposition

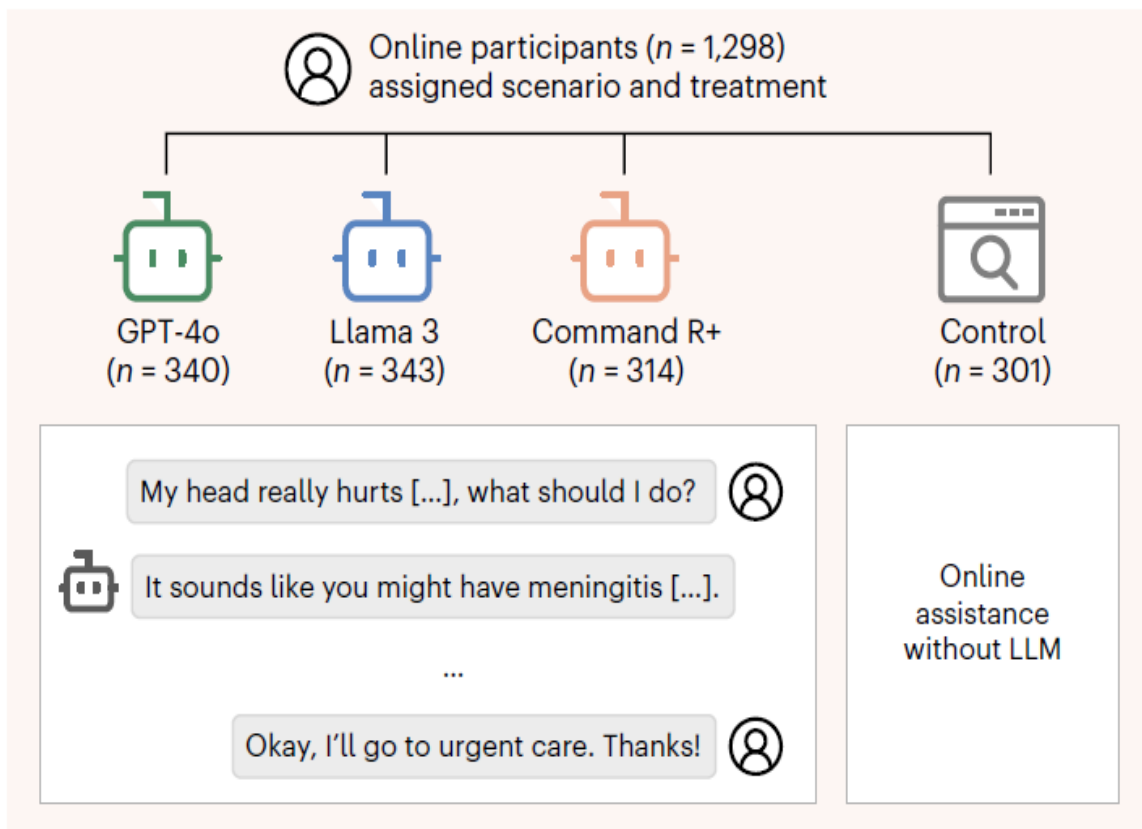
- Ambulance
- A&E
- Urgent Primary Care
- Routine GP
- Self-care

Relevant conditions

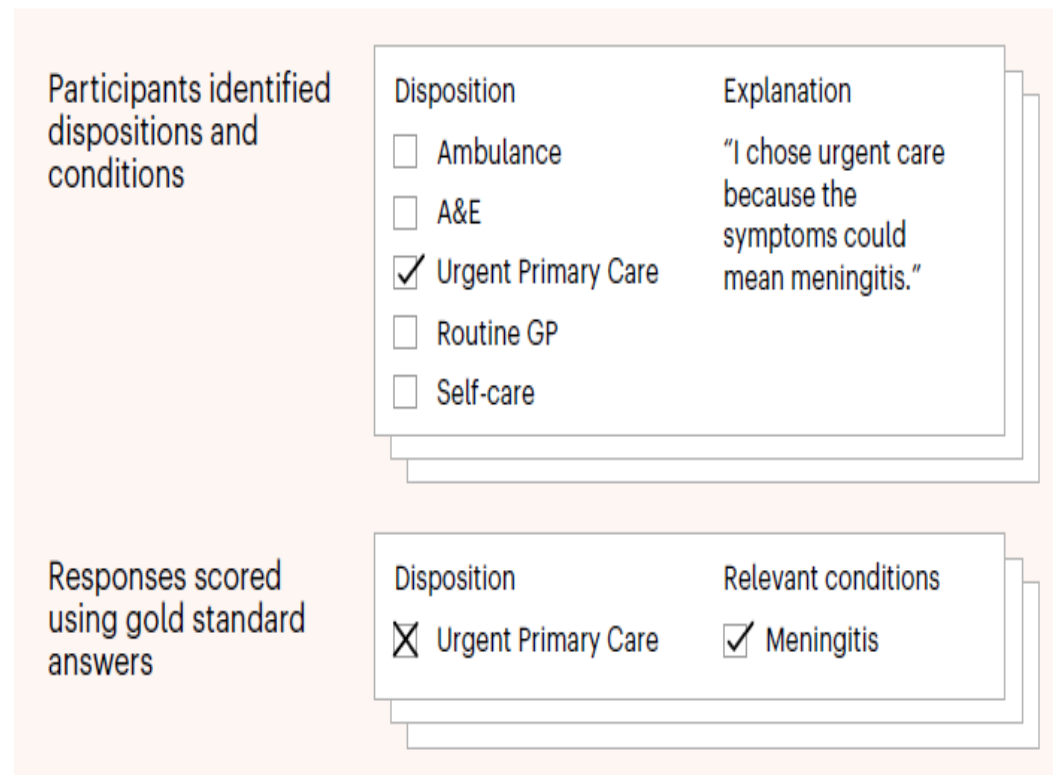
- Subarachnoid hemorrhage
- Meningitis
- Cluster headache
- ...

Reliability of LLMs as medical assistants for the general public: a randomized preregistered study

c



d



Extended Data Table 2 | Interaction Transcript: Inconsistent Responses

Participant 1	Participant 2
User: I have a terrible headache and my neck is stiff so I can't look down. In addition to that the light hurts my eyes.	User: I have suddenly developed the worst headache ever. It came on suddenly. I can't concentrate and my neck is stiff. Also, I cannot stand the light. What action should I take?
Model: I'm sorry to hear that. You might be experiencing a migraine or tension headache. Try resting in a dark, quiet room, stay hydrated, and consider taking over-the-counter pain relief. If symptoms persist or worsen, please consult a healthcare professional.	Model: Seek immediate medical attention. Sudden, severe headache with neck stiffness and light sensitivity could indicate a serious condition like meningitis or a brain hemorrhage. Go to the emergency room or call emergency services.

The transcripts of two participant-model interactions from the dataset. In both of these conversations, the model is GPT-4o and the scenario is 'subarachnoid haemorrhage'. Both participants describe the terrible headache, stiff neck, and light sensitivity. Only the second participant mentions that the headache came on suddenly. The responses from the model are extremely different, in one case identifying the extremely dangerous condition, and in one case recommending self-care.

How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals

A comprehensive overview of medical AI devices approved by the US Food and Drug Administration sheds new light on limitations of the evaluation process that can mask vulnerabilities of devices when they are deployed on patients.

Eric Wu, Kevin Wu, Roxana Daneshjou, David Ouyang, Daniel E. Ho and James Zou

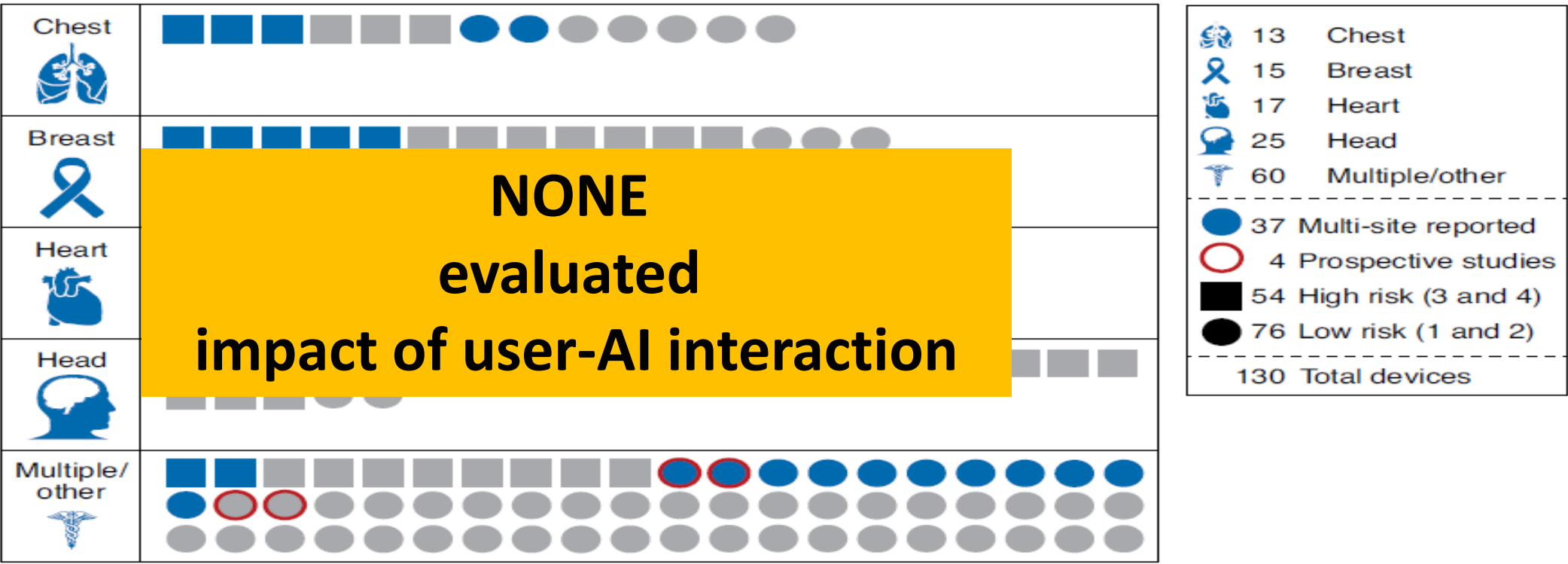


Fig. 1 | Breakdown of 130 FDA-approved medical AI devices by body area. Devices are categorized by risk level (square, high risk; circle, low risk). Blue indicates that a multi-site evaluation was reported; otherwise, symbols are gray. Red outline indicates a prospective study (key, right margin). Numbers in key indicate the number of devices with each characteristic.

Taylorism

(scientific management)

Standardisation (reduction of variability)

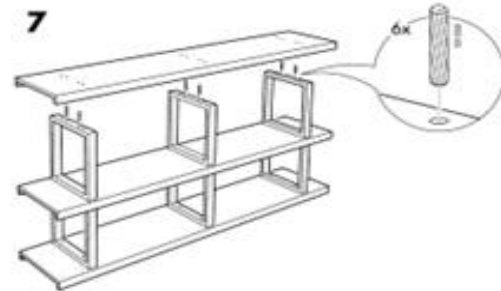
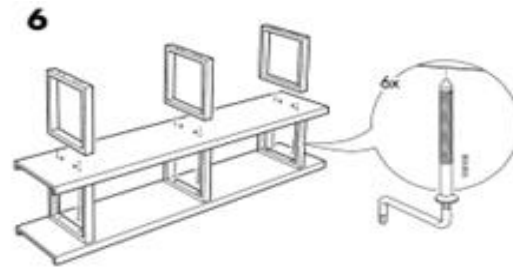
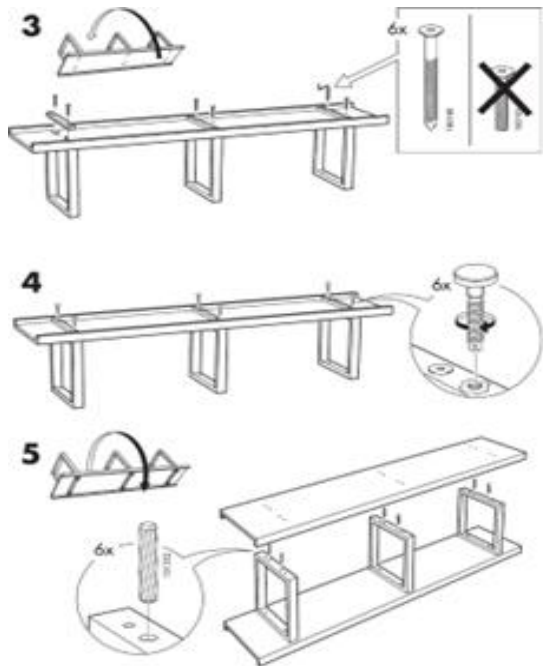
“one best way” to do things

Strict separation between “thinkers” and “do-ers”

Taylorism

(scientific management)

CONTROL



Taylorism

(scientific management)

“advantages”

- **Increase productivity**
- **Increase and standardize quality**
- **Increase profit**
- **No need for experts => exchangeability**

“disadvantages”

- **Deskilling**
- **Reduce variability and flexibility**
- **Decrease meaning of work**

HOOFDSTUK 12.

AI als transformator van geneeskunde: Een pleidooi voor gezamenlijke betrokkenheid

Wim Van Biesen en Sigrid Sterckx

“And it dawned on me that I might have to change my inner thought patterns... that I would have to start believing in possibilities that I wouldn't have allowed before ...”

(Bob Dylan)

Leidraad voor kwalitatieve en duurzame implementatie van AI in een zorgcontext

