AI evidence for use in clinical practice

The good, the bad and the ugly – and what now for the EBHC community?

Per Olav Vandvik
CEO, MAGIC Evidence Ecosystem Foundation, Oslo
Professor, Faculty of Medicine, University of Oslo
Acting consultant, Department of Medicine, Lovisenberg Diaconal Hospital Trust,
Senior researcher, Norwegian Institute for Public Health

Declarations of interest: CEO of MAGIC, no AI expert (and perhaps an EBM dinosaur). No financial COI
Agenda

- Perspective of my talk*
  - Advances in EBHC and role of guidelines
  - An encounter with AI, what is it and why the fuzz?
  - Where does AI fit in the evidence ecosystem?

- AI evidence for use in clinical practice
  - The good, the bad and the ugly
  - Pandemic lessons learned: living AI evidence?
  - Making the evidence ecosystem work for AI

- What now for the EBHC community?
  - Critical appraisal of AI evidence; Ready?
  - AI in EBHC education; Keeping up?
  - A call for urgent engagement and collaboration

*Slides will be shared, including a glossary of AI terms and links to useful podcasts and papers
Perspective of my talk: Use of AI evidence in clinical practice

Great advances in standards, methods, tools and processes in the evidence ecosystem

EBHC shift towards use of pre-appraised evidence and tools: guidelines are critical

Declarations of interest: CEO of MAGIC, no AI expert (and perhaps an EBM dinosaur). No financial COI
Meet myself, with bodily symptoms and concerns spring 2023

Asked [www.healthily.com](http://www.healthily.com) offering AI platform for self-care in the NHS, keen on Norway

Response after 10 minutes of queries:
”You may have pancreatic cancer or sacroileitis. You should see a doctor within 2 days!”

My GP ordered lots of blood tests, MRI of my back + colonoscopy
I am still a bit shaken; Is this a case of the ugly AI?
Colonoscopy did not detect any disease (I still have IBS)

Should I have met AI Genius? Computer Assisted Detection (CADe) of polyps is popular now.

My gastroenterologist apologized: “This AI device simply too expensive...”

How can clinicians, citizens and patients make well-informed decisions, based on AI evidence (aligned with EBHC principles)?
Breakthrough for AI 2023: the Large Language Models (LLM)*
How many have ChatGPT in your pockets? 180 million users, fastest growing app ever
Can it answer my overarching question today? It takes 1 minute, I have spent 6 weeks

*LLM: A type of Natural Language Processing (NLP) model comprising large neural networks trained over large amounts of text, usually to output continuations of texts from corresponding prefixes.+
What is Artificial Intelligence (AI) and how is it progressing?
The science of developing computer systems that can perform complex
tasks approximating human intelligence.
Where and how will AI enhance the evidence ecosystem?

AI acceleration across steps (a topic for another day)

Explosion of AI models, products & services, already implemented in practice

• Accelerated SRs
  AI knowledge products

• Accelerated R&D
  • AI-based diagnostics, prediction, devices

• Accelerated program evaluation
  • AI-based evaluation data analytics

• Accelerated implementation
  • AI-based personalized decision support

• Accelerated dissemination
  • AI decision support products

• Accelerated guidance & HTA
  • Guidance & HTA of AI-based products

Agenda

• Setting the scene
  • Progress in EBHC and guidelines; clinical practice perspective
  • What is AI and why the fuzz?
  • Where does AI fit in the evidence ecosystem?

• AI evidence for use in clinical practice
  • Examples of the good, the bad and the ugly
  • Lessons learned in the pandemic, time for living AI evidence?
  • Does the evidence ecosystem work for AI?

• What now for the EBHC community?
  • Critical appraisal methods up for the challenge?
  • AI in EBHC education; are we keeping up?
  • A call for urgent engagement and collaboration
Use of AI evidence in clinical practice
to answer questions on diagnosis, prognosis, and treatment
Examples of the good, the bad and the ugly (from evidence producers)

*India Fights Diabetic Blindness With Help From A.I.*

A technician screening a patient at the Aravind Eye Hospital in Madurai, India. The hospital is using a Google system that relies on artificial intelligence to diagnose a retinal problem from such a scan. Atul Loke for The New York Times
AI-assisted diagnosis: The good

Excellent performance (similar to specialists) and worked real-time in Thailand national screening programme: a prospective cohort study

Diabetic retinopathy is a leading cause of preventable blindness, especially in low-income and middle-income countries (LMICs). Deep-learning systems have the potential to enhance diabetic retinopathy screenings in LMICs, yet prospective studies assessing their usability and performance are scarce.

I conducted a prospective interventional cohort study to evaluate the real-world performance and feasibility of deep-learning systems in the health-care system of Thailand. Patients with diabetes and listed on the registry, aged 18 years or older, able to have their fundus photograph taken for at least one eye, and whose data was included in the Thai registry of diabetic retinopathy programme. Patients with a previous diagnosis of diabetic macular oedema, severe non-proliferative retinopathy, or proliferative diabetic retinopathy; previous laser treatment of the retina or retinal surgery; diabetic retinopathy eye disease requiring referral to an ophthalmologist; or inability to have fundus photograph taken of both eyes for any reason were excluded. Deep-learning system-based interpretations of patient images and referral recommendations were provided in real-time. As a safety mechanism, regional retina specialist read each image. Performance of the deep-learning system (accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV)) were measured against an adjudicated reference standard.

Findings Between Dec 12, 2018, and March 29, 2020, 7940 patients were screened for inclusion. 7651 (96.3%) patients were eligible for study analysis, and 2412 (31.5%) patients were referred for diabetic retinopathy, diabetic macular oedema, ungradable images, or low visual acuity. For vision-threatening diabetic retinopathy, the deep-learning system had an accuracy of 94.7% (95% CI 93.0–96.2), sensitivity of 91.4% (87.1–95.0), and specificity of 95.4% (94.1–96.7). The retina specialist over-readers had an accuracy of 93.5 (91.7–95.0; p=0.17), a sensitivity of 84.8% (79.4–90.0; p=0.024), and specificity of 95.5% (94.1–96.7; p=0.98). The PPV for the deep-learning system was 79.2 (95% CI 73.8–84.3) compared with 75.6 (69.8–81.1) for the over-readers. The NPV for the deep-learning system was 95.5 (92.8–97.9) compared with 92.4 (89.3–95.5) for the over-readers.

Interpretation A deep-learning system can deliver real-time diabetic retinopathy detection capability similar to retina specialists in community-based screening settings. Socioenvironmental factors and workflows must be taken into consideration when implementing a deep-learning system within a large-scale screening programme in LMICs.

Machine learning: computer systems able to learn from data without following explicit instructions
Deep learning: use of multiple layers in an artificial neural network
Neural networks: a collection of connected nodes, loosely modeling neurons in a biological brain
AI ECG results shared with physicians

11,573

Positive ECG results
692 (6.0%)

Echo 343 (49.6%)

LVEF ≤ 40%
102 (4.6% of 2,222)
or (0.9% of 11,573)

No Echo 349 (51.4%)

Negative ECG results
10,881 (94.0%)

Echo 1,879 (17.2%)

No Echo 9,002 (82.8%)

Positive ECG results
664 (6.0%)

Echo 253 (38.1%)

No Echo 411 (61.9%)

Negative ECG results
10,404 (94.0%)

Echo 1,764 (17.0%)

No Echo 8,640 (83.0%)

AI ECG results not shared with physicians

11,068

Positive ECG results
664 (6.0%)

Echo 1,764 (17.0%)

No Echo 8,640 (83.0%)

Negative ECG results
10,404 (94.0%)

Echo 411 (61.9%)

No Echo 253 (38.1%)

* Sau A, Ng FS- The emerging role of artificial intelligence enabled electrocardiograms in healthcare
AI to inform treatment decisions: The ugly?
Can not reliably tell what works (relative effects of interventions)
With heterogenous effects; can AI reliably predict responses to treatment?
I found lots of studies in oncology, no systematic reviews so I asked ChatGPT

IBM’s Watson supercomputer recommended ‘unsafe and incorrect’ cancer treatments, internal documents show

The reliability of AI in predicting responses to treatment in the presence of heterogeneous effects (differences in treatment effects across individuals or subgroups) is a nuanced and multi-faceted issue. The application of AI and machine learning can indeed be powerful in understanding and predicting treatment responses, yet several challenges and considerations persist.

In conclusion, while AI has the potential to reliably predict responses to treatment considering heterogeneous effects, the success of such endeavors relies on addressing the mentioned challenges and considerations.
LLM answering clinical questions at the point of care: The good?

Very rapid progress, but still need validation in clinical settings


¹Google Research, ²DeepMind,
Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum

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

**RESULTS** Of the 195 questions and responses, evaluators preferred chatbot responses to physician responses in 78.6% (95% CI, 75.0%-81.8%) of the 585 evaluations. Mean (IQR) physician responses were significantly shorter than chatbot responses (52 [17-62] words vs 211 [168-245] words; t = 25.4; P < .001). Chatbot responses were rated of significantly higher quality than physician responses (t = 13.3; P < .001). The proportion of responses rated as *good* or *very good* quality (≥ 4), for instance, was higher for chatbot than physicians (chatbot: 78.5%, 95% CI, 72.3%-84.1%; physicians: 22.1%, 95% CI, 16.4%-28.2%). This amounted to 3.6 times higher prevalence of *good* or *very good* quality responses for the chatbot. Chatbot responses were also rated significantly more empathetic than physician responses (t = 18.9; P < .001). The proportion of responses rated *empathetic* or *very empathetic* (≥4) was higher for chatbot than for physicians (physicians: 4.6%, 95% CI, 2.1%-7.7%; chatbot: 45.1%, 95% CI, 38.5%-51.8%; physicians: 4.6%, 95% CI, 2.1%-7.7%). This amounted to 9.8 times higher prevalence of *empathetic* or *very empathetic* responses for the chatbot.

**CONCLUSIONS** In this cross-sectional study, a chatbot generated quality and empathetic responses to patient questions posed in an online forum. Further exploration of this technology is warranted in clinical settings, such as using chatbot to draft responses that physicians could then edit. Randomized trials could assess further if using AI assistants might improve responses, lower clinician burnout, and improve patient outcomes.
ChatGPT and revisiting role of trustworthy guidelines revisited

The robot can not write trustworthy guidelines
Perhaps they are optimal EBM point of care resources? Now, or in next version?

Thank you. Since I am not a US citizen, I wonder if there are other trustworthy guidelines to inform this question? Maybe even living guidelines?

The World Health Organization (WHO) has provided a living guideline, which is continually updated as new evidence emerges, regarding the use of therapeutics in the treatment of COVID-19. According to the guideline updated on March 3, 2022, there is a conditional recommendation for the use of a combination of neutralizing monoclonal antibodies (casirivimab and imdevimab) in the following...
Lessons learned during the pandemic: Living evidence

From trigger trials to updated guidelines at record speed, perfect for AI?

Risk prediction remains a challenge; living prognosis review not helped by AI

### For patients with non-severe COVID-19 at highest risk of hospitalization

<table>
<thead>
<tr>
<th>Strong recommendation for</th>
</tr>
</thead>
<tbody>
<tr>
<td>We recommend treatment with nirmatrelvir-ritonavir (strong recommendation for).</td>
</tr>
</tbody>
</table>

- See Section 6.1 for help to identify patients at highest risk.
- Several therapeutic options are available: see [decision support tool](https://bmj.com) that displays benefits and harms of nirmatrelvir-ritonavir, molnupiravir and remdesivir.
- The GDG concluded that nirmatrelvir-ritonavir represents a superior choice because it may have greater efficacy in preventing hospitalization than the alternatives; has fewer concerns with respect to harms than does molnupiravir; and is easier to administer than intravenous remdesivir and the antibodies.
- Clinicians should review all medications and not consider nirmatrelvir-ritonavir in patients with possible dangerous drug interactions (note: many drugs interact with nirmatrelvir-ritonavir).
- Fully informed shared decision-making should determine whether nirmatrelvir-ritonavir should be used in pregnant or breast-feeding women, given possible benefit and residual uncertainty regarding potential undesirable effects.
- Nirmatrelvir-ritonavir should be administered as soon as possible after onset of symptoms, ideally within 5 days.

### References

<table>
<thead>
<tr>
<th>Reference</th>
<th>Evidence</th>
<th>Recommendations</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMJ 2020: 370</td>
<td>175</td>
<td>39</td>
</tr>
</tbody>
</table>

| Research evidence (3) | Evidence to decision | Justification | Practical info | Decision Aids |

### For patients with non-severe COVID-19 at low risk of hospitalization

<table>
<thead>
<tr>
<th>Conditional recommendation against</th>
</tr>
</thead>
<tbody>
<tr>
<td>We suggest not to use treatment with nirmatrelvir-ritonavir (conditional recommendation against).</td>
</tr>
</tbody>
</table>

- In the GDG's assessment, only a minority of low-risk patients will choose to consider using nirmatrelvir-ritonavir.
- Trials on antivirals included patients with some risk factors for hospital admission, resulting in a baseline risk of 3% that the GDG applied to generate the recommendation. The risk of hospitalization is likely to be lower in the general population.
Making the evidence ecosystem loop work for AI

Moving to living guidelines for AI, exemplified by AI Genius (CADe, CADx to come)

Current standards and methods works just fine, also from Evidence to Decisions (EtD)

What should the recommendation be?

Would health care systems invest in CADe, if they were well-informed, given costs + implementation challenges?
Looking beyond individual examples for clinical application of AI
84 RCTs by now, mostly bad (if not ugly)
129 systematic reviews, mostly bad (2022): We need high-quality umbrella living SR

Randomized Controlled Trials Evaluating AI in Clinical Practice: A Scoping Evaluation

Ryan Han, Julián N. Acosta, Zahra Shakeri, John P.A. Ioannidis, Eric J. Topol, Pranav Rajpurkar

doi: https://doi.org/10.1101/2023.09.12.23295381

Review article

Artificial intelligence and its impact on the domains of universal health coverage, health emergencies and health promotion: An overview of systematic reviews

Antonio Martínez-Millana, Aida Saez-Saez, Roberto Tornero-Costa, Natasha Azzopardi-Muscat, Vicente Traver, David Novillo-Ortíz

* Instituto Universitario de Investigación de Aplicaciones de las Tecnologías de la Información y de las Comunicaciones Avanzadas (ITACA), Universitat Politécnica de València, Camino de Vera S/N, València 46022, Spain

** Division of Country Health Policies and Systems, World Health Organization, Regional Office for Europe, Copenhagen, Denmark
Agenda

• Setting the scene*
  • Advances in clinical practice guidelines and EBHC
  • What is AI and why the fuzz?
  • Where does AI fit in the evidence ecosystem?

• AI evidence for use in clinical practice
  • The good, the bad and the ugly
  • Pandemic lessons learned: living AI evidence?
  • Making the evidence ecosystem work for AI

• What now for the EBHC community?
  • Critical appraisal of AI evidence; Ready?
  • AI in EBHC education; keeping up?
  • A call for urgent engagement and collaboration
Are we ready to deal with the flood of AI publications?

Nature survey 2023: Optimisms and concerns from scientists

The number of AI publications worldwide more than doubled from 2010 to 2021, growing from 200,000 to nearly 500,000.

“The main problem is that AI is challenging our existing standards for proof and truth,” said Jeffrey Chuang, who studies image analysis of cancer at the Jackson Laboratory in Farmington, Connecticut.

**NEGATIVE IMPACTS OF AI**

Q: Considering machine-learning methods, what do you think are negative impacts of AI in research? (Choose all that apply.)

- Leads to more reliance on pattern recognition without understanding
- Results can entrench bias or discrimination in data
- Makes fraud easier
- Ill-considered use leads to irreproducible research
- Exacerbates power imbalances: only scientists at well-resourced universities or firms can be at the cutting edge
- Expensive or energy-intensive tool
LLM most impressive and most concerning for scientists
proliferation of misinformation, mistakes, fraud, and entrenched with bias

Respondents added that they were worried about faked studies, false information and perpetuating bias if AI tools for medical diagnostics were trained on historically biased data. Scientists have seen evidence of this: a team in the United States reported, for instance, that when they asked the LLM GPT-4 to suggest diagnoses and treatments for a series of clinical case studies, the answers varied depending on the patients’ race or gender (T. Zack et al. Preprint at medRxiv https://doi.org/ktdz; 2023) – probably reflecting the text that the chatbot was trained on.

Conclusion (T. Zack et al):
”Urgent need for comprehensive and transparent bias assessments of tools like GPT-4 for every intented use case before integrated into clinical care”
Can we adequately review AI papers? Any reporting standards?
Many of us probably lack skills to appraise/peer-review, checklists are emerging

When asked if journal editors and peer reviewers could adequately review papers that used AI, respondents were split. Among the scientists who used AI for their work but didn’t directly develop it, around half said they didn’t know, one-quarter thought reviews were adequate, and one-quarter thought they were not. Those who developed AI directly tended to have a more positive opinion of the editorial and review processes.

QUALITY OF AI REVIEW IN RESEARCH PAPERS
Q: Do you think that journal editors and peer-reviewers, in general, can adequately review papers in your field that use AI?

Yes □ No □ Don’t know/cannot tell

Respondents who study AI

Respondents who use AI in research

Respondents who don’t use AI in research

“Reviewers seem to lack the required skills and I see many papers that make basic mistakes in methodology, or lack even basic information to be able to reproduce the results,” says Duncan Watson-Parris, an atmospheric physicist who uses machine learning at the Scripps Institution of Oceanography in San Diego, California. The key, he says, is whether journal editors are able to find referees with enough expertise to review the studies.
Blackbox problem* and the Explainability of AI

- Lack of transparency raises challenges with bias, accountability and responsibility leading to also ethical and legal problems
- Explainable AI (XAI) aims to address these issues by developing models that are more interpretable and transparent
- **XAI has 3 main problems** (thought provoking from Dr. Ghassemi @NEJM AI Grand Rounds podcast)
  - Squishy definition
  - Too simple methods to explain, turns of critical thinking (see preprint below)
  - Medicine has lots of black boxes, we need to know well calibrated, how to use in clinical contexts
- **A key critical appraisal challenge for the EBHC community?**

* Blackbox problem: The challenge of understanding how AI systems and machine learning models operate, especially in processing data and making predictions or decisions
AI in EBHC education; are we keeping up?
Inevitable that health care professionals need to learn, why not link to EBM?
2 week elective AI course for medical students at University of Oslo lots of fun;-)

Medisinstudenter kan nå lære om kunstig intelligens, stor-data og innovasjon

KURS_KOMITEEN: (f.v) Medisinstudent Birk Hunskaar, professor Per Olav Vundvik, lege i spesialisering Isbita Barma og universitetslektor Anja Fog Heen utgjør kurs-komiteen til MED3065 – AI, innovasjon, big data og beslutningsstøtte. Foto: Anita Aalby

»All should understand how to best use AI tools, their limitations and evidence-base that surrounds them»
NEJM AI Grand Rounds Podcast
Dr. Alan Karthikesalingam
Research lead at Google
Summary: AI evidence for use in clinical practice 2023

What now for the EBHC community?

- AI will increasingly inform (and accelerate) the evidence ecosystem
- Current AI evidence mostly the bad and the ugly but likely to change rapidly
- Start dancing with AI folks, right now
- If my talk did not make sense; ask ChatGPT and be positively surprised;-)
Glossary of key terms for AI

• **AI**: The science of developing computer systems that can perform complex tasks approximating human intelligence

• **Machine learning**: computer systems able to learn from data without following explicit instructions

• **Deep learning**: use of multiple layers in an artificial neural network

• **Neural networks**: a collection of connected nodes, loosely modeling neurons in a biological brain

• **Generative AI**: Can generate text, images, or other media, using patterns of input training data

• **Natural Language Processing (NLP)**: A branch of artificial intelligence that seeks to enable computers to interpret and manipulate human text

• **Large Language Model (LLM)**: A type of NLP model comprising large neural networks trained over large amounts of text, usually to output continuations of texts from corresponding prefixes.

• **GPT**: Generative Pre-trained Transformer
Useful resources for learning more

• NEJM AI Grand Rounds