Does Artificial Thought Have a Ceiling? Lessons from Fetal Weight Prediction

Or Degany MD[1] and Itamar Ben Shitrit MD MPH[2,3]

[1]Gray Faculty of Medical and Health Sciences, Tel Aviv University, Tel Aviv, Israel
[2]Ben Gurion Faculty of Health Sciences, Beer Sheva, Israel
[3]Clinical Research Center, Soroka University Medical Center, Faculty of Health Sciences, Ben Gurion University of the Negev, Beer Sheva, Israel

KEY WORDS: artificial intelligence (AI), explainable artificial intelligence, fetal weight, prediction models, transparency

IMAJ 2026; 28: 262–264

Artificial intelligence (AI) and machine learning have moved to the forefront of scientific discourse and clinical medicine, offering improved accuracy and efficiency while raising concerns about transparency, accountability, and unintended consequences. Recent developments, particularly large-scale and generative models, have fueled these debates. However, efforts to mimic aspects of human intelligence long predate ChatGPT. These efforts range from early rule-based systems to Weizenbaum's ELIZA program, which humorously simulated a Rogerian psychotherapist in its DOCTOR script [1]. For clinicians, the real test is not whether predictions become marginally more accurate on average, but whether they improve the identification of high-risk patients and meaningfully change management.

In this issue of the *Israel Medical Association Journal* (*IMAJ*), Shomron and Yogev [2] reported on a recent multicenter study evaluating fetal weight estimation methods [3]. Using machine learning models trained on approximately 10,000 births, they tested whether these approaches could outperform the Hadlock formula. Although the models demonstrated superior statistical accuracy in birthweight prediction, this improvement did not translate into better accuracy in identifying fetuses at risk. Gains were largely confined to fetuses within the normal range, whereas performance was limited for small-for-gestational-age (SGA) and large-for-gestational-age (LGA) fetuses.

Beyond the technical challenge of fetal weight estimation, the authors raised fundamental aspects of the nature of AI: its performance at the extremes, the trade-off between sophistication and interpretability, the continued role of clinical judgement, and the question of *machine thought.*

Still, there are several open questions. Why did the system fail to improve SGA and LGA detection? Even if a definitive explanation is not possible, it is reasonable to suggest that results might differ with a substantially larger training dataset. Alternatively, the data may have included too few SGA and LGA cases to reliably learn the extremes. Last, an endpoint misalignment may have had an impact. Training models to predict birthweight as a continuous value can improve average predictions without improving discrimination at the tails, where clinical risk is concentrated.

Earlier research has suggested that AI-based algorithms cannot only improve prediction within normal ranges more reliably but can also enhance the identification of outliers at the extremes in many disciplines. For example, a recent study by Mikołaj and colleagues [4] reported that a deep learning model, trained on a dataset of more than 65,000 patients, detected both SGA and LGA significantly better than the Hadlock formula. Houri et al. [5] showed similar results for LGA fetuses. Across other domains, machine learning models demonstrated strong performance not only in improving precision in detecting normal medical conditions but also in identifying pathological outliers. These specialties included cardiology, radiology, and critical care [6-9].

Study design, data composition, and algorithmic approaches differed across the studies. Nevertheless, the broader implication is clear: the inability to outperform the Hadlock formula in SGA/LGA detection in the study discussed by Shomron and Yogev [2] should not be interpreted as a general limitation of AI, but rather as a case-specific or design-dependent outcome. However, the central message remains highly relevant: better statistical performance (e.g., lower error) does not necessarily translate into better clinical decision making.

Although more modest, this finding is important for clinical practice.

Moreover, the authors highlight the value of transparent, explainable prediction tools, using the Hadlock formula as a paradigm. In the present analysis, however, Hadlock was applied as a standalone estimator, without clinical interpretation or context. As such, the relationship between transparency and clinical utility should be interpreted cautiously, as it may extend beyond what was directly examined. Transparency remains a key advantage of such simple models, as it enhances trust and communication; however, this benefit should be distinguished from the empirical outcomes used to evaluate them.

Medical AI often forces a trade-off between explainability and predictive accuracy [10,11]. As physicians, our ethical priority is always the patient's overall good; therefore, when a less interpretable model is validated, well-calibrated, and shown to provide clinically meaningful benefit, we may need the humility to consider using it without fully understanding all aspects of it, provided that appropriate clinician oversight and ongoing monitoring are present. As Aristotle observed, human understanding of causation is inherently limited, and the ability to produce accurate and beneficial results may at times matter more than our capacity to fully explain how they are achieved [10].

Explainable AI is often proposed as a bridge between performance and trust, typically by adding post-hoc models intended to clarify how decisions are produced by otherwise opaque systems. Yet explainability is not always feasible; post-hoc explanations may not reflect a model's true decision drivers, and the existence and magnitude of an accuracy–explainability trade-off remain debated [12-15].

As the authors noted, AI technologies are known to "see patterns invisible to the human eye." Indeed, AI systems often exhibit emergent properties that cannot be explained by their individual components, as previously discussed by Sorin and Klang [16]. This concept has been widely explored. Just as complex structures built by ant colonies are not encoded in a single ant, the outputs of a computer program cannot be explained by individual bits, and human thought cannot be simply reduced to the activity of a single neural pathway.

The paradigm that machines cannot *understand* context may hold for many traditional machine learning and image-recognition systems; however, it may be less definite as a universal characteristic of AI, particularly in the era of large language models and multimodal systems [17]. The nature of *understanding* has been discussed extensively, from Hofstadter's *Gödel, Escher, Bach* [18] to a more recent perspective by Hinton (e.g., What Is Understanding? IASEAI 2025) [19].

Last, it is important to acknowledge the limitations of human judgement and understanding. Aside from the challenge to consider more than a few variables, humans are largely influenced by cognitive bias and random variability in judgement (i.e., noise), as previously discussed by Tversky and Kahneman [20]. Thus, cautious adoption should mean neither reflexive rejection nor uncritical enthusiasm, but disciplined evaluation against endpoints that matter to patients, such as outcomes, values, preferences, and equity. AI may help us to be better physicians, compensating for our inherent weaknesses.

CONCLUSIONS

Shomron and Yogev's article [2] focuses on fetal weight estimation but raises broader questions about integrating AI into medicine. At the most immediate level, the study illustrates that better statistical accuracy does not ensure improved clinical decision-making. Building on this observation, the editorial discussion highlights a recurring challenge in medical AI. Although transparency has clear advantages, the trade-off between explainability and accuracy also reveals transparency limitations. Extending further, the comparison between machine prediction and clinical judgement invites a discussion on how human understanding is defined and constrained and whether machines can achieve comparable contextual understanding. AI adoption, therefore, calls for cautious humility and awareness of human cognitive limitations, as well as a sustained focus on patient benefit, backed by strong evidence.

Correspondence

Dr. O. Degany
Gray Faculty of Medicine, Tel Aviv University, Tel Aviv 6997801, Israel
Email: ordegany@gmail.com

References

1. Weizenbaum J. ELIZA—a computer program for the study of natural language communication between man and machine. *Commun ACM* 1966; 9 (1): 36-45.
2. Shomron N, Yogev Y. Artificial intelligence does not always win. *IMAJ* 2026; 28 (4): 260-1.
3. Dor O, Ashwal E, Cohen M, et al. Machine learning versus traditional formulas for fetal weight estimation: An international multicenter study evaluating prediction accuracy across birth weight percentiles. *Int J Gynecol Obstet* 2025. 2026; 173 (1): 456-62.
4. Mikołaj KW, Christensen AN, Taksøe-Vester CA, et al. Predicting abnormal fetal growth using deep learning. *NPJ Digit Med* 2025; 8 (1): 318.

5. Houri O, Romano A, Walfisch A, et al. Machine learning-based prediction of large-for-gestational-age neonates in diabetic and non-diabetic pregnancies. *Int J Gynecol Obstet* 2025; 00: 1-10.
6. Bachtiger P, Petri CF, Scott FE, et al. Point-of-care screening for heart failure with reduced ejection fraction using artificial intelligence during ECG-enabled stethoscope examination in London, UK: a prospective, observational, multicentre study. *Lancet Digit Health* 2022; 4 (2): e117-e125.
7. Storey M, Packer J, Grandal NS, et al. Early clinical evaluation of AI triage of chest radiographs: time to diagnosis for suspected cancer and number of urgent CT referrals. *N Engl J Med AI* 2025: AIcs2500539.
8. Boussina A, Shashikumar SP, Malhotra A, et al. Impact of a deep learning sepsis prediction model on quality of care and survival. *NPJ Digit Med* 2024; 7 (1): 14, 153.
9. Rakers MM, van Buchem MM, Kucenko S, et al. Availability of evidence for predictive machine learning algorithms in primary care: a systematic review. *JAMA Netw Open* 2024; 7 (9): e2432990.
10. London AJ. Artificial intelligence and black-box medical decisions: accuracy versus explainability. *Hastings Cent Rep* 2019; 49 (1): 15-21.
11. Quer G, Arnaout R, Henne M, Arnaout R. Machine learning and the future of cardiovascular care: JACC state-of-the-art review. *J Am Coll Cardiol* 2021; 77 (3): 300-13.
12. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* 2019; 1 (5): 206-15.
13. Klauschen F, Dippel J, Keyl P, et al. Toward explainable artificial intelligence for precision pathology. *Annu Rev Pathol* 2024; 19: 541-70.
14. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health* 2021; 3 (11): e745-e750.
15. Bell A, Solano-Kamaiko I, Nov O, Stoyanovich J. It's just not that simple: an empirical study of the accuracy-explainability trade-off in machine learning for public policy. Proc ACM Conf Fairness Account Transparency 2022: 248-66).
16. Sorin V, Klang E. The emergence phenomenon in artificial intelligence: a warning sign on the path to artificial general intelligence. *IMAJ* 2024; 26 (2): 120-1.
17. Tu T, Azizi S, Driess D, et al. Towards generalist biomedical AI. *N Engl J Med AI* 2024; 1 (3): AIoa2300138.
18. Hofstadter DR. Gödel, Escher, Bach: an eternal golden braid. Basic books. 1999.
19. Hinton G. What Is Understanding? IASEAI 2025. Paris: International Association for Safe & Ethical AI [Available from https://www.youtube.com/watch?v=6fvXWG9Auyg&t=600s].
20. Tversky A, Kahneman D. Judgment under uncertainty: heuristics and biases. *Science* 1974; 185 (4157): 1124-31.

Capsule

A neuroimmune link in skin inflammation

Although psychological stress is thought to aggravate atopic dermatitis, the cellular and molecular mechanisms are not well established. **Tian** et al. performed a retrospective analysis of patients with this condition and found correlations between elevated stress and increased accumulation of eosinophils in the skin. In mice, a distinct type of sympathetic neuron that innervated the skin relayed stress responses from the brain to eosinophils and exacerbated inflammation. These neurons released a chemokine, CCL11, that recruited eosinophils. The ability of eosinophils to promote stress-induced inflammation was dependent on their expression of the adrenergic receptor β2. These results suggest that, in combination with other treatments, managing stress or blocking stress-dependent signaling between neurons and eosinophils may help to alleviate dermatitis.

Science 2026; 391: 1269

Eitan Israeli

Capsule

Predicting onset of symptomatic Alzheimer's disease with plasma p-tau217 clocks

Predicting not just if, but also when, cognitively unimpaired individuals are likely to develop the onset of Alzheimer's disease (AD) symptoms would be useful to clinical trials and, eventually, clinical practice. Although clock models based on amyloid and tau positron emission tomography have shown promise in predicting the onset of AD symptoms, a model based on plasma biomarkers would be more accessible**. Petersen** and colleagues used longitudinal plasma %p-tau217 (the ratio of phosphorylated to non-phosphorylated tau at position 217) from two independent cohorts (n=258 and n=345), clock models and estimated the age at plasma %p-tau217 positivity. The estimated age at plasma %p-tau217 positivity was associated with the age at onset of AD symptoms (adjusted R^2 of 0.337−0.612) with a median absolute error of 3.0−3.7 years. Notably, the time from %p-tau217 positivity to onset of AD symptoms was markedly shorter in older individuals. Similar models were constructed with data from one p-tau217/Aβ42 immunoassay and four plasma p-tau217 immunoassays. These findings suggest that the time until onset of AD symptoms can be estimated using a single blood test within a margin of error that is acceptable for use in clinical trials.

Nature Medicine 2026; 32: 1085

Eitan Israeli