Journal of Clinical Epidemiology 142 (2022) 252–257

Journal of Clinical Epidemiology

COMMENTARY

Believing in black boxes: machine learning for healthcare does not need explainability to be evidence-based

Liam G. McCoy[a,*], Connor T.A. Brenna[b,c], Stacy S. Chen[c,d], Karina Vold[e,f,g,h], Sunit Das[g,i]

[a] *Temerty Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada*
[b] *Department of Anesthesiology & Pain Medicine, University of Toronto, Toronto, Ontario, Canada*
[c] *Department of Philosophy, University of Toronto, Toronto, Ontario, Canada*
[d] *Joint Centre for Bioethics, University of Toronto, Toronto, Ontario, Canada*
[e] *Institute for the History and Philosophy of Science and Technology, University of Toronto, Toronto, Ontario, Canada*
[f] *Schwartz Reisman Institute for Technology and Society, University of Toronto, Toronto, Ontario, Canada*
[g] *Centre for Ethics, University of Toronto, Toronto, Ontario, Canada*
[h] *Leverhulme Centre for the Future of Intelligence, University of Cambridge, Cambridge, United Kingdom*
[i] *Division of Neurosurgery, University of Toronto, Toronto, Ontario, Canada*

Accepted 1 November 2021; Available online 5 November 2021

Abstract

Objective: To examine the role of explainability in machine learning for healthcare (MLHC), and its necessity and significance with respect to effective and ethical MLHC application.

Study Design and Setting: This commentary engages with the growing and dynamic corpus of literature on the use of MLHC and artificial intelligence (AI) in medicine, which provide the context for a focused narrative review of arguments presented in favour of and opposition to explainability in MLHC.

Results: We find that concerns regarding explainability are not limited to MLHC, but rather extend to numerous well-validated treatment interventions as well as to human clinical judgment itself. We examine the role of evidence-based medicine in evaluating inexplicable treatments and technologies, and highlight the analogy between the concept of explainability in MLHC and the related concept of mechanistic reasoning in evidence-based medicine.

Conclusion: Ultimately, we conclude that the value of explainability in MLHC is not intrinsic, but is instead instrumental to achieving greater imperatives such as performance and trust. We caution against the uncompromising pursuit of explainability, and advocate instead for the development of robust empirical methods to successfully evaluate increasingly inexplicable algorithmic systems.

Keywords: Machine Learning; Explainability; Evidence-Based Medicine; Mechanistic Reasoning; Algorithms; Artificial Intelligence

1. Introduction

The incredible growth of the machine learning for healthcare (MLHC) field has spurred both optimism and concern in the medical community. Proponents are drawn to the prospect of rapid algorithmic analysis of patient data on a massive scale enabling cheaper, more efficient, and more accurate medical care. Purported uses include both those which overlap with physician expertise (such as reading CT scans with similar accuracy to a radiologist [1]), and those which will extend beyond it (such as performing analyses of voice waveform data to predict the onset of dementia [2]). A number of machine learning techniques central to this recent progress, such as "deep learning", involve the analysis of large amounts of data to generate outputs based on highly complex sets of and interactions between data features [3]. Such techniques have been characterized as "black boxes" [4] as they can perform with a high degree of empirical accuracy while being unable to indicate to human observers specifically *why* and *how* a specific output has been attained.

This "black box" inexplicability has led to significant disquiet amongst critics of MLHC – broadly divided into the categories of performance-related (e.g., inexplicable MLHC may be ineffective, or may be unable to adapt

Conflict of interest: Dr. Sunit Das declares roles as: Advisory board for Subcortical Surgery Group and Xpan Medical. Consultant for Medexus. Speaker's Bureau for Congress of Neurological Surgeons, American Association of Neurological Surgery, Society for NeuroOncology. Grant funding from CIHR, Medicenna, Alkermes. Clinical trialist for Agios. The other authors declare no potential conflicts of interest.

* Corresponding author.

https://doi.org/10.1016/j.jclinepi.2021.11.001
0895-4356/

What is new?

Key findings:

- Explainability is instrumental to important aims in machine learning for healthcare (MLHC), but a closer analysis in its full context reveals that explanability cannot be considered intrinsically essential to MLHC.

What this adds to what is known?

- This work explores the connection between explainability in MLHC and its analogue of mechanistic reasoning in evidence-based medicine.

What is the implication, what should change now?

- MLHC should not be held to a unique standard of explainability beyond that required of other medical interventions.
- Established tools used to validate evidence-based medicine should be adapted to enable empirical evaluation of even completely inexplicable MLHC technologies.

to changes in clinical circumstances) and ethical (e.g., inexplicable MLHC may interfere with a patient's right to understand their medical care) concerns. Explainability is commonly cited as an important principle in MLHC guidelines [5,6], with some authors going so far as to identify explainability as an essential prerequisite for MLHC models to be ethical and effective in clinical practice[7]. Subsequently, the pursuit of explainability has become an area of significant importance in contemporary MLHC research, emphasized by academics in computer science [4,8–12] medicine[12–14], and bioethics [7], as well as major regulatory bodies[15] and healthcare technology companies [16].

Within this paper we explore these concerns, the practical and ethical reasoning behind them, and the potential trade-offs that arise between explainability and other desired characteristics of MLHC such as accuracy or efficiency. We seek to situate concerns about explainability in MLHC within the context of concerns about explainability in medicine more broadly, and highlight that the challenges posed to explicability by deep learning technology are far from unique. Rather, we find that medicine has a long and ongoing history of harnessing technologies (in a broad sense, including pharmaceuticals, procedures, and diagnostic aids) for which physicians lack clear mechanistic explanations. In other words, medicine has not traditionally held "mechanistic reasoning" (explanations forged by way of inference from proposed mechanisms or physiologic rationale) as strictly necessary for the implementation of a new clinical tool. We turn to an examination of the practice of evidence-based medicine (EBM), which includes the development of tools designed to allow empirical validation of even inexplicable interventions and thereby privileges empirical evidence above mechanistic proof. Ultimately, we conclude that, while explainability confers a positive value to MLHC, this value is instrumental rather than intrinsic. We caution against the single-minded pursuit of explainability at the expense of the other opportunities presented by MLHC, and advocate for a moderated approach which relies on empiricism to guide the validation of new technologies in the same way that it already validates new medicines and procedures.

2. The desire for explainability in MLHC

Proponents of explainability argue that systemic opacity will interfere with the ability of human oversight to identify and address errors arising from naive misinterpretation of data without contextual clinical understanding [7,10]. A prototypical example of this is an MLHC model which erroneously identified asthma as a protective factor against pneumonia severity, when in reality the "protective effect" was a manifestation of the aggressive use of intensive care for asthmatic patients [17]. Similar concerns arise regarding the inability of models to properly identify and highlight findings which are rare but highly consequential, such as an aggressive subtype of lung cancer on radiographic imaging [18]. In light of these failures, explainable MLHC may be expected to facilitate a sort of partnership between the physician and the algorithm, with human oversight preventing errors and inaccuracies.

The pursuit of explainability can, however, exert cost on performance. Extracting information from models which may have millions of parameters and presenting this information in a way understandable to the human mind is an inherently reductive process [19]. Trade-offs may arise between accuracy and explainability [20,21], as reducing opacity may motivate the use of more simplistic models, or the evaluation of smaller and more comprehensible pools of data. However, MLHC models are valued precisely because they have the potential to process information in ways—and at speeds—that are impossible for human brains to match. Even if such insight is achieved, relying upon human oversight into the algorithmic prediction system raises a broad new range of human factors challenges [22]. Explainability is therefore not costless, and from the perspective of MLHC performance it must be viewed as one among several means rather than an ultimate end in itself.

There are also significant concerns surrounding the ability of inexplicable models to fulfil the ethical duties of medicine and achieve the trust of patients and providers; informed by the sense that the frameworks established for artificial intelligence (AI) should echo the principles and standards to which physicians are held [23]. Guidelines for ethical production and the use of artificial intelligence systems, such as the one published by the European Commission, require AI to uphold principles of explica-

bility, prevention of harm, fairness, and human autonomy [24]. These overarching requirements parallel the *prima facie* principles of medical practice and biomedical ethics: beneficence, non-maleficence, justice, and autonomy [25].

From this perspective, it becomes clear that the desire for explainability in AI parallels the desire for transparency from human physicians. The expectation in both cases is that the patient will be provided with adequate and coherent information needed to make informed choices which result in predictable and ideal outcomes [26]. From the provider perspective, clinicians have identified explainability as an important element of transparency, and a requisite to justify their reliance upon a given MLHC model[27]—particularly in the context of informed consent [28], legal liability concerns, and a lack of regulatory clarity [29].

Patients and providers alike seek the assurance that an MLHC algorithm is making appropriate recommendations, and explainability provides a way to better understand, engage with, or control MLHC. Nonetheless, we propose that it is difficult to argue that an interest in explainability is intrinsically morally valuable, let alone foundational to ethical acceptance of a new technology given that trust in predecessor technologies in healthcare has been established through empirical means, typically without the clarity of understanding mechanisms. Just as in the case of performance, explainability is an instrumental means of establishing and maintaining trust and control, but is not a critical end in and of itself.

3. The reality of (Un)explainability in healthcare

In discussions around the need for explainability in MLHC, such as Char and colleagues' declaration that "ML systems in medicine must have an explainable architecture, designed to align with human cognitive decision-making processes" [7], there is irony in recognizing that these human cognitive processes themselves often escape explainability [20,30–32]. One sees this empirically in the finding that, in contrast with the rules-based approaches of beginners, the inferences of expert radiologists tend to be so holistic that their underlying reasoning cannot be explained in natural language [33,34]. What are presented as 'explanations' in such cases may in fact be post-hoc rationalizations. Indeed, it has been argued that every diagnosis is, to some degree, a "clinical diagnosis"—relying on subtleties of clinical judgment which extend beyond a clearly explainable or rules-based framework[35].

Beyond the cognitive domain, inexplicability is also pervasive in the various tools at the clinician's disposal. From acetaminophen [36] to metformin [37], or many antidepressants and mood stabilizers [38], numerous medications are prescribed regularly—and to great effect—despite the fact that their mechanisms of action are partially or even entirely unclear. Even certain surgical procedures, such as gastric bypass for obesity [39–41], are performed despite their mechanisms of action not being fully understood. This lack of explainability does not, however, diminish the utility of such therapies.

Ultimately, the primary goals of medicine are pragmatic: to relieve suffering and promote health [6]. The elucidation of mechanisms comes secondary to this goal, insofar as understanding may enable better intervention, may support informed consent, or may provide greater comfort to patients and families. Some have, in fact, contested the very concept of explainability in medicine: for this reason, explainability lacks a consensus definition within the medical context [42]. Instead, modern medicine has constructed a scientific edifice to evaluate the pragmatic impact of interventions whose mechanisms are not fully understood [43]—the robust infrastructure of EBM, which privileges outcomes as its measure of success.

4. Empirically validating the inexplicable

The history of tools used in medical practice which have resisted mechanistic explanation reaches its apotheosis in the paradigm of EBM which, in exploring the epistemological tensions surrounding these longstanding therapies, has already precipitated the development of a relevant concept: mechanistic reasoning. Mechanistic reasoning refers to the inferential use of a proposed process of action to rationalize the expectation of a given outcome, which is broadly analogous to the role invoked for explainability in the MLHC context [44] (though indeed it is worth noting that, in both contexts, this reasoning is often performed *a posteriori* working backwards from an observed result).

At its core, EBM revolves around using the best available evidence (e.g., randomized controlled trials rather than anecdotes), in combination with individual clinical expertise and patient values, to guide decision-making [45–47]. However, discourse surrounding the role of mechanistic reasoning in EBM has not yet reached consensus [44,48]. Traditionally, mechanistic reasoning has been considered lesser than correlative or statistical reasoning in hierarchies of evidence [48,49]. This view is supported by the unpredictable nature of human physiology—biological mechanisms tend to have high complexity, and a probabilistic disposition which challenges mechanistic reasoning[44]. Consequently, mechanistic explanations can lead to false conclusions, and mechanistic reasoning alone has been shown to have a high degree of fallibility. At times empirical results can be entirely contrary to mechanistic expectations, as in the case of prophylactic antiarrhythmic drugs actually acting to *increase* mortality from arrhythmia after recurrent acute myocardial infarction [50]. Mechanistic reasoning is not, therefore, sufficient to guide evidence-based practice.

Plausible mechanistic suggestions are critical in the development of new hypotheses [51], and may be useful in extrapolating the results of a statistical finding to a new population [52,53]. Biological plausibility features as one of nine Bradford Hill criteria for defending any suggestion of causality (although with the explicit caveat that it is

wholly dependent on contemporaneous biological knowledge, and therefore not always available) [54]. Biological plausibility has also gathered traction in more recent interpretations of causality like the Russo-Williamson Thesis, which suggests that causality in the health sciences requires both mechanistic and probabilistic (or "difference-making") evidence, and that the former can complement the latter [55,56]. Mechanistic reasoning does have demonstrable utility as a form of evidence, but considering EBM's progress in its absence, we argue that such reasoning cannot be considered strictly necessary for evidence-based practice.

At its core, the issue of explicability in MLHC becomes a question about appropriate rationale: what type and amount of evidence is compelling enough for the implementation of a new intervention in clinical practice? In the case of mood stabilizers, effective interventions with undetermined mechanisms were embraced rather than discarded—empiricism was used to circumvent the demand for mechanistic reasoning. In other words, clinicians and policy makers effectively chose to tolerate mechanistic uncertainty in exchange for well-evidenced utility. Rigorous clinical testing is able to validate such interventions, characterizing their full range of positive and negative effects in the absence of any knowledge of their mechanisms. Thus, clinical practice under the EBM framework is privileged towards statistically sound evidence that an intervention *does* work, even in the absence of mechanistic evidence as to *why* it would.

Similarly, the benefits and consequences of MLHC can be discovered and validated consequentially, through robust empirical analysis of even its most inexplicable outputs. Empirical evidence is also relevant to the various ethical dimensions in which a need for explainability in MLHC has been invoked. With fairness, for example, the relevant dimension is the *performance* of the algorithm for different groups, which depends on the (empirically visible) outputs of the model, rather than any underlying reasoning. Indeed, prominent case studies have demonstrated precisely how explainable (and not intentionally biased) models can nonetheless deliver biased results that are only apparent upon empirical analysis [57,58]. Similarly, the trust of both patients and physicians might be established through empirical evaluation and regulatory approval similar to that of any pharmaceutical or other medical technology.

It is important to highlight that awareness of the potential costs of explainability, however, does not imply that inexplicable models are superior in all cases. Indeed, in certain medical contexts (such as predicting heart failure outcomes from claims data) explainable logistic regression models have demonstrated similar efficacy when compared to more complex MLHC systems [59]. Small disparities may not be clinically meaningful, and in certain contexts explainability may be considered worthwhile in exchange for a small degree of reduced performance [60]. Further, significant work is being done toward the development of methods to improve the explainability of complex methods, offering a hope of "opening the black box" and enabling presently inscrutable deep learning architectures to become explainable [61–63]. As with other healthcare interventions whose mechanisms were better clarified over time, increasing explainability may arise in parallel with or subsequent to the implementation of black-box models through evidence-based evaluation.

There are specific and valid reasons for concern regarding the challenges of empirically validating MLHC tools, such as concerns that shifts in the underlying population may render a model inaccurate over time [64], or concerns that a continuously updating model may develop aberrant characteristics. These concerns can and should drive the adaptation of rigorous empirical methods to the specific needs of the MLHC context. Yet as with numerous other tools used in medicine, from pharmaceuticals to clinical reasoning aids to the very cognition of clinicians, the utility of MLHC is largely independent of our ability to fully explain its actions. Explainability has its value in this context, but to regard it as essential for effective and ethical MLHC is to apply a unique burden of proof to this particular class of emerging technology and to underestimate the capabilities of well-validated EBM concepts to be adapted to succeed in this novel context.

5. Conclusion

In examining the reasons that explainability tends to be a desirable feature of contemporary MLHC, we observe that these reasons are instrumental rather than intrinsic. Explainability is sought as a means toward ensuring model effectiveness and developing trust in both patients and providers. While explainability in MLHC may indeed advance these interests, an examination of other domains reveals that the scientific community has robust and well-established mechanisms for evaluating the effectiveness of (and developing trust in) tools that are not mechanistically explained or explainable. It must not be forgotten that significant trade-offs may arise between explainability and other interests, such as accuracy and overall performance. To narrowly emphasize the importance of explainability for the use of a medical tool would be to reject not only certain types of MLHC tools but also the well-validated and trusted edifice of EBM.

Continued work must be done to explore the trade-offs inherent in the pursuit of explainability in MLHC in any given domain. Tensions are likely to arise between raw predictive performance and the abstractive simplification necessary for an algorithm to be explainable, particularly in the bedside context. Achieving an optimal balance will depend on the specifics of the clinical context, and the strength of the validation procedures in place. Ultimately, however, we must not forget that artificially intelligent tools are attractive precisely because they are able to perform tasks of data synthesis and analysis at a scale and

speed not achievable by human cognition. In limiting machines to reasoning as humans do, we may rob them—and ourselves—of their unique potential to solve problems which we cannot.

Final Contributions Statement

Liam G. McCoy: Conceptualization, Investigation, Writing - Original Draft, Writing - Review & Editing. *Connor T. A. Brenna*: Conceptualization, Investigation, Writing - Original Draft, Writing - Review & Editing. *Stacy S. Chen*: Conceptualization, Investigation, Writing - Original Draft, Writing - Review & Editing. *Karina Vold*: Conceptualization, Writing - Review & Editing. *Sunit Das*: Conceptualization, Writing - Review & Editing, Supervision

References

[1] Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. Nat Rev Cancer 2018;18(8):500–10. doi:10.1038/s41568-018-0016-5.

[2] Fraser KC, Meltzer JA, Rudzicz F. Linguistic features identify alzheimer's disease in narrative speech. Journal of Alzheimer's Disease 2016;49(2):407–22. doi:10.3233/JAD-150520.

[3] Chassagnon G, Vakalopolou M, Paragios N, Revel M-P. Deep learning: definition and perspectives for thoracic imaging. Eur Radiol 2020;30(4):2021–30. doi:10.1007/s00330-019-06564-3.

[4] Bhatt U, Xiang A, Sharma S, et al. Explainable machine learning in deployment. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* '20. Association for Computing Machinery; 2020. p. 648–57. doi:10.1145/3351095.3375624.

[5] Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. Nat Mach Intelligence 2019;1(9):389–99. doi:10.1038/s42256-019-0088-2.

[6] Centre for Ethics. *Juliette Ferry-Danini, What Is the Problem with the Opacity of Artificial Intelligence in Medicine?*; 2021. Accessed March 28, 2021. https://www.youtube.com/watch?v=xNWe3PsfNng

[7] Char DS, Abràmoff MD, Feudtner C. Identifying ethical considerations for machine learning healthcare applications. Am J Bioeth 2020;20(11):7–17. doi:10.1080/15265161.2020.1819469.

[8] Meske C, Bunde E. Transparency and trust in human-ai-interaction: the role of model-agnostic explanations in computer vision-based decision support. *arXiv:200201543 [cs]*. 2020;12217:54-69. doi:10.1007/978-3-030-50334-5_4

[9] Holzinger A, Biemann C, Pattichis CS, Kell DB. What do we need to build explainable AI systems for the medical domain?*arXiv:171209923 [cs, stat]*. Published online December 28, 2017. Accessed November 24, 2020. http://arxiv.org/abs/1712.09923

[10] Adadi A, Berrada M. Explainable AI for healthcare: from black box to interpretable models. Embedded Syst Artificial Intelligence 2020:327–37 Published online. doi:10.1007/978-981-15-0947-6_31.

[11] E T, C G. A survey on explainable artificial intelligence (XAI): toward medical XAI. IEEE Trans Neural Netw Learn Syst 2020 PP. doi:10.1109/tnnls.2020.3027314.

[12] Cutillo CM, Sharma KR, Foschini L, Kundu S, Mackintosh M, Mandl KD. Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency. npj Digital Medicine 2020;3(1):1–5. doi:10.1038/s41746-020-0254-2.

[13] Gordon L, Grantcharov T, Rudzicz F. Explainable artificial intelligence for safe intraoperative decision support. JAMA Surg 2019;154(11):1064–5. doi:10.1001/jamasurg.2019.2821.

[14] Morley J, Machado CCV, Burr C, et al. The ethics of AI in health care: a mapping review. Social Sci Med 2020;260:113172. doi:10.1016/j.socscimed.2020.113172.

[15] Cohen IG, Evgeniou T, Gerke S, Minssen T. The European artificial intelligence strategy: implications and challenges for digital health. Lancet Digital Health 2020;2(7):e376–9. doi:10.1016/S2589-7500(20)30112-6.

[16] Current Health. A response to the FDA's new artificial intelligence proposals. Accessed 2020. https://currenthealth.com/response-to-fda-artificial-intelligence-proposals

[17] Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '15. Association for Computing Machinery; 2015. p. 1721–30. doi:10.1145/2783258.2788613.

[18] Oakden-Rayner L, Dunnmon J, Carneiro G, Ré C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. arXiv:190912475 [cs, stat]. Published online November 15, 2019. Accessed November 30, 2020. http://arxiv.org/abs/1909.12475

[19] Edwards L, Veale M. Enslaving the algorithm: from a "right to an explanation" to a "right to better decisions"? IEEE Secur Privacy 2018;16(3):46–54. doi:10.1109/MSP.2018.2701152.

[20] London AJ. Artificial intelligence and black-box medical decisions: accuracy versus explainability. Hastings Center Report 2019;49(1):15–21. doi:10.1002/hast.973.

[21] Nanayakkara S, Fogarty S, Tremeer M, et al. Characterising risk of in-hospital mortality following cardiac arrest using machine learning: A retrospective international registry study. PLOS Medicine 2018;15(11):e1002709. doi:10.1371/journal.pmed.1002709.

[22] Zerilli J, Knott A, Maclaurin J, Gavaghan C. Algorithmic decision-making and the control problem. *minds & machines*. 2019;29(4):555-578. doi:10.1007/s11023-019-09513-7

[23] Beil M, Proft I, van Heerden D, Sviri S, van Heerden PV. Ethical considerations about artificial intelligence for prognostication in intensive care. Intens Care Med Experiment 2019;7(1):70. doi:10.1186/s40635-019-0286-6.

[24] High Level Expert Group on AIEthics guidelines for trustworthy AI. European Commission 2019. Published onlinehttps://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf.

[25] Gillon R. Defending the four principles approach as a good basis for good medical practice and therefore for good medical ethics. J med ethics 2015;41(1):111–16.

[26] Kass NE, Faden RR. Ethics and Learning Health Care: The Essential roles of engagement, transparency, and accountability. Learn health syst 2018;2(4):e10066.

[27] Tonekaboni S, Joshi S, McCradden MD, Goldenberg A. What clinicians want: contextualizing explainable machine learning for clinical end use. *arXiv:190505134 [cs, stat]*. Published online August 7, 2019. Accessed November 1, 2020. http://arxiv.org/abs/1905.05134

[28] LeBlang TR. Informed consent and disclosure in the physician-patient relationship: expanding obligations for physicians in the United States. Med & L 1995;14:429.

[29] Price WN, Gerke S, Cohen IG. Potential liability for physicians using artificial intelligence. JAMA 2019;322(18):1765–6. doi:10.1001/jama.2019.15064.

[30] Zerilli J, Knott A, Maclaurin J, Gavaghan C. Transparency in algorithmic and human decision-making: is there a double standard? Philos Technol 2019;32(4):661–83. doi:10.1007/s13347-018-0330-6.

[31] Lipton ZC. The mythos of model interpretability. arXiv:160603490 [cs, stat] 2020. Published online March 6, 2017. Accessed March 3 http://arxiv.org/abs/1606.03490 .

[32] Zerilli J. Explaining machine learning decisions. Philosophy Sci 2021. Published online Forthcominghttp://philsci-archive.pitt.edu/19096/.

[33] Sevilla J, Hegde J. Deep visual patterns are informative to practicing radiologists in mammograms in diagnostic tasks. Journal of Vision 2017;17(10) 90-90. doi:10.1167/17.10.90.

[34] Hegdé J, Bart E. Making expert decisions easier to fathom: on the explainability of visual object recognition expertise. Front Neurosci 2018;12. doi:10.3389/fnins.2018.00670.
[35] Dhaliwal G. Clinical diagnosis—is there any other type? JAMA Intern Med 2020;180(10):1304–5. doi:10.1001/jamainternmed.2020.3048.
[36] Toussaint K, Yang XC, Zielinski MA, et al. What do we (not) know about how paracetamol (acetaminophen) works? Journal of Clinical Pharmacy and Therapeutics 2010;35(6):617–38. doi:10.1111/j.1365-2710.2009.01143.x.
[37] Rena G, Hardie DG, Pearson ER. The mechanisms of action of metformin. Diabetologia 2017;60(9):1577–85. doi:10.1007/s00125-017-4342-z.
[38] Lenox RH, Frazer A. Mechanism of action of antidepressants and mood stablizers. Neuropsychopharmacology: The Fifth Generation of Progress. Philadelphia: Lippincott Williams & Wilkins Citeseer; 2002.
[39] Pérez-Pevida B, Escalada J, Miras AD, Frühbeck G. Mechanisms underlying type 2 diabetes remission after metabolic surgery. Front Endocrinol (Lausanne) 2019;10. doi:10.3389/fendo.2019.00641.
[40] Pucci A, Batterham RL. Mechanisms underlying the weight loss effects of RYGB and SG: similar, yet different. J Endocrinol Invest 2019;42(2):117–28. doi:10.1007/s40618-018-0892-2.
[41] Ionut V, Bergman RN. Mechanisms Responsible for Excess Weight Loss after Bariatric Surgery. J Diabetes Sci Technol 2011;5(5):1263–82. Accessed October 17, 2021 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3208891/ .
[42] Lemoine M. Explanation in medicine. In: The Routledge Companion to Philosophy of Medicine. Routledge; 2016. p. 310–23. doi:10.4324/9781315720739-36.
[43] Fuller J. The new medical model: a renewed challenge for biomedicine. CMAJ 2017;189(17):E640–1. doi:10.1503/cmaj.160627.
[44] Howick J, Glasziou P, Aronson JK. Evidence-Based Mechanistic Reasoning. London, England: SAGE Publications Sage UK; 2010.
[45] Mayer D. Evidence-based Medicine. Epilepsia 2006;47:3–5.
[46] Gaeta R, Gentile N. Evidence, discovery and justification: the case of evidence-based medicine. J Evaluation Clin Pract 2016;22(4):550–7.
[47] Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ 1996;312(7023):71–2. doi:10.1136/bmj.312.7023.71.
[48] Clarke B, Gillies D, Illari P, Russo F, Williamson J. The evidence that evidence-based medicine omits. Prevent Med 2013;57(6):745–7.
[49] Nardini C, Annoni M, Schiavone G. Mechanistic understanding in clinical practice: complementing evidence-based medicine with personalized medicine. J evaluation clin pract 2012;18(5):1000–5.
[50] Echt DS, Liebson PR, Mitchell LB, et al. Mortality and morbidity in patients receiving encainide, flecainide, or placebo. the cardiac arrhythmia suppression trial. N Engl J Med 1991;324(12):781–8. doi:10.1056/NEJM199103213241201.
[51] Contopoulos-Ioannidis DG, Ntzani E, Ioannidis JP. Translation of highly promising basic science research into clinical applications. Am J med 2003;114(6):477–84.
[52] Marchionni C, Reijula S. What is mechanistic evidence, and why do we need it for evidence-based policy? Studies History Philosophy Sci Part A 2019;73:54–63.
[53] Aronson J, La Caze A, Kelly M, Parkkinen V-P, Williamson J. The use of evidence of mechanisms in drug approval. J Evaluation Clin Pract 2018 Published online.
[54] Hill AB. The Environment and Disease: Association or Causation? Sage Publications; 1965.
[55] Russo F, Williamson J. Interpreting causality in the health sciences. Int studies philosophy sci 2007;21(2):157–70.
[56] Illari PM. Mechanistic evidence: disambiguating the Russo–Williamson thesis. Int Studies Philosophy Sci 2011;25(2):139–57.
[57] Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019;366(6464):447–53. doi:10.1126/science.aax2342.
[58] Benjamin R. Assessing risk, automating racism. Science 2019;366(6464):421–2. doi:10.1126/science.aaz3873.
[59] Desai RJ, Wang SV, Vaduganathan M, Evers T, Schneeweiss S. Comparison of machine learning methods with traditional models for use of administrative claims with electronic medical records to predict heart failure outcomes. JAMA Network Open 2020;3(1):e1918962. doi:10.1001/jamanetworkopen.2019.18962.
[60] 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. doi:10.1038/s42256-019-0048-x.
[61] Hicks SA, Isaksen JL, Thambawita V, et al. Explaining deep neural networks for knowledge discovery in electrocardiogram analysis. Sci Rep 2021;11(1):10949. doi:10.1038/s41598-021-90285-5.
[62] Lundberg S, Lee S-I. A unified approach to interpreting model predictions. *arXiv:170507874 [cs, stat]*. Published online November 24, 2017. Accessed October 17, 2021. http://arxiv.org/abs/1705.07874
[63] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Grad-CAM Batra D. Visual explanations from deep networks via gradient-based localization. Int J Comput Vis 2020;128(2):336–59. doi:10.1007/s11263-019-01228-7.
[64] Nestor B, McDermott MBA, Boag W, et al. Feature robustness in non-stationary health records: caveats to deployable model performance in common clinical machine learning tasks. *arXiv:190800690 [cs, stat]*. Published online August 1, 2019. Accessed August 12, 2020. http://arxiv.org/abs/1908.00690