The need to strengthen the evaluation of the impact of Artificial Intelligence-based decision support systems on healthcare provision

Cresswell, Kathrin; Rigby, Michael; Magrabi, Farah; Scott, Philip; Brender, Jytte; Craven, Catherine K; Wong, Zoie Shui-Yee; Kukhareva, Polina; Ammenwerth, Elske; Georgiou, Andrew; Medlock, Stephanie; De Keizer, Nicolette F; Nykänen, Pirkko; Prgomet, Mirela; Williams, Robin

Published in:
Health Policy

DOI (link to publication from Publisher):
10.1016/j.healthpol.2023.104889

Publication date:
2023

Document Version
Publisher's PDF, also known as Version of record

Health policy 136 (2023) 104889

Contents lists available at ScienceDirect

Health policy

journal homepage: www.elsevier.com/locate/healthpol

Policy Comment

The need to strengthen the evaluation of the impact of Artificial Intelligence-based decision support systems on healthcare provision☆

Kathrin Cresswell[a,*], Michael Rigby[b], Farah Magrabi[c], Philip Scott[d], Jytte Brender[e], Catherine K. Craven[f], Zoie Shui-Yee Wong[g], Polina Kukhareva[h], Elske Ammenwerth[i], Andrew Georgiou[c], Stephanie Medlock[j,k], Nicolette F. De Keizer[j,k], Pirkko Nykänen[l], Mirela Prgomet[m], Robin Williams[n]

[a] *The University of Edinburgh, Usher Institute, Edinburgh, United Kingdom*
[b] *Keele University, School of Social, Political and Global Studies and School of Primary, Community and Social Care, Keele, United Kingdom*
[c] *Macquarie University, Australian Institute of Health Innovation, Sydney, Australia*
[d] *University of Wales Trinity Saint David, Swansea, United Kingdom*
[e] *Department of Health Science and Technology, Aalborg University, Aalborg, Denmark*
[f] *University of Texas Health Science Center at San Antonio, San Antonio, TX, United States*
[g] *St. Luke's International University, Graduate School of Public Health, Tokyo, Japan*
[h] *Department of Biomedical Informatics, University of Utah, United States of America*
[i] *UMIT TIROL, Private University for Health Sciences and Health Informatics, Institute of Medical Informatics, Hall in Tirol, Austria*
[j] *Amsterdam UMC location University of Amsterdam, Department of Medical Informatics, Meibergdreef 9, Amsterdam, the Netherlands*
[k] *Amsterdam Public Health research institute, Digital Health and Quality of Care Amsterdam, the Netherlands*
[l] *Tampere University, Faculty for Information Technology and Communication Sciences, Finland*
[m] *Faculty of Medicine, Health and Human Sciences, Macquarie University, Sydney, Australia*
[n] *The University of Edinburgh, Institute for the Study of Science, Technology and Innovation, Edinburgh, United Kingdom*

A R T I C L E I N F O

Keywords:
Artificial Intelligence (AI)
Health information technology
eHealth
Evidence
Evaluation

A B S T R A C T

Despite the renewed interest in Artificial Intelligence-based clinical decision support systems (AI-CDS), there is still a lack of empirical evidence supporting their effectiveness. This underscores the need for rigorous and continuous evaluation and monitoring of processes and outcomes associated with the introduction of health information technology.

We illustrate how the emergence of AI-CDS has helped to bring to the fore the critical importance of evaluation principles and action regarding all health information technology applications, as these hitherto have received limited attention. Key aspects include assessment of design, implementation and adoption contexts; ensuring systems support and optimise human performance (which in turn requires understanding clinical and system logics); and ensuring that design of systems prioritises ethics, equity, effectiveness, and outcomes.

Going forward, information technology strategy, implementation and assessment need to actively incorporate these dimensions. International policy makers, regulators and strategic decision makers in implementing organisations therefore need to be cognisant of these aspects and incorporate them in decision-making and in prioritising investment. In particular, the emphasis needs to be on stronger and more evidence-based evaluation surrounding system limitations and risks as well as optimisation of outcomes, whilst ensuring learning and contextual review. Otherwise, there is a risk that applications will be sub-optimally embodied in health systems with unintended consequences and without yielding intended benefits.

☆ The authors are members of the International Medical Informatics Association Working Group on Technology Assessment and Quality Development and the European Federation for Medical Informatics Working Group on Evaluation.

* Corresponding author.
E-mail address: kathrin.cresswell@ed.ac.uk (K. Cresswell).

https://doi.org/10.1016/j.healthpol.2023.104889
Received 27 July 2023; Accepted 4 August 2023
Available online 12 August 2023

1. Introduction

The use of Artificial Intelligence (AI) in medicine has great potential to help achieve the quintuple aims of healthcare [1–3]. AI-based computer systems can perform tasks that normally require elements of human cognitive skills such as visual perception, pattern recognition, speech recognition, rapid data comparisons and projections, translation between languages, and decision-making between set options. We here focus on AI-based clinical decision support (AI-CDS) systems supporting decisions by human healthcare professionals (e.g. through image analysis, establishing clinical diagnoses, proposing the best course of treatment, or identifying key deviations in vital or other signs), and in shared decision making together with patients [4]. Traditional CDS is based on encoded human expertise or authoritative clinical guidelines, whereas the knowledge base in AI-CDS draws on statistical calculations or pattern recognition, not evidence synthesis.

Although AI-CDS has a long history, recent years have seen an explosive renewal of interest in AI in medicine stimulated by advances in deep learning, and increased computational power [5–15]. This has been accompanied by heavy governmental and private sector investment in the development and implementation of AI-based systems.

Despite intense commercialisation [16,17], there is still limited empirical evidence behind existing claims of improved patient outcomes, healthcare effectiveness, and efficiency [18]. In addition, evaluation of AI-CDS has focused on demonstrating performance of systems in laboratory or trial implementation settings [19,20], and on measuring immediate outcomes [21,22]. There is a lack of focus on longer-term impacts, potential disbenefits and unintended consequences (e.g. de-skilling, possible increase in unnecessary referrals or tests, bias against specific groups or conditions) [22,23].

In order to inform evidence-based decision making on selection and implementation of AI-based systems, there is a need to assess and build on existing frameworks and standards to evaluate the introduction of AI-CDS in healthcare in everyday use [24,25]. This should go hand-in-hand with increased emphasis on the importance of evidence-based systems and policy [26,27].

We will here explore what evaluation dimensions the literature surrounding AI-CDS has highlighted and extract lessons to inform decision-making for health policy internationally, nationally, and locally. To date, most AI-based applications in healthcare have been developed and implemented in high-income settings, and therefore, we focus on these.

1.1. The importance of contextual sensitivity

The implementation of AI-CDS across healthcare settings has been difficult [28,29]. Systems cannot be dependably transferred from one setting to another (e.g. from a research lab into clinical use or from an initial adoption site to other settings). At present, measurement of performance and publishing of studies is not frequently done. A review of measurement practices in health informatics, showed for example a lack of validity of instruments used in many studies [30]. Underlying reasons include, amongst others, differences in needs, existing work processes, health information infrastructures, health and care practices, inter-organisational and transactional relationships, socio-demographic and ethnic characteristics, and organisational cultures [31–33]. Moreover, many current studies of AI in healthcare do not include components that enable clinicians to understand how algorithms may be incorporated effectively in their workflow, even though differences in work organisation between sites (and changes in practices as a result of the use of new tools) may impact on the performance of the algorithm [32].

A key consideration here are the characteristics of the training data set and how these relate to targeted patient populations [34]. For example, if a model was trained on data from one specific hospital with specific demographic characteristics, then it may not be readily transferable to a different hospital with different target populations. This is known as dataset shift. In a recent paper Finlayson and colleagues give an example of the decommissioning of an AI-based sepsis alert system due to the Coronavirus pandemic, which changed the use patterns of antibiotics, meaning that the alerts were spurious and therefore ignored by clinicians [35].

Organizational, technological and user contexts need to be key components of evaluation studies as they can help inform the generalisability of the results and highlight aspects that may need to be reformulated when implementing systems across contexts. Formative approaches to evaluation have incorporated these requirements, often beginning with an assessment of existing systems, structures and processes before technology implementation, and following changes introduced by technology through in-depth study across a range of settings [36,37]. It is encouraging that new reporting guidelines specifically designed for AI increasingly incorporate such approaches [38].

There are also recent attempts to develop integrative AI evaluation frameworks with attention to wider processes in healthcare settings [39, 40]. These highlight the unique features of AI beyond the immediate context of implementation and the importance of wider macro-environmental considerations shaping technology adoption and use. Some have, for example, emphasised important but potentially perverse political and commercial drivers associated with economic success through big data surrounding the introduction of AI in healthcare settings [41–44]. Others have highlighted the dynamic nature of the market and regulatory environments surrounding AI internationally and their role in shaping technology implementation and use [45].

Unfortunately, context-related issues surrounding commercial, economic, regulatory, market and legal issues have to date received far too little attention in HIT evaluation.

1.2. System logic and assistive tools

AI has further highlighted the importance of clinician and patient users' understanding of and trust in systems [46]. There are currently many different assumptions and understandings of what AI is and how it operates [47,48]. Previous work in high-income countries has shown that if users of a system understand how decisions are made, then they are more likely to adopt it [49]. A lack of such an understanding can lead to limited adoption/use of a system, or to workarounds, which may in turn have adverse effects for the safety and quality of care. Particular problems may arise where users lack the information and expertise required to assess the model and the evidence it is based on, and adopt its recommendations uncritically. In these situations, patient and clinician users may find it hard to compensate for known shortcomings of systems. Prospective users therefore need to develop AI competencies to understand how an application operates, and the data sets upon which it is based [50].

Unfortunately, there are enduring political and commercial pressures for implementation and scale-up of AI-CDS whilst bypassing investment in evaluation [51]. Application of the Precautionary Principle (involving up-front risk assessment and mitigation, and continuing this scrutiny in an iterative ongoing manner e.g. through post-market surveillance) [52–54], and Evidence-Based Health Informatics Principles are essential going forward [26,27]. In AI-CDS, these may help to ensure that advice is presented in a way that is consistent with the level of evidence and path of deduction behind it. Otherwise, the fast-evolving nature of these systems, although potentially beneficial in the ability to respond to changing circumstances, may have unintended consequences emerging from algorithmic bias.

AI-CDS has further highlighted issues surrounding levels of autonomy of systems but the issue of how machine and human capabilities may most effectively complement each other has to date been neglected [55]. For example, AI-CDS can process large volumes of data consistently and at speed, but has difficulties in dealing with ambiguous settings which may be readily understood by human experts.

Notwithstanding that AI-based systems may have more autonomy in the future, work has shown that, in order to promote adoption, systems need to be conceptualised as assistive tools and not as autonomous entities [56–58]. Here, algorithmic outputs need to be interpreted by humans who understand their strengths and limitations. For example, algorithms can help to compensate for the tendency of humans towards optimistic predictions (e.g. concerning life expectancy) [56]. There are currently different levels of autonomy for AI-CDS ranging from assistive devices to autonomous machine decision-making [59]. Greater reliance on algorithmic recommendations may depend upon the complexity of the clinical problem, the fit between model performance and task, the levels of user trust and confidence in model performance, the availability of (good quality) data relevant for the problem, the match between the training population and the target population, the transferability of the model to other contexts, and the degree of clinician review at the point of decision making. Evaluations need to take these dimensions into account, as their evidence can help to prevent bias and unintended adverse consequences and is likely to determine patterns of use and outcomes.

1.3. Designing and optimising systems in the interest of ethics and equity

Algorithmic bias, privacy/security and data drift have highlighted ethical complexities surrounding the implementation of systems. These include trustworthiness, transparency, justice, fairness, accountability, equity and consent [41,60-64]. For example, work has shown that systems are often designed from certain socio-economic and cultural viewpoints (e.g. associated with the lack of ethnic diversity in the AI workforce) [65,66], while other studies show that in many instances health and care technologies are not used by those who would benefit most from them, potentially inadvertently contributing to increased health disparities [67,68].

Ethical issues require consideration of complex trade-offs and should therefore be an essential part of any HIT evaluation. These trade-offs are particularly apparent when considering AI-CDS. For example, there are tensions between data protection, consent, and exploitation of data. Whilst data protection is governed in most countries through privacy and security law [69], it can be a potential barrier to beneficial secondary uses of data including building AI models (e.g. when models require data sharing for training) [70,71].

Internationally developed evaluation frameworks that focus on ethical considerations surrounding AI in healthcare exist, but are not widely adopted [72]. Emphasis should be on increasing patient engagement and for widespread community-based participatory research to understand views on using systems and data. Co-creation approaches have significant potential in this respect and can help to negotiate complex ethical tensions [73]. The use of synthetic data generation techniques to preserve data privacy and increase the volume of data is promising [74].

Unintended consequences caused by HIT include issues due to algorithmic bias, data incongruent provenance, and inadequate data quality [75–80]. In many instances, AI systems trained on specific datasets do not perform well when applied to other datasets and in different contexts, diminishing the transportability of the model. This may then lead to loss of predictive capability/reliability for under-represented segments and algorithmic bias. A classic example is Google's dermatology app, which was trained on Caucasian skin and did not detect melanoma in darker skin [81]. For evaluation, this means that it is critically important to evaluate how a system performs on local data used for 'training' and construction [82], ascertain that such performance has been validated, and match training populations (and treatment options) to the characteristics of the potential transfer site [83].

There are existing international frameworks accounting for algorithmic bias in healthcare settings [84,85], and increasing efforts to utilise incident reporting systems of AI aiming to learn from adverse events [86–88]. These highlight the large untapped potential of automated approaches for incident analysis and quantification. However, their application has to date been limited and each must itself be validated. Routine evaluation practices now need to incorporate such approaches in order to proactively mitigate for potential biases and ethical risks.

Table 1
Implications of AI-CDS and HIT evaluation for health policy in high-income countries.

- The area surrounding AI-CDS is dynamic and constantly evolving – any policy and strategy therefore needs to incorporate a degree of dynamic review and flexibility
- Empirical evidence surrounding the effectiveness and efficacy of particular applications is limited – strategic decisions need to draw on, and seek to synthesise, existing empirical evidence
- Concurrent formative evaluation of implementation and adoption of systems needs to be factored in from the start in order to monitor and mitigate any potential adverse consequences (e.g. for work practices and equity)
- Careful risk assessment needs to be made, navigating the tensions and trade-offs surrounding benefits and risks of AI system implementation (e.g. around confidentiality, trust)
- Training in understanding system limitations is crucial for effective implementation. There may be scope for developing guidance around training requirements for potential users.
- Recognising the importance to build a community of practice, and repositories of evaluation methods and outcomes, while respecting both patient data, intellectual property and patient confidentiality is important. Trusted third party agencies and methods may be part of this.

2. Conclusions

Considering the rapid proliferation of AI in healthcare, and the multiple pressures for roll-out, there is an urgent need for rigorous evaluation. This calls in turn for establishing networks of experience in effective application of evaluation tools, and for building an accessible verified evidence base. These will help to ensure that procurement and implementation decisions are evidence informed.

As health systems and contexts are constantly evolving - through the introduction of novel HIT, including AI-CDS, as well as better understanding of illness, new treatments, and treatment responses - it is vital that evaluations include a longitudinal component that accounts for these changes and surfaces emerging risks (e.g. degradation of model performance) over time. Such continuous systemic evaluation can promote learning health systems [89,90], but is lacking in evaluation practice [91,92]. The emergence of AI-CDS has also helped to illustrate the need for continuous post-market surveillance [93,94].

There are many well-established frameworks relevant for AI-CDS, but routine evaluation practices now need to take these into account. Key considerations include attention to contexts, focusing on helping users to understand system logics and designing assistive tools, designing and optimising systems in the interest of ethics and equity, and continuous evaluation and monitoring of processes and outcomes. These dimensions are likely to be important irrespective of health systems and existing health information technology infrastructures. We summarise implications of this work for international health policy in high-income countries in Table 1.

References

[1] Bates DW, Levine D, Syrowatka A, Kuznetsova M, Craig KJ, Rui A, Jackson GP, Rhee K. The potential of artificial intelligence to improve patient safety: a scoping review. NPJ Digit Med 2021;4(1):1–8. Mar 19.

[2] Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019;25(1):44–56.

[3] Itchhaporia D. The evolution of the quintuple aim: health equity, health outcomes, and the economy. J Am Coll Cardiol 2021;78(22):2262–4. PMCID:PMC8608191.

[4] Davenport T., Kalakota R. The potential for artificial intelligence in healthcare. Fut Healthc J 6 (2): 94.

[5] The state of AI in 2021. Available from: https://www.mckinsey.com/business-functions/quantumblack/our-insights/global-survey-the-state-of-ai-in-2021 (last accessed: 31/07/2022).

[6] Artificial Intelligence. What it is and why it matters. Available from: https://www.sas.com/en_us/insights/analytics/what-is-artificial-intelligence.html#:~:text=The%20term%20artificial%20intelligence%20was,in%20computing%20power%20and%20storage (last accessed: 22/07/22).
[7] Turing AM. Computing machinery and intelligence. Netherlands: Springer; 2009.
[8] Moor J. The Dartmouth College artificial intelligence conference: the next fifty years. AI Mag 2006;27(4):87. Dec 15.
[9] History of AI winters. Available from: https://www.actuaries.digital/2018/09/05/history-of-ai-winters/(last accessed: 21/06/22).
[10] Britannica. Available from:https://www.britannica.com/technology/artificial-intelligence (last accessed: 17/08/2022).
[11] The Centre for Evidence-Based Medicine. Available from: https://www.cebm.net/(last accessed: 31/07/2022).
[12] Meta-analysis: what, why, and how. Available from: https://uk.cochrane.org/news/meta-analysis-what-why-and-how (last accessed: 31/07/2022).
[13] Lexico. Available from: https://www.lexico.com/definition/machine_learning (last accessed: 31/07/2022).
[14] Wikipedia. Available from: https://en.wikipedia.org/wiki/Precautionary_principle (last accessed: 13/08/2022).
[15] The SAGE Encyclopedia of action research. Available from: https://methods.sagepub.com/reference/encyclopedia-of-action-research/n320.xml (last accessed: 13/08/2022).
[16] Global AI in Healthcare Market (2021 to 2027) - by Sections, Diagnosis, End-user and geography. Available from: https://www.prnewswire.com/news-releases/global-ai-in-healthcare-market-2021-to-2027-by-sections-diagnosis-end-user-and-geography-301465668.html (last accessed: 22/07/22).
[17] Emerging Start-ups 2022. Available from: https://tracxn.com/d/emerging-startups/top-big-data-analytics-startups-2022 (last accessed: 22/07/22).
[18] Vollmer S, Mateen BA, Bohner G, Király FJ, Ghani R, Jonsson P, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics and effectiveness. BMJ 2020:368.
[19] Yin J, Ngiam KY, Teo HH. Role of artificial intelligence applications in real-life clinical practice: systematic review. J Med Internet Res 2021;23(4):e25759. Apr 22.
[20] Wong A, Otles E, Donnelly JP, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Intern Med 2021;181(8):1065–70. PMCID:PMC8218233.
[21] Nsoesie EO. Evaluating artificial intelligence applications in clinical settings. JAMA Netw Open 2018;1(5):e182658. Sep 7.
[22] Magrabi F, Ammenwerth E, McNair JB, De Keizer NF, Hyppönen H, Nykänen P, Rigby M, Scott PJ, Vehko T, Wong ZS, Georgiou A. Artificial intelligence in clinical decision support: challenges for evaluating AI and practical implications. Yearb Med Inform 2019;28(01):128. Aug–34.
[23] Krass M, Henderson P, Mello MM, Studdert DM, Ho DE. How US law will evaluate artificial intelligence for covid-19. BMJ 2021:372.
[24] Evidence standards framework for digital health technologies. Available from: https://www.nice.org.uk/corporate/ecd7/chapter/update-information (last accessed: 25/01/2023).
[25] Tools such as ChatGPT threaten transparent science; here are our ground rules for their use. Available from: https://www.nature.com/articles/d41586-023-00191-1 (last accessed: 29/01/2023).
[26] Rigby M, Ammenwerth E, Beuscart-Zephir MC, Brender J, Hyppönen H, Melia S, et al. Evidence Based Health Informatics: 10 years of efforts to promote the principle. Yearb Med Inform 2013;22(01):34–46.
[27] Rigby M, Ammenwerth E, Talmon J. Forward outlook: the need for evidence and for action in health informatics. In: Evidence-Based Health Informatics. IOS Press; 2016. p. 355–63.
[28] Morrison K. Artificial intelligence and the NHS: a qualitative exploration of the factors influencing adoption. Future Healthc J 2021;8(3):e648. Nov.
[29] Coiera E. The last mile: where artificial intelligence meets reality. J Med Internet Res 2019;21(11):e16323. Nov 8.
[30] Schloemer T, Schröder-Bäck P. Criteria for evaluating transferability of health interventions: a systematic review and thematic synthesis. Implement Sci 2018;13(1):1–7. Dec.
[31] Coiera E, Ammenwerth E, Georgiou A, Magrabi F. Does health informatics have a replication crisis? J Am Med Inform Assoc 2018;25(8):963. Aug–8.
[32] Scott PJ, Brown AW, Adedeji T, Wyatt JC, Georgiou A, Eisenstein EL, Friedman CP. A review of measurement practice in studies of clinical decision support systems 1998–2017. J Am Med Inform Assoc 2019;26(10):1120. Oct–8.
[33] Wong ZS, Rigby M. Identifying and addressing digital health risks associated with emergency pandemic response: problem identification, scoping review, and directions toward evidence-based evaluation. Int J Med Inform 2022;157:104639. Jan 1.
[34] Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med 2019;17(1):1–9. Dec.
[35] Finlayson SG, Subbaswamy A, Singh K, Bowers J, Kupke A, Zittrain J, Kohane IS, Saria S. The clinician and dataset shift in artificial intelligence. N Engl J Med 2021;385(3):283. Jul 15–6.
[36] Cresswell K, Sheikh A, Franklin BD, Krasuska M, Nguyen H, Hinder S, Lane W, Mozaffar H, Mason K, Eason S, Potts H. Formative independent evaluation of a digital change programme in the English National Health Service: study protocol for a longitudinal qualitative study. BMJ Open 2020;10(10):e041275. Oct 1.
[37] Prgomet M, Georgiou A, Callen J, Westbrook J. Fit between individuals, tasks, technology, and environment (FITTE) framework: a proposed extension of FITT to evaluate and optimise health information technology use. MEDINFO 2019: Health and Wellbeing e-Networks for All. IOS Press; 2019. p. 744–8.
[38] Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med 2022;28(5):924–33.
[39] Gama F, Tyskbo D, Nygren J, Barlow J, Reed J, Svedberg P. Implementation frameworks for artificial intelligence translation into health care practice: scoping review. J Med Internet Res 2022;24(1):e32215. Jan 27.
[40] Alami H, Lehoux P, Auclair Y, de Guise M, Gagnon MP, Shaw J, Roy D, Fleet R, Ahmed MA, Fortin JP. Artificial intelligence and health technology assessment: anticipating a new level of complexity. J Med Internet Res 2020;22(7):e17707. Jul 7.
[41] Shaw J, Rudzicz F, Jamieson T, Goldfarb A. Artificial intelligence and the implementation challenge. J Med Internet Res 2019;21(7):e13659. Jul 10.
[42] Choudhury A, Asan O, Mansouri M. Role of artificial intelligence, clinicians & policymakers in clinical decision making: a systems viewpoint. In: 2019 International Symposium on Systems Engineering (ISSE). IEEE; 2019. p. 1–8.
[43] Rachel Clarke: Why Matt Hancock's promotion of Babylon worries doctors. Available from: https://blogs.bmj.com/bmj/2018/12/04/rachel-clarke-why-matt-hancocks-promotion-of-babylon-worries-doctors/(last accessed: 21/06/22).
[44] Shareholders of firm backed by Matt Hancock have donated to the Tories. Available from: https://www.theguardian.com/politics/2021/jun/22/shareholders-of-firm-backed-by-matt-hancock-have-donated-to-the-tories (last accessed: 21/06/22).
[45] van Leeuwen KG, Schalekamp S, Rutten MJ, van Ginneken B, de Rooij M. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. Eur Radiol 2021;31(6):3797–804. Jun.
[46] U.S. Food and Drug Administration (FDA). Intended use of imaging software for intracranial large vessel occlusion - letter to health care providers. Available from: https://www.fda.gov/medical-devices/letters-health-care-providers/intended-use-imaging-software-intracranial-large-vessel-occlusion-letter-health-care-providers (last accessed: 22/07/22).
[47] Bélisle-Pipon JC, Couture V, Roy MC, Ganache I, Goetghebeur M, Cohen IG. What makes artificial intelligence exceptional in health technology assessment? Front Artif Intell 2021;4:153.
[48] Ji M, Genchev GZ, Huang H, Xu T, Lu H, Yu G. Evaluation framework for successful artificial intelligence–enabled clinical decision support systems: mixed methods study. J Med Internet Res 2021;23(6):e25929. Jun 2.
[49] Yusof MM. A case study evaluation of a critical care information system adoption using the socio-technical and fit approach. Int J Med Inform 2015;84(7):486. Jul 1–99.
[50] Machleid F, Kaczmarczyk R, Johann D, Balčiūnas J, Atienza-Carbonell B, von Maltzahn F, Mosch L. Perceptions of digital health education among European medical students: mixed methods survey. J Med Internet Res 2020;22(8):e19827. Aug 14.
[51] Rigby M. Evaluation: 16 powerful reasons why not to do it - and 6 over-riding imperatives. In: Patel V, Rogers R, Haux R, editors. Medinfo 2001: Proceedings of the 10th. World Congress on Medical Informatics. Amsterdam: IOS Press; 2001. p. 1198–202.
[52] The precautionary principle: decision-making under uncertainty. Available from: https://ec.europa.eu/environment/integration/research/newsalert/pdf/precautionary_principle_decision_making_under_uncertainty_FB18_en.pdf (last accessed: 21/06/22).
[53] Ashford NA. Implementing the Precautionary Principle: incorporating science, technology, fairness, and accountability in environmental, health, and safety decisions. Int J Occup Med Environ Health 2004;17(1):59–67.
[54] Commission of the European Community. Communication from the commission on the precautionary principle. Available from: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52000DC0001&from=EN (last accessed: 07/08/22).
[55] Cresswell K, Cunningham-Burley S, Sheikh A. Health care robotics: qualitative exploration of key challenges and future directions. J Med Internet Res 2018;20(7):e10410. Jul 4.
[56] Verghese A, Shah NH, Harrington RA. What this computer needs is a physician: humanism and artificial intelligence. JAMA 2018;319(1):19–20. Jan 2.
[57] Loftus TJ, Tighe PJ, Filiberto AC, Efron PA, Brakenridge SC, Mohr AM, et al. Artificial intelligence and surgical decision-making. JAMA Surg 2020;155(2):148–58.
[58] Van Cauwenberge D, Van Biesen W, Decruyenaere J, Leune T, Sterckx S. Many roads lead to Rome and the Artificial Intelligence only shows me one road": an interview study on physician attitudes regarding the implementation of computerised clinical decision support systems. BMC Med Ethics 2022;23(1):1–4. Dec.
[59] Lyell D, Coiera E, Chen J, Shah P, Magrabi F. How machine learning is embedded to support clinician decision making: an analysis of FDA-approved medical devices. BMJ Health Care Inform 2021;28(1).
[60] Beil M, Proft I, van Heerden D, Sviri S, van Heerden PV. Ethical considerations about artificial intelligence for prognostication in intensive care. Intensiv Care Med Exp 2019;7(1):70.
[61] Asan O, Bayrak AE, Choudhury A. Artificial intelligence and human trust in healthcare: focus on clinicians. J Med Internet Res 2020;22(6):e15154. Jun 19.
[62] Zhang Z, Citardi D, Wang D, Genc Y, Shan J, Fan X. Patients' perceptions of using artificial intelligence (AI)-based technology to comprehend radiology imaging data. Health Inform J 2021;27(2):14604582211011215. Apr.
[63] Ethics guidelines for trustworthy AI Available from: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (last accessed: 21/06/22).

[64] Siala H, Wang Y. SHIFTing artificial intelligence to be responsible in healthcare: a systematic review. Soc Sci Med 2022;296:114782.
[65] Showell C, Turner P. The PLU problem: are we designing personal ehealth for people like us? Stud Health Technol Inform 2013;183:276.
[66] Showell C, Turner P. Personal health records are designed for people like us. In MEDINFO. IOS Press; 2013. p. 1037.
[67] Cresswell K, Rigby M, Georgiou A, Wong ZS, Kukhareva P, Medlock S, De Keizer NF, Magrabi F, Scott P, Ammenwerth E. The role of formative evaluation in promoting digitally-based health equity and reducing bias for resilient health systems: the case of patient portals. Yearbook of medical informatics. 2022. Jun 2.
[68] Wong MC, Almond H, Cummings E, Roehrer E, Showell C, Turner P. Patient centred systems: techno-anthropological reflections on the challenges of'meaningfully engaging'patients within health informatics research. Stud Health Technol Inform 2015;215:52–66. Jan 1.
[69] Data Protection in Japan: all You Need to Know about APPI. Available from: https://www.endpointprotector.com/blog/data-protection-in-japan-appi/(last accessed: 21/06/22).
[70] OpenAI A.P.I. Available from: https://openai.com/blog/openai-api/(last accessed: 21/06/22).
[71] Pathways Language Model (PaLM): Scaling to 540 Billion Parameters for Breakthrough Performance. Available from: https://ai.googleblog.com/2022/04/pathways-language-model-palm-scaling-to.html (last accessed: 21/06/22).
[72] Reddy S, Rogers W, Makinen VP, Coiera E, Brown P, Wenzel M, et al. Evaluation framework to guide implementation of AI systems into healthcare settings. BMJ Health Care Inform 2021;28(1).
[73] Liaw ST, Zhou R, Ansari S, Gao J. A digital health profile & maturity assessment toolkit: cocreation and testing in the Pacific Islands. J Am Med Inform Assoc 2021;28(3):494–503. Mar 1.
[74] Bhanot K, Qi M, Erickson JS, Guyon I, Bennett KP. The problem of fairness in synthetic healthcare data. Entropy 2021;23(9):1165. Sep 4.
[75] NESTcc publishes data quality maturity model self-assessment tools to complement initial data quality framework. Available from: https://nestcc.org/nestcc-publishes-data-quality-maturity-model-self-assessment-tools/(last accessed: 21/06/22).
[76] Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med 2022;28(1):31. Jan–8.
[77] Zozus M.H.W, Green B., Kahn M., Richesson R., Rusinkovich S., et al. Assessing Data Quality for Healthcare Systems Data Used in Clinical Research (Version 1.0) 2014 [cited 2022 July 19, 2022]. Available from: https://www.researchgate.net/profile/Meredith_Zozus/publication/283267713_Data_Quality_Assessment_Recommendations_for_Secondary_ise_of_EHR_Data/links/562f9d3908aeb1709b6000af.pdf.
[78] Richesson RL, Horvath MM, Rusincovitch SA. Clinical research informatics and electronic health record data. Yearb Med Inform 2014;9(1):215. –23.
[79] Manoharan L, Cattrall JWS, Harris C, Newell K, Thomson B, Pritchard MG, et al. Evaluating clinical characteristics studies produced early in the Covid-19 pandemic: a systematic review. PLoS One 2021;16(5):e0251250.
[80] Zozus MN, Richesson RL, Walden A, Tenenbaum JD, Hammond WE. Research reproducibility in longitudinal multi-center studies using data from electronic health records. AMIA Jt Summits Transl Sci Proc 2016;2016:279. –85.
[81] Google's new dermatology app wasn't designed for people with darker skin. Available from: https://www.vice.com/en/article/m7evmy/googles-new-dermatology-app-wasnt-designed-for-people-with-darker-skin (last accessed: 21/06/22).
[82] Keikes L, Medlock S, van de Berg DJ, Zhang S, Guicherit OR, Punt CJA, van Oijen MGH. The first steps in the evaluation of a "black-box" decision support tool: a protocol and feasibility study for the evaluation of Watson for Oncology. J Clin Transl Res 2018;3(Suppl 3):411–23. Jul 27PMID: 30873490; PMCID: PMC6412599.
[83] Ebrahimian S, Kalra MK, Agarwal S, Bizzo BC, Elkholy M, Wald C, Allen B, Dreyer KJ. FDA-regulated AI algorithms: trends, strengths, and gaps of validation studies. Acad Radiol 2022;29(4):559. Apr 1–66.
[84] Hernandez-Boussard T, Bozkurt S, Ioannidis JP, Shah NH. MINIMAR (MINimum Information for Medical AI Reporting): developing reporting standards for artificial intelligence in health care. J Am Med Inform Assoc 2020;27(12):2011. Dec–5.
[85] Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK, Ashrafian H, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Lancet Digital Health 2020;2(10):e537–48.
[86] Friedman CP, Wong AK, Blumenthal D. Achieving a nationwide learning health system. Sci Transl Med 2010;2(57):57cm29.
[87] Kim MO, Coiera E, Magrabi F. Problems with health information technology and their effects on care delivery and patient outcomes: a systematic review. J Am Med Inform Assoc JAMIA 2017;24(2):246. –50.
[88] Chai KE, Anthony S, Coiera E, Magrabi F. Using statistical text classification to identify health information technology incidents. J Am Med Inform Assoc JAMIA 2013;20(5):980. –5.
[89] Friedman C, Rigby M. Conceptualising and creating a global learning health system. Int J Med Inform 2013;82:e63–71.
[90] Grossmann C., Powers B., McGinnis J.M. Digital infrastructure for the learning health system: the foundation for continuous improvement in health and health care.
[91] Catwell L, Sheikh A. Evaluating eHealth interventions: the need for continuous systemic evaluation. PLoS Med 2009;6(8):e1000126. Aug 18.
[92] van Gennip E.M., Talmon J.L. The conception of assessment and evaluation of information technologies in medicine. 1995;17:45. Available from: https://books.google.co.uk/books?hl=en&lr=&id=smmVyRKNMx8C&oi=fnd&pg=PA45&dq=van+Gennip+life-cycle+perspective+on+evaluation+of+medical+DSS&ots=SZ5zZ6Mb8C&sig=U9YEiJIJQLHJFSt-BwE8cGImeOc&redir_esc=y#v=onepage&q&f=false (last accessed: 21/06/22).
[93] FDA guidance on premarket and postmarket data collection. Available from: https://www.regdesk.co/fda-guidance-on-premarket-and-postmarket-data-collection/(last accessed: 21/06/22).
[94] Park Y, Jackson GP, Foreman MA, Gruen D, Hu J, Das AK. Evaluating artificial intelligence in medicine: phases of clinical research. JAMIA Open 2020;3(3):326–31. Oct.