Artificial Intelligence in Clinical Decision Support: Challenges for Evaluating AI and Practical Implications

A Position Paper from the IMIA Technology Assessment & Quality Development in Health Informatics Working Group and EFMI Working Group for Assessment of Health Information Systems

Farah Magrabi[1], Elske Ammenwerth[2], Jytte Brender McNair[3], Nicolet F. De Keizer[4], Hannele Hyppönen[5], Pirkko Nykänen[6], Michael Rigby[7], Philip J. Scott[8], Tuulikki Vehko[5], Zoie Shui-Yee Wong[9], Andrew Georgiou[1]

[1] Macquarie University, Australian Institute of Health Innovation, Sydney, Australia
[2] UMIT, University for Health Sciences, Medical Informatics and Technology, Institute of Medical Informatics, Hall in Tyrol, Austria
[3] Department of Health Science and Technology, Aalborg University, Aalborg, Denmark
[4] Amsterdam UMC, University of Amsterdam, Department of Medical Informatics, Amsterdam Public Health research institute, The Netherlands
[5] National Institute for Health and Welfare, Information Department, Helsinki, Finland
[6] Tampere University, Faculty for Information Technology and Communication Sciences, Tampere, Finland
[7] Keele University, School of Social Science and Public Policy, Keele, United Kingdom
[8] University of Portsmouth, Centre for Healthcare Modelling and Informatics, Portsmouth, United Kingdom
[9] St. Luke's International University, Tokyo, Japan

Summary

Objectives: This paper draws attention to: i) key considerations for evaluating artificial intelligence (AI) enabled clinical decision support; and ii) challenges and practical implications of AI design, development, selection, use, and ongoing surveillance.

Method: A narrative review of existing research and evaluation approaches along with expert perspectives drawn from the International Medical Informatics Association (IMIA) Working Group on Technology Assessment and Quality Development in Health Informatics and the European Federation for Medical Informatics (EFMI) Working Group for Assessment of Health Information Systems.

Results: There is a rich history and tradition of evaluating AI in healthcare. While evaluators can learn from past efforts, and build on best practice evaluation frameworks and methodologies, questions remain about how to evaluate the safety and effectiveness of AI that dynamically harness vast amounts of genomic, biomarker, phenotype, electronic record, and care delivery data from across health systems. This paper first provides a historical perspective about the evaluation of AI in healthcare. It then examines key challenges of evaluating AI-enabled clinical decision support during design, development, selection, use, and ongoing surveillance. Practical aspects of evaluating AI in healthcare, including approaches to evaluation and indicators to monitor AI are also discussed.

Conclusion: Commitment to rigorous initial and ongoing evaluation will be critical to ensuring the safe and effective integration of AI in complex sociotechnical settings. Specific enhancements that are required for the new generation of AI-enabled clinical decision support will emerge through practical application.

Keywords

Artificial intelligence; machine learning; clinical decision support, evaluation studies; program evaluation

Yearb Med Inform 2019
http://dx.doi.org/10.1055/s-0039-1677903

1 Introduction

Artificial intelligence (AI) promises to transform clinical decision-making processes as AI has the potential to harness the vast amounts of genomic, biomarker, and phenotype data that is being generated across the health system including from health records and delivery systems, to improve the safety and quality of care decisions [1]. Today AI has been incorporated successfully into decision support systems (DSSs) for diagnosis in data-intensive specialties like radiology, pathology, and ophthalmology [2]. Future systems are expected to be increasingly more autonomous, going beyond making recommendations about possible clinical actions to autonomously performing certain tasks such as triaging patients and screening referrals [3, 4].

In this paper, we focus on the present role of AI in supporting clinical decisions. Evaluation of these machine-learning-based systems has tended to focus on examining the performance of algorithms in laboratory settings. There are a few observational studies which have tested systems in clinical settings providing a safe environment while patients continue to receive standard care [5]. However, little is known about the effects of AI on care delivery and patient outcomes which need to be carefully considered to ensure that it is appropriately applied. Like any technology, AI will also come with its unintended effects that may disrupt care delivery and pose risks to patients [6]. It is clear that the many benefits of AI cannot be realized safely unless AI is responsibly and effectively integrated into clinical decision-making and care delivery processes, and risks are effectively identified and mitigated.

While we have witnessed the beginnings of mainstream adoption of AI in only the last five years or so, academic researchers within the informatics community have studied AI for almost half a century [7-9]. However, an enhanced risk is now presented by a shift in the environment from draw-down of innovation based on positive evidence to political and commercial pressures for speedy adoption based on promise, with the call for evidence being presented as self-interested commitment to the status quo [10]. Clearly this disregard of evidence puts patients at risk.

The objective of this paper is to draw attention to key considerations for evaluating AI in clinical decision support; and to examine challenges and practical implications of AI design, development, selection, use, and ongoing surveillance. The paper begins with a historical perspective about the evaluation of AI in healthcare. It then examines key challenges of evaluating AI-enabled clinical decision-support. The final section deals with the practical aspects of evaluating AI, including approaches to evaluation and indicators to monitor AI. It will show that evaluation and ongoing surveillance is central to safe and effective integration of AI in complex sociotechnical settings.

2 Evaluation of AI in Healthcare – a Historical Perspective

AI in healthcare is not new, and neither is AI evaluation; the dedicated journal, *Artificial Intelligence in Medicine* dates back to the early 1990's. AI as a term was used for a collection of established technologies that were called decision support technologies, knowledge-based systems, or expert systems in earlier years. The health informatics community has a rich history and tradition of studying and evaluating the application of AI to solve problems in care delivery. Early applications were focussed on performing diagnosis and making therapy recommendations in medical settings. The early versions of AI mostly used symbolic approaches based on rules and knowledge, whereas present day AI uses statistical methods along with symbolic approaches to disease representation. In the late 1950's, McCarthy and colleagues developed the language LISP which was particularly well suited to symbolic approaches [11]. The PROLOG language for logic programming to manage reasoning and decision-making processes with logical structures was developed in the 1960's [12]. In the 1970's, there was a big wave of the so called first generation of AI in medicine [13], such as Shortliffe's work with MYCIN [14], Kulikowski's individualized clinical decision models [15], and de Dombal's computer-aided diagnosis of acute abdominal pain [16] which incorporated newer statistical reasoning methods such as probabilistic reasoning and neural networks.

Within the period covered above, there were extensive discussions about the role of AI in medicine, ranging from expert systems replacing doctors, to AI supporting doctors in their clinical decision-making processes, and AI embedded in medical devices. The latter has been successfully applied for several decades, e.g. in ECG machines and ICU equipment such as pumps, and insulin pens.

The need for rigorous evaluation of the quality and impact of AI was recognised in the 1980s. While the first publications focussed mainly on methodologies for evaluating performance in a laboratory environment, later papers addressed field-testing in clinical settings to examine effects on the structure, process, and outcome of care delivery [17-21]. The 1998 Helsinki workshop of our IMIA Working Group (then called the WG on *Technology Assessment and Quality Development, and Organizational and Social Issues*) focussed on the importance of addressing broader organisational issues and the need to focus on constructive evaluation within the whole lifecycle of information technology (IT) development [22-24]. Two books from 2006 which serve as textbooks for health IT evaluation methodologies specifically address techniques for AI as well as potential biases that may compromise an evaluation [25, 26]. Biases introduced during the development of algorithms due to differences between training and real-world populations are particularly important for machine learning-based systems and are discussed in section 3.

Later, following implementations of AI in clinical practice, studies shifted to the clinical impacts of AI [18] and on the related methodological challenges to find adequate clinical endpoints [19]. Building upon evidence about the benefit of AI in medicine, research then focussed on reviewing the impact of AI on patient outcomes in inpatient settings [20], in psychiatry [21], and medication safety [22].

The challenges recognised in the early days of applying AI were among others: 1) the legal (e.g., who is responsible when the system makes an error?) and ethical issues (algorithms dealing with discriminative investigations that are highly unethical, see for example [27]); 2) the context including informal information that is not documented in the medical record but is nevertheless part of a doctor's mental image about the patient [28]; 3) the transferability of algorithms from one setting to another both with respect to patient groups and clinical setting (i.e. local technologies/methodologies, local or regional professional culture [29]); 4) brittleness (inability of AI to reason at the boundaries or outside own application range, resulting in "sense in – garbage out" [30]; and 5) the dynamic nature of professional knowledge development in health care that needs to be captured in dynamic AI.

Today, there is an increasing and renewed interest in using AI, where a new generation

of clinical decision support is facilitated by the availability of powerful computing tools to manage big data and to analyse and generate new knowledge. Big data offers a challenge to connect molecular and cellular biology to the clinical world and to consider individual variations, as well as large volume responses and outcomes to similar presenting problems, rather than using population averages [31] or the restricted population of clinical trials. While such advances in technology may provide solutions to the problem of generalisation, they cannot solve all the other issues mentioned above. AI now incorporates a diverse range of computational reasoning methods to support clinicians in making decisions. Evaluation of these kinds of applications is even more critical as they start to penetrate healthcare on a product purchase or turn-key basis, where they may have an impact on the individual treatment of patients and care delivery, as well as on the liability of the individual healthcare professional. There are also risks related to these systems, they may be ill-functioning or may even have negative impacts on the outcome of care in a specialised unit or a different populations or health systems [32]. We look at some of these challenges in the next section.

3 Challenges in Evaluating AI for Clinical Decision Support

As with all interventions in healthcare we need to understand and evaluate the effects of AI on care delivery and patient outcomes. Evaluation is needed at each distinct stage in the IT lifecycle including design, development, selection, use, and ongoing surveillance.

Design and development of AI: Historically, evaluation of AI was limited to the design and development phase, as implementation and use of AI systems in routine clinical practice was rare. During design and development, evaluation concentrates upon the performance of the algorithms in terms of discrimination, accuracy, and precision. Depending on the use case, one performance measure might be more important than the other. For example, an algorithm used for triage needs high discrimination, while an algorithm that predicts mortality or complication risks in shared decision-making needs to be highly accurate and precise for all types of patients. Even at this stage of performance evaluation, we need to recognise that the algorithm may be "mathematically optimal but ethically problematic" [33]. A fundamental challenge is that AIs built using machine learning do not necessarily generalise well beyond the data upon which they are trained. Even in restricted tasks like image interpretation, AI can make erroneous diagnoses because of differences in the training and real-world populations, including new 'edge' cases, as well as differences in image capture workflows [4]. Another challenge is that algorithms need to reflect up-to-date knowledge given the dynamic changes of professional knowledge. Based on which knowledge do the algorithms work? What level of evidence is sufficient? How is regular adaptation to new professional knowledge organized?

Designers also need to consider if the computational finding is actionable in an ethical way or if it might disadvantage people from a particular socio-demographic background. For example, an algorithm to prioritise which patient should receive an organ donated for transplant might include expected longevity as a predictor variable. This might seem sensible on face value, but would ignore the implicit correlation with social determinants of health and lifespan [33]. Furthermore, the inferential logic should be put in context – is the reasoning behind the algorithm meaningful in the real world or only a statistical artefact? This question has been raised about the so-called "week-end effect" of apparent increases in hospital mortality [34]. Many studies have relied upon administrative data to investigate the association between time or day of admission and hospital mortality, which has resulted in inadequate adjustment for case mix difference between weekdays and weekends and consequently has produced inconsistent findings and conclusions.

Evaluation challenges also arise from the nature of AI and the manner in which it is developed. Some computational reasoning methods in AI such as neural networks are considered black boxes to end-users. Auditing has been proposed as a pragmatic approach to evaluating opaque algorithms that were devised autonomously. This follows an analogy to human judgement; typically we measure outcomes, not problem-solving style or cognitive process [35]. However, given the fundamental healthcare ethic of "first do no harm", we suggest more effort is needed in the design phases to explain the principles of a computational model to allow transparent assessment. This would help to keep clinicians and patients engaged and avoid conflict between practitioners and commercial algorithm developers (e.g. [36]). Algorithm developers, including those who operate on a proprietary basis, need to consider how to open the black box (even if partially) and work within a framework for shareable biomedical knowledge so that clinicians can judge the merits of AI models [37].

Selection and use of AI: Widespread availability of clinical data, easy to use AI development environments (e.g. Tensorflow, DeepLearning4J and Keras [38]) and online communities (e.g. healthcare.ai) have resulted in rapidly growing numbers of algorithms that have become available to clinicians. When multiple algorithms are available and one must be selected, it is important to evaluate any risks of data quality issues, and poor fit of the foundational data to a new situation, such as different population and morbidity patterns. What does the provenance of an algorithm tell us about its generalizability? For example, a computation that finds an association between blood pressure and complications in ICU patients where blood pressure is measured continuously might not be applicable to other hospitalised patients where blood pressure is measured perhaps only once a day; or between a coronary unit where pressure may be controlled in most patients and a general surgical ward with a different and non-controlled hypertensive population. Data generation is often a "social phenomenon" [39]: data may flow from unreliable processes [40] or human workarounds to mandatory entry fields [41].

We are also interested in the decision-making performance of humans with

and without AI assistance. Once an algorithm is developed, clinical validation of its utility is needed. The algorithm may be correct but is it operationally meaningful and useful? Does it fit clinical workflow? Does it still represent up-to-date clinical knowledge? Will it change clinical decisions? What level of confidence can be given? A significant concern here is that when humans are assisted by DSSs, they tend to over-rely and delegate full responsibility to the DSS rather than continuing to be vigilant. This is known as automation bias and can have dangerous consequences when the DSS is wrong or fails, or the presenting problem is subtly unique. Previous work has demonstrated the effects of automation bias with a prescribing DSS that was based on simple rules of logic [42]. In this study, the presence of the DSS reduced the verification of prescription safety and increased prescribing errors when the DSS was incorrect [43]. With AI, these effects of automation bias are likely to be exacerbated because of the black box nature of many machine learning approaches which are not conducive to verification. Methods are thus needed to mitigate automation bias, for example by improving the interpretability of machine learning models, explaining how an AI came to a conclusion, and training clinicians to maintain vigilance.

Consideration also needs to be given to the various stakeholders involved in processes to select AI. Evaluation findings need to be communicated in a manner that is understandable and meaningful to clinicians, but also to administrators and policymakers. One possible approach to support better selection of AI systems is by labelling. Neither patient nor clinician can be fully comfortable in the use of AI 'black box' decision support unless they know the system is appropriately created and relevant for their situation. Based on the transferability and quality schemas, label structures need to be developed which identify the training population type including demography and treatment setting, the nature of the decisions supported, the clinical environment, and any problems identified. It should be possible to devise a standard population characteristics data set that is verifiable, yet at the same time does not reveal technical design details and thus protects commercial confidence. Indeed population information may become mandatory as health systems are increasingly expecting AI developers to be transparent about the limitations and ethical examination of the population data used to develop algorithms, how data performance was validated, and how clinicians integrate into care delivery effective treatment and value for money [44]. Reporting to trusted third parties of any suspected adverse outcomes would also be valuable.

Ongoing surveillance of AI: AI will have systemic effects in complex sociotechnical settings, including aspects such as usage, usability, interpretability, up-to-date capabilities, safety, unintended effects, and ethical issues. Algorithms may affect resource allocation and prioritisation, so the risk of reinforcing bias in care delivery must be considered. Of course, the impact of AI on patient outcomes and experience and on clinical experience, as well as both organisational and social impacts, all need to be monitored. Over time, the context, treatment possibilities, and patient population might change. Therefore, once implemented, ongoing surveillance is needed to monitor and recalibrate AI algorithms [45]. This surveillance is also needed for dynamic algorithms that continuously update themselves based on practice data and published clinical evidence. Another aspect is in detecting and evaluating "hidden" errors – subtle flaws in inferred data or a misconfiguration that may not be as obvious as the financial "flash crash" of 2010 [46] which may take months to be revealed (as with hand-crafted algorithms [47]). Thus, evaluation is likely to shift from a one-off activity to a continuous process to ensure that the use of AI, including those incorporating dynamic algorithms, is meeting expectations.

4 Practical Aspects of Evaluating AI-enabled Clinical Decision Support

This section examines some of the practical aspects of evaluating AI-enabled clinical decision support. It begins by considering broad paradigms and frameworks to approach such evaluations. We reviewed existing guidelines for conducting and reporting evaluation studies, and concluded with exemplar indicators to monitor AI. The aim is to highlight existing approaches that can be readily applied and to identify areas where further work is required to meet the specific needs of current AI.

Approaching AI evaluation: As we discussed in the previous section, evaluating AI as a one-off activity will not be sufficient, therefore continuous evaluation or surveillance might become necessary to monitor the emergent behaviour of AI in complex sociotechnical settings, and the response of its users. Indeed, it may become an ethical imperative as the complexity of interventions and their effects on sociotechnical interactions become impossible to predict in advance. One paradigm that might be particularly relevant for approaching the evaluation of AI on an ongoing basis is the Learning Healthcare System [48]. This paradigm usefully incorporates the notion of continuous system improvement using locally generated evidence to inform practice changes. Such learning can occur at different levels including institutional, national, or international. While aspects like algorithm performance may be addressed at institutional level, safety governance might be considered at a national level and the ethical implications of using AI for clinical decision support could be tackled at an international level.

Evaluation guidelines and models: Existing resources such as the guidelines for Good Evaluation Practice in Health Informatics (GEP-HI) provide a solid foundation for planning and executing evaluation projects [49]. Validity of GEP-HI for AI and its completeness will become apparent through application and use. Similarly, the Statement on Reporting of Evaluation Studies in Health Informatics (STARE-HI) provides useful principles for publication of evaluation studies [50, 51]. These guidelines support design, execution, and reporting of an evaluation study irrespective of the type of the system under study. The evaluation question and the purpose of the study guide the selection of evaluation methods and their application.

Very important evaluation criteria for AI applications include validity of the system, i.e. correctness in reasoning, usefulness for the clinical practice and effects and impacts on clinical work, care processes, and patient outcomes. With AI applications that are based on big data sets, especially genomic, biomarker, and phenotype data from across the health system, it is important to ensure sufficient coverage, specificity, and validity of the data. This will be an additional step in the evaluation process with AI systems i.e. to assess and collect evidence on data quality, and is the key point to prevent unintended consequences and to avoid risks of harmful results.

Understanding of the comparability of the design and development situation with the potential implementation situation is crucial. Aspects such as the algorithm training data set (clinical situation, physiology, and demography) are critical, as are treatment options and user population (including professional and e-health skills). A lot can be learned from Implementation science; for instance Schloemer and Schröder-Bäck have created the useful generic PIET model, in which Population, Intervention, (clinical) Environment, and Transfer methods are decomposed [52]. Adaptation of the sub-categories of each to fit the AI situation would yield a valuable assessment tool.

Indicators to monitor AI: Implementation of AI in an organisation can have an impact on several aspects of care quality, depending on the clinical area where implemented. Existing methodologies for defining indicators for continuous monitoring of health IT help not only to pinpoint AI-specific care quality indicators, but also to define AI-specific indicators for AI-quality monitoring and surveillance, for example, the methodology for the Nordic eHealth Indicators which has four major phases including definition of the context of measurement, monitoring goals (e.g. care quality, efficiency, technology quality …etc.), methods for indicator selection, and the collection, analysis, and feedback of data from relevant user groups (Box 1) [53].

Governing AI in healthcare may involve trade-offs between risk and rewards. Questions on ethics, safety, and human values at stake depend on the type and purpose of use of the technologies, emphasizing the importance of the first two phases of the methodology. Peer-reviewed literature on AI for specific situations remains one solid source for practical indicator selection and grouping.

Box 1 Defining indicators for monitoring AI.

1	**Define context for measurement**
1.1	Define AI application • Identify AI components and functions within operational system • Define outputs (which may be inputs to another component) and risks
1.2	Identify service context and boundaries (on micro-, meso- and macro levels).
1.3	Identify key stakeholders.
2	**Define monitoring goals** (e.g. safety, usability, transparency, up-to-dateness, reliability, patient safety, work process changes).
3	**Define methods for indicator selection and categorisation** e.g. by reviewing expert knowledge, peer-reviewed literature, or existing indicator work.
4	**Define data for repeated measurements, analysis, and feedback from different user groups** • Produce objective reports.

Availability (state of the art) of custom AI applications for specific purposes in different contexts of use for different user groups provides the first level of monitoring, important for policy makers for local and national steering. If these data are collected systematically from a whole country, they can also be used as benchmarking data by policy makers to steer development and by healthcare providers to find best practices and learn about different AI solutions.

The information system (IS) success model also offers an option for grouping indicators to examine AI quality, focussing on system quality, information quality, service quality, use, user satisfaction, and outcomes [54-56]. Each of these are examined below:

System quality: Once implemented, AI system requires ongoing surveillance based on a set of measures to recalibrate AI algorithms. Apart from surveillance targeted directly at AI development, *IT System quality* monitoring includes indicators like (IT) accessibility, reliability, flexibility, integration, response time, ease of learning [55]. Measures and indicators for these are targeted at monitoring and development of the entire socio-technical system where AI is implemented. Function quality is an important part of system quality, particularly when AI components feed other interim processes. AI applications with direct impact on patient health can be approached by the EU directive for medical devices, including the requirements for CE marking with tight clinical evidence requirements for Class II and III devices according to Regulation 2017/745 [57]. System quality is directly associated with patient safety, and existence of CE marking could be one indicator for AI quality and a proxy for patient safety.

Information quality may be the most important surveillance and monitoring domain for AI algorithms alongside with system quality. It covers quality of data used as input for the AI as well as quality of the output information (e.g. accuracy, reliability, relevance, completeness) [55]. In addition to these, reproducibility, sensitivity and specificity of the results, type and severity of errors, observed versus expected error rates, causes of errors, how algorithms respond and what indicators to use if input data characteristics start vary, and the availability of data, code, and workflows need to be focussed on, since the algorithms are

trained with certain data, and their soundness needs to be confirmed in the context of use in relation to the quality of actual clinical data to be used [58].

Service quality in the IS success model refers to help desk-type support available for users as well as long-term feedback from users for both immediate updates as well as for the entire system development.

System use refers to utilization (e.g. use/non-use, frequency of use) of the system output [55]. AI in healthcare can be used for decision support as well as automating actions of care (e.g., medication administration). Existing OECD and EU eHealth monitoring surveys focus on the availability and use of HIS [59-61]. Functionalities for diagnostic or treatment support (for professionals or patients) are not covered in the current surveys, and need revisions to encompass AI-enabled functionalities.

User satisfaction/acceptance refers to user perceptions about system output [55]. Validated satisfaction scales miss indicators for trust and competence of the personnel to deal with the decisions made by AI (e.g. to identify false results, and to record override and its contemporaneous justification). The competence of professionals to use commonly agreed documentation standards, structures, and classifications may also be essential, since it impacts the quality of documentation (input for AI). From the viewpoint of patients, barriers to use, including trust and difficult instructions are likely to become increasingly important with growing numbers of AI implementations and wider settings. User competence to evaluate decisions offered by AI would gain crucial importance in future.

Outcomes refer to the individual and organizational impacts [55]. AI promises to enhance healthcare effectiveness and efficiency, but requires due diligence to validate actions, optimise their use, and minimize risks. Outcome indicators need to reflect the purpose of the algorithm, and can include appropriateness of AI-assisted decisions, efficiency in care process, as well as patient acceptance and adherence. Patient outcomes are increasingly important to monitor along with patient safety incidents. Monitoring risks/adverse effects/trade-offs is an essential part of monitoring AI outcomes, including equity and inclusion, data privacy and security, algorithmic bias, and representativeness of data [62].

A broader outlook is also essential, rather than solely focussing on the quality of algorithms or quality of their outcomes. The presence of governance structures (e.g. transparency of processes and policies to ensure reproducibility for large scale computational models, regulation related to safe and secure use of data, liability and accountability questions, data ownership), education and training, as well as national development and validation procedures can be used as indicators of system maturity on national level.

5 Conclusions

Technological developments are outpacing our ability to predict the effects of AI on the practice of medicine, the care received by patients, and the impact on their life. In the immediate future, we can expect AI to support clinical decisions with humans in the decision loop. The dynamics of decision-making processes are likely to be altered with clinicians needing to weigh AI-generated advice against other evidence and patient preferences [1]. To ensure the safe and effective integration of AI in care delivery, a robust commitment to evaluation is critical. Evaluators can draw important lessons from past efforts and should build upon current best practice frameworks, evaluation guidelines, and methods to guide evidence-based design, development, selection, use, and ongoing surveillance of AI in clinical decision support. Labels will be important for defining source-training data so as to identify optimal transferable use patterns. Specific enhancements required for evaluating dynamic algorithms incorporating vast amounts of genomic, biomarker, and phenotype data will emerge through practical application.

References

1. Coiera E. The fate of medicine in the time of AI. Lancet 2018 Dec 1;392(10162):2331-2.
2. Yu KH, Kohane IS. Framing the challenges of artificial intelligence in medicine. BMJ Qual Saf 2019;28:238-41.
3. Yu K-H, Beam AL, Kohane I.S. Artificial intelligence in healthcare. Nat Biomed Eng 2018;2(10):719-31.
4. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. BMJ Qual Saf 2019;28:231-7.
5. Kanagasingam Y, Xiao D, Vignarajan J, Preetham A, Tay-Kearney M-L, Mehrotra A. Evaluation of Artificial Intelligence–Based Grading of Diabetic Retinopathy in Primary CareArtificial Intelligence–Based Grading of Diabetic Retinopathy in Primary Care Artificial Intelligence–Based Grading of Diabetic Retinopathy in Primary Care. JAMA Netw Open 2018;1(5):e182665-e.
6. 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 2017;24(2):246-50.
7. Shortliffe EH. The adolescence of AI in medicine: will the field come of age in the '90s? Artif Intell Med 1993;5(2):93-106.
8. Coiera EW. Artificial intelligence in medicine: the challenges ahead. J Am Med Inform Assoc1996;3(6):363-6.
9. Patel VL, Shortliffe EH, Stefanelli M, Szolovits P, Berthold MR, Bellazzi R, Abu-Hanna A. The coming of age of artificial intelligence in medicine. Artif Intell Med 2009;46(1):5-17.
10. Benber B, Lay K. Health secretary Matt Hancock endorses untested medical app; The Times, 17 September 2018, available at https://www.thetimes.co.uk/article/matt-hancock-endorses-untested-health-app-3xq0qcl0x
11. McCarthy J. Recursive Functions of Symbolic Expressions and Their Computation by Machine. Commun ACM 1960;3(4):184–95.
12. Colmerauer A, Roussel P. The birth of Prolog. ACM SIGPLAN Notices. 1993;28(3):37.
13. Kulikowski, C.A. An Opening Chapter of the First Generation of Artificial Intelligence in Medicine: The First Rutgers AIM Workshop. Yearb Med Inform 2015;10(1):227-33.
14. Shortliffe EH, Davis R, Axline SG, Buchanan BG, Green CC, Cohen SN. Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the MYCIN system. Comput Biomed Res 1975 8(4):303-20.
15. Sonnenberg FA, Hagerty CG, Kulikowski CA. An architecture for knowledge-based construction of decision models. Med Decis Making 1994;14(1):27-39.
16. de Dombal FT, Leaper DJ, Staniland JR, McCann AP, Horrocks JC. Computer-aided diagnosis of acute abdominal pain. Br Med J 1972;1(2(5804)):9-13.
17. Nykänen P, Chowdhury S, Wigertz O. Evaluation of decision support systems in medicine. Comput Methods Programs Biomed 1991;34(2/3):229 - 38.
18. Wyatt JC, Spiegelhalter D. Evaluating medical expert systems: What to test and how? Int J Med Inform 1990;15:205-17.
19. Clarke K, O'Moore R, Smeets R, Talmon J, Brender J, McNair P, et al. A Methodology for Evaluation of Knowledge-Based Systems in Medicine.

In: van Bemmel J, McCray AT, editors. Yearbook of Medical Informatics1995. p. 513-27.
20. Yu VL, Buchanan BG, Shortliffe EH, Wraith SM, Davis R, Scott AC, et al. Evaluating the performance of a computer-based consultant. Comput Programs Biomed 1979 9(1):95-102.
21. van Gennip EM, Talmon JL, Bakker AR. ATIM, accompanying measure on the assessment of information technology in medicine. Comput Methods Programs Biomed 1994;45(1-2):5-8.
22. Brender J. Methodology for constructive assessment of IT-based systems in an organisational context. Int J Med Inform 1999;56:67-86.
23. Nykänen P, Enning J, Talmon J. Inventory of validation approaches in selected health telematics projects. Int J Med Inform 1999;56:87-96.
24. van Gennip E, Lorenzi NM. Results of discussions at the IMIA WG 13 and 15 working conference. Int J Med Inform 1999;56:177-80.
25. Brender J. Handbook of evaluation methods for health informatics. Burlington, MA: Elsevier Academic Press; 2006.
26. Friedman CP, Wyatt JC. Evaluation Methods in Medical Informatics. 2nd ed. New York: Springer; 2006.
27. Karthaus V, Thygesen H, Egmont-Petersen M, Talmon J, BrenderJ, McNair P. User-requirements driven learning. Comput Methods Programs Biomed.1995;48(1-2):39-44.
28. Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. BMJ 2018;361:k1479.
29. Nolan J, McNair P, Brender J. Factors influencing the transferability of medical decision support systems. Int J Biomed Comput 1991;27(1):7-26.
30. Feigenbaum EA, editor. Autoknowledge: from file server to knowledge servers. MEDINFO; 1986; Amsterdam: Elsevier Science Publishers BV; 1986.
31. Nykänen P, Zvarova J. Big data challenges for personalised medicine. Editorial. International Journal of Biomedicine and Healthcare 2015;3(1):1.
32. Ammenwerth E, Shaw N. Bad health informatics can kill - is evaluation the answer? Methods Inf Med 2005;44:1-3.
33. DeDeo S. Wrong side of the tracks: Big Data and Protected Categories. arXiv preprint arXiv:14124643. 2014.
34. Bray BD, Steventon A. What have we learnt after 15 years of research into the 'weekend effect'? BMJ Qual Saf 2017;26(8):607-10.
35. Sandvig C, Hamilton K, Karahalios K, Langbort C. Auditing algorithms: Research methods for detecting discrimination on internet platforms. Data and discrimination: converting critical concerns into productive inquiry; 2014. p. 1-23.
36. Crouch H. RCGP chair says GPs are not 'technophobic dinosaurs'. Digital Health; 2018 [Available from: https://www.digitalhealth.net/2018/10/rcgp-chair-technophobic-dinosaurs/.
37. Lehmann HP, Downs SM. Desiderata for sharable computable biomedical knowledge for learning health systems. Learn Health Syst 2018:e10065.
38. Bhattacharya S, Czejdo B, Agrawal R, Erdemir E, Gokaraju B, editors. Open Source Platforms and Frameworks for Artificial Intelligence and Machine Learning. SoutheastCon 2018; 2018 19-22 April 2018.
39. Osoba OA, Welser IV W. An intelligence in our image: The risks of bias and errors in artificial intelligence. Rand Corporation; 2017.
40. Burnett S, Franklin BD, Moorthy K, Cooke MW, Vincent C. How reliable are clinical systems in the UK NHS? A study of seven NHS organisations. BMJ Qual Saf 2012;21(6):466-72.
41. Ser G, Robertson A, Sheikh A. A qualitative exploration of workarounds related to the implementation of national electronic health records in early adopter mental health hospitals. PLoS One 2014;9(1):e77669.
42. Lyell D, Magrabi F, Raban MZ, Pont LG, Baysari MT, Day RO, et al. Automation bias in electronic prescribing. BMC Med Inform Decis Mak 2017;17(1):28.
43. Lyell D, Magrabi F, Coiera E. Reduced Verification of Medication Alerts Increases Prescribing Errors. Appl Clin Inform 2019;10(1):66-76.
44. NHS code of conduct for data-driven health and care technology, 19 February 2019: https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology/initial-code-of-conduct-for-data-driven-health-and-care-technology
45. Minne L, Eslami S, de Keizer N, de Jonge E, de Rooij SE, Abu-Hanna A. Effect of changes over time in the performance of a customized SAPS-II model on the quality of care assessment. Intensive Care Med 2012;38(1):40-6.
46. Treleaven P, Galas M, Lalchand V. Algorithmic trading review. Commun ACM 2013;56(11):76-85.
47. Crouch H. East and North Herts could face £7m bill to fix Lorenzo issue: Digital Health; 2018 [Available from: https://www.digitalhealth.net/2018/10/east-and-north-herts-lorenzo-it-issue/.
48. Friedman C, Rigby M. Conceptualising and creating a global learning health system. Int J Med Inform 2013;82(4):e63-71.
49. Nykänen P, Brender J, Talmon J, de Keizer NF, Rigby M, Beuscart-Zephir M, et al. Guideline for good evaluation practice in health informatics (GEP-HI). Int J Med Inform 2011;80(12):815-27.
50. Talmon J, Ammenwerth E, Brender J, de Keizer N, Nykanen P, Rigby M. STARE-HI -statement on reporting of evaluation studies in health informatics. Yearb Med Inform 2009:23-31.
51. Brender J, Talmon J, de Keizer N, Nykanen P, Rigby M, Ammenwerth E. STARE-HI - Statement on Reporting of Evaluation Studies in Health Informatics: explanation and elaboration. Appl Clin Inform 2013;4(3):331-58.
52. Schloemer T, Schroder-Back P. Criteria for evaluating transferability of health interventions: a systematic review and thematic synthesis. Implement Sci 2018;13(1):88.
53. Hyppönen H, Faxvaag A, Gilstad H, Hardardottir GA, Jerlvall L, Kangas M, et al. Nordic eHealth Indicators: Organisation of research, first results and the plan for the future [Internet]. Copenhagen: Nordic Council of Ministers; 2013. Available from: http://urn.kb.se/resolve?urn=urn:nbn:se:norden:org:diva-675.
54. Canada Health Infoway Benefits Evaluation Indicators Technical Report version 2.0. 2012.
55. DeLone WH, McLean ER. Information systems success: The quest for the dependent variable. Inf Syst Res 1992;3(1):60–95.
56. DeLone WH, McLean ER. The DeLone and McLean Model of Information Systems Success: A Ten-Year Update. J Manage Inf Syst 2003;19(4):9-30.
57. Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC: https://publications.europa.eu/en/home.
58. Artificial Intelligence for Health and Health Care, JSR-17-Task-002, JASON, The MITRE Corporation 2017: https://fas.org/category/artificial-intelligence/.
59. Draft OECD Guide for Measuring ICTs in the Health Sector, Paris: OECD. COM/DELSA/DSTI(2013)3/FINAL; available at: http://www.oecd.org/health/health-systems/Draft-oecd-guide-to-measuring-icts-in-the-health-sector.pdf. 2013.
60. Adler-Milstein J, Ronchi E, Cohen GR, Winn LA, Jh, AK. Benchmarking health IT among OECD countries: better data for better policy. J Am Med Inform Assoc 2014;21(1):111-6.
61. Codagnone C, Lupiañez-Villanueva F. Benchmarking Deployment of eHealth among General Practitioners Final report; 2013.
62. Lannquist Y. Ethical & Policy Risks of Artificial Intelligence in Healthcare. The Future Society; 2018.

Correspondence to:
A/Prof. Farah Magrabi
Centre for Health Informatics
Australian Institute of Health Innovation
Macquarie University
Level 6, 75 Talavera Road
Macquarie University NSW 2109
Australia
Tel: +61 2 9850 2429
E-mail: farah.magrabi@mq.edu.au