Mind the gap! Overcoming the lack of evidence to support the innovation journey of AI in healthcare

Milou Evelien Wilhelmijn Marie Silkens, Maura Leusder, Nicholas Woznitza, Harry Scarbrough

Submitted to: Journal of Medical Internet Research
on: June 10, 2024

Table of Contents

Mind the gap! Overcoming the lack of evidence to support the innovation journey of AI in healthcare

Milou Evelien Wilhelmijn Marie Silkens[1, 2] PhD; Maura Leusder[2] MSc; Nicholas Woznitza[3, 4, 5] PhD; Harry Scarbrough[1] PhD

[1]Centre for Healthcare Innovation Research City University of London London GB
[2]Erasmus School of Health Policy and Management Erasmus University Rotterdam NL
[3]Lungs for Living University College London London GB
[4]University College London Hospitals London GB
[5]School of Allied and Public Health Canterbury Christ Church University Kent GB

Corresponding Author:
Nicholas Woznitza PhD
Lungs for Living
University College London
Division of Medicine, Rayne Building
5 University Street
London
GB

Abstract

The rapid development of artificial intelligence (AI) applications for the healthcare setting confronts providers and practitioners with the challenge of choosing those applications that have the best chance to reduce the burden of care in their context. This is challenging due to a general lack of evaluation metrics and because the evidential claims provided by AI vendors are not always in line with the forms of evidence needed by healthcare providers and practitioners. This evidence gap currently harms the development of trust in and acceptability of AI, and thereby hampers the successful implementation and adoption of AI in healthcare. In this viewpoint, we argue that closing this evidence gap is crucial to helping AI achieve its full potential in the healthcare context and we provide practical guidance towards this objective.

(JMIR Preprints 10/06/2024:63087)
DOI: https://doi.org/10.2196/preprints.63087

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Viewpoint

Mind the gap! Overcoming the lack of evidence to support the innovation journey of AI in healthcare

Authors & affiliations:

Milou E.W.M. Silkens[1,2], Maura Leusder[1], Nick Woznitza[3,4,5], Harry Scarbrough[2]

[1] Erasmus School of Health Policy and Management, Erasmus University, Rotterdam, the Netherlands

[2] Centre for Healthcare Innovation Research, City University of London, London, the United Kingdom

[3]Lung for Living, UCL Division of Medicine, London, the United Kingdom

[4]University College London Hospitals, London, the United Kingdom

[5] School of Allied and Public Health, Canterbury Christ Church University, Kent, the United Kingdom

Corresponding author:

Nicholas Woznitza

n.woznitza@ucl.ac.uk

Classification: Internal

Abstract

The rapid development of artificial intelligence (AI) applications for the healthcare setting confronts providers and practitioners with the challenge of choosing those applications that have the best chance to reduce the burden of care in their context. This is challenging due to a general lack of evaluation metrics and because the evidential claims provided by AI vendors are not always in line with the forms of evidence needed by healthcare providers and practitioners. This evidence gap currently harms the development of trust in and acceptability of AI, and thereby hampers the successful implementation and adoption of AI in healthcare. In this viewpoint, we argue that closing this evidence gap is crucial to helping AI achieve its full potential in the healthcare context and we provide practical guidance towards this objective.

Keywords: artificial intelligence, evidence, machine learning, digital health, healthcare innovation

Introduction

A large and growing number of artificial intelligence (AI) applications have been developed and launched for healthcare in recent years[1]. These innovations claim to offer significant advances over current services in healthcare relating to the quadruple aim[2], namely cost reduction, reduced workforce pressures, improved productivity, and safety and quality of care. Such advances are much needed as healthcare systems face significant workforce shortages and consistently rising, economically unsustainable resource demands[3]. These shortages, and overall resource scarcity in the healthcare landscape, represent a wicked problem for managers facing complex investment decisions with uncertain outcomes under financial and political pressures[4]. Consequently, healthcare providers and practitioners face the 'huge challenge' of needing to evaluate and choose among many new AI applications to reduce the burden of care in their specific context, with insufficient clinical and system validation evidence[5]. This is particularly challenging because the relevant evaluation metrics are largely unknown[6], and the evidential claims provided by AI vendors currently fall short of the evidence forms demanded by healthcare providers and practitioners[5]. In this viewpoint, we define and evaluate the evidence gap for AI applications in healthcare and explore this gap by comparing the evidence required and the evidence currently available for AI applications in healthcare settings. We then discuss how to start closing this evidence gap by identifying lessons learned from the development and implementation of previous disruptive technologies that faced similar unique challenges, such as audit and feedback applications like dashboards[7]. Closing AI's evidence gap can help to reduce the 'chasm' between AI-development and AI-implementation and adoption in healthcare[8], reducing global spending and time-to-market for relevant applications that can make a difference. Although closing this gap may seem easier said than done due to AI's complex nature and particular challenges, there are parallels to be drawn between the nature and implementation of AI applications and previous disruptive technologies that faced similar unique challenges as AI. We will therefore finish with recommendations for future work that emphasize the importance of the adopting context in generating the embedded clinical and system evidence needed to evaluate whether particular AI applications are capable of addressing the wicked problem of resource scarcity highlighted above.

The AI evidence gap in healthcare

AI applications are considered one of the most disruptive technologies to date[9]. The development, but

especially the implementation and adoption of AI in a complex context such as healthcare, is therefore a time-consuming, potentially expensive, and challenging process[6,10] that requires trans-disciplinary solutions[4]. Some of these challenges are well described in the literature and can, amongst others, be of a regulatory, ethical, technical, societal, psychological, structural and financial nature[11]. They can be a significant barrier to the implementation and use of AI, causing perceived uncertainty and ambiguity of how AI applications function or interact. Furthermore, the commercial organisations that are often in the lead of the development and initial testing of AI applications for healthcare can be reluctant to reveal too many details as this could jeopardize their competitive advantage over similar companies[1]. As a result, AI applications currently face greater mistrust from prospective adopters and users than conventional applications (e.g. predictive analytics, dashboards). This puts an increased pressure on healthcare managers, practitioners, and providers to use accepted forms of evidence to overcome this mistrust, to select appropriate AI applications for their context[10], and to build a strong case for the deployment of AI applications in healthcare[5].

Based on the work of Mathews et al.[5], we can distinguish three types of validity evidence that, if acquired during a transparent, standardized, and rigorous process, can contribute to creating the aforementioned necessary insight into the real-life embedded performance of AI applications in healthcare. Firstly, *technical validation* provides evidence on the accuracy and precision of an AI application, including information on security and interoperability[5]. Secondly, *clinical validation* provides insights in whether the AI application contributes to improved clinical outcomes when used in the real-world clinical setting[5]. Thirdly, *system validation* refers to the extent to which the AI application is successfully integrated into clinical pathways, patients' lives, and broader healthcare systems[5]. This includes real-world utility, which means a reduction in overall morbidity, costs or workload upon deployment[2].

Technical, clinical, and system validation evidence are three critical success factors for effective implementation and adoption of AI in healthcare[5]. Yet, technical validity evidence is thus far the most common type of evidence provided by AI manufacturers[10], and much less attention has been paid to the other evidence types. There seem to be multiple factors explaining this predominant role for technical validation. Firstly, such evidence is relatively easy to acquire once an AI-algorithm is developed as its technological properties and performance can be tested on an existing dataset[5]. Secondly, the innovation process of AI – ranging from the development of the technology to its embedding and acceptance into clinical pathways[12,13] – is still predominantly conceptualized as a

linear process in which the technical validation is most prominent at the start and provides the initial evidence required to facilitate market-launch (Figure 1). Thirdly, the rapid technological developments underlying AI applications have outrun the development of accompanying regulatory frameworks with the result that AI manufacturers have been able to produce and market AI applications while fulfilling only the bare minimum requirement of validation[5,14]. This has enabled AI developers motivated by commercial concerns to exaggerate the technical effectiveness of AI[15] or to market their AI applications as effective, whereas the AI might not (yet) work as promised, or only under conditions representative of the training data.

Figure 1. The innovation-journey of AI (blue) and the corresponding evidence necessary at each stage of the journey (orange).

Compared to technical evidence, clinical and system validity evidence is much harder to gather and is therefore only available for a small subset of AI applications in healthcare[16]. Acquiring this kind of evidence is particularly challenging as it requires the collection of real-world data relating to actual AI usage in context. This means the AI application needs to be implemented in the clinical context in which it is expected to perform, and real-time data must be collected to assess the application's performance and to determine if the AI will improve patient care, outcomes, or processes. This is not only time-consuming, but also challenging when upfront investment costs are high, and mistrust due to the lack of evidence prevents sufficient use of the technology. Furthermore, local differences in patient case-mix as well as practice pose significant contextual nuances that can heavily impact the performance of AI applications, especially when trained on data originating from another context[1,16]. Overall, this can result in negative first experiences and calibration issues that pose significant hurdles to implementation for the purpose of evidence generation. The wider context in which clinical decisions are made, including how information is integrated and combined by healthcare

practitioners, the available systems, pathways, and resources to act on information and available treatments mean there is much more to AI improving outcomes than its technically validated "accuracy".

This lack of available evidence on how AI applications interact with the clinical context in which they are implemented, represents the so-called evidence gap. This evidence gap suggests that current AI applications marketed today are not always accurate, robust or useful from a clinical practice perspective[14]. Moreover, without accepted clinical and system validation evidence, the implementation and adoption of AI applications is more likely to fail[5,17]. Healthcare managers, providers and practitioners thus need transparent, real-world testing of AI applications in situ to generate situational and external validity evidence[18], which can help them to choose suitable AI for their needs. This will enable them to trust AI[19], to better understand how the AI works and to understand the impact that AI applications have on healthcare pathways and clinical outcomes[8,16].

Closing the evidence gap

To theorize how future research can close the AI-evidence gap, we draw parallels between the nature and implementation of AI and audit-and-feedback technologies like performance dashboards[20,21]. Such technologies, like AI applications, evaluate and visualize data on local processes or outcomes, and provide situated guidance intended to steer clinical action[7,22].

Both types of applications share their 'black boxed' nature, meaning that the underlying mechanisms through which (effective) outcomes are achieved can be obscure, as staff can e.g. ignore or override the advice provided by the application if they think it is unsuitable to the given circumstances. Lack of trust, testing, or embedding of these technologies have all been found to act as barriers to use, preventing causal changes in outcomes, processes, or performance[7]. For these reasons, we argue that the implementation of non-AI digital performance applications can provide significant insights for the production of relevant evidence relating to AI implementation.

One key lesson from the implementation of audit and feedback applications is that the commonly conducted quantitative studies, designed to establish sought-after validity evidence, were often insufficient to gain a good understanding of the performance of the applications in situ[7]. One reason for this was that the performance of these applications varied across contexts, and quantitative investigations were not sensitive enough to the role of these contexts[23]. As a result, qualitative and mixed-method approaches gained popularity, and supplemented quantitative studies to establish a

holistic understanding of the performance of and mechanisms underlying audit and feedback applications. Furthermore, realist evaluations were used to understand *how, why* and *under what circumstances* the applications worked or not[21], rather than only trying to establish *whether* the audit and feedback applications were effective, on average, in a technical sense. Whilst large-scale, quantitative studies were able to establish average effects, they were unable to explain why some implementations were highly effective and others failed.

In closing the evidence gap for AI applications, it is crucial that we take these lessons forward to encourage the development of a grounded understanding of how the AI functions within a local context[15]. We should embrace and adopt approaches that emphasize context and circumstance in generating embedded clinical and system evidence to evaluate if, how and why AI might address the wicked problem of resource scarcity in the current healthcare landscape[24]. This is particularly relevant for AI applications as they interact with the healthcare system in which they are implemented and continuously evolve together with these systems, rather than being stand-alone technologies that can be imposed onto healthcare contexts to achieve desired effects. This suggests that the innovation-journey needs to be seen as iterative and continuous, instead of linear, and as being affected by the context (i.e. the system) in which the AI is implemented.

Based on these insights, we propose a new way of viewing the role of validity evidence within this iterative and continuous innovation-journey of AI (Figure 2). As Figure 2 suggests, both the innovation-journey of AI and the acquisition of relevant evidence are inextricably linked via a cyclical and interconnected process in which each stage can inform and impact the other stages. For example, a clinical validation may clarify that AI's training data does not match real-world patient populations in particular contexts, thereby hampering implementation and adoption. This could subsequently inform and improve a following technical validation, increasing the accuracy, feasibility, and accuracy of the AI in various clinical contexts and thus improving the chances of successful implementation.

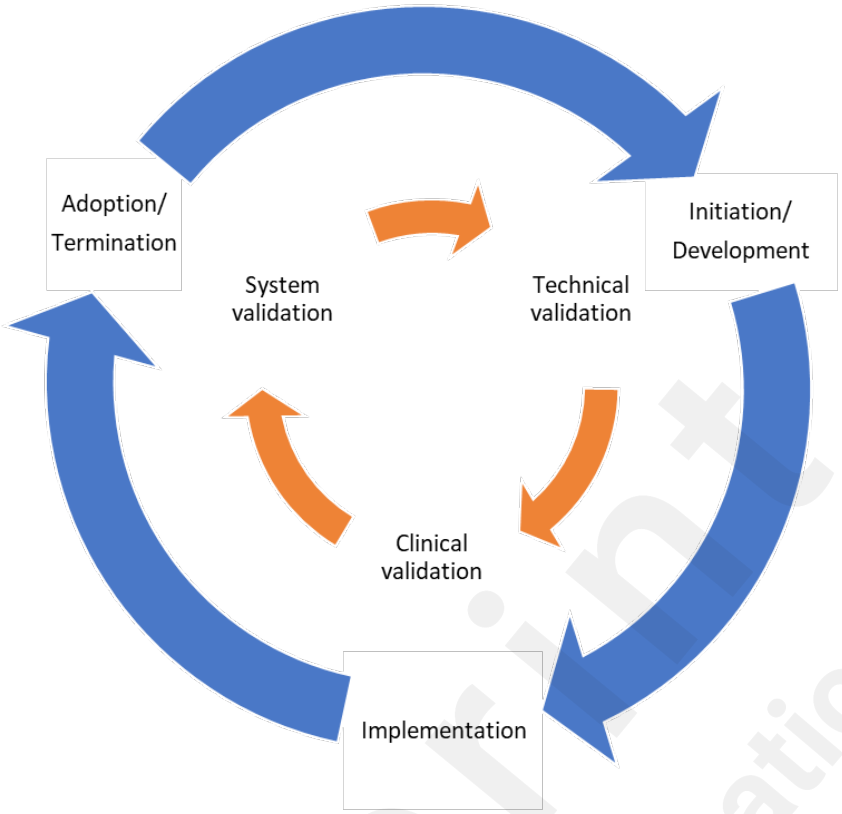


Figure 2. The iterative innovation-journey of AI (blue) and the corresponding evidence necessary throughout the innovation-journey (orange).

We see an opportunity to take these lessons learned forward by integrating them into governance and legislation frameworks for AI applications in healthcare. We are in dire need of standardized regulatory frameworks that hold AI applications in healthcare to the same or – considering that AI applications might access sensitive personal medical information – even stricter standard of evidence than other diagnostic and therapeutic tools and applications used in medicine, such as medical devices[16]. This means that AI applications would need to be subjected to technical and clinical investigations and then registered before bringing them to the market. In agreement with Vasey et al.[8], we argue for the introduction of early and small-scale clinical evaluations that sit between the in-silico algorithm development and the larger-scale clinical trials. The focus of such evaluations would be to assess the real-world impact of an AI application on its users' decisions at an early stage when changes to the application are still feasible, as this has been found to facilitate the implementation of performance dashboards[7,8]. Furthermore, this could help to identify the application's safety profile therefore avoiding exposing large groups of patients to harm[8]. Regulatory frameworks should also dictate the route to legal approval for clinical use, such as a CE marking, which would most likely depend on the risk associated with the AI. Such frameworks would hold the AI-manufacturers' accountable for tracking the performance of the AI longitudinally within the various implementation

contexts, to handle complaints, and to improve the AI, thereby delivering the much-needed system level validation evidence. By making sure upcoming regulatory frameworks require ongoing technical, clinical and systems validity evidence on a continuous basis, we not only gain a clearer understanding of what AI applications can realistically achieve, but also make it more difficult for AI developers to embellish their AI's performance for profitability.

Conclusion

The current evidence gap of AI applications in healthcare poses a significant challenge to healthcare providers and practitioners who have to choose among new applications with uncertain efficacy. Without accepted and appropriate technical, clinical, and system validation the implementation and adoption of these applications is unlikely to succeed. To close AI’s evidence gap, we should draw on key lessons learned from the implementation of previous disruptive technologies that bear semblance to AI applications. This includes encouraging the gathering of contextually specific evidence to help us understand how, why and under what circumstances AI applications are fully successful or not. Implementing standardized regulatory frameworks mandating the longitudinal collection of technical, clinical, and system validity evidence is imperative to ensure robust accountability for AI manufacturers. Such frameworks not only foster accountability but also enable practitioners and society to develop a more comprehensive and realistic understanding of AI's potential in healthcare.

References

1. Gilbert FJ, Smye SW, Schönlieb C-. Artificial intelligence in clinical imaging: a health system approach. *Clin Radiol*. 2020;75(1):3-6. doi:10.1016/j.crad.2019.09.122

2. Seneviratne MG, Shah NH, Chu L. Bridging the implementation gap of machine learning in healthcare. *BMJ Innov*. 2020;6(2):45. doi:10.1136/bmjinnov-2019-000359

3. Torbay R. The Backbone Of Our Health System Is Breaking. *Health Aff*. 2023;42(6):880. doi:10.1377/hlthaff.2023.00427

4. Maguire W, Murphy L. Enhancing value in healthcare: towards a trans-disciplinary approach. *Accounting, Auditing & Accountability Journal*. 2023;36(2):494-519. doi:10.1108/AAAJ-06-2016-2596

5. Mathews SC, McShea MJ, Hanley CL, Ravitz A, Labrique AB, Cohen AB. Digital health: a path to validation. *npj Digital Medicine*. 2019;2(1):38. doi:10.1038/s41746-019-0111-3

6. Sun TQ, Medaglia R. Mapping the challenges of Artificial Intelligence in the public sector: Evidence from public healthcare. *Government Information Quarterly*. 2019;36(2):368-383. doi:10.1016/j.giq.2018.09.008

7. Ahumada-Canale A, Jeet V, Bilgrami A, Seil E, Gu Y, Cutler H. Barriers and facilitators to implementing priority setting and resource allocation tools in hospital decisions: A systematic review. *Soc Sci Med*. 2023;322:115790. doi:10.1016/j.socscimed.2023.115790

8. Vasey B, Clifton DA, Collins GS, et al. DECIDE-AI: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. *Nat Med*. 2021;27(2):186-187. doi:10.1038/s41591-021-01229-5

9. Păvăloaia V, Necula S. Artificial Intelligence as a Disruptive Technology - A Systematic Literature Review.

Electronics. 2023;12(5). doi:10.3390/electronics12051102

10. Reddy S, Rogers W, Makinen V, et al. Evaluation framework to guide implementation of AI systems into healthcare settings. *BMJ Health Care Inform*. 2021;28(1):e100444. doi: 10.1136/bmjhci-100444. doi:10.1136/bmjhci-2021-100444

11. Aung YY, Wong DC, Ting DS. The promise of artificial intelligence: a review of the opportunities and challenges of artificial intelligence in healthcare. *Br Med Bull*. 2021;139(1):4-15

12. Scarbrough H, Sanfilippo KRM, Ziemann A, Stavropoulou C. Mobilizing pilot-based evidence for the spread and sustainability of innovations in healthcare: The role of innovation intermediaries. *Soc Sci Med*. 2024;340:116394. doi:10.1016/j.socscimed.2023.116394

13. Van de Ven AH. *The Innovation Journey*. Oxford University Press; 1999. https://books.google.nl/books?id=B4OJHnZMnfcC

14. Recht MP, Dewey M, Dreyer K, et al. Integrating artificial intelligence into the clinical practice of radiology: challenges and recommendations. *Eur Radiol*. 2020;30(6):3576-3584. doi:10.1007/s00330-020-06672-5

15. Anthony C, Bechky BA, Fayard A. “Collaborating” with AI: Taking a System View to Explore the Future of Work. *Organization Science*. 2023;34(5):1672-1694. doi:10.1287/orsc.2022.1651

16. Siontis GCM, Sweda R, Noseworthy PA, Friedman PA, Siontis KC, Patel CJ. Development and validation pathways of artificial intelligence tools evaluated in randomised clinical trials. *BMJ Health \& Care Informatics*. 2021;28(1). doi:10.1136/bmjhci-2021-100466

17. Marshall CR, Uchegbu I. Artificial intelligence for detection of Alzheimer's disease: demonstration of real-world value is required to bridge the translational gap. *The Lancet Digital Health*. 2022;4(11):e768-e769. doi:10.1016/S2589-7500(22)00190-X

18. Higgins D, Madai VI. From Bit to Bedside: A Practical Framework for Artificial Intelligence Product Development in Healthcare. *Adv Intell Syst*. 2020;2(10):2000052. doi:10.1002/aisy.202000052

19. Gaube S, Suresh H, Raue M, et al. Non-task expert physicians benefit from correct explainable AI advice when reviewing X-rays. *Scientific Reports*. 2023;13(1):1383. doi:10.1038/s41598-023-28633-w

20. Ivers N, Jamtvedt G, Flottorp S, et al. Audit and feedback: effects on professional practice and healthcare outcomes. *Cochrane Database Syst Rev*. 2012;(6):CD000259. doi(6):CD000259. doi:10.1002/14651858.CD000259.pub3

21. Ivers NM, Grimshaw JM, Jamtvedt G, et al. Growing literature, stagnant science? Systematic review, meta-regression and cumulative analysis of audit and feedback interventions in health care. *J Gen Intern Med*. 2014;29(11):1534-1541. doi:10.1007/s11606-014-2913-y

22. Dowding D, Randell R, Gardner P, et al. Dashboards for improving patient care: review of the literature. *Int J Med Inform*. 2015;84(2):87-100. doi:10.1016/j.ijmedinf.2014.10.001

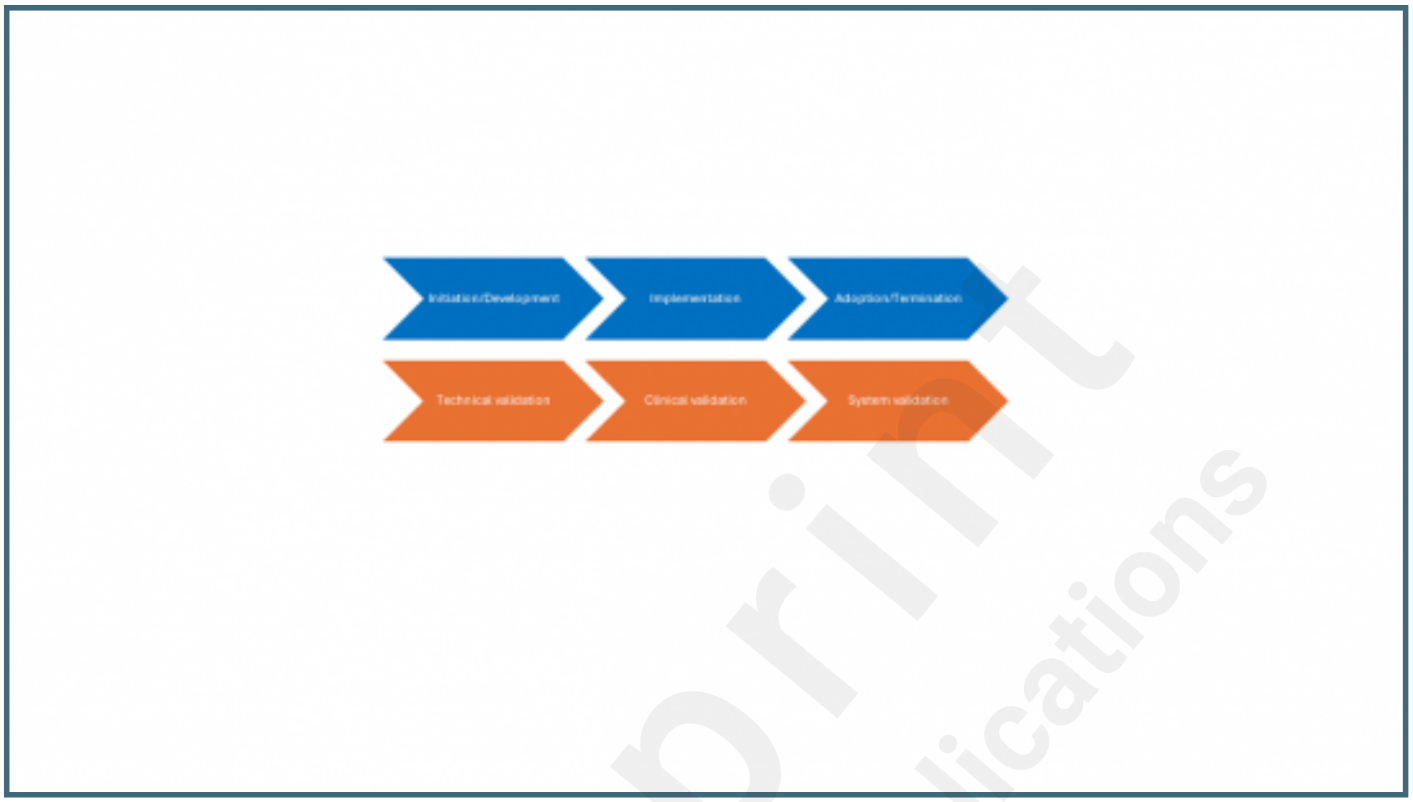
23. Randell R, Alvarado N, McVey L, et al. Requirements for a quality dashboard: Lessons from National Clinical Audits. *AMIA Annu Symp Proc*. 2020;2019:735-744

24. Begkos C, Antonopoulou K, Ronzani M. To datafication and beyond: Digital transformation and accounting technologies in the healthcare sector. *The British Accounting Review*. 2023:101259. doi:10.1016/j.bar.2023.101259

Supplementary Files

Figures

The innovation-journey of AI (blue) and the corresponding evidence necessary at each stage of the journey (orange).



The iterative innovation-journey of AI (blue) and the corresponding evidence necessary throughout the innovation-journey (orange).