Prova Health

Trust through evidence

Evidence generation for AI solutions in healthcare

Executive summary

In recent years, the rapid evolution of artificial intelligence (AI) has propelled innovation across multiple industries. In healthcare, **novel AI solutions are being created to address some of the biggest challenges in the prognosis, diagnosis, and treatment of disease, as well as clinician workflows and service improvement**. The emergence of **generative AI** presents a **compelling opportunity** to revolutionise diagnostics, treatment planning, medical research, and patient engagement. The hypothesised uses of generative AI are broad, ranging from medical education, providing information to patients, generating synthetic patient data for the validation of AI tools, to the analysis of continuous data from wearables to detect early signs of disease.

Adoption of digital solutions and AI in healthcare is slower than in other industries. The majority of clinicians don't have direct experience with AI technologies, only a quarter have recommended a digital therapeutic, and less than a fifth have prescribed one.[1,2] There are several reasons for this, including the complexity of the healthcare industry; the limited availability of high quality, integrated data on which to train and deploy models; the lack of skills to develop, manage and use AI solutions; the lack of clarity on pathways for regulatory approval and payment for AI solutions; and a lack of patient and clinician confidence.

Safety and quality are critical factors in the deployment of any AI solution in healthcare to ensure that they perform as intended, deliver the desired benefit, and do not cause harm. These factors are also crucial to ensure that clinicians and patients have trust and confidence in AI solutions (and, often, in digital health solutions more broadly).

Safety, quality, and confidence can be built through **appropriate governance**, **testing, careful implementation,** and **appropriate clinical use.**[2] A key ingredient that can help change attitudes is **evidence** – the focus of our white paper.

AI developers should ensure they generate evidence and validate their models and solutions throughout the product's lifecycle, so that the resulting information can be used to reassure and convince professionals and patients. Health systems and providers, as well as regulators, have a role to play in creating clear expectations for evidence – individual professional users and patients don't necessarily need to look at the evidence themselves, but can be confident that robust standards have been met.

The requirements for evidence and approaches to generate it are not fully established (although increasingly being clarified) for digital health solutions in general, let alone for the rapidly evolving world of AI. **This white paper clarifies the unique evidence generation requirements of AI solutions** and is aimed at both AI developers and health systems and providers. The key messages of the paper are summarised in the box on the next page.

Key messages

- What is unique to AI solutions is that they require both **model evidence** (evidence for the underlying algorithm) and **solution evidence** (evidence for the product in which the algorithm is embedded). Models will need validation first on internal and then on external datasets (**internal** and **external validation**). This will indicate how accurate and reliable the model is but will be limited to the dataset that the client has access to.
- The AI model will be part of a digital solution/product. Once a model has been internally and externally validated from a data perspective, **the solution as a whole needs to be evaluated**.
- Given their likely use to support decision making in healthcare, it is critical for AI developers to show that their **solution does not reinforce or exacerbate existing biases and inequities**, including cognitive biases present in the augmented decision-making process.
- **Evidence is required throughout the product life cycle** (product development; regulatory approval; market access/payment; post-market surveillance): different types of evidence for different aims and to inform different stakeholders.
- The **evidence required to enter markets** (e.g., FDA/CE/UKCA marking) **will differ depending on the class of the tool** and may require additional evidence generation steps.
- AI developers will face an **evidence limbo** between the early-stage evidence that is sufficient for product development and the evidence needed at later stages of the product life cycle to demonstrate specific outcomes. **Innovative evidence generation methodologies like clinical simulation can help bridge that gap**.
- Once the solution is in use, there are **significant opportunities to generate real-world evidence** to demonstrate its value. This evidence needs translation to determine whether it can be considered generalisable to other populations/workflows or should be considered only applicable to a single context.
- Ultimately, commercial success will depend on whether a product can deliver value for money for a client. **Economic modelling and analysis evaluating the benefits against the costs of a solution is essential.**

Clinicians' perspectives on AI

As promising as the prospects of AI and generative AI in healthcare are, these technologies come with their own challenges. These include the need for clinician and patient engagement, the risk of not having access to enough data or data of an appropriate accuracy to base decisions on, as well as the need for robust quality and regulatory frameworks to ensure the safe use of AI and generative AI in healthcare, as outlined in Figure 1.[3]

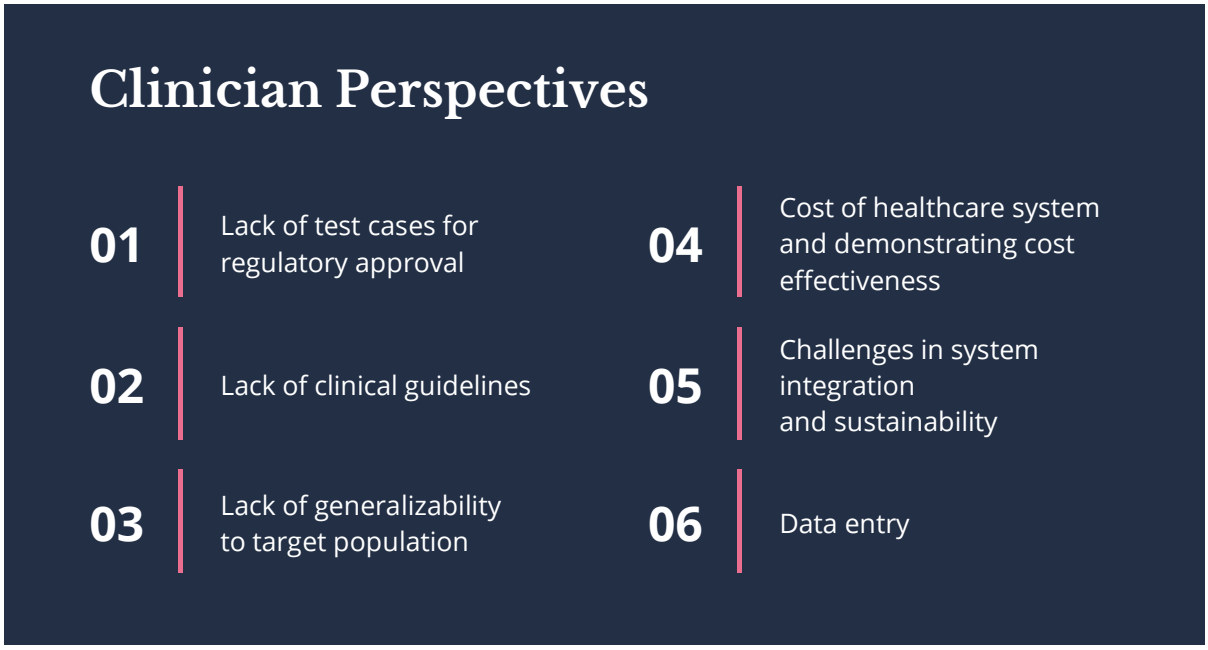


Figure 1: Clinician perspectives on AI.

Clinicians expect sufficient levels of regulation and robust validation of AI-based medical devices. A recent study by Health Education England found that clinician confidence in AI is largely dependent on how AI solutions are governed and highlighted the need for a robust, AI-specific regulatory guidance, and guidance on the safe and effective use of AI tools.[2,4] In addition, it highlighted the need for formal evidence requirements in the validation of such tools, and specific pathways for the prospective clinical studies focused on these new technologies and how to conduct ongoing monitoring.[3] Another challenge highlighted by the study was the need for mitigating model bias and ensuring fairness in AI-driven diagnostics and treatments by ensuring training data are representative of real-world populations.[2]

There are also **concerns from clinicians about liability and accountability when using AI based tools** as there is a lack of guidance from regulators with regards to this. Whilst liability remains with the clinician, the uptake of AI tools in clinical decision making will strongly depend on evidence quality, translation, and education for users.

While a majority (68%) of clinicians are excited about the potential of AI in healthcare, less than a third have used it in practice, according to a recent global survey by IPSOS Mori.[1] Their key concerns are a lack of training (62%), doubts about efficacy (48%), and a lack of clinical evidence validating these tools (45%).[1]

AI is a complex new application of technology, difficult to fully understand even for experts. **Investing in education**, both to build the right skills within health systems and a basic level of understanding in clinicians, **is imperative to build trust**. Any successful implementation of AI into clinical practice will require the digital literacy of clinicians and ensure that training is accounted for in any deployment.[5,6] This education needs to start at an undergraduate level and be considered in postgraduate curricula as clinicians progress.[2] Commercial companies should also consider providing product-specific education to teach clinicians about the scope, performance, and limitations of an AI solution, and how best to use it for maximum efficiency, benefit and safety.

Evidence for AI solutions

Trust and confidence are core factors in adopting artificial intelligence (AI) in healthcare.[7,8,9,10] **Robust evidence is a fundamental part of fostering this trust and confidence in any digital health solution, particularly one that is AI-based**. However, **producing strong evidence for digital health solutions, including AI solutions, is a significant challenge for developers, and remains a hurdle to the wider adoption of potentially impactful products**. As of 2022, 44% of the leading digital health companies in the USA had no regulatory filings or published clinical trials for their solution (Day et al., 2022).[11]

Unlike other types of digital health solutions, for AI products there are two levels of evidence and validation: **the algorithm ("model evidence") and the product in which the algorithm is embedded ("solution evidence")**. The AI model will be part of a digital solution/product. Once a model has been internally and externally validated from a data perspective, **the solution as a whole needs to be evaluated**.

The methodologies to do this would be similar to those used to evaluate other types of digital health technology. **Clinical pathways need to be scrutinised pre and post implementation of a solution to ensure that any improvements are being captured.**

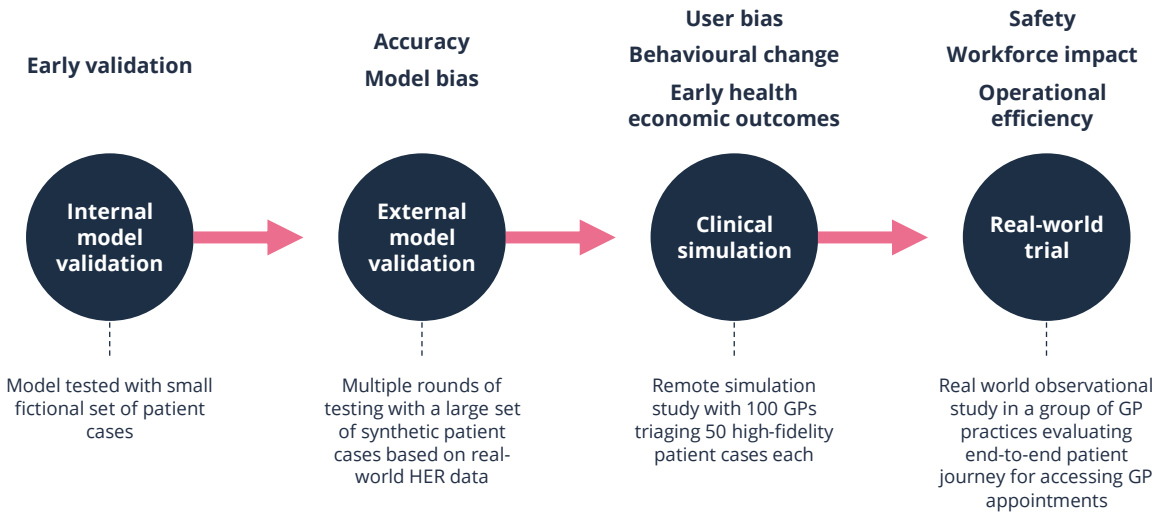


Figure 2: AI evidence roadmap for a hypothetical patient triage tool.

In simple terms, developers of AI solutions (and those wishing to adopt them) should answer with robust evidence at least the following questions:

- **Is the solution addressing a real clinical or operational problem?**
- **What is the scope of use, and what are the limitations and exclusion criteria?**
- **Does the model perform well on the developer's own datasets?**
- **Does the model perform well on external datasets?**
- **What are the characteristics of the external test dataset, and how do they correlate with real-world scenarios that the product is likely to be used in?**
- **Does the solution that includes the model address the problem effectively in a real-world context?**
- **Does the solution reduce or eliminate existing biases and inequalities?**
- **Does the solution deliver value for money to the client?**

This section of the white paper summarises the key considerations on how evidence for AI solutions should be generated.

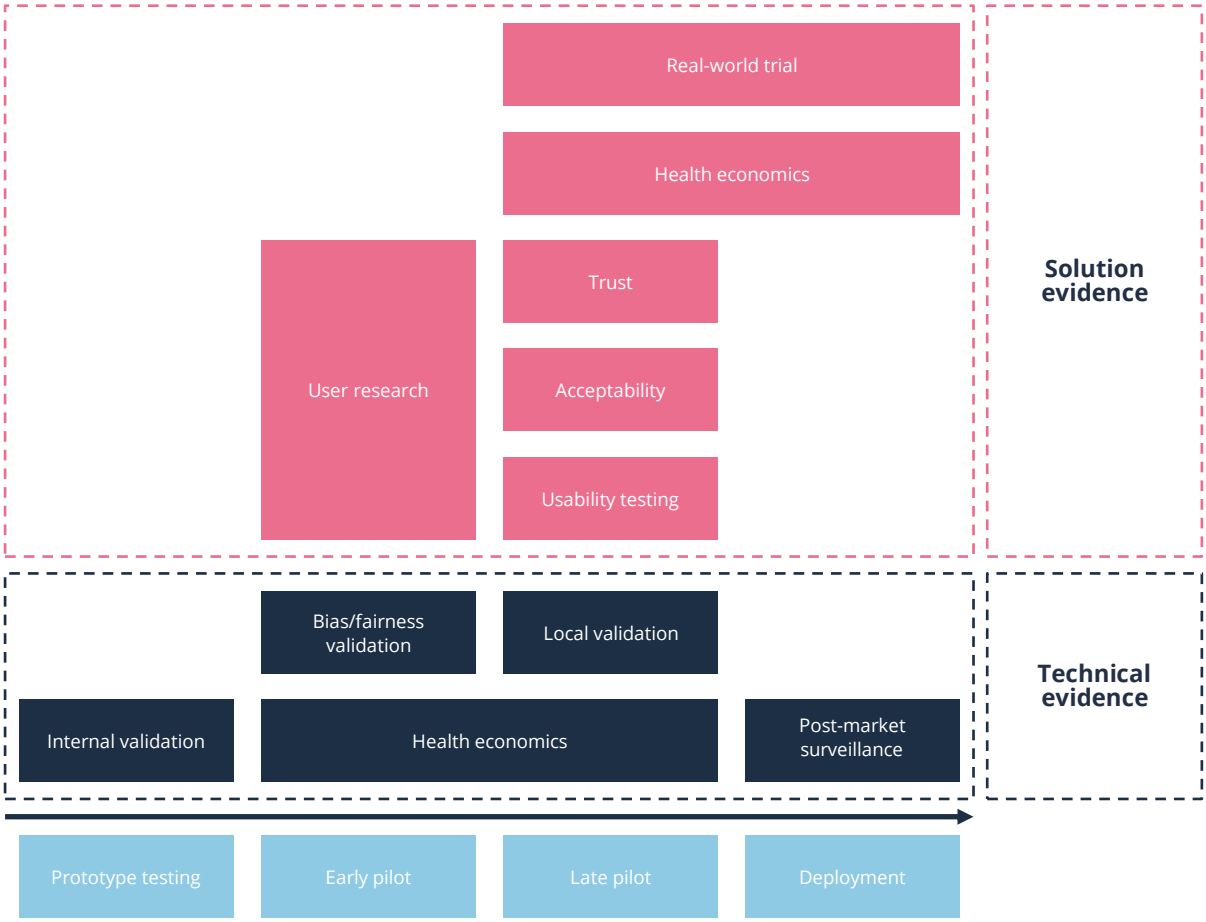


Figure 3: Factors contributing to technical and solution evidence.

Model evidence consists of internal and external validation

An innovator will develop a model from a given dataset and conduct **internal validation.** This indicates **how the model performs against the criteria defined in the scope of use but is limited to the dataset that the developer has access to. Testing the model on an external dataset (external validation) assesses its 'generalisability'.** It is important to know the characteristics of the internal and external validation datasets – they should be different, broad, and representative – and to make these characteristics available to prospective users so they can establish whether further local validation is needed prior to deployment, or the existing external validation is sufficient. It is vital to consider how training and testing data have been curated, quality assured and filtered, and to understand any limitations and exclusions that have been applied. Any labels used should represent a ground truth or robust gold standard.

It is important to note that **valuable evidence can be generated early on in a model's development that is not limited to purely model validation**. Utilising methods like clinical simulation at this early stage can facilitate the assessment of factors that impact trust and confidence in a model. For example, **evaluating how the presentation of auxiliary data in addition to the AI generated outputs impacts the user's clinical decision making.** This can be done in parallel with the purely technical model validation.

Various frameworks have been developed to help standardise expectations and reporting outcomes for specific types of AI technologies. These are helpful for developers building a product that matches one of these use cases (see Table 1).

Guideline	Framework	Area
CONSORT-AI	Consolidated Standards of Reporting Trials-AI	Reporting guideline
DECIDE-AI	Developmental and Exploratory Clinical Investigation of DEcision-support systems driven by Artificial Intelligence	Clinical Decision Support
PROBAST-AI	Prediction model Risk Of Bias ASsessment Tool-AI	Diagnostic and prediction models
QUADAS-AI	QUality Assessment tool for artificial intelligence-centered Diagnostic test Accuracy Studies	Diagnostic accuracy
SPIRIT-AI	Standard Protocol Items: Recommendations for Interventional Trials-AI	Reporting guideline
STANDING Together	STANdards for Data INclusivity and Generalisability	Representative data
STARD-AI	Standards for Reporting of Diagnostic Accuracy Study-AI	Diagnostic accuracy
TRIPOD-AI	Transparent Reporting of a multivariable prediction model of Individual Prognosis Or Diagnosis-Artificial Intelligence	Prediction models

Table 1: AI evaluation and clinical trial reporting guidelines and tools[12]

From an understanding and explainability perspective, a technical challenge is that AI models can be considered a "black box" due to **difficulty in understanding how the model is producing an output**.[13]

Another major challenge to overcome is the **lack of sufficient data to train and validate AI models** on, or ensuring the data is labelled accurately, unbiased, and from a population that is representative of that in which the solution is going to be implemented.[11] This is due to the large datasets required to train AI models. In some specialties, such as ophthalmology, there exists a large library of anonymised imaging data for training AI models, however, this isn't the case for others, making it difficult to optimise and train algorithms.

Data privacy and security are critical factors in the adoption of any software for healthcare, including AI tools.[14] AI systems rely heavily on large datasets. Ensuring that patient information remains confidential and secure is paramount. Breaches in data privacy and security can erode patient trust and have significant legal implications. The move towards secure data environments in the NHS in England is an example of how data can be better safeguarded.[14]

Key challenges for the translation of AI systems in healthcare include those intrinsic to the science of machine learning, logistical difficulties in implementation, and consideration of the barriers to adoption as well as of the necessary sociocultural or pathway changes.

Biases and inequities should be minimised

The data on which algorithms are based should be representative of the populations that they will be deployed in and should have sufficient breadth and depth to capture the multitude of clinically important associations between ethnicity, demographic, social and clinical features that may exist[13]. Developers need to build in sensitivity checks to reduce bias. Such checks may include simulated data sets and running counterfactual simulations, during the design and subsequent phases of AI development.[13]

At a minimum, the following points should be documented when thinking of bias and inequities in AI algorithms[15]:

- A description of the data set and its representativeness of different minority ethnic populations.
- What measures were taken to prevent and address bias across different minority groups in: 1. data used, 2. defined outcome, and 3. modelling.
- Whether the performance of the algorithm has been tested and validated on different demographic subgroups.
- Ethical considerations regarding how the algorithm will be used once deployed, and whether there are risks of this creating or perpetuating disparities in health and healthcare.

Evidence is required across the product life cycle

In the context of digital health solutions, "evidence generation" is a broad term for the process by which various types of evidence are produced to support **product development and validation.**[16] Many types of evidence can be generated at any stage; however, certain types of evidence are more strongly associated with specific stages. For example, product development is informed from the earliest stages by techniques such as **secondary research** (reviewing existing research, which helps innovators to better understand a clinical problem), **user research** (which can help validate a problem and solution concept), and **A/B testing** (which allows comparison of different versions or features). Later, when developers have a well-defined product, **clinical data demonstrating safety and clinical performance** will be critical for regulatory certification, and **outcome and economic analyses** will be important for showing evidence of value to health systems to sell a solution.

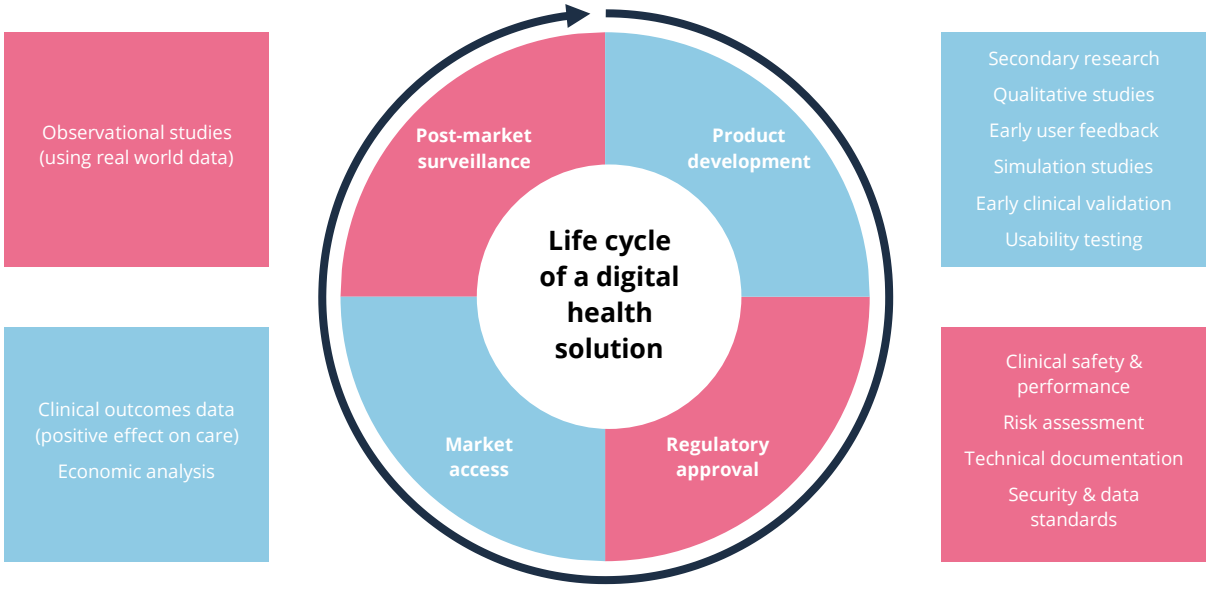


Figure 4: Examples of evidence generation at different stages of the product life cycle.

Innovative approaches can help overcome the “evidence limbo”

Innovators in any area of health technology face significant hurdles to arrive at the generation of real-world evidence. These barriers are even higher with AI solutions, in part as many will qualify as medical devices.

External model validation can be difficult without access to good external datasets. Establishing and maintaining good partnerships and ensuring that there are strong clinical and technical teams in place during the implementation phase are crucial necessities. The multidisciplinary team approach is critical to avoid potential safety issues and unintended consequences. A further challenge exists in proving the model's efficacy, safety, and benefit on a localised population.

Such challenges become even more significant when innovators seek to enter the market and generate real-world solution evidence. To achieve deployment, they may be required to generate evidence before their solution is actually adopted in the real world. The next step is to build an evidence base through translation, reusing evidence where appropriate (rather than generating this in each local setting). This approach can provide a pathway to mature evidence and long-term market access.

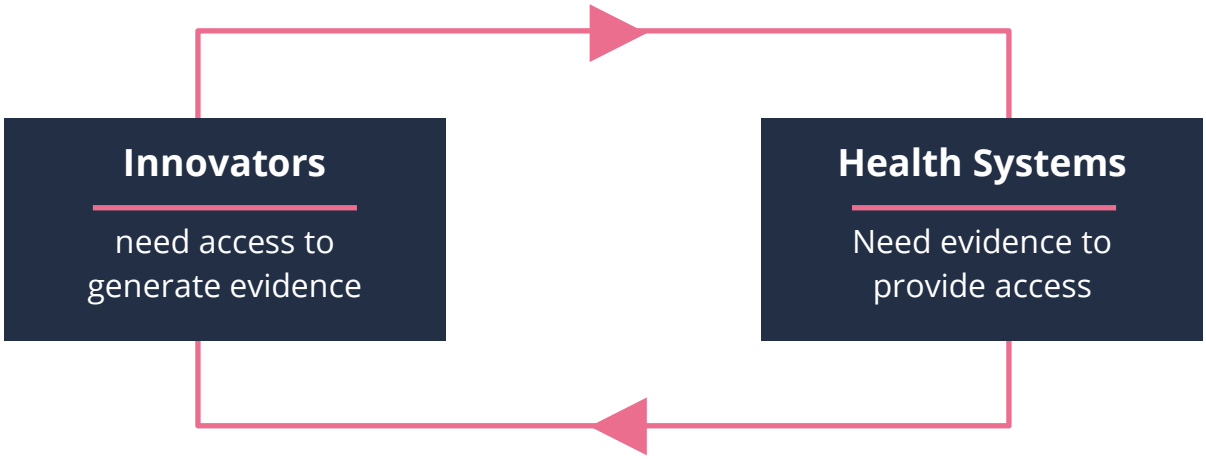


Figure 5: The evidence paradox.

Again, **close partnership with healthcare providers** to deliver small early real-world pilots, and translate the generated evidence to a general context, is a potential solution. These partnerships can be hard to come by, but are crucial for the safe testing and deployment of any AI based tool. In addition, innovators can rely on innovative methodologies like clinical simulation to generate evidence in a way that addresses the “evidence limbo”. When the solution is finally adopted, there are significant opportunities to generate further real-world evidence and reach a position of evidence-base maturity.

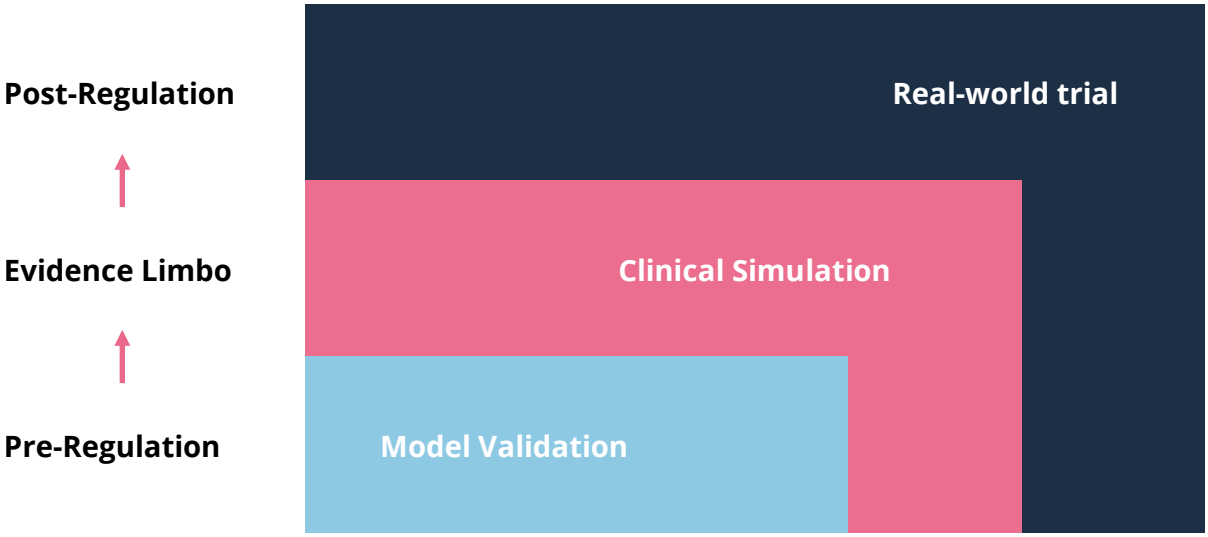


Figure 6: Evidence generation steps for AI tools.

Simulation has emerged as a promising tool that enables safe, efficient, and cost-effective evaluation of digital health solutions in a controlled environment.[17,18] Simulation studies can help break the evidence generation deadlock, most often encountered in the early-mid stages of development of new AI and digital solutions.[18]

The use of clinical simulation has been extended to evaluate digital health/AI solutions. This is achieved by closely replicating their intended real-world use in a controlled environment. Remote, multi-site trials can be conducted at relatively low cost using virtual communication platforms.[17] **Simulation studies are highly scalable and flexible and study designs can easily be adapted** to keep up with the frequent updates to digital solutions.

An additional benefit of simulation is the **ability to use synthetic patient data**. Realistic, synthetic datasets can be modelled using real data in a way that minimises privacy concerns while preserving the complexities of the data. It is important to note that synthetic data cannot make up for a lack of data, as it can only be generated from data that exists.

Simulation studies may also allow researchers to test solutions with data representative of higher-risk patients, who are often excluded from traditional trials because of safety risks.[17,18] It also allows for more extensive testing of subpopulation data helping to alleviate the risk of biases (e.g., ethnicity, gender) and can also help uncover any potential biases in an AI tool. Simulation studies can also help better understand and study the human-AI interaction and the potential for AI to impact on decision making and clinical behaviours.

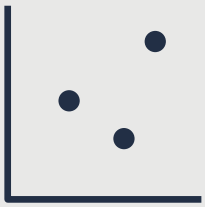
Ultimately, evidence generated in simulation studies is unlikely to be sufficient on its own to support decisions around regulatory approval for higher-risk solutions. However, it is a pragmatic adjunct to established methods that can be used to generate evidence of reasonable strength.[17]

AI solutions are well-suited to generate real-world evidence

Demonstrating that AI can improve patient outcomes, enhance diagnostics, or streamline workflows through well-conducted implementation trials is crucial to improve adoption. Solutions require evidence supporting their function in a healthcare setting and their impact on incumbent workflows and how clinicians adopt and interact with such tools. This is where real-world evidence studies are key to informing the true clinical utility of an AI solution in practice.[19]

Significantly, if real-world benefit can be demonstrated as linked to technical performance, then showing that technical performance is equivalent in new settings, and that the implementation, clinical situation and use case are equivalent, should be sufficient to justify use in new settings.

The majority of AI studies have been retrospective in more tightly controlled conditions and have relied on comparing clinical expert performance vs algorithm performance[20]. Ideally an AI tools performance should be compared to the performance of a pathway pre-implementation. **Real world studies relying on prospective data are key to informing the true clinical utility of an AI solution in practice in any given workflow.**[20]



Real-world data (RWD)

= data relating to patient health status and/or the delivery of health care routinely collected data form a variety of sources

Real-world evidence (RWE)

= clinical evidence about the usage and potential benefits or risks of a medical product derived from analysis of RWD

Figure 7: RWE and RWD definitions according to the US Food and Drug Administration.

In comparison with randomised controlled trials (RCTs), which may inform researchers about how an intervention performs for a specific group (under tightly controlled conditions), RWE can provide more certainty about how effective a solution is when

deployed in the real world. These studies can also illuminate potential unintended consequences of technologies on different population groups, for example minority ethnic groups, identifying potential bias and other negative outcomes, which can then be addressed in a timely manner.[19,20]

Regulators in several countries have begun to encourage the use of RWE to inform regulatory decisions in the post market phase.[17] As regulated digital solutions are constantly updated, it is necessary to conduct ongoing postmarked surveillance and clinical validation of these solutions to ensure that they remain safe and effective despite any iterations.

Demonstrating value for money is essential for success

Another key piece of evidence required for AI solutions is **economic evidence.**[21] As AI solutions are being launched to the market at a rapid speed, there is frequently insufficient data to support their efficacy. Traditional health technology assessment (HTA) approaches, which rely on published research, can be time-consuming and may not be compatible with the quick development cycles of digital health technologies (DHTs).[22]

Economic evidence is critical because it assesses the cost-effectiveness and economic impact of these technologies especially in health systems with constrained resources.[23] This evidence will assist decision makers in determining if investing in a dedicated solution will provide value in terms of better patient outcomes relative to the expenditure involved, and in comparison to the incumbent pathway

References

1. Ipsos finds doctors remain wary over patient use of health data, but are excited about AI in diagnosis [Internet]. Ipsos; 2023 [cited 2023 Dec 5]. Available from: https://www.ipsos.com/en-uk/ipsos-finds-doctors-remain-wary-over-patient-use-health-data-are-excited-about-ai-diagnosis
2. Nix M, Onisiforou G, Painter S. Understanding healthcare workers confidence in AI. NHS AI Lab & Health Education England; 2022.
3. Huang JD, Wang J, Ramsey E, Leavey G, Chico TJ, Condell J. Applying artificial intelligence to wearable sensor data to diagnose and predict cardiovascular disease: a review. Sensors. 2022 Oct 20;22(20):8002.
4. Software and AI as a medical device change programme - roadmap [Internet]. Medicines & Healthcare products Regulatory Agency; 2023 [cited 2023 Dec 5]. Available from: https://www.gov.uk/government/publications/software-and-ai-as-a-medical-device-change-programme/software-and-ai-as-a-medical-device-change-programme-roadmap#:~:text=Last%20year%2C%20the%20MHRA%20announced,clear%20and%20patients%20are%20protected
5. Hardie T, Horton T, Willis M, Warburton W. Switched on: How do we get the best out of automation and AI in health care? The Health Foundation; 2021 (https://doi.org/10.37829/HF-2021-I03).
6. Topol E. The Topol Review. Preparing the healthcare workforce to deliver the digital future. Health Education England; 2019 Feb.
7. Asan O, Bayrak AE, Choudhury A. Artificial intelligence and human trust in healthcare: focus on clinicians. Journal of Medical Internet Research. 2020 Jun 19;22(6):e15154.
8. Wenjuan F, Liu J, Shuwan Z, Pardalos PM. Investigating the impacting factors for the healthcare professionals to adopt artificial intelligence-based medical diagnosis support system (AIMDSS). Annals of Operations Research. 2020 Nov 1;294(1-2):567-92.
9. Gille F, Jobin A, Ienca M. What we talk about when we talk about trust: Theory of trust for AI in healthcare. Intelligence-Based Medicine. 2020 Nov 1;1:100001.
10. Wysocki O, Davies JK, Vigo M, Armstrong AC, Landers D, Lee R, Freitas A. Assessing the communication gap between AI models and healthcare professionals: Explainability, utility and trust in AI-driven clinical decision-making. Artificial Intelligence. 2023 Mar 1;316:103839.
11. Day S, Shah V, Kaganoff S, Powelson S, Mathews SC. Assessing the clinical robustness of digital health startups: cross-sectional observational analysis. Journal of Medical Internet Research. 2022 Jun 20;24(6):e37677.
12. Understanding healthcare workers confidence in AI. Chapter 3: Governance [Internet]. NHS England; 2023 [cited 2023 Dec 5]. Available from: https://digital-transformation.hee.nhs.uk/building-a-digital-workforce/dart-ed/horizon-scanning/understanding-healthcare-workers-confidence-in-ai/chapter-3-governance/evaluation-and-validation
13. O'Brien N, Van Dael J, Clarke J, Gardner C, O'Shaughnessy J, Darzi A, Ghafur S. Addressing racial and ethnic inequities in data-driven health technologies. Institute of Global Health Innovation, Imperial College London; 2022 (https://spiral.imperial.ac.uk/bitstream/10044/1/94902/2/Imperial_IGHI_AddressingRacialandEthnicInequities%20_Report.pdf)
14. Ghafur S, O'Brien N, Howitt P, Painter A, O'Shaughnessy J, Darzi A. NHS data: Maximising its impact for all. Institute of Global Health Innovation, Imperial College London; 2023 (https://spiral.imperial.ac.uk/bitstream/10044/1/103404/8/NHS%20Data%20-%20Maximising%20Impact%20for%20All.pdf)
15. Review into bias in algorithmic decision making. Centre for Data Ethics and Innovation; 2020 (https://assets.publishing.service.gov.uk/media/60142096d3bf7f70ba377b20/Review_into_bias_in_algorithmic_decision-making.pdf)
16. Digital Health Solutions and Evidence Generation [Internet]. Healthcare Transformers; 2023 Mar 15 [cited 2023 Dec 5]. Available from: https://healthcaretransformers.com/digital-health/current-trends/digital-health-solutions-evidence-generation/
17. Lau K, Halligan J, Fontana G, Guo C, O'Driscoll FK, Prime M, Ghafur S. Evolution of the clinical simulation approach to assess digital health technologies. Future Healthcare Journal. 2023 Jul;10(2):173.
18. Guo C, Ashrafian H, Ghafur S, Fontana G, Gardner C, Prime M. Challenges for the evaluation of digital health solutions—A call for innovative evidence generation approaches. NPJ digital medicine. 2020 Aug 27;3(1):110.
19. Stern AD, Brönneke J, Debatin JF, Hagen J, Matthies H, Patel S, Clay I, Eskofier B, Herr A, Hoeller K, Jaksa A. Advancing digital health applications: priorities for innovation in real-world evidence generation. The Lancet Digital Health. 2022 Mar 1;4(3):e200-6.
20. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC medicine. 2019 Dec;17:1-9.
21. Gomes M, Murray E, Raftery J. Economic evaluation of digital health interventions: methodological issues and recommendations for practice. Pharmacoeconomics. 2022 Apr;40(4):367-78.
22. Kolasa K, Kozinski G. How to value digital health interventions? A systematic literature review. International Journal of Environmental Research and Public Health. 2020 Mar;17(6):2119.
23. Evidence standards framework (ESF) for Digital Health Technologies [Internet]. National Institute for Health and Care Excellence; 2022 [cited 2023 Dec 5]. Available from: https://www.nice.org.uk/about/what-we-do/our-programmes/evidence-standards-framework-for-digital-health-technologies

Acknowledgements

The development of this white paper benefited significantly from the input provided by the following people. Their feedback and insights challenged us to improve the final output, and we would like to thank each of them for their time. While the white paper has significantly benefited from their guidance, the views it contains are solely those of the authors and may not necessarily reflect those of other contributors.

Dr Dominic King, Dr Annabelle Painter, Dr Umang Patel, Dr Mike Nix, Dr Haris Shuaib

Recommended citation:

Patel M, Popescu M, Jangam S, Conroy D, Fontana G, Ghafur S. Trust through evidence: Evidence generation for AI solutions in healthcare. Prova Health; 2023