

Learning at the Speed of AI

USING PLATFORM TRIALS TO TEST AI INNOVATION

 Anthony McDonnell

Abstract

Artificial intelligence has the potential to transform healthcare, but to see this upside we need to be able to know which products work best. Clinical trials were built to evaluate technologies that are stable over time. AI updates frequently, meaning that if we rely on old evaluation approaches, evidence will be outdated before a study is complete. Head-to-head clinical trials that continuously run in multiple countries would allow us to generate evidence that keep up with the pace of innovation. This needs to be combined with an approach that makes it easier to collect evidence because if standards are too high, promising technologies will never reach patients.

This paper argues that standing research infrastructure through platform trials or clinical trial networks would greatly reduce both the cost and time needed to evaluate AI, making it easier to run comprehensive studies. Further work is now needed to determine where these models add most value, how they should be funded and governed, and whether they are acceptable to regulators, developers, and health systems.

Learning at the Speed of AI: Using Platform Trials to Test AI Innovation

Anthony McDonnell

Center for Global Development

I would like to thank Tim Ohlenburg for all of his feedback on this paper and the idea more broadly. It is a much stronger piece of work thanks to his influence. I am also very grateful to Han Sheng Chia for all of his support and feedback. I received excellent comments on the paper from Akhil Bansal, Ravi Nirmal, and the external peer reviewer, and the paper is much stronger because of this. Thank you to Sara Viglione for editing this work.

Anthony McDonnell. 2026. "Learning at the Speed of AI: Using Platform Trials to Test AI Innovation."

CGD Working Paper 758. Washington, DC: Center for Global Development. <https://www.cgdev.org/publication/learning-speed-ai-using-platform-trials-test-ai-innovation>

CENTER FOR GLOBAL DEVELOPMENT

2055 L Street, NW Fifth Floor
Washington, DC 20036

1 Abbey Gardens
Great College Street
London
SW1P 3SE

www.cgdev.org

Center for Global Development. 2026.

The Center for Global Development works to reduce global poverty and improve lives through innovative economic research that drives better policy and practice by the world's top decision makers. Use and dissemination of this Working Paper is encouraged; however, reproduced copies may not be used for commercial purposes. Further usage is permitted under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License.

The views expressed in CGD Working Papers are those of the authors and should not be attributed to the board of directors, funders of the Center for Global Development, or the authors' respective organizations.

Contents

Introduction	1
Why existing frameworks will struggle	2
Constantly updating interventions	2
Comparing different AI interventions	3
Contextual challenges	4
Overview of main challenges and how they are impacted by AI	5
Using standing infrastructure to enable better trials	6
Practically difficult	6
Platform trials	6
Clinical trial network	10
Embedded randomisation	11
Choosing between a platform trial and a clinical trial network	11
Challenges that remain	14
Limitations	16
Next steps	17
References	19

Tables

1. How AI impacts clinical evaluation	6
2. Difference between a platform trial and a clinical trial network	12

Boxes

1. The FDA's approach to regulating AI models that change with time	3
2. STAMPEDE: An 18-year platform trial in prostate cancer	7
3. The RECOVERY trial: rapid learning in a fast-moving environment	8
4. ADVANCE-ID: Building regional clinical trial infrastructure for antimicrobial resistance	11
5. Illustrative use cases for platform trials or clinical trial networks	13

Introduction

Artificial intelligence (AI) has seen vast technical progress over recent years. Across a wide range of areas such as reviewing medical images, diagnostics, triage, and mental health care, AI systems are beginning to perform tasks that until recently required humans with deep expertise, and they are beginning to outperform humans on some narrow tasks (Weissman, 2025). As models and technology improve, the range and scope of tasks that AI models can undertake will expand too, and they will likely be integrated further into existing healthcare systems. If managed well, the benefits for human health could be huge (Silcox et al., 2024).

However, there are challenges on how to evaluate these models. Existing approaches for clinical evaluation and comparative effectiveness were designed for stable, human-delivered interventions and appear poorly suited for a technology that evolves continuously. Patient-facing diagnostic and treatments face high regulatory thresholds before they can be deployed at scale, with regulators generally requiring demonstrations of safety and for treatments they also require efficacy. These requirements reflect the fact that even modest or rare issues caused by a new product, can have a large impact when rolled out across a wide population. Because of this, evaluation and approval processes have evolved to be deliberately demanding, slow-moving, and geared towards relatively stable products with clearly defined indications and mechanisms of action (Rosenthal, Beecy, and Sabuncu, 2025, WHO, 2021a). Health systems and clinicians rely on clinical trial evidence to assess which products to procure and prescribe, so even if products don't need to show efficacy to be approved for use, they will need a way to demonstrate efficacy before there's widespread adoption (van de Sande et al., 2024).

There are two familiar risks when it comes to evaluating new AI models, often known as the “Goldilocks Problem.” The first is that the evidentiary standards are too high, and promising technologies that could greatly improve people's lives do not get used, either for lack of approval or because procurers are not convinced of their value. A second worry is that AI models are not treated as medical devices, but rather as general wellness products, treated similarly to an app that tracks people's calorie intake. In this instance, unsafe products might get rolled out, with consumers and healthcare systems lacking a good understanding of which products work best (Hou and Gattinger, 2024).

The goal of this paper is to assess how evaluation and regulatory frameworks should assess AI-enabled diagnostics and treatments before such products become entrenched in clinical practice. Regulators and health technology assessment (HTA) agencies should design their approaches ahead of deployment so that systems are ready to test AI interventions as they emerge, rather than improvising oversight after adoption.

This paper will focus on three challenges specific to testing AI. The need for timely comparisons of tools designed to perform the same job, how to regulate the many iterations of AI devices, and how to

make it easy and affordable for AI devices to be tested in the many different markets and populations that they'll be useful for. It will then look at the role that standing infrastructure can play in enabling this research by exploring two ideas: a platform trial where different products that are seeking to achieve the same end point are compared in one trial that has a single process, and second, a clinical trial network, where different hospital and primary care settings are connected in a network that makes it easier for researchers to build a clinical trial.

Why existing frameworks will struggle

The evaluation of new AI-enabled interventions in healthcare poses measurement challenges that differ from those associated with conventional drugs, devices, or clinical programmes. Existing approaches are designed to evaluate interventions that are stable, well-defined, and comparable across settings and over time. Many AI systems violate these assumptions, complicating both the generation and interpretation of evidence (Rosenthal, Beecy, and Sabuncu, 2025, FDA, 2019).

Constantly updating interventions

Most AI models are updated frequently in response to new data, performance monitoring, safety concerns, or improvements in underlying architecture. These changes might each be small, however many incremental updates can alter behaviour in meaningful ways that are difficult to observe. Even if a clinical application itself is unchanged, updates to the underlying foundation model or software on which it depends may also affect its performance. This creates a challenge for measurement: evidence generated about one version of an AI system may not apply to subsequent versions that are deployed in practice (Rosenthal, Beecy, and Sabuncu, 2025, Meijerink et al., 2025).

Randomised control trials estimate the effect for a specific intervention, at a specific point in time. These results are then used to approve a product for use long into the future. This assumes that the intervention being tested remains materially the same throughout a study and after the study concludes. When applied to continuously evolving AI systems, this logic doesn't hold.

Attempts to “freeze” models during evaluation may improve internal validity but reduce external validity because the frozen system that is tested is going to be less like the product that is ultimately deployed. This problem could be avoided by building models that are static, which is essentially “freezing” them indefinitely. However, this would mean losing one of AI's main advantages—the ability to improve over time through learning from new data and real-world use (Rosenthal, Beecy, and Sabuncu, 2025, Meijerink et al., 2025).

In other regulated technologies, regulators have typically allowed incremental modifications after approval, such as manufacturing adjustments or label changes, often without requiring entirely new trials. But for AI systems this might be more complicated. Even small changes to model architecture or training data could have potentially wide-ranging and difficult-to-predict effects on performance

across patient groups and settings (FDA, 2019). As a result, while freezing models may simplify evaluation, they could reduce the quality of products available to the public. Regulators have begun to develop approaches to manage the trade-offs between allowing model improvements whilst ensuring safety and efficacy is maintained. One example is the U.S. Food and Drug Administration's Predetermined Change Control Plan (PCCP) outlined in Box 1.

BOX 1. The FDA's approach to regulating AI models that change with time

The U.S. FDA has created its Predetermined Change Control Plan (PCCP) to overcome some of the challenges of evaluating AI devices that can update regularly. A PCCP allows developers to specify, at the time of approval, the types of modifications they expect to make (for example, retraining with new data), alongside the methods they will use to evaluate these changes. If the FDA agrees with the proposed changes and judges the monitoring approach to be appropriate, the model can be updated without the need for a new regulatory submission. This provides an important way of enabling some iterative improvements whilst maintaining necessary oversight. However, challenges remain. It can be difficult to estimate ex-anti the impact of a given modification, and the impact of multiple smaller modifications might be difficult to predict. So even a well-designed PCCP will require careful post-market monitoring, as well as clear thresholds for when new clinical trials are needed (FDA, 2024).

Potential solution: continuously running trials

One answer to frequent updates is a trial that continuously runs after a product is approved, with a much smaller number of patients. This would enable researchers to identify when results drift—either because the model changed, clinicians started using it differently, or the patient population shifted. This could be done on quite a small scale, with enrolment expanded when something warrants a closer look, either when results signal that safety or efficacy results might have changed, or when there's a known major model revision. These triggers can be agreed in advance. Once the question is settled, trial size can be reduced to background level.

This builds on the approach taken by adaptive platform trials but stretches the logic across a product's whole life. The same infrastructure could test an updated version alongside the current one before it's rolled out widely. This would identify any risks from a newer model. As well as potential clinical benefits, detailed information about products being created in almost real-time will also help researchers identify what works and what doesn't in their product, improving learning and quality.

Comparing different AI interventions

When assessing treatment options, patients and doctors don't just want to know if something works, but also whether it is the best treatment available. For this reason, treatments are not evaluated by regulators in isolation but are considered against available alternatives. As standards of care improve, the bar needed for new interventions rise accordingly. For example, most of

the cancer drugs approved in the 1980s are not widely used today and would be unlikely to meet current standards of approval, not because they lacked clinical value in their original context, but because they would now be judged against far more effective existing treatments (Djulbegovic and Guyatt, 2017).

Alongside regulation, comparative studies are also key for clinicians and health system decision-makers to know what treatments to procure and use and are widely used in HTAs. HTAs are studies that combine clinical and economic information about an intervention to determine clinical and cost effectiveness, and in many countries these assessments inform decisions about what treatments should be provided (Drummond et al., 2008).

Comparison studies will be as useful for informing decisions about approval and use in AI products as they are for other areas of health, but the timelines will be far shorter. Innovation in AI is often measured in months rather than decades for pharmaceuticals, so, the “next best alternative” against which an AI intervention should be compared may often change. This will reduce the amount of time any comparison remains valid, and complicate decisions about adoption and scale-up (Goetz et al., 2024).

This will make it particularly challenging for researchers to undertake HTAs and write clinical guidelines, both of which rely on synthesising evidence from multiple sources that have stable comparisons, so that they can make durable recommendations (Goetz et al., 2024, Drummond et al., 2008).

Potential solution: comparison studies

Comparative studies directly assess two or more interventions in the same trial. Rather than asking whether a product performs better than a placebo or an old best practice, the comparative design assesses how well an intervention performs against the best available alternative. For pharmaceuticals, this often takes the form of a head-to-head random control trial. These studies often happen after a product has already been approved to inform clinical guidelines or aid a HTA (Institute of Medicine, 2009).

Most areas of medicine would benefit from more timely comparative studies, but they would be particularly useful for AI-based products because of the speed at which innovation takes place. Without strong comparative evaluations, each study risks generating fragmented and quickly outdated evidence. Governments, either through regulation or procurement, could require comparative studies in areas where it would otherwise be difficult to make comparisons.

Contextual challenges

Traditionally, many low- and middle-income countries (LMICs) have relied heavily on regulatory approvals from regulators in high-income countries, particularly the U.S. FDA, as signals of safety

and effectiveness. This approach reduces duplications and allows less well-funded regulators to benefit from the expertise elsewhere (WHO, 2021b). For pharmaceuticals and conventional devices, this model can make sense. Biological responses to pharmacological treatments are usually relatively stable across populations.

However, many AI-based products are much more likely to be sensitive to local contexts. Their effectiveness will depend on local culture, health system organisation, and different ways of working. For example, an algorithm trained on US electronic health data might build in assumptions about access to treatment, or note taking, that do not hold in other settings. Tools trained primarily using English-language data may not perform as well when applied to other languages or even other dialects of English (Yang et al., 2024, WHO, 2024, Goetz et al., 2024).

Potential solution: multi-country trials

To counter this, AI tools should be evaluated across a wide range of settings. This will reduce the risk of adopting tools that underperform or behave unpredictably in certain settings. It is also likely that LMICs will gain the most from this, as this is where research budgets are smallest, and usually differ more from HICs than HICs differ from each other. For LMICs to fully benefit from AI, there needs to be investments in trial sites, data infrastructure, and regulatory capabilities that enable meaningful local evaluation rather than passive reliance on external approval.

These multi-country trials would help identify where a product performs well and where adaptation or retraining is required. Doing this in a single study, rather than multiple individual studies in different countries, can accelerate evidence generation by allowing concurrent recruitment across sites, producing faster and more representative results than sequential, single-country studies.

Coordinated multi-country trials are already common in drug development, but they are likely to be more important for AI tools as there are more reasons to think results from one country won't extend elsewhere.

Overview of main challenges and how they are impacted by AI

AI has created some new challenges that did not really exist before, such as interventions changing during or after evaluation. However, most of the challenges discussed in this paper existed already. Evidence becoming outdated is not a new problem and will continue to be an issue in clinical research as long as new evidence continues to be generated. However, the speed at which evidence is generated because of AI, greatly exacerbates this issue. Table 1 outlines the key challenges identified in this paper, and whether they are new problems or existing ones made more difficult.

TABLE 1. How AI impacts clinical evaluation

Challenge	Is this a New Challenge or One Exacerbated by AI?
Interventions change during and after evaluation	Predominantly new
Evidence becomes outdated quickly	Existing problem, substantially exacerbated
Need for comparative effectiveness evidence	Existing problem, exacerbated
Performance varies across settings and populations	Existing problem, likely to be exacerbated
Need for continuous post-market evaluation	Existing problem, substantially exacerbated
Determining when a modified product requires new evidence	Predominantly new
Inference when the intervention changes during a trial	Predominantly new

Note: Challenges are judged to be new or existing by the author, each one of these topics is discussed in more detail earlier in the paper.

Using standing infrastructure to enable better trials

Practically difficult

There are strong reasons to want to create a multi-country, head-to-head continuous study. However, there are two challenges with this approach: cost and time. It costs money to set up every new site in a clinical study, and this is more difficult when sites are in different countries. Continuous evaluation requires funding to maintain recruitment, monitoring, and analysis long after initial approval. For many developers, particularly smaller firms and academic groups, these costs may be prohibitive.

Raising evidentiary standards therefore risks reducing competition if only the largest companies working with affluent patients can afford to do it. A way to reduce these burdens is to use platform trials and clinical trial networks to generate robust evidence more cheaply and quickly than standalone trials

Platform trials

Platform trials (PTs) essentially create a trial infrastructure that can test multiple different products, using the same hospitals or sites. Instead of building a new trial from scratch each time researchers want to test a product, they place that product into a PT. This can remove the need for standalone trials that have their own governance arrangements, data systems, recruitment pathways, or clinical sites, and usually share a master trial protocol.

Sharing infrastructure reduces the time and cost of starting new studies, while making it easier to compare different interventions because multiple products are being tested using the same criteria.

Platform trials are particularly useful when:

1. Several products are trying to solve the same clinical problem, making head-to-head comparisons valuable (Roustit et al., 2023)
2. Technologies or standards of care are changing quickly, so evidence can become outdated within a short period (Pericàs, Derde, and Berry, 2023)
3. Trials can share infrastructure, reducing costs and avoiding duplication. This is particularly true for areas like antimicrobial resistance where a large amount of infrastructure is needed (see Box 4) (Wellcome Trust, 2016; McDonnell et al., 2016)
4. There is a lot of iteration of products, allowing ineffective interventions to be identified early (Roustit et al., 2023)
5. Evidence is needed from multiple hospitals or countries to show that results are generalisable (Australian Clinical Trials Alliance, 2019)

PTs are a well-established tool in medicine and have been used in oncology, cardiovascular disease, and infectious diseases. The STAMPEDE trial for prostate cancer (Box 2) shows how PTs can operate in the long term. The RECOVERY trial (Box 3) shows how this approach can generate reliable evidence during a rapidly changing public health emergency.

BOX 2. STAMPEDE: An 18-year platform trial in prostate cancer

The STAMPEDE (Systemic Therapy in Advancing or Metastatic Prostate Cancer: Evaluation of Drug Efficacy) was one of the first and longest running platform trials. Launched in the UK in 2005, it was designed to evaluate multiple treatments for high-risk and advanced prostate cancer.

STAMPEDE was a multi-armed clinical trial, which allowed new treatment arms to be added as promising therapies became available and ineffective interventions to be dropped following interim analyses. These were tested against the same control group and using the same master protocol (Sydes et al., 2012).

Over almost two decades, the platform recruited nearly 12,000 participants from 117 hospitals in the UK and Switzerland, with 10 different interventions tested. Several findings changed international standards of care for men with advanced prostate cancer (STAMPEDE, 2026; Park et al., 2022).

Rather than building a new trial for every promising treatment, researchers were able to use shared governance, participating sites, and trial procedures to evaluate successive innovations more rapidly (James et al., 2016).

BOX 3. The RECOVERY trial: rapid learning in a fast-moving environment

The speed at which researchers learnt about COVID-19 in the early days of the pandemic was comparable, in important respects, to the pace at which innovation around artificial intelligence is now occurring. Clinical understanding of the disease evolved rapidly, with new treatments appearing almost weekly, and the virus evolving. Researchers responded by generating a large amount of research. There were more than 320 studies of COVID-19 therapies launched in January 2020 alone, and during the first 10 months of the pandemic hundreds more followed. But many of these studies were too small to provide reliable answers. The median interventional trial enrolled only 100 participants, and 349 interventional trials enrolled 24 patients or fewer, making it difficult to distinguish genuine treatment effects from chance (McDonnell, 2022).

Small trials usually struggle to distinguish genuine treatment effects from random variation. COVID-19 had a roughly 2 percent mortality rate. In a trial of 100 patients, there is a reasonable chance that no deaths would occur in the control arm. Even if a treatment prevented every death, such a study of this size would not be able to reliably demonstrate the benefit. Researchers use power calculations to estimate the required trial size; these estimate the sample size needed to detect a given effect. In this example, about 770 participants would be required to have an 80 percent chance of detecting a treatment that eliminated all deaths, while more than 4,600 would be needed if the treatment reduced mortality by only half. During the first 10 months of the pandemic, there were 3,481 interventional studies started, but only 87 or 2.5 percent were large enough to identify a 50 percent reduction in mortality. Smaller treatment effects require larger studies still.¹ These underpowered trials had consequences. Several trials of the drug ivermectin had statistical powers of under 10 percent (Kow and Hasan, 2021). More generally, a recent meta-epidemiological analysis of 221 COVID-19 randomised control trials found that small trials systematically produced larger estimated treatment effects than large trials, had wider confidence intervals, and were substantially more fragile (Kim et al., 2026).

Rather than relying on a large number of underpowered studies, researchers needed a system that could keep generating robust evidence, whilst keeping pace with the changing clinical environment. The RECOVERY trial was designed explicitly to meet this challenge. Established near the start of the pandemic and embedded in the UK's NHS, RECOVERY was a large, adaptive platform trial. Multiple prospective treatments for COVID-19 were evaluated at the same time using a shared control group. Eligibility criteria and data collection requirements were minimized to enable rapid enrolment at scale. Whilst researchers at the University of Oxford coordinated the trial, it was delivered across hundreds of hospitals.

1 The use of surrogate endpoints can reduce the required sample size, as would studying a higher risk demographic such as people admitted to hospital with COVID-19, which is the group the RECOVERY trial focused on.

Interventions could be quickly added or removed from RECOVERY as evidence emerged. Treatments demonstrating clear effectiveness were added to clinical guidelines within months. Researchers continuously incorporated new hypotheses as the clinical and epidemiological context changed, so that evidence generation could keep pace with the changing COVID-19 landscape. At its peak, one in four people going into hospital with COVID-19 in England were being enrolled into this trial (Duncan, Halim, and El Kholy, 2022, Park et al., 2022).

COVID-19 was a clear medical emergency and the demand to find answers quickly made it quicker and easier to set up a platform trial than it will be for AI, but otherwise there are strong parallels. The large number of people working on COVID-19 led to innovation at pace. The huge increase in thinking power that is being created by AI is likely to speed up products and development in a similar way. Moreover, new variants of the virus altered the effectiveness of certain interventions over time. This has similarities to how AI models are frequently updated, retrained, or replaced, changing their performance characteristics.

RECOVERY demonstrates how speed does not have to come at the expense of rigour. By standardising processes and governance across sites, this PT was able to generate high-quality, policy-relevant evidence quickly enough to influence national and international guidelines. The ability to combine adaptability and scale is what makes trial networks a useful way to test products in a rapidly changing environment.

How platform trials could reduce the cost of trials

One of the benefits of PTs is that they are likely to reduce both the cost and time needed to set up and run a clinical trial. This is particularly true for long-running multi-country studies.

The largest savings come from sharing infrastructure across multiple studies, this would include sharing the cost of running trial sites, data systems, and agreeing governance arrangements. Processes such as recruitment pathways, safety monitoring, and statistical oversight can be reused for multiple products. Further to this, sharing control groups reduce recruitment requirements.

PTs can also reduce transaction costs. The 2023 O'Shaughnessy Review into how to improve clinical trial research in the UK (Department of Health and Social Care, 2023) found that much of the delay in commercial clinical research comes from repeated contracting and site set-up across NHS organisations, rather than from the scientific aspects of the trial itself. A standing platform with pre-agreed governance arrangements would avoid repeating much of this work.

PTs would also reduce indirect costs, by allowing researchers to conduct clinical trials more quickly. Because site infrastructure is already in place, research can begin much more quickly than if it had to be established for each trial, and there is a large commercial value to private companies in having their products approved quickly.

While savings will depend in part on trial use, evidence from the STAMPEDE trial, Box 2, demonstrates that PTs can greatly reduce the cost of research. Park and colleagues took STAMPEDE and modelled three ways of evaluating the same 10 interventions: as a single PT, as one multi-arm trial followed by five two-arm trials, or as a series of 10 separate two-arm trials. Cost and timing inputs were elicited from a survey of international experts with published work on PTs. Running the ten evaluations as separate two-arm trials cost a median 57.5 percent more and took 311.9 percent longer in cumulative duration than running them through a single platform. They found that STAMPEDE took longer and cost more money to set up initially compared to testing a two-armed study, highlighting that there is an up-front investment cost to getting a PT started. However, in the long run it was cost saving (Park et al., 2022).

Clinical trial network

A clinical trial network (CTN) is a network of trial-ready sites. This could include hospitals, clinics, and community health facilities that are able to support many different types of study. Participating sites would agree common standards on data security, research standards, and ethical reviews. Unlike a PT, each study would have its own protocol and eligibility criteria. Different studies will also seek to use a different subset of sites, depending on the question they are trying to answer.

This model would essentially operate as a network of “sandboxes” for testing AI-enabled interventions. Different studies could run at the same time across different groups of sites, depending on the technology and patient population being evaluated. For example, there could be a study testing an AI diagnostic tool in emergency departments, while another evaluates an AI mental health tool that works with primary care clinics to identify patients. Having a list of sites that might participate in the study will speed up early recruitment, as will the fact that these sites are used to running trials. This would be particularly helpful when researchers are looking for places to trial their product outside of their home market, and could reduce the challenges of running studies in LMICs. Box 4 illustrates how this kind of research infrastructure has been used to test treatments for antimicrobial resistance across Asia.

BOX 4. ADVANCE-ID: Building regional clinical trial infrastructure for antimicrobial resistance

Antimicrobial resistance poses a particular challenge for clinical research. There are no centres of excellence due to the need to treat patients quickly and the risk of resistant bacteria spreading. Instead, patients with resistant infections are spread across many hospitals and countries, making it difficult for individual centres to recruit enough participants for robust clinical trials. Companies trialling new antibiotics thus need to set up numerous trial sites each time a promising antibiotic needs to be tested, which can take many months.

ADVANCE-ID (Advancing Clinical Evidence in Infectious Diseases) was established to address this problem. This research programme is led by the National University of Singapore and supported by Wellcome. It is a network that provides a standing clinical trial infrastructure across Asia; there are trained investigators and trial-ready hospitals that can rapidly participate in multi-country studies. This means developers do not have to go out and recruit many hospitals every time they want to test a new product. This network has expanded rapidly and now includes 60 active sites across 20 countries, with more than 10,900 patients enrolled into its observational and clinical research programmes (ADVANCE-ID, 2026).

ADVANCE-ID highlights how shared research infrastructure can make clinical evaluation faster, cheaper, and more representative. A permanent platform allows evidence to be generated across multiple countries using consistent methods, without each developer starting from scratch. The same principle could be applied to AI-enabled health technologies, where products need to be evaluated across diverse clinical settings and populations (Wellcome Trust, 2026).

Embedded randomisation

Embedded pragmatic randomisation offers another promising approach for evaluating AI-enabled interventions. By integrating randomisation and outcome collection into routine care systems, such as electronic health records, these trials can generate evidence quickly and at relatively low cost. This may be particularly well suited to AI tools already integrated into clinical workflows. However, embedded randomisation does not by itself solve the need for comparisons across competing products or evaluation across different health systems and populations. Rather than being competing approaches, embedded randomisation could therefore form part of the infrastructure used by CTNs or PTs (van Staa et al., 2014, Sarkisian et al., 2026).

Choosing between a platform trial and a clinical trial network

PTs and CTNs both aim to make it cheaper, easier, and quicker to test new products. When choosing between them there is a trade-off between efficiency and flexibility: PTs are more efficient, whilst CTNs are more adaptable. The greater the alignment between the products being trialled, the more infrastructure can be shared—and as a result, the more efficiency gains there will be. With lots of

similar products being tested, a PT makes more sense. However, such an alignment can mean that a PT doesn't work well for a more unusual tool, and where greater flexibility is needed, a CTN is likely to make more sense.

For areas of medicine where there are multiple companies trying to produce products—for example, X-rays, triaging patients, or providing AI-enabled mental health therapy—a PT is likely to make the most sense. This allows products to be compared using the same outcomes, and where appropriate, shared control groups. This will make it easier to compare products, making results more useful for users, regulators, clinicians, and HTA.

CTNs become more attractive when products differ. If there are only one or two products that are likely to be tested, or if interventions differ substantially in their intended use, requiring every study to fit within a common protocol will not be feasible. However, a CTN will still make it easier to identify trial sites and get a trial up and running quickly. Table 2 outlines the differences between these two approaches.

TABLE 2. Difference between a platform trial and a clinical trial network

	Platform Trial	Clinical Trial Network
Best for	Similar competing products	Diverse or unusual products
Main advantage	Efficiency	Flexibility
Shared protocol	Yes	No
Shared controls	Usually	No
Main use	Comparative effectiveness	Independent evaluation
Example	STAMPEDE	ADVANCE-ID

CTNs and PTs aren't discreet. As happened during the RECOVERY trial, a CTN can also be used to establish a PT more easily. Where there are two products in a market, it might make sense for them to run a joint head-to-head trial conducted using a CTN. This would create a study that had many aspects of a PT, but without the standing infrastructure. Similarly, within PTs, some trials will be able to share more infrastructure than others, depending on the similarities between the products being tested. When designing a PT, researchers will need to think about the trade-offs between adaptability and efficiency. Box 5 outlines some illustrative examples of when it might be useful to use a PT and when a CTN could make more sense.

BOX 5. Illustrative use cases for platform trials or clinical trial networks

1. LLM-based therapy systems (e.g., AI-delivered Cognitive Behavioural Therapy)

There aren't enough trained therapists in the world to provide the mental health support to everyone who needs it. Even if there were, the costs of providing support in this way would be prohibitive for many of the world's poorest. LLM-based therapies offer the potential to overcome this problem and provide mental health support to people. The low marginal cost of delivering AI therapy makes it likely that a competitive market with multiple candidates emerges (Stade et al., 2024). Outcomes such as symptom scores, relapse rates, and treatment adherence can be measured consistently, making them good candidates for a PT. A shared platform would allow competing systems to be compared directly and make it easier to reassess them as models are updated (Rosenthal, Beecy, and Sabuncu, 2025).

2. AI rehabilitation therapy (e.g., stroke recovery or musculoskeletal rehabilitation)

The constraints on physiotherapy are similar to those for mental health therapies, in that both rely on repeated one-to-one interactions with trained professionals. This makes rehabilitation difficult to scale, leaving patients recovering from a stroke, joint replacement, or muscle injury without the level of support they ideally need. Even where services are available, patients often receive relatively little supervised therapy, as sessions are expensive. This leads to people doing most of their repetitive exercises on their own, which is difficult. AI tools could make sessions more interactive, as well as provide personalised rehabilitation programmes, monitor a patient's progress, and give patients real-time feedback (Abedi et al., 2024). This could either supplement the work of a physio or fill in where one is absent. The marginal cost of providing this support might be very low (unless specialised hardware is needed), meaning that it's likely a competitive market with multiple developers will develop. If this is the case, a PT would be well suited to test these technologies. This might be more complicated if some tools only focus on specific injuries, whilst others are more general. However, it should still be possible to create a PT that randomises people between any tool that is available for their condition.

3. AI advice tools for clinicians (e.g., diagnostic or prescribing support)

Clinicians often have little time to make decisions on patient care or on topics that they don't see very often. AI systems are increasingly able to support these decisions, saving time, reducing errors, and improving consistency. If we end up in a world where multiple tools exist to inform decision-making, and their endpoints are similar enough to allow direct comparison, then a PT would make sense. However, if there are a small number of these tools or they differ in scope, then a CTN might be more useful. It is also unclear whether these clinician-focused support tools need to be clinically tested in the same way, so the right answer might be neither.

In practice, we might end up in a world where there are a small number of very large CTNs. These could be adaptable and test first-of-kind products and those in areas where competition is low. These CTNs could then form the basis of identifying sites and setting up PTs when they are needed.

Challenges that remain

Inferencing

Networks and platforms do not solve a general inference problem. If during a clinical trial of a new tool, the foundational model that the tool is based on goes from v3.5 to v4.0, the options are to continue the trial and pool the patients in both groups, restart the arm, or treat it as a new arm—and each has a different bias and cost. Frameworks need to be agreed on how to solve this problem, and the exact approach will depend on the nature of any trial and product update.

This will be more manageable for interventions that are used repeatedly over time, such as AI-delivered therapies or clinical decision-support systems. In real world use, users would likely be switched to the latest version of a product shortly after its release, and a trial could mirror this. Rather than treating every update as a new intervention, researchers could compare outcomes before and after major model revisions, to ensure outcomes are similar.

Where trials need to be adapted to deal with updates and questions around inference, PTs and CTNs should be able to make this easier. Their ability to introduce new trial arms and recruit participants quickly, and to maintain long-term evaluation, makes them better suited than repeatedly establishing conventional standalone trials. That they keep running against the same control group also makes it possible to measure the treatment effect before and after any update.

Incentives

Encouraging industry participation in head-to-head studies may prove difficult. Developers often favour single-arm or non-comparative studies because direct comparisons create the possibility that a competing product will outperform theirs. Developers of products already selling well will be particularly wary about a study that might undermine their market share, which is problematic—as these are the tools it is most useful to test. The fast nature of AI development also makes it harder to do these studies without their support than it would be in other areas of medicine.

There are three ways this can be addressed:

1. Lower the cost of participation. If these trial designs make it significantly cheaper or quicker to generate evidence, participation may become commercially attractive despite the risks of comparison and high upfront costs.

2. Demand for comparative evidence. HTA bodies, procurement agencies, and clinicians could demand head-to-head, multi-country evidence before procuring a product. Products lacking comparative evidence may therefore struggle to gain market share. This already happens in some areas of medicine where the absence of direct comparative evidence can delay reimbursement, procurement, or widespread adoption while additional evidence is generated (Guo, Gehchan, and Hartzema, 2025; Drummond et al., 2008).
3. Require comparative evidence. Regulators could require head-to-head, continuous multi-country studies, PTs are likely the best way of undertaking these.

Costs

Both PTs and CTNs will require long-term coordination and investment. Infrastructure and staffing will need to be maintained, even when recruitment is low. More work is needed to understand how this infrastructure can be funded.

One option would be for contract research organisations (CROs) to establish and operate networks on behalf of multiple sponsors, recovering their costs through study fees. However, the commercial case may initially be weak. Building standing infrastructure will require substantial upfront investment, while demand from developers may be uncertain until the value of the model has been demonstrated.

PTs and CTNs could greatly improve the standard of evidence on AI products, and by extension, improve people's health. If the private sector is unwilling to fund the creation of this infrastructure, governments or philanthropic organisations could underwrite some of the fixed costs. Should this infrastructure become widely used, then costs could be recouped from developers using it.

The exact funding nature might vary per trial, with publicly motivated funders paying for platforms or networks that tackle problems they see as priorities, and CROs setting up studies in more lucrative markets.

Data and technical infrastructure

AI tools may need access to electronic health records or other patient data, raising questions about data ownership, security, interoperability, and whether data and models can be transferred across borders or need to be hosted locally. These challenges are likely to be greatest in LMICs, where many trial sites have limited digital infrastructure and technical capacity. A CTN would therefore need to invest not only in conventional trial capacity, but also in common data standards, secure computing infrastructure, and local technical expertise.

There are also important questions around patient control of data. AI evaluation may involve retaining clinical data for purposes that are difficult to specify when consent is first obtained, particularly where models are subsequently updated or data are reused for continued evaluation. There will need to be well designed rules on who can access trial data, how long these data can be

retained for, what secondary uses are permitted, and whether patients can withdraw consent for their data to be used. Obtaining meaningful informed consent may be particularly challenging where digital or health literacy is low. Consent processes will therefore need to explain potential future uses of data in accessible terms without relying on patients to understand the technical details of AI systems.

Access to innovation

There are also questions around post-trial access to products. If an AI intervention is found to be effective, it is important to make sure that participants and people from the population from which they are recruited have access to the product once the trials are completed. This is particularly important in lower-income settings. Otherwise, there is a risk that these population will contribute to the generation of evidence without sharing in its benefits. Rules around post-trial access should be agreed before any product is tested in a PT or CTN. This should be in line with the Council for International Organizations of Medical Sciences' guidelines, which says that communities should benefit from research they are part of. Stating that "the community or population will not benefit from successful research unless the knowledge and interventions that it produces are made available to them and products are reasonably priced" (CIOMS, 2016).

Unit of randomisation

The level of randomisation needed for a study depends on the type of intervention. Pharmaceutical treatments impact the patient directly, so randomisation happens at the patient level, and the same would be true for AI tools that are patient-facing, like therapists. However, randomisation becomes more challenging for interventions that are clinician-facing or operate on healthcare centres (many of these interventions probably won't need clinical trials). It often won't be possible to test a decision support tool or triage system on some of a clinician's patients but not others, as exposure to the tool might change their behaviour in ways that persist when treating other patients. To counter this, many evaluations will require randomisation at the level of the clinician, practice, or hospital rather than the individual patient.

Cluster-randomised trials require many more sites than patient-randomised studies because statistical power depends on the number of independent clusters rather than patients. CTNs would make it easier to recruit the larger number of sites needed for a study like this.

Limitations

There are several limitations to this work that need to be acknowledged, most of which can be addressed by follow-up research.

It is not clear what level of evidence will be demanded for different categories of AI systems—whether from regulators or clinicians approving them, or health systems purchasing a given product. Any trial system will need to be designed around meeting these evidentiary needs.

This paper focuses on technologies that are only emerging, there are thus few empirical examples to inform the discussion. However, frameworks for how to evaluate products should be ready by the time they come to market, so it makes sense to start these discussions now.

Trial sites in a network may not always be representative of standard clinical practice. If evaluation becomes concentrated in a relatively small number of research-active hospitals or clinics, participating sites may differ meaningfully from where products are deployed. Staff may be more research-oriented, infrastructure more advanced, or clinical workflows more standardised. PTs and CTNs thus need to include a broad range of sites and update the sites available to ensure they remain representative.

Next steps

AI is likely to lead to a large number of new products that are fundamentally very different from anything currently available. This leads to important questions about how these tools are evaluated, both to satisfy regulators and generate the evidence needed to support use. This paper has attempted to set out some of the challenges measuring these tools and the role that PTs and CTNs could play in overcoming them.

These are presented as promising ideas which might play a useful role in future evaluations. However, this paper is intended as a thought piece to encourage, rather than to put forward a fully formed plan. More work is needed to test this idea, assess how this infrastructure would work, and whether there is the support for taking them forward.

There are six next steps that could be taken to test and progress this idea:

- 1. Identify where each model adds value**

Develop a framework that determines which categories of AI are best suited to PTs and CTNs, based on factors such as the pace of innovation, degree of competition, and whether it is a patient or clinician facing intervention.

- 2. Assess the costs, benefits, and feasibility**

Produce detailed cost estimates for different network models in different disease areas. Quantify the expected savings compared with conventional trials. Then assess the potential health and economic benefits of improving the quality and speed of evidence generation.

- 3. Funding models**

Evaluate different funding models for sustaining a PT and a CTN. This should include looking at a CRO-based model, and one underwritten by governments or philanthropy.

4. Test the idea with stakeholders

Engage key stakeholders (including regulators, HTA agencies, health systems, researchers, AI developers, CROs, and patient groups) to socialise these ideas, whether this is an approach they support, how it should be designed, and what incentives should be put in place.

5. Design a pilot

Develop a detailed implementation plan for one or more pilot studies, and identify the most suitable tools for an initial study.

6. Run and evaluate the pilot

Establish and evaluate a pilot PT or CTNs to assess how these trial designs would work in practice.

References

- Abedi, A., Colella, T.J.F., Pakosh, M. and Khan, S.S. (2024) 'Artificial intelligence-driven virtual rehabilitation for people living in the community: a scoping review', *npj Digital Medicine*, 7(1), 25.
- ADVANCE-ID (2026) ADVANCE-ID: clinical trial network. Available at: <https://www.advance-id.network/> (Accessed: 11 August 2026).
- Australian Clinical Trials Alliance (2019) Establishment of new clinical trial networks. Melbourne: Australian Clinical Trials Alliance. Available at: <https://clinicaltrialsalliance.org.au/> (Accessed: 4 February 2026).
- Council for International Organizations of Medical Sciences (2016) International ethical guidelines for health-related research involving humans. 4th edn. Geneva: Council for International Organizations of Medical Sciences. Available at: <https://cioms.ch/wp-content/uploads/2017/01/WEB-CIOMS-EthicalGuidelines.pdf> (Accessed: 29 September 2026).
- Department of Health and Social Care (2023) Commercial clinical trials in the UK: the Lord O'Shaughnessy review – final report. London: Department of Health and Social Care. Available at: <https://www.gov.uk/> (Accessed: 10 August 2026).
- Djulbegovic, B. and Guyatt, G.H. (2017) 'Progress in evidence-based medicine: a quarter century on', *The Lancet*, 390(10092), pp. 415–423.
- Drummond, M.F., Schwartz, J.S., Jönsson, B., Luce, B.R., Neumann, P.J., Siebert, U. and Sullivan, S.D. (2008) 'Key principles for the improved conduct of health technology assessments for resource allocation decisions', *International Journal of Technology Assessment in Health Care*, 24(3), pp. 244–258.
- Duncan, A., Halim, D. and El Kholy, K. (2022) 'The RECOVERY trial: an analysis and reflection two years on', *European Journal of Internal Medicine*, 105, pp. 111–112.
- Food and Drug Administration (2019) Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD): discussion paper and request for feedback. Silver Spring, MD: US Food and Drug Administration.
- Food and Drug Administration (2024) Predetermined change control plans for medical devices: guidance for industry and Food and Drug Administration staff. Silver Spring, MD: US Food and Drug Administration. Available at: <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/predetermined-change-control-plans-medical-devices> (Accessed: 21 February 2026).
- Goetz, L., Seedat, N., Vandersluis, R. and van der Schaar, M. (2024) 'Generalization – a key challenge for responsible AI in patient-facing clinical applications', *npj Digital Medicine*, 7(1), 126.

- Guo, J.D., Gehchan, A. and Hartzema, A. (2025) 'Selection of indirect treatment comparisons for health technology assessments: a practical guide for health economics and outcomes research scientists and clinicians', *BMJ Open*, 15(3), e091961.
- Hou, K. and Gattinger, G. (2024) 'The evolving regulatory paradigm of AI in MedTech: a review of perspectives and where we are today', *Therapeutic Innovation & Regulatory Science*, 58.
- Institute of Medicine (2009) Initial national priorities for comparative effectiveness research. Washington, DC: The National Academies Press.
- James, N.D., Sydes, M.R., Clarke, N.W., Mason, M.D., Dearnaley, D.P., Spears, M.R. et al. (2016) 'Addition of docetaxel, zoledronic acid, or both to first-line long-term hormone therapy in prostate cancer (STAMPEDE): survival results from an adaptive, multiarm, multistage, platform randomised controlled trial', *The Lancet*, 387(10024), pp. 1163–1177.
- Kim, D.H., Lim, S., Eisenhut, M., Kronbichler, A., Kim, E., Kim, M.S. et al. (2026) 'The impact of study size on COVID-19 treatment outcomes: a meta-epidemiological study comparing large and small randomized controlled trials', *Reviews in Medical Virology*, 36(2), e70125.
- Kow, C.S. and Hasan, S.S. (2021) 'Pitfalls in reporting sample size calculation across randomized controlled trials involving ivermectin for the treatment of COVID-19', *American Journal of Therapeutics*, 28(5), pp. e616–e619.
- McDonnell, A., Rex, J.H., Goossens, H., Bonten, M., Fowler, V.G. and Dane, A. (2016) 'Efficient delivery of investigational antibacterial agents via sustainable clinical trial networks', *Clinical Infectious Diseases*, 63(suppl_2), pp. S57–S59.
- McDonnell, A. (2022) We need better clinical trial systems before the next pandemic: the World Health Assembly's resolution is a good first step. Washington, DC: Center for Global Development. Available at: <https://www.cgdev.org/blog/we-need-better-clinical-trial-systems-next-pandemic-world-health-assemblys-resolution-good> (Accessed: 6 August 2026).
- Meijerink, L.M., Dunias, Z.S., Leeuwenberg, A.M., de Hond, A.A.H., Jenkins, D.A., Martin, G.P., Sperrin, M., Peek, N., Spijker, R., Hooft, L., Moons, K.G.M., van Smeden, M. and Schuit, E. (2025) 'Updating methods for artificial intelligence–based clinical prediction models: a scoping review', *Journal of Clinical Epidemiology*, 178, 111636.
- Park, J.J.H., Sharif, B., Harari, O., Dron, L., Heath, A., Meade, M., Zarychanski, R., Lee, R., Tremblay, G., Mills, E.J., Jemai, Y., Mehta, C. and Wathen, J.K. (2022) 'Economic evaluation of cost and time required for a platform trial vs conventional trials', *JAMA Network Open*, 5(7), e2221140.
- Pericàs, J.M., Derde, L.P.G. and Berry, S.M. (2023) 'Platform trials as the way forward in infectious disease clinical research: the case of coronavirus disease 2019', *Clinical Microbiology and Infection*, 29(3), pp. 277–280.
- Rosenthal, J.T., Beecy, A. and Sabuncu, M.R. (2025) 'Rethinking clinical trials for medical AI with dynamic deployments of adaptive systems', *npj Digital Medicine*, 8, 252.

- Roustit, M., Demarcq, O., Laporte, S., Barthélémy, P., Chassany, O., Cucherat, M. et al. (2023) 'Platform trials', *Thérapie*, 78(1), pp. 29–38.
- Sarkisian, C., Ibrahim, K., Vangala, S., Villaflores, C.W., Cheng, E.M., Turner, W. et al. (2026) 'Embedded point of care stratified block randomization: demonstration of the point of care randomization (POCR) engine with an electronic health record pragmatic clinical trial', medRxiv [Preprint]. doi: [10.64898/2026.01.26.26344847](https://doi.org/10.64898/2026.01.26.26344847).
- Silcox, C., Zimlichmann, E., Huber, K., Rowen, N., Saunders, R., McClellan, M., Kahn, C.N., Salzberg, C.A. and Bates, D.W. (2024) 'The potential for artificial intelligence to transform healthcare: perspectives from international health leaders', *npj Digital Medicine*, 7(1), 88.
- Stade, E.C., Wiltsey Stirman, S., Ungar, L.H., Boland, C.L., Schwartz, H.A., Yaden, D.B., Sedoc, J., DeRubeis, R.J., Willer, R. and Eichstaedt, J.C. (2024) 'Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation', *npj Mental Health Research*, 3(1), 12.
- STAMPEDE (2026) About STAMPEDE. Available at: <https://www.stampedetrial.org/about-stampede/> (Accessed: 11 August 2026).
- Sydes, M.R., Parmar, M.K.B., Mason, M.D., Clarke, N.W., Amos, C., Anderson, J. et al. (2012) 'Flexible trial design in practice – stopping arms for lack-of-benefit and adding research arms mid-trial in STAMPEDE: a multi-arm multi-stage randomized controlled trial', *Trials*, 13, 168.
- van de Sande, D., Fung Fen Chung, E., Oosterhoff, J., van Bommel, J., Gommers, D. and van Genderen, M.E. (2024) 'To warrant clinical adoption AI models require a multi-faceted implementation evaluation', *npj Digital Medicine*, 7(1), 58.
- van Staa, T.P., Dyson, L., McCann, G., Padmanabhan, S., Belatri, R., Goldacre, B. et al. (2014) 'The opportunities and challenges of pragmatic point-of-care randomised trials using routinely collected electronic records: evaluations of two exemplar trials', *Health Technology Assessment*, 18(43), pp. 1–146.
- Weissman, G.E. (2025) 'Evaluation and regulation of artificial intelligence medical devices for clinical decision support', *Annual Review of Biomedical Data Science*, 8(1), pp. 81–99.
- Wellcome Trust (2016) Clinical trial networks for antibiotic development. London: Wellcome Trust.
- Wellcome Trust (2026) Phase 2 clinical trial of a new antimicrobial for drug-resistant infection. London: Wellcome. Available at: <https://wellcome.org/> (Accessed: 12 August 2026).
- World Health Organization (2021a) Global strategy on digital health 2020–2025. Geneva: World Health Organization.
- World Health Organization (2021b) WHO publishes new guidance to promote strong, efficient and sustainable regulatory systems. Geneva: World Health Organization. Available at: <https://www.who.int/news/item/29-04-2021> (Accessed: 10 August 2026).

World Health Organization (2024) Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. Geneva: World Health Organization.

Yang, J., Dung, N.T., Thach, P.N., Phong, N.T., Phu, V.D., Phu, K.D., Yen, L.M., Thy, D.B.X., Soltan, A.A.S., Thwaites, L. and Clifton, D.A. (2024) 'Generalizability assessment of AI models across hospitals in a low-middle and high income country', *Nature Communications*, 15, 8270.