MAYO CLINIC

MAYO CLINIC PROCEEDINGS: DIGITAL HEALTH

SPECIAL ARTICLE

Assessing Artificial Intelligence Solution Effectiveness: The Role of Pragmatic Trials

Mauricio F. Jin, MD; Peter A. Noseworthy, MD; and Xiaoxi Yao, PhD

Abstract

The emergence of artificial intelligence (AI) and other digital solutions in health care has considerably altered the landscape of medical research and patient care. Rigorous evaluation in routine practice settings is fundamental to the ethical use of AI and consists of 3 stages of evaluations: technical performance, usability and acceptability, and health impact evaluation. Pragmatic trials often play a key role in the health impact evaluation. The current review introduces the concept of pragmatic trials, their role in AI evaluation, the challenges of conducting pragmatic trials, and strategies to mitigate the challenges. We also examined common designs used in pragmatic trials and highlighted examples of published or ongoing AI trials. As more health systems advance into learning health systems, where outcomes are continuously evaluated to refine processes and tools, pragmatic trials embedded into everyday practice, leveraging data and infrastructure from delivering health care, will be a critical part of the feedback cycle for learning and improvement.

 ■ Mayo Clin Proc Digital Health 2024;2(4):499-510

Over the past several years, artificial intelligence (AI), large language models (LLM), and other digital solutions have begun to be adopted in health care. These advancements present both new challenges and opportunities and may change many fundamental operations of health systems. These innovations, either developed by embedded scientists or available as vended products, include technologies such as AI-guided disease screening and decision support tools, AI-based support tools to facilitate novel care delivery models (ie, remote care), and LLM to assist with clinical note-writing or record aggregation/summarization.[1–3] Furthermore, unlike drugs, devices, or other traditional interventions that are presented to the health system as static entities, digital interventions may evolve over time as they are used within the health system, and their evaluation and implementation require a much shorter timeline and additional considerations from technical, ethical, and legal perspectives.

Indeed, harnessing this disruptive power of AI to transform health care inevitably surfaces a collision of 2 cultures: the technology innovation and entrepreneurship ethos of move fast and break things and the health care culture grounded in the imperatives to both deliver high-quality care and safeguard patient safety, that is, first, do no harm. Pragmatic trials, traditionally used to assess the effectiveness of medical interventions in routine clinical practice, have shown a promising role in evaluating of health impact of AI in large populations, leveraging cutting-edge technologies and emerging data.[4] The current article aims to provide an overview of the role of pragmatic trials in AI evaluation and illustrate their utility with a few examples.

From the Department of Internal Medicine (M.F.J.), Department of Cardiovascular Medicine (P.A.N., X.Y.), and Robert D and Patricia E Kern Center for the Science of Health Care Delivery (X.Y.), Mayo Clinic, Rochester, MN.

AI EVALUATION

AI evaluation consists of 3 stages—technical performance evaluation, usability and acceptability evaluation, and health impact evaluation. The evaluation of technical performance remains under the data scientists' expertise, where the algorithms are being evaluated externally in other populations,[5,6] and in different patient subgroups, such as those with different races and ethnicities.[7] In many cases, validation in a prospective study is also encouraged.[8,9]

The usability and acceptability evaluation of AI is where the AI is translated to inform clinical decision-making through an interface, such as data and alerts through dashboards, apps, or electronic health records (EHRs).

The evaluation of usability and acceptability falls under the realm of human-AI interaction and often requires mixed-methods studies, stakeholder engagement, and iterative revisions of the interface. This stage is often neglected as investigators are excited to jump on large-scale health impact evaluation. However, if the users, eg, clinicians and patients, do not find the AI tool useful or easy to use, then even a well-designed randomized controlled trial (RCT) cannot detect an impact that does not exist due to low adoption in the population. In comparison to other interventions, AI faces particular challenges related to trust,[10] so engaging stakeholders early and resolving usability and acceptability is highly recommended before starting the health impact evaluation.

When evaluating the health impact of AI, the methods are largely the same as any population-level evaluation of other medical interventions, which can be grouped into as follows: (1) RCTS and (2) nonrandomized studies. Clinical trials, by definition, can be either randomized or nonrandomized. Per National Institutes of Health definition, a study is considered a clinical trial if the answers to the following 4 questions are all yes:[11]

- Does the study involve human participants?
- Are the participants prospectively assigned to an intervention?
- Is the study designed to evaluate the effect of the intervention on the participants?
- Is the effect being evaluated a health-related biomedical or behavioral outcome?

A clinical trial can be used for a diverse range of strategies for prevention, screening, diagnosis, and treatment of medical conditions, and interventions to improve care delivery, reduce health disparities, and improve medical communications and decision-making processes. A nonrandomized trial is common and can be a viable option for evaluating AI as well, especially when borrowing real-world controls.[12] In fact, the incorporation of real-world data as the external controls to either enrich or replace the control arm has attracted increasing interest, especially in oncology and rare diseases.[13,14]

Among the nonrandomized studies, other than the nonrandomized clinical trials discussed above, prospective, and retrospective observational studies can be used for AI evaluation. A key difference between a prospective observational study and a nonrandomized clinical trial is that a nonrandomized clinical trial assigns interventions to the participants although the assignment does not include randomization, whereas a prospective observational study does not assign intervention to participants. However, as with any nonrandomized studies, confounding needs to be carefully managed, using tools such as propensity score, interrupted time series or regression discontinuity, and instrumental variable.[15–17]

PRAGMATIC TRIALS AND AI

Clinical trials and AI have a bidirectional relationship—clinical trials can be used to evaluate AI, while AI and other digital solutions can facilitate the conduct of clinical trials. Large RCTs have long been the gold standard for obtaining evidence on whether an intervention works, as all the above mentioned methods to control confounding in nonrandomized studies have limitations.

The RCTs can be developed into 2 categories: (1) explanatory trials that aim to evaluate the efficacy of an intervention and (2) pragmatic trials that aim to evaluate the effectiveness of an intervention.[18,19] In other words, explanatory trials aim to investigate the mechanism of the intervention and answer the question—"Can the intervention work under ideal conditions?", whereas pragmatic trials seek to inform real-world decisions by answering, "Can the intervention work under usual care?" Explanatory trials strive to achieve high internal validity, whereas pragmatic trials strive to increase generalizability (Figure 1).

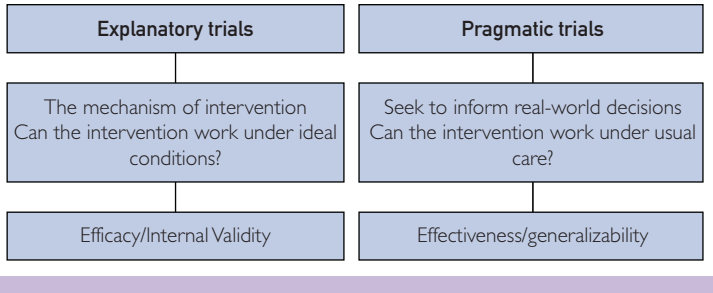

FIGURE 1. Explanatory trials vs pragmatic trials.

Mayo Clin Proc Digital Health ■ December 2024;2(4):499-510 ■ https://doi.org/10.1016/j.mcpdig.2024.06.010
www.mcpdigitalhealth.org

The concept of pragmatic trials was mentioned in the 1960s.[20] Over the decades, an increasing number of pragmatic trials have been published (Figure 2). This is somewhat driven by an increasing awareness that clinicians and patients lack high-quality evidence to guide decisions. For example, only 10% of the American College of Cardiology and American Heart Association's guideline recommendations were supported by multiple RCTs or meta analyses; nearly half of the guideline recommendations were based on expert opinions, case studies, or standards of care; and this situation has not improved much over time.[21–23] As such, many clinical decisions have been made through the influence of anecdotal experiences,[24,25] exposure to marketing messages,[26,27] and financial incentives.[28] Furthermore, health care systems often implement quality improvement projects and other innovations wholesale because they appear to be good ideas, but few were rigorously evaluated,[29] thereby wasting the resources that could be potentially used to address other health care priorities.

The pragmatic trial is a suitable tool and has played an increasingly important role in the evaluation of AI and its health impact. First, explanatory trials focus on the efficacy of medical interventions, but the efficacy of AI has already been reported during the technical performance evaluation before progressing to the health impact evaluation. Few people at the health impact stage would be interested in questions like "Can AI work under ideal conditions?" More stakeholders would be interested in whether the AI tool works in routine practice, so clinicians and patients can decide on prevention, diagnosis, or treatment strategies, and health systems can decide on whether to implement the AI tool.

Second, because new drugs or devices can incur substantial harm to patients, their efficacy and safety need to be reported in explanatory phase 1-3 trials before being tested in pragmatic trials embedded in routine practice. By contrast, many AI tools are used to support clinical decision-making by providing predictions or increasing operational efficiency, so they are used for lower or minimal risk scenarios. Therefore, many of the AI tools might not need the phase 1-3 trials. In this aspect concerning safety, AI is similar to other interventions commonly evaluated in pragmatic trials, such as screening and quality improvement programs.[30,31]

Of interest, the medical and research communities have witnessed a shift in the type of trials published. In a review of 134,144 completed trials from 2000 to 2019, the authors found in early years such as 2000-2004, over 70% of the interventions in trials were drugs, and over 80% of the trials were phase 1-4 trials. However, in recent years, such as 2015-2019, the proportion of trials evaluating drugs dropped to 40%, and nearly 60% of the trials were Phase NA (Table).[32] Some pragmatic trials can be counted as phase 4, but many pragmatic trials are Phase NA trials because they do not have any phase 1-3 trials before the current trial. Going forward, pragmatic trials will be increasingly used to assess AI and other digitally enabled innovations.

Pragmatic trials have their own advantages and challenges as study designs. Contrary to

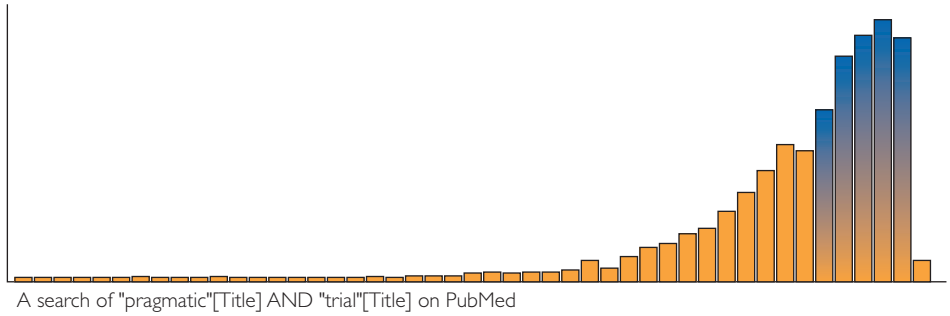

FIGURE 2. Increased N of pragmatic trials published in recent years.

TABLE. Design Characteristics of Completed Trials by Lead Sponsor and Start Year

Lead sponsor	Government	Industry	Other	Total
2000-2004				
Intervention: drug	61.4%	87.6%	59.2%	70.5%
Phase 1-4	80.5%	95.8%	71.3%	81.8%
Phase NA	19.5%	4.2%	28.7%	18.2%
2015-2019				
Intervention: drug	30.4%	69.8%	26.2%	40.0%
Phase 1-4	51.2%	77.4%	26.7%	42.4%
Phase NA	48.8%	22.6%	73.3%	57.6%

Abbreviation: NA, not applicable.
Redrawn from Table 2. *JAMA Netw Open*,[32] with permission.

the notion that they are inherently better or easier than explanatory trials, the research community recognizes that the choice between a pragmatic trial and an explanatory trial hinge on the trial's intent. Specifically, one must consider whether the goal is to understand the mechanism of action of an intervention (explanatory trial) or to inform real-world decisions (pragmatic trial).[33]

Pragmatic trials are well-suited for evaluating the health impact of AI technologies in large populations. They offer the advantage of informing real-world decisions with a focus on patient-centered outcomes and can be generalized to diverse routine practice settings. Many AI technologies are inherently embedded in routine clinical practice, and their efficacy has often been reported through technical performance, usability, and acceptability evaluations. Consequently, the primary intent during the final stage of AI evaluation is to inform real-world decisions rather than understand the mechanism of action.

Furthermore, the effectiveness of AI in routine practice depends on how it is implemented. Implementation trials, ie, those trials that generate scientific knowledge to improve the uptake of evidence-based interventions in practice, are, by nature, pragmatic trials.[34] There are also hybrid trials with dual aims to test both the effectiveness of the intervention and the outcome of implementation strategies.[35]

CHALLENGES WITH PRAGMATIC TRIALS AND STRATEGIES TO OVERCOME THEM

It is now nearly 60 years since the conceptualization of explanatory and pragmatic trials, and the enthusiasm for pragmatic trials has increased substantially since the turn of the century, likely because of the demand for real-world evidence at all levels, eg, patients, caregivers, clinicians, health systems, payers, and policy-makers, to inform decision-making.[20] The rise of the learning health system (LHS) and the emergence of digital tools also help overcome the traditional barriers to conducting pragmatic trials.

Challenges of Conducting Pragmatic Trials

To design a pragmatic trial, investigators use the PRagmatic Explanatory Continuum Indicator Summary-2 tool to help design trials that are fit for purpose. The PRagmatic Explanatory Continuum Indicator Summary-2 tool rates the extent of pragmatism from 9 dimensions:[36]

(1) Eligibility: To what extent are participants in the trial similar to those in usual care? A highly pragmatic approach would be to include anyone who is likely to be a candidate for the intervention if it were being provided in usual care.
(2) Recruitment: How much extra effort is made to recruit participants over and above what would be used in the usual care setting to engage patients? A highly pragmatic approach would be to recruit through usual care at a diverse range of clinics and hospitals.
(3) Setting: How different are the settings of the trial from the usual care setting? A highly pragmatic approach would be to perform the trial in a similar or identical setting to which you intend the results to be applied.
(4) Organization: What expertise and resources are needed to deliver the intervention? A highly pragmatic design would not

Mayo Clin Proc Digital Health ■ December 2024;2(4):499-510 ■ https://doi.org/10.1016/j.mcpdig.2024.06.010
www.mcpdigitalhealth.org

require a lot of additional clinical expertise or resources that are difficult to obtain or do not exist in that setting.
(5) Flexibility (delivery): How much flexibility is there to deliver the intervention? In a highly pragmatic trial, how to deliver an intervention is not rigidly prescriptive, leaving the details up to providers.
(6) Flexibility (adherence): What measures are in place to ensure participants adhere to the intervention? A highly pragmatic design would allow for full flexibility in how end users engage with the intervention.
(7) Follow-up: How closely are participants followed up? The most pragmatic approach would be to have no more follow-up of recipients than would be the case in usual care.
(8) Primary outcome: How relevant is it to participants? A pragmatic approach would be to select an outcome that is of obvious importance from the participants' perspective.
(9) Primary analysis: To what extent are all data included? A pragmatic approach would be an intention-to-treat analysis using all available data.

Although there is no consensus regarding above which threshold a trial should be considered pragmatic rather than explanatory, most would agree that a pragmatic trial should be rated as rather or very pragmatic in most of the 9 dimensions,[33,37] which leads to the challenges with conducting pragmatic trials.

One accepted limitation of the pragmatic trial is that the increased generalizability or external validity comes at the cost of the reduced internal validity. Few pragmatic trials are able to use blinding or placebo controls. When allowing flexibility in care delivery and adherence, cross-over between randomized arms and lack of adherence are common; requiring primary analysis, such as all data in an intent-to-treat analysis, often leads to an estimate of treatment effect toward null. When requiring no more follow-up than routine practice and relying on existing data collected from care delivery to ascertain outcomes, it runs into the risk that different sites might obtain the measurement using different approaches or definitions, and the baseline characteristics and outcomes can be misclassified. Therefore, the choice of outcomes in a pragmatic trial is often limited to those that are both readily available in EHR and least likely to be subject to site-level or clinician-level variation. For example, all-cause mortality would be more commonly used than cause-specific mortality.[38]

A less accepted or more challenging limitation of pragmatic trials is the requirement of large sample sizes and diverse practice settings. First, the patient-centered outcomes in pragmatic trials are often less sensitive to an intervention than the outcomes in explanatory trials. For example, all-cause mortality would be less sensitive to an intervention than cause-specific mortality or a biomarker. Second, the treatment effect from explanatory trials, which select patients who are most likely to benefit and managed in academic medical centers by clinicians with extensive training and experience with the intervention, would be greater than that in pragmatic trials. Third, cluster RCT, a common design of pragmatic trials, requires a larger sample size due to the intracluster correlation of the outcomes. Furthermore, cluster RCT requires a sufficient number of clusters, eg, clinicians, hospital floors, and sites. As a result, to achieve high pragmatism, trials require a large number of participants and sites. Everything else being equal, the greater the sample size, the more resources and time it requires. Similarly, everything else being equal, performing a pragmatic trial from a large number of sites, including community practices, is more challenging than conducting an explanatory trial within a few academic medical centers. In the following sections, we will discuss how the rise of LHS and the emergence of digital tools help mitigate these challenges.

Rise of LHS

Traditional clinical trials rely on a parallel research infrastructure outside routine clinical practice to recruit patients, deliver interventions, and measure outcomes. As such, many people believe that pragmatic trials embedded into routine clinical practice could reduce the cost of conducting trials. However, this perspective oversimplifies the reality. Instead of straightforward cost reduction, it often represents a transfer of costs from research to

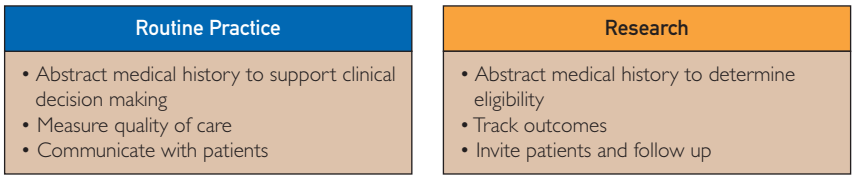

FIGURE 3. Same set of digital tools applied to practice and research within a learning health system.

practice. Frontline staff—physicians, nurses, and administrative personnel—become integral to the trial process. Their involvement necessitates time, effort, and resources, which may offset any initial cost savings. Moreover, trialists encounter resistance, especially when clinicians are already burdened with clinical tasks and lack extensive research training or interest. Explaining concepts like randomization to patients and introducing research studies can be uncomfortable for clinicians primarily focused on patient care. Historically, health systems have provided limited incentives for staff, particularly nonphysician personnel and those outside academic medicine, to engage in research. In addition, relying on clinical staff without extensive research training and experience might not be efficient as additional training, monitoring, and quality assurance are needed. Neglecting these procedures can jeopardize human subject protection and compromise the validity of the research.

The rise of LHS over the past 15 years has mitigated this challenge to some extent. Per the Agency for Health Care Research and Quality definition, an LHS is a health system in which internal data and experience are systematically integrated with external evidence, and that knowledge is put into practice.[39] A critical role of LHS lies in shifting the mindset of health system leaders and staff. There has been growing recognition that a health system's responsibility extends beyond merely delivering clinical services and generating evidence is an integral part of health care delivery. Consequently, health system leadership is more inclined to accept the costs associated with pragmatic trials. Clinical units, which traditionally focused solely on practice, are now more willing to accommodate additional requests from research teams. Moreover, many health systems provide incentives for staff—especially those outside academic medicine—to engage in research. There is also a greater understanding that these costs and burdens from embedded research represent necessary investments in providing high-quality care, positioning health systems competitively, and reaping financial benefits tied to performance-based reimbursement.[40]

In addition to changing the mindsets of the health systems and staff, another characteristic of LHS is the investment in building a robust data and IT infrastructure to support clinical decision-making and measure the quality of care. The same set of tools can be used in research, especially embedded pragmatic trials (Figure 3). For example, both the EAGLE and BEAGLE trials[41–43] used both structured and unstructured data to abstract patients' medical history to determine patient eligibility and follow-up on outcomes. These were the same set of tools used to power quality dashboards and other measurements. In the BEAGLE trial,[43] patients were invited either electronically through the patient portal or postal mail based on their indicated communication preference in the EHR, which is the same workflow used to communicate with patients in routine practice.

Digital Tools and Decentralized Trials

One way to overcome the challenges of conducting pragmatic trials is to shift the mindset, engage staff, and more efficiently use the information from routine practice with the data and IT infrastructure, which is facilitated by the rise of LHS. The other method involves bypassing the clinical and administrative staff and reducing their burden through digital tools and decentralized trial design.[44] Traditionally, clinical trials are centralized in that all trial procedures are conducted at research

Mayo Clin Proc Digital Health ■ December 2024;2(4):499-510 ■ https://doi.org/10.1016/j.mcpdig.2024.06.010
www.mcpdigitalhealth.org

sites; a fully decentralized trial will be a trial where all procedures will be conducted virtually or remotely without requiring participants to visit a research site. Fully decentralized trials are still rare, whereas most trials going forward will be hybrid in that some procedures will be conducted onsite and some will be conducted virtually using telemedicine, mobile health apps, wearable devices, and electronic consent to collect data and monitor participants. This allows for more accessibility and convenience for a broader range of participants, but there exist limitations, such as challenges related to technology access and literacy and maintaining participant engagement over time without face-to-face interactions.

Both pragmatic and explanatory trials can be decentralized. but pragmatic trials are often easier to decentralize. First, pragmatic trials are mostly minimal risk research; often, participants can read the materials and consent remotely without the presence of a clinician or research staff. Second, routine clinical practice is already decentralized in many ways, so pragmatic trials embedded into routine practice can leverage the decentralized features in everyday practice. For instance, patients seeking subspecialty care at academic medical centers can follow-up with their primary care providers in community practice and fill medications at local drug stores. In addition, routine practices increasingly adopt remote home monitoring that directly transmits data back to the providers. Pragmatic trials can readily incorporate these routine practice features, whereas in explanatory trials, the intervention, and the outcome measures (eg, a biomarker) might not be available outside a few academic medical centers.

Among the first wave of decentralized trials is the batch enrollment for an AI-guided intervention to lower neurologic events in patients with undiagnosed atrial fibrillation (BEAGLE). The BEAGLE trial is a decentralized, pragmatic trial utilizing EHR, digital phenotyping algorithms, and remote monitoring technologies to evaluate an AI algorithm designed to detect previously undiagnosed atrial fibrillation in patients at risk of stroke.[43]

Another Mayo Clinic team is also building a digital biobank, collecting data from digital wearable devices, such as the Apple watch. In a recent study published in Nature Medicine, the authors enrolled 2454 unique participants from 46 US states and 11 countries, who downloaded a Mayo Clinic iPhone application that sends Apple watch electrocardiograms (ECGs) to the study's data platform. These participants sent 125,610 ECGs to the data platform. On average, patients used the study application 2.1 times each month, and increasing use positively correlated with patient age.[45] Other decentralized trials utilizing this infrastructure are also under way.[46]

EXAMPLES BY RCT DESIGN

Conceptually, clinical trials can be divided into randomized trials and nonrandomized trials; in the meantime, clinical trials can also be categorized as explanatory trials or pragmatic trials. In other words, explanatory trials can be either randomized or nonrandomized; and similarly, pragmatic trials can be either randomized or nonrandomized as well. Here, when we present the examples below, we focus on the most salient pragmatic trials—the pragmatic RCT.

Depending on the unit of randomization, RCTs can be divided into randomization at the individual-level and randomization at the group level, and the latter is also called cluster randomized trial (CRT). Stepped wedged trial is a subtype of CRT. Individual-level randomization should always be used unless a strong rationale exists for group level randomization because of the logistic and analytical challenges with CRT, eg, increased risk of bias and reduced statistical power.[47]

The most common reasons for using a cluster randomization design are the intervention is a group level intervention or there exists a high risk of contamination, ie, the treatment effect can spill over to participants in the same group. In these cases, due to the nature of the intervention, the purpose of the trial is often to evaluate the effectiveness of the intervention, eg, a new tool to promote cancer screening or a quality improvement program, so group level randomization is more prevalent in the pragmatic trials than in the explanatory trials, but there exist explanatory CRTs with a focus on the efficacy.[48] In other words, there are pragmatic RCTs using cluster randomization, but there are many other pragmatic RCTs using individual-level randomization; similarly,

most explanatory RCTs use individual-level randomization, but there also exist explanatory RCTs that use cluster randomization.

Individual-Level Randomization

Individual-level randomization trials are RCTs where the unit of randomization is an individual participant. Individual-level randomization facilitates a straightforward comparison between treatment and control groups, allowing for precise assessments of the intervention's effect.

One example of an individual-level randomized trial evaluating the effectiveness of AI in clinical practice is the Screening for Peripartum Cardiomyopathies using Artificial Intelligence in Nigeria (SPEC-AI Nigeria). Nigeria has the highest reported incidence of peripartum cardiomyopathy worldwide. This trial is currently evaluating an AI-enabled ECG's ability to detect cardiomyopathy in individuals seeking routine obstetrics care at 6 sites in Nigeria. They aim to enroll 1000 pregnant and postpartum women randomized to routine care with the AI-ECG cardiomyopathy screening, with the primary outcome being a new diagnosis of cardiomyopathy during pregnancy or within 12 months postpartum.[49]

The effect of LLM in assisting discharge summary notes writing for hospitalized patients is another ongoing individual-level randomized trial evaluating whether discharge summary writing assisted by a LLM, checker for unvalidated response errors, improves care delivery without adversely affecting patient outcomes. Participating clinicians are randomized to the control group, who will continue with standard practice for discharge summary writing, or the intervention group, who will have access to checker for unvalidated response errors to assist with discharge summary writing. This is a pilot study, aiming to assess the feasibility of later carrying out a full-scale RCT.[50]

Cluster Randomized Trial

The CRTs are randomized trials where the unit of randomization is a cluster, or group of individuals, such as participants managed by the same clinician, on the same hospital floor, or within the same hospital. This approach is useful when evaluating interventions typically delivered at the group level or when the intervention's effects are likely to spill over to other participants, ie, contamination.[51] Limitations of CRTs include the possible need for larger sample sizes and large numbers of clusters,[52] which possibly increases the cost and complexity of the trial.

One salient example of a CRT assessing AI's role in clinical practice is the EAGLE trial, which assessed an ECG-based AI-powered clinical decision support tool's ability to enable early diagnosis of low ejection fraction in routine primary care. A primary care team was considered a cluster, and 120 primary care teams from 45 clinics or hospitals were 1:1 randomized to either the intervention arm, with access to AI results, or usual care. The trial encompassed 22,641 adults without prior heart failure diagnoses and found a significant increase in detection of low ejection fraction in the intervention.[42,53–55]

The AI-enhanced ECG to detect cardiac amyloidosis: protocol for a pragmatic cluster randomized clinical trial (PREDICT-AMY) is another example of an ongoing CRT. This trial evaluates an AI-based tool's ability to detect cardiac amyloidosis from 12-lead ECGs in everyday practice. The trial randomizes providers to an interventional arm or standard-of-care arm. Providers in the interventional arms are notified about patients whose AI-determined scores yield above predetermined cut-offs for cardiac amyloidosis risk. Providers are then provided with links to cardiac amyloid education and an amyloid order set. If additional results identified by the electronic retrieval programs permit additional risk models to be calculated, an enhanced risk score is also calculated, which is communicated to the provider.[56]

A study to detect advanced liver disease by AI-enabled ECG (ADVANCE) is another ongoing cluster randomized study evaluating the effectiveness of the AI-cirrhosis-ECG 2.0 (ACE2.0) model in its ability to detect advanced liver fibrosis in its early stages using AI analysis of ECG data. Care teams within 45 Mayo Clinic practices are randomized in a 1:1 ratio to usual care or the intervention arm, where providers are alerted regarding their patients' likelihood of advanced liver disease. Providers with patients identified as at risk by the ACE2.0 model are then given recommendations to order a FibroTest-ActiTest.[57,58]

Mayo Clin Proc Digital Health ■ December 2024;2(4):499-510 ■ https://doi.org/10.1016/j.mcpdig.2024.06.010
www.mcpdigitalhealth.org

Stepped Wedge Trials

Stepped wedge trial is a subtype of CRTs and an alternative to parallel CRTs. In a stepped wedge trial, there is an initial period in which no clusters are exposed to the intervention; subsequently, at regular intervals (so-called steps), one cluster or a group of clusters is randomized to cross from the control to the intervention arm. This process continues until all clusters have crossed over to be exposed to the intervention, and there is a period when all clusters are exposed at the end of the study.[59] In general, stepped wedge trials are most appropriate for interventions when introducing the intervention simultaneously across participants is not feasible, due to financial, logistical, or other practical constraints.[60] Stepped wedge trials are, by nature, pragmatic trials involving randomization. To the best of our knowledge, we have not seen any nonrandomized or explanatory trials that are stepped wedge trials. Limitations on stepped wedge trials include their design and analysis complexity, their longer duration when compared with traditional RCTs due to their phased rollout, and the potential for temporal trends to confound the effects of the intervention.

One example of a stepped wedge trial evaluating AI's effectiveness is an ongoing study by Heinzen et al,[61] which aims to evaluate the utility of a machine learning algorithm in identifying patients with a high likelihood of palliative care need and thereby reducing time to palliative care consultation in outpatient primary care. Forty-two care team units in 9 clusters are randomized to 7 wedges, with treatment wedges integrating the algorithm into clinical practice. Weekly, 40 patients with the highest predicted palliative care scores as defined by the algorithm are identified and presented to palliative medicine specialists to determine the need for palliative care referral. New care teams enter the treatment arm every 42 days, with the first wedge including all units on the control arm and the last wedge including all units on the intervention arm.[61]

In the inpatient setting, another stepped wedge trial assessed an AI model's ability to predict the need for palliative care consultation. Focusing on hospitalized patients who could benefit most from timely palliative care consultations, the study evaluated the AI model's predictive accuracy and its impact on clinical outcomes. Over the trial period, patient clusters randomized to the intervention arm were 40% more likely to receive timely palliative care, demonstrating the model's capability to identify patient needs.[62] Another similar pragmatic, cluster randomized, stepped wedge trial found that the use of an AI-powered decision support tool for predicting patient needs for inpatient palliative care services increased the palliative consultation rate among hospitalized patients.[63]

CONCLUSION

A pragmatic trial is a study design used for evaluating the effectiveness of an intervention in contrast to the efficacy of an intervention in an explanatory trial. Pragmatic trials are well-suited for evaluating the health impact of AI technologies in large populations. They offer the advantage of informing real-world decisions with a focus on patient-centered outcomes and can be generalized to diverse routine practice settings. Many AI technologies are inherently embedded in routine clinical practice, and their efficacy has often been found through technical performance, usability, and acceptability evaluations. Consequently, the primary intent during the final stage of AI evaluation is to inform real-world decisions rather than understanding the mechanism of action. With the rise of LHS and the ability to leverage digital health technologies to facilitate and improve the efficiency of trial conduct, some of the challenges of conducting pragmatic trials have been overcome. As the expansion of AI's role in health care, pragmatic trials will remain pivotal in producing the evidence needed to guide decision-making for patients, clinicians, and health systems.

POTENTIAL COMPETING INTERESTS

Dr. Noseworthy reports in the past 36 months, that he receives research funding from National Institutes of Health (NIH, including the National Heart, Lung, and Blood Institute [NHLBI] and the National Institute on Aging [NIA]), Agency for Healthcare Research and Quality (AHRQ), Food and Drug Administration (FDA), and the American Heart Association (AHA); study investigator in an ablation trial sponsored by Medtronic; involved in potential equity/royalty

relationship with AliveCor along with Mayo Clinic; served on an expert advisory panel for Optum; have filed patents related to the application of AI to the ECG for diagnosis and risk stratification along with Mayo Clinic. Dr. Yao reports in the past 36 months, that he received research support through Mayo Clinic from NIA (R01AG062436) and the Food and Drug Administration (FDA, U01FD005938). The other author reports no competing interests.

Abbreviations and Acronyms: **AI,** artificial intelligence; **CRT,** cluster randomized trial; **ECG,** electrocardiogram; **EHR,** electronic health records; **LHS,** learning health system; **LLM,** large language models; **RCT,** randomized controlled trial

Grant Support: This study was funded by Mayo Clinic Robert D and Patricia E Kern Center for the Science of Health Care Delivery. Study contents are the sole responsibility of the authors and do not necessarily represent the official views of Mayo Clinic. The study sponsor had no role in the design or conduct of the study; collection, management, analysis, or interpretation of the data; preparation, review, or approval of the manuscript; or the decision to submit the manuscript for publication.

Correspondence: Address to Xiaoxi Yao, PhD, Robert D and Patricia E Kern Center for the Science of Health Care Delivery, Mayo Clinic, 200 First Street SW, Rochester, MN 55905 (yao.xiaoxi@mayo.edu; Twitter: @xiaoxiyao).

ORCID

Xiaoxi Yao: https://orcid.org/0000-0001-9906-7106

REFERENCES

1. Attia ZI, Kapa S, Lopez-Jimenez F, et al. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. *Nat Med*. 2019;25(1):70-74. https://doi.org/10.1038/s41591-018-0240-2.
2. Attia ZI, Noseworthy PA, Lopez-Jimenez F, et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. *Lancet*. 2019;394(10201):861-867. https://doi.org/10.1016/S0140-6736(19)31721-0.
3. Yao X, Paulson M, Maniaci MJ, et al. Effect of hospital-at-home vs. traditional brick-and-mortar hospital care in acutely ill adults: study protocol for a pragmatic randomized controlled trial. *Trials*. 2022;23(1):503. https://doi.org/10.1186/s13063-022-06430-6.
4. Harmon DM, Noseworthy PA, Yao X. The digitization and decentralization of clinical trials. *Mayo Clin Proc*. 2023;98(10):1568-1578. https://doi.org/10.1016/j.mayocp.2022.10.001.
5. Attia IZ, Tseng AS, Benavente ED, et al. External validation of a deep learning electrocardiogram algorithm to detect ventricular dysfunction. *Int J Cardiol*. 2021;329:130-135. https://doi.org/10.1016/j.ijcard.2020.12.065.
6. Mondo CK, Attia ZI, Benavente ED, et al. External validation of an electrocardiography artificial intelligence-generated algorithm to detect left ventricular systolic function in a general cardiac clinic in Uganda. *Eur Heart J*. 2020;41(suppl 2). https://doi.org/10.1093/ehjci/ehaa946.1013.
7. Noseworthy PA, Attia ZI, Brewer LC, et al. Assessing and mitigating bias in medical artificial intelligence: the effects of race and ethnicity on a deep learning model for ECG analysis. *Circ Arrhythm Electrophysiol*. 2020;13(3):e007988. https://doi.org/10.1161/CIRCEP.119.007988.
8. Attia ZI, Kapa S, Yao X, et al. Prospective validation of a deep learning electrocardiogram algorithm for the detection of left ventricular systolic dysfunction. *J Cardiovasc Electrophysiol*. 2019;30(5):668-674. https://doi.org/10.1111/jce.13889.
9. Gottesman O, Johansson F, Komorowski M, et al. Guidelines for reinforcement learning in healthcare. *Nat Med*. 2019;25(1):16-18. https://doi.org/10.1038/s41591-018-0310-5.
10. Richardson JP, Curtis S, Smith C, et al. A framework for examining patient attitudes regarding applications of artificial intelligence in healthcare. *Digit Health*. 2022;8:20552076221089084. https://doi.org/10.1177/20552076221089084.
11. *NIH's definition of a clinical trial*. United States Government; 2024. https://grants.nih.gov/policy/clinical-trials/definition.htm. Accessed August 10, 2024.
12. Noseworthy PA, Attia ZI, Behnken EM, et al. Artificial intelligence-guided screening for atrial fibrillation using electrocardiogram during sinus rhythm: a prospective non-randomised interventional trial. *Lancet*. 2022;400(10359):1206-1212. https://doi.org/10.1016/S0140-6736(22)01637-3.
13. Izem R, Buenconsejo J, Davi R, Luan JJ, Tracy L, Gamalo M. Real-world data as external controls: practical experience from notable marketing applications of new therapies. *Ther Innov Regul Sci*. 2022;56(5):704-716. https://doi.org/10.1007/s43441-022-00413-0.
14. Kim T-E, Park S-I, Shin K-H. Incorporation of real-world data to a clinical trial: use of external controls. *Transl Clin Pharmacol*. 2022;30(3):121-128. https://doi.org/10.12793/tcp.2022.30.e14.
15. Haukoos JS, Lewis RJ. The propensity score. *JAMA*. 2015;314(15):1637-1638. https://doi.org/10.1001/jama.2015.13480.
16. Maciejewski ML, Basu A. Regression discontinuity design. *JAMA*. 2020;324(4):381-382. https://doi.org/10.1001/jama.2020.3822.
17. Yao X, Noseworthy PA. Finding order in chaos: can instrumental variables help us understand observed treatment effects? *Circ Cardiovasc Qual Outcomes*. 2020;13(4):e006650. https://doi.org/10.1161/CIRCOUTCOMES.120.006650.
18. Ford I, Norrie J. Pragmatic trials. *N Engl J Med*. 2016;375(5):454-463. https://doi.org/10.1056/NEJMra1510059.
19. Zuidgeest MGP, Goetz I, Groenwold RHH, et al. Series: pragmatic trials and real world evidence: Paper 1. Introduction. *J Clin Epidemiol*. 2017;88:7-13. https://doi.org/10.1016/j.jclinepi.2016.12.023.
20. Schwartz D, Lellouch J. Explanatory and pragmatic attitudes in therapeutical trials. *J Chronic Dis*. 1967;20(8):637-648. https://doi.org/10.1016/0021-9681(67)90041-0.
21. Tricoci P, Allen JM, Kramer JM, Califf RM, Smith SC Jr. Scientific evidence underlying the ACC/AHA clinical practice guidelines. *JAMA*. 2009;301(8):831-841. https://doi.org/10.1001/jama.2009.205.
22. Fanaroff AC, Califf RM, Windecker S, Smith SC Jr, Lopes RD. Levels of evidence supporting American College of Cardiology/American Heart Association and European Society of Cardiology guidelines, 2008-2018. *JAMA*. 2019;321(11):1069-1080. https://doi.org/10.1001/jama.2019.1122.
23. DuBose-Briski V, Yao X, Dunlay SM, et al. Evolution of the American College of Cardiology and American Heart Association cardiology clinical practice guidelines: A 10-year assessment. *J Am Heart Assoc*. 2019;8(19):e012065. https://doi.org/10.1161/JAHA.119.012065.
24. Schoenborn NL, Massare J, Park R, Boyd CM, Choi Y, Pollack CE. Assessment of clinician decision-making on cancer screening cessation in older adults with limited life expectancy.

JAMA Netw Open. 2020;3(6):e206772. https://doi.org/10.1001/jamanetworkopen.2020.6772.

25. Cowley LE, Maguire S, Farewell DM, Quinn-Scoggins HD, Flynn MO, Kemp AM. Factors influencing child protection professionals' decision-making and multidisciplinary collaboration in suspected abusive head trauma cases: A qualitative study. *Child Abuse Negl*. 2018;82:178-191. https://doi.org/10.1016/j.chiabu.2018.06.009.
26. Grundy Q, Bero L, Malone R. Interactions between non-physician clinicians and industry: a systematic review. *PLOS Med*. 2013;10(11):e1001561. https://doi.org/10.1371/journal.pmed.1001561.
27. Lublóy Á. Factors affecting the uptake of new medicines: a systematic literature review. *BMC Health Serv Res*. 2014;14:469. https://doi.org/10.1186/1472-6963-14-469.
28. Hillman AL, Pauly MV, Kerstein JJ. How do financial incentives affect physicians' clinical decisions and the financial performance of health maintenance organizations? *N Engl J Med*. 1989;321(2):86-92. https://doi.org/10.1056/NEJM198907133210205.
29. Horwitz LI, Kuznetsova M, Jones SA. Creating a learning health system through rapid-cycle, randomized testing. *N Engl J Med*. 2019;381(12):1175-1179. https://doi.org/10.1056/NEJMsb1900856.
30. Hillier TA, Pedula KL, Ogasawara KK, et al. A pragmatic, randomized clinical trial of gestational diabetes screening. *N Engl J Med*. 2021;384(10):895-904. https://doi.org/10.1056/NEJMoa2026028.
31. DeVore AD, Granger BB, Fonarow GC, et al. Effect of a hospital and postdischarge quality improvement intervention on clinical outcomes and quality of care for patients with heart failure with reduced ejection fraction: the CONNECT-HF randomized clinical trial. *JAMA*. 2021;326(4):314-323. https://doi.org/10.1001/jama.2021.8844.
32. Gresham G, Meinert JL, Gresham AG, Meinert CL. Assessment of trends in the design, accrual, and completion of trials registered in ClinicalTrials.gov by sponsor type, 2000-2019. *JAMA Netw Open*. 2020;3(8):e2014682. https://doi.org/10.1001/jamanetworkopen.2020.14682.
33. Nicholls SG, Zwarenstein M, Hey SP, Giraudeau B, Campbell MK, Taljaard M. The importance of decision intent within descriptions of pragmatic trials. *J Clin Epidemiol*. 2020;125:30-37. https://doi.org/10.1016/j.jclinepi.2020.04.030.
34. Wolfenden L, Foy R, Presseau J, et al. Designing and undertaking randomised implementation trials: guide for researchers. *BMJ*. 2021;372:m3721. https://doi.org/10.1136/bmj.m3721.
35. Curran GM, Bauer M, Mittman B, Pyne JM, Stetler C. Effectiveness-implementation hybrid designs: combining elements of clinical effectiveness and implementation research to enhance public health impact. *Med Care*. 2012;50(3):217-226. https://doi.org/10.1097/MLR.0b013e3182408812.
36. Loudon K, Treweek S, Sullivan F, Donnan P, Thorpe KE, Zwarenstein M. The PRECIS-2 tool: designing trials that are fit for purpose. *BMJ*. 2015;350:h2147. https://doi.org/10.1136/bmj.h2147.
37. Dal-Ré R, Janiaud P, Ioannidis JPA. Real-world evidence: how pragmatic are randomized controlled trials labeled as pragmatic? *BMC Med*. 2018;16(1):49. https://doi.org/10.1186/s12916-018-1038-2.
38. Jones WS, Mulder H, Wruck LM, et al. Comparative effectiveness of Aspirin dosing in cardiovascular disease. *N Engl J Med*. 2021;384(21):1981-1990. https://doi.org/10.1056/NEJMoa2102137.
39. Agency for Healthcare Research and Quality. About learning health systems. https://www.ahrq.gov/learning-health-systems/about.html. Accessed January 1, 2024.
40. Centers for Medicare & Medicaid Services (CMS). HHS. Medicare program; merit-based incentive payment system (MIPS) and alternative payment model (APM) incentive under the physician fee schedule, and criteria for physician-focused payment models. *Fed Regist*. 2016;81(214):77008-77831.
41. Yao X, McCoy RG, Friedman PA, et al. Clinical trial design data for electrocardiogram artificial intelligence-guided screening for low ejection fraction (EAGLE). *Data Brief*. 2020;28:104894. https://doi.org/10.1016/j.dib.2019.104894.
42. Yao X, McCoy RG, Friedman PA, et al. ECG AI-Guided Screening for Low Ejection Fraction (EAGLE): rationale and design of a pragmatic cluster randomized trial. *Am Heart J*. 2020;219:31-36. https://doi.org/10.1016/j.ahj.2019.10.007.
43. Yao X, Attia ZI, Behnken EM, et al. Batch enrollment for an artificial intelligence-guided intervention to lower neurologic events in patients with undiagnosed atrial fibrillation: rationale and design of a digital clinical trial. *Am Heart J*. 2021;239:73-79. https://doi.org/10.1016/j.ahj.2021.05.006.
44. Goodson N, Wicks P, Morgan J, Hashem L, Callinan S, Reites J. Opportunities and counterintuitive challenges for decentralized clinical trials to broaden participant inclusion. *npj Digit Med*. 2022;5(1):58. https://doi.org/10.1038/s41746-022-00603-y.
45. Attia ZI, Harmon DM, Dugan J, et al. Prospective evaluation of smartwatch-enabled detection of left ventricular dysfunction. *Nat Med*. 2022;28(12):2497-2503. https://doi.org/10.1038/s41591-022-02053-1.
46. Yao X, Attia ZI, Behnken EM, et al. Realtime Diagnosis from electrocardiogram artificial intelligence-guided screening for atrial fibrillation with long follow-up (REGAL): rationale and design of a pragmatic, decentralized, randomized controlled trial. *Am Heart J*. 2024;267:62-69. https://doi.org/10.1016/j.ahj.2023.10.005.
47. Taljaard M, Goldstein CE, Giraudeau B, et al. Cluster over individual randomization: are study design choices appropriately justified? Review of a random sample of trials. *Clin Trials*. 2020;17(3):253-263. https://doi.org/10.1177/1740774519896799.
48. Mitjà O, Corbacho-Monné M, Ubals M, et al. A cluster-randomized trial of hydroxychloroquine for prevention of Covid-19. *N Engl J Med*. 2021;384(5):417-427. https://doi.org/10.1056/NEJMoa2021801.
49. Adedinsewo DA, Morales-Lara AC, Dugan J, et al. Screening for peripartum cardiomyopathies using artificial intelligence in Nigeria (SPEC-AI Nigeria): clinical trial rationale and design. *Am Heart J*. 2023;261:64-74. https://doi.org/10.1016/j.ahj.2023.03.008.
50. Mayo Clinic. Effect of large language model in assisting discharge summary notes writing for hospitalized patients. https://ctv.veeva.com/study/effect-of-large-language-model-in-assisting-discharge-summary-notes-writing-for-hospitalized-patient. Accessed January 1, 2024.
51. Hemming K, Taljaard M, Weijer C, Forbes AB. Use of multiple period, cluster randomised, crossover trial designs for comparative effectiveness research. *BMJ*. 2020;371:m3800. https://doi.org/10.1136/bmj.m3800.
52. Hemming K, Eldridge S, Forbes G, Weijer C, Taljaard M. How to design efficient cluster randomised trials. *BMJ*. 2017;358:j3064. https://doi.org/10.1136/bmj.j3064.
53. Yao X, Rushlow DR, Inselman JW, et al. Artificial intelligence-enabled electrocardiograms for identification of patients with low ejection fraction: a pragmatic, randomized clinical trial. *Nat Med*. 2021;27(5):815-819. https://doi.org/10.1038/s41591-021-01335-4.
54. Barry B, Zhu X, Behnken E, et al. Provider perspectives on artificial intelligence—guided screening for low ejection fraction in primary care: qualitative study. *JMIR AI*. 2022;1(1):e41940. https://doi.org/10.2196/41940.
55. Rushlow DR, Croghan IT, Inselman JW, et al. Clinician adoption of an artificial intelligence algorithm to detect left ventricular systolic dysfunction in primary care. *Mayo Clin Proc*. 2022;97(11):2076-2085. https://doi.org/10.1016/j.mayocp.2022.04.008.
56. Grogan M, Lopez-Jimenez F, Cohen-Shelly M, et al. Artificial intelligence-enhanced electrocardiogram for the Early Detection of cardiac amyloidosis. *Mayo Clin Proc*. 2021;96(11):2768-2778. https://doi.org/10.1016/j.mayocp.2021.04.023.
57. Ahn JC, Attia ZI, Rattan P, et al. Development of the AI-cirrhosis-ECG score: an electrocardiogram-based deep learning model in cirrhosis. *Am J Gastroenterol*. 2022;117(3):424-432. https://doi.org/10.14309/ajg.0000000000001617.

58. Simonetto DA. A study to detect advanced liver disease via AI-enabled electrocardiogram. https://www.mayo.edu/research/clinical-trials/cls-20547081. Accessed January 1, 2024.
59. Hemming K, Haines TP, Chilton PJ, Girling AJ, Lilford RJ. The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. *BMJ*. 2015;350:h391. https://doi.org/10.1136/bmj.h391.
60. Federico CA, Heagerty PJ, Lantos J, et al. Ethical and epistemic issues in the design and conduct of pragmatic stepped-wedge cluster randomized clinical trials. *Contemp Clin Trials*. 2022;115:106703. https://doi.org/10.1016/j.cct.2022.106703.
61. Heinzen EP, Wilson PM, Storlie CB, et al. Impact of a machine learning algorithm on time to palliative care in a primary care population: protocol for a stepped-wedge pragmatic randomized trial. *BMC Palliat Care*. 2023;22(1):9. https://doi.org/10.1186/s12904-022-01113-0.
62. Morgan A, Karow J, Olson E, et al. Randomized trial of a novel artificial intelligence/machine learning model to predict the need for specialty palliative care. *J Pain Symptom Manag*. 2022;63(5):879-880. https://doi.org/10.1016/j.jpainsymman.2022.02.078.
63. Wilson PM, Ramar P, Philpot LM, et al. Effect of an artificial intelligence decision support tool on palliative care referral in hospitalized patients: A randomized clinical trial. *J Pain Symptom Manage*. 2023;66(1):24-32. https://doi.org/10.1016/j.jpainsymman.2023.02.317.