HHS Public Access

Author manuscript

*Lancet Digit Health*. Author manuscript; available in PMC 2024 May 03.

Published in final edited form as:

*Lancet Digit Health*. 2024 May ; 6(5): e367–e373. doi:10.1016/S2589-7500(24)00047-5.

# Randomised controlled trials evaluating artificial intelligence in clinical practice: a scoping review

**Ryan Han**,

**Julián N Acosta**,

**Zahra Shakeri**,

**John P A Ioannidis**,

**Eric J Topol***,

**Pranav Rajpurkar***

Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA (R Han MS, P Rajpurkar PhD); Department of Computer Science, Stanford University, Stanford, CA, USA (R Han); University of California Los Angeles–Caltech Medical Scientist Training Program, Los Angeles, CA, USA (R Han); Department of Neurology, Yale School of Medicine, New Haven, CT, USA (J N Acosta MD); Rad AI, San Francisco, CA, USA (J N Acosta); Institute of Health Policy, Management and Evaluation, Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada (Z Shakeri PhD); Stanford Prevention Research Center, Department of Medicine (Prof J P A Ioannidis MD DSc), and Meta-Research Innovation Center at Stanford (Prof J P A Ioannidis), Stanford University, Stanford, CA, USA; Scripps Research Translational Institute, Scripps Research, La Jolla, CA, USA (Prof E J Topol MD)

## Abstract

This scoping review of randomised controlled trials on artificial intelligence (AI) in clinical practice reveals an expanding interest in AI across clinical specialties and locations. The USA and China are leading in the number of trials, with a focus on deep learning systems for medical imaging, particularly in gastroenterology and radiology. A majority of trials (70 [81%] of 86) report positive primary endpoints, primarily related to diagnostic yield or performance; however, the predominance of single-centre trials, little demographic reporting, and varying reports of operational efficiency raise concerns about the generalisability and practicality of these results. Despite the promising outcomes, considering the likelihood of publication bias and the need for more comprehensive research including multicentre trials, diverse outcome measures, and improved reporting standards is crucial. Future AI trials should prioritise patient-relevant outcomes to fully understand AI's true effects and limitations in health care.



Correspondence to: Prof Eric J Topol, Scripps Research Translational Institute, Scripps Research, La Jolla, CA 92037, USA, etopol@scripps.edu.

Contributors

RH and PR conceptualised the scoping review. JPAI, PR, and EJT supervised the scoping review. RH, JPAI, and PR contributed to the design of the scoping review. RH, JNA, and PR did the screening of the search results and data extraction. RH drafted the manuscript. ZS contributed to the presentation of data. All authors had access to the data, interpreted the analyses, and critically revised and edited the manuscript.

*Contributed equally

See **Online** for appendix

## Introduction

The use of artificial intelligence (AI) in health care has seen remarkable growth in the past 5 years, with several publications reporting that medical AI models can perform as well as or better than clinicians across a number of tasks and specialties;[1–3] however, many of these models have only been tested retrospectively, using surrogate endpoints, and outside of real-world clinical settings. Of nearly 300 AI-enabled medical devices approved or cleared by the US Food and Drug Administration, only a few have undergone evaluation using prospective randomised controlled trials (RCTs).[4]

The scarcity of real-world evaluation of AI systems contributes to substantial uncertainty, including in terms of the possibility of meaningful risk to patients and clinicians. One example of this risk is a widely used sepsis model that was found to have "substantially worse" performance than was reported by its developer, leading to "a large burden of alert fatigue" due to incorrect or irrelevant alerts.[5] It might not be uncommon for AI to perform worse when deployed prospectively, and the difficulty of adopting AI systems in a clinical setting can further impede any potential benefits in terms of important outcomes.[6,7] Additionally, without real-world evaluation, AI models' bias could remain undetected, which could inadvertently contribute to disparities in health outcomes.[8–10]

To provide a clearer understanding of the AI landscape in health care, this scoping review aims to examine the state of RCTs for AI algorithms being used in clinical practice. Although several systematic reviews[11–14] have been conducted on this topic, our scoping review updates the evidence with many new trials published up to the end of 2023, as the number of trials published has more than doubled since 2021. Our scoping review also introduces new inclusion criteria. Specifically, we require that the AI intervention reflects current advancements in machine learning and is integrated into actual patient management done by clinical teams. This stringent focus on clinically significant AI applications ensures that our review is acutely relevant to informing medical practice. Furthermore, our review uniquely examines detailed analyses that highlight the diversity in algorithms, comparisons of various groups, differences in modalities, and the nature of trial endpoints. This distinction sets this scoping review apart from earlier systematic reviews that have primarily concentrated on evaluating overall evidence, methodological quality, or statistical rigour. Our analysis examines the potential of AI to improve care management, patient behaviour and symptoms, and clinical decision-making efficiency, and identifies areas that require more research. We aim to help stakeholders better comprehend the clinical relevance and readiness of AI and guide future research in this rapidly evolving domain.

## Methods

### Search strategy and selection criteria

We systematically searched PubMed, SCOPUS, CENTRAL, and the International Clinical Trials Registry Platform for relevant studies published between Jan 1, 2018, and Nov 14, 2023. This timeline was selected to coincide with the era when modern AI models began to play an important role in trials. We used free-text search terms such as "artificial

intelligence", "clinician", and "clinical trial". The detailed search strategy can be found in the appendix (pp 3–7). Additionally, we manually scrutinised the references of pertinent publications to find more articles.

Our inclusion criteria were specific to RCTs that met the following conditions: the intervention incorporated a substantial AI component, which we defined as a non-linear computational model (ie, machine learning components including, but not limited to, decision trees, neural networks, etc); the intervention was integrated into clinical practice, thereby influencing a patient's health management by a clinical team; and the results were published as a full-text article in a peer-reviewed English-language journal. We excluded studies that evaluated linear risk scores, such as logistic regression, secondary studies, abstracts, and interventions that were not integrated into clinical practice. This scoping review follows the PRISMA extension for scoping reviews guidelines (appendix pp 8–9), and the protocol for this scoping review was registered with PROSPERO (CRD42022326955).[15]

### Data analysis

To ensure the quality of our search results, we used Covidence Review software to screen publication titles and abstracts. Two independent investigators (RH and JNA) conducted the initial screening, followed by a full-text review of screened papers. Data extraction of eligible papers was done in Google Sheets by a single investigator and then verified by a second investigator (RH or JNA). Any discrepancies were resolved through discussion with a third reviewer (PR).

We extracted study-level information, including study location, participant characteristics, clinical task, primary endpoint, time efficiency endpoint, comparator, and result, as well as the type and origin of the AI used. Additionally, we classified studies by primary endpoint group (diagnostic yield or performance, clinical decision making, patient behaviour and symptoms, and care management), clinical area or speciality, and data modality used by the AI.

We did not attempt to contact study authors for additional or uncertain information. Due to the expected heterogeneity in tasks and endpoints, we did not conduct formal meta-analyses. Instead, we present simple descriptive statistics to provide an overview of the features of the eligible trials.

## Results

Our electronic search retrieved 6219 study records and 4299 trial registrations, resulting in 10 484 records after deduplication (figure 1). After title and abstract screening, 133 articles were retained for full-text review. Of these, 60 were excluded, leaving 73 studies after the primary screening. An additional 13 articles were identified through secondary reference screening, resulting in a total of 86 unique RCTs included in our scoping review. The references and characteristics for all the included studies are available in the appendix (p 2).

Of 86 RCTs, 37 (43%) were related to gastroenterology, 11 (13%) to radiology, five (6%) to surgery, and five (6%) to cardiology. Gastroenterology trials were notable for their uniformity, with all trials testing video-based deep learning algorithms in an assistive setup supporting clinicians, and all but one trial measuring a primary endpoint relating to diagnostic yield or performance (detection rate, miss rate, etc). 24 (65%) of the 37 gastroenterology trials were conducted by only four groups (eight trials from Wuhan University, six from Wision AI, six from Medtronic, and four from Fujifilm).

79 (92%) of 86 RCTs were conducted in a single country, with the USA conducting the most trials (27 [31%]), followed by China (26 [30%]). Trials conducted in the USA were distributed across various specialties, whereas 21 (81%) of the 26 trials conducted in China predominantly related to gastroenterology. Trials conducted in multiple countries primarily involved European nations (6 [86%] of 7). Figure 2 highlights the distribution of trials across countries and specialties.

Trials were predominantly conducted in a single centre (54 [63%] of 86) and included a median of 359 patients (IQR 150–1050) in their final analysis. Of the 86 trials, 83 (97%) reported mean or median participant age, with the median age being 57·3 years (range 0·0034–78; IQR 49·9–62·0). Similarly, sex was reported in 83 (97%) of 86 trials, with a median of 48·9% of participants being male (range 0–89·2; IQR 45·4–54·2). Race or ethnicity was reported in 22 trials, of which 18 (82%) were from the USA. Among these trials, the median percentage of White (non-Hispanic or Latino) participants was 70·5% (range 0–98·4; IQR 35·0–81·8). Only three trials in China and one in South Korea explicitly reported on a single ethnicity: Han Chinese and Asian, respectively.

Of the 63 trials published since the start of 2021, 12 (19%) cited the 2020 CONSORT-AI reporting guidelines for clinical trials assessing AI interventions.[16]

Approximately half (46 [54%] of 86) of the trials had primary endpoints relating to diagnostic yield or performance, such as detection rate or mean absolute error. Other primary endpoints were grouped according to care management (18 [21%]), patient behaviour and symptoms (15 [17%]), and clinical decision making (7 [8%]). Table 1 summarises the distribution of results and endpoint types.

18 RCTs have assessed the effect of AI interventions on care management quality metrics, providing an outcome-oriented view of the use of AI in clinical practice. For example, AI systems for insulin dosing and hypotension monitoring have been shown to improve the average time that patients spend within target ranges for glucose and blood pressure, respectively.[17–20] Similarly, trials assessing AI systems for radiation therapy and prostate brachytherapy have been evaluated by their ability to reduce rates of acute care and the volume of the prostate tumour.[21,22]

15 AI systems have also been evaluated in terms of their effect on patient behaviour and symptoms. For example, one trial reported that making AI-generated predictions for diabetic retinopathy risk immediately available to patients increased referral adherence compared with having patients wait for grading by clinicians.[23] Another trial reported that the adoption of a nociception monitoring system was able to decrease postoperative pain scores in

patients when compared with unassisted clinicians.[24] These trials highlight the potential for AI interventions to have a direct impact on patient experience.

Seven trials have also measured the ability of AI systems to influence clinical decision making. For example, the availability of AI mortality predictions for cancer patients was reported to increase the number of serious illness conversations had between oncologists and patients.[25] In contrast, the adoption of an AI system for identifying atrial fibrillation patients at high risk of stroke did not increase new anticoagulant prescriptions.[26] These studies explore the potential for AI predictions to inform clinicians' judgement collaboratively.

59 (69%) of 86 trials evaluated deep learning systems for medical imaging. Notably, the medical imaging systems under evaluation were predominantly video based (42 [71%] of 59) rather than image based (17 [29%] of 59). This effect was primarily driven by the large number of endoscopy trials (34 [81%] of 42). Outside of imaging, AI systems operated on structured data, such as from the Electronic Health Record (14 [52%] of 27), waveform data (ten [37%] of 27), and free text (three [11%] of 27). These systems use a mix of decision trees (six [22%] of 27), neural networks (two [7%] of 27), reinforcement learning (two [7%] of 27), case-based reasoning (two [7%] of 27), Bayesian classifiers (one [4%] of 27), and unspecified machine learning (14 [52%] of 27).

Most systems operating on medical imaging (50 [85%] of 59) were evaluated in an assistive setup with a clinician, whereas models based on structured data tended to be compared with routine care (12 [86%] of 14). Models were developed primarily in industry (47 [55%] of 86) followed by academia (35 [41%] of 86), with the remaining four models having mixed or unstated origins.

Table 2 summarises the distribution of results and group comparisons. Of the 86 trials, 81 attempted to show improvement and five used non-inferiority designs. 65 (80%) of the 81 trials that aimed to show improvement have reported significant improvement for their primary endpoint. 46 (71%) of these trials noted improvements for AI-assisted clinicians compared with unassisted clinicians, 16 (25%) noted improvements for AI systems compared with routine care, and three (5%) reported superior performance from standalone AI systems compared with clinicians.

Of the five trials with non-inferiority designs, three established non-inferiority between standalone AI systems and clinicians and two established non-inferiority between assisted and unassisted clinicians.[17,21,27–29] Hence, 70 (81%) of 86 trials reported a favourable result for their primary endpoint. A similar success rate was observed for the gastroenterology subset, with 28 (76%) of the 37 trials reporting significant improvement and one (3%) showing non-inferiority, for an overall 78·4% success rate.

16 RCTs with a negative result for their primary endpoint included ten trials that did not show an improvement of assisted clinicians compared with unassisted clinicians, four trials that did not show an improvement of AI systems compared with routine care, and one trial that did not show an improvement of standalone AI systems compared with clinicians. One trial also reported standalone AI systems to have significantly worse performance than

clinicians;[30] however, eight (50%) of these 16 trials reported a significant improvement for a secondary endpoint.[30–37]

52 (60%) of 86 trials also reported on operational time measurements with varying results. Approximately a third of the trials (18 [35%] of 52) reported a significant decrease concerning operational time ($p<0·05$); however, approximately a quarter (13 [25%] of 52) reported a significant increase in operational time ($p<0·05$). The remaining 21 (40%) of the 52 trials found no significant changes in operational time measurements.

Gastroenterology was the primary contributor to these results, with 32 trials involving operational time measurements. These results were varied with two trials (6%) noting a decrease in operational time, 12 trials (38%) reporting increased operational time, and the remaining 18 (56%) observing no significant effect. All five radiology trials and all three ophthalmology trials reported a significant reduction in operational time. In other specialties, two or fewer trials usually considered the aspect of operational time.

## Discussion

This scoping review of AI RCT publications reveals several noteworthy trends and implications for the development and implementation of AI systems in clinical practice. The distribution of trials across clinical specialities and locations highlights a concentration of AI RCTs in gastroenterology, radiology, surgery, and cardiology. Notably, there is less focus on primary care than specialty care, indicating a potential area for future research. The geographical distribution of trials reveals a dominance of single-country studies, with most trials from the USA, followed by China. A 2023 systematic review of AI and machine learning-enabled device trial registrations found a similar distribution of specialities and geographies, and also noted the predominance of national trials.[38] This scoping review also found different trends, however, with China leading in trial registrations and radiology being the most common speciality. This finding suggests a need for more international collaboration and multicentre trials to ensure the generalisability of AI systems across various populations and health-care systems.

The predominance of single-centre trials, with a median of 359 patients, suggests smaller, controlled environments are often chosen for AI health-care trials; however, little demographic reporting, particularly on race and ethnicity, raises concerns about the representativeness of these studies. The infrequency of citation of the CONSORT-AI reporting guidelines further underscores the need for greater transparency in trial methods. This transparency would enhance under standing of the trial's applicability to broader populations, as factors such as inclusion criteria, setting, and follow-up duration substantially influence the generalisability of results. Future trials should prioritise comprehensive reporting and participant diversity to bolster the external validity of their findings.

The use of deep learning systems for medical imaging, particularly in video-based systems, is a prevalent trend in AI applications evaluated in RCTs. This trend is evident in the large number of trials assessing video-based gastroenterology interventions, in contrast

with the dominance of image-based radiology algorithms in academic literature and regulatory clearances.[39–42] For image-based radiology algorithms, other designs besides RCTs might be most suitable for addressing diagnostic accuracy. Paired design studies allow for comparison of diagnostic performance in the same individuals, removing all confounding;[43,44] however, in gastroenterology applications, such as adenoma detection, paired designs are not feasible because the detected lesions are typically removed.[43,44] This trend appears to be driven by a few groups that account for most video-based gastroenterology trials, indicating that the field of clinical AI trials is still homogeneous in terms of investigators, trial designs, and outcome measures. Systems using structured data such as electronic health records and waveform data, however, have used a mix of decision trees, neural networks, reinforcement learning, and other machine learning techniques. This variety of models and data sources shows the adaptability of AI to address different health-care challenges. More research is needed to evaluate the effect of AI systems that incorporate clinical context (multiple modalities) or clinical priors (multiple timepoints) into their decision making, as these factors are crucial to many clinical tasks.[45,46]

The discrepancy between our success rate and success rates of historical reviews of RCTs for medical interventions and for AI systems in health care[11–14] can be attributed to our specific definitions of AI and clinical practice, which excluded studies that did not have clinical integrations and non-linear AI, and our updated search strategy that included several new and previously overlooked trials.[12–14,47] Our review extends the window of consideration to 2023, thus capturing more than a year of advancements and a large number of recent trials in this rapidly progressing field compared with previous reviews. Despite these favourable results, the generalisability of AI applications remains uncertain. Specifying whether the AI training data were sourced from the same or diverse institutions is crucial for trials. Furthermore, analyses comparing RCTs conducted in internal versus external testing settings could provide valuable insights into AI performance generalisability. Furthermore, interpretation of this success rate should be viewed in light of the infancy of the field and the likeliness of publication bias. A 2023 systematic review identified 627 AI-enabled technology trials registered on ClinicalTrials.gov, but only nine (1%) were readily identified as published.[48] Of the trials that are listed as ongoing or that have no posted results, the number with negative results is unknown, which leads to a delay in their completion or in the posting and publication of results. Therefore, publication bias poses a substantial threat to the valid interpretation of the overall effect and effectiveness of AI in clinical practice.

Most trials evaluated interventions on endpoints related to diagnostic yield or performance. Although such trials offer convincing evidence of the prospective technical performance of clinical AI systems, this evidence might not accurately reflect the overall effect of AI systems on patient care, as high sensitivity and specificity do not necessarily translate to improved patient outcomes. For example, a 2023 systematic review of 21 colonoscopy trials found that although AI assistance helped increase polyp detection, it did not yield significant increases in the detection of clinically critical advanced adenomas.[49] More generically, statistically favourable results in both diagnostic performance and other AI trials might not necessarily translate to clinically meaningful benefits. Some trials have assessed the effect of AI systems on care management quality metrics, patient behaviour

and symptoms, and clinical decision making. These diverse outcome measures reflect the various ways that AI systems can influence clinical practice, from improving care quality to enhancing patient experience and informing clinical judgement. To better assess the true value of AI algorithms in health care, it is crucial for real-world evidence to focus on clinically meaningful endpoints such as symptoms and need for treatment, as well as longer-term outcomes such as survival.[48,50] Furthermore, larger-scale evidence would allow a better appreciation of whether the absolute magnitude of the benefits of these outcomes is substantive or not.

In terms of operational efficiency, the results varied across specialities, with a large number of trials reporting increases or decreases in operational time. This finding highlights the potential of AI systems to either streamline or complicate clinical workflows, depending on the specific application and context. Given this complexity, successful adoption of AI tools will depend on factors such as operational efficiency, cost-effectiveness, and the level of training required, as much as performance. Therefore, future research should not only focus on clinical outcomes, but also on these multifaceted aspects of implementation, to provide a more comprehensive understanding of AI's effect on health-care delivery.

In conclusion, the existing landscape of RCTs on AI in clinical practice shows an expanding interest in applying AI across a range of clinical specialties and locations. Most trials report favourable outcomes, highlighting AI's potential to enhance care management, patient behaviour and symptoms, and clinical decision making, but this early success should be tempered by the likelihood of publication bias. The true success of AI applications ultimately depends on their generalisability to their target patient populations and settings, a subject upon which efforts like the STANDING Together initiative offer valuable guidance.[51] To understand AI's true effects and limitations more comprehensively in health care, more research is essential, including a focus on multicentre trials and the incorporation of diverse endpoint measures, especially patient-relevant outcomes.

This scoping review has two important limitations. First, the search for relevant studies was conducted in English only. This language restriction might have excluded relevant trials published in other languages, potentially limiting the comprehensiveness and generalisability of our findings. Second, despite extending the window of consideration to 2023, our review does not address updated trends in trial risk of bias. Future systematic reviews should address trends in trial risk of bias (eg, using Cochrane risk of bias and other related tools) and provide a deeper analysis of reporting transparency (CONSORT-AI), given the constantly rising influx of RCTs.[15,50]

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

RH pursued this work while supported by the Summer Institute for Biomedical Informatics at Harvard Medical School.

### Declaration of interests

EJT receives funding from the National Center for Advancing Translational Sciences/National Institutes of Health (grant number UL1TR002550). JNA is an employee of Rad AI, outside of the submitted work. RH receives funding from the National Institute of General Medical Sciences (grant number T32 GM008042), and was formerly employed at Quadrant Health, outside of the submitted work. All other authors declare no competing interests.

## References


1. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med 2022; 28: 31–38. [PubMed: 35058619]
2. Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health 2019; 1: e271–97. [PubMed: 33323251]
3. Rajpurkar P, Lungren MP. The current and future state of AI interpretation of medical images. N Engl J Med 2023; 388: 1981–90. [PubMed: 37224199]
4. Wu E, Wu K, Daneshjou R, Ouyang D, Ho DE, Zou J. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. Nat Med 2021; 27: 582–84. [PubMed: 33820998]
5. Wong A, Otles E, Donnelly JP, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Intern Med 2021; 181: 1065–70. [PubMed: 34152373]
6. Mallick A, Hsieh K, Arzani B, Joshi G. Matchmaker: data drift mitigation in machine learning for large-scale systems. 2022. https://proceedings.mlsys.org/paper_files/paper/2022/file/069a002768bcb31509d4901961f23b3c-Paper.pdf (accessed March 30, 2024).
7. Beede E, Baylor E, Hersch F, et al. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. April 23, 2020. https://dl.acm.org/doi/abs/10.1145/3313831.3376718 (accessed March 30, 2024).
8. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019; 366: 447–53. [PubMed: 31649194]
9. Ganapathi S, Palmer J, Alderman JE, et al. Tackling bias in AI health datasets through the STANDING Together initiative. Nat Med 2022; 28: 2232–33. [PubMed: 36163296]
10. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med 2021; 27: 2176–82. [PubMed: 34893776]
11. Ospina-Tascón GA, Büchele GL, Vincent JL. Multicenter, randomized, controlled trials evaluating mortality in intensive care: doomed to fail? Crit Care Med 2008; 36: 1311–22. [PubMed: 18379260]
12. Lam TYT, Cheung MFK, Munro YL, Lim KM, Shung D, Sung JJY. Randomized controlled trials of artificial intelligence in clinical practice: systematic review. J Med Internet Res 2022; 24: e37188. [PubMed: 35904087]
13. Plana D, Shung DL, Grimshaw AA, Saraf A, Sung JJY, Kann BH. Randomized clinical trials of machine learning interventions in health care: a systematic review. JAMA Netw Open 2022;5: e2233946. [PubMed: 36173632]
14. Shahzad R, Ayub B, Siddiqui MAR. Quality of reporting of randomised controlled trials of artificial intelligence in healthcare: a systematic review. BMJ Open 2022; 12: e061519.
15. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018; 169: 467–73. [PubMed: 30178033]
16. Liu X, Cruz Rivera S, Moher D, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med 2020; 26: 1364–74. [PubMed: 32908283]
17. Nimri R, Battelino T, Laffel LM, et al. Insulin dose optimization using an automated artificial intelligence-based decision support system in youths with type 1 diabetes. Nat Med 2020; 26: 1380–84. [PubMed: 32908282]

18. Wijnberge M, Geerts BF, Hol L, et al. Effect of a machine learning-derived early warning system for intraoperative hypotension vs standard care on depth and duration of intraoperative hypotension during elective noncardiac surgery: the HYPE randomized clinical trial. JAMA 2020; 323: 1052–60. [PubMed: 32065827]
19. Tsoumpa M, Kyttari A, Matiatou S, et al. The use of the hypotension prediction index integrated in an algorithm of goal directed hemodynamic treatment during moderate and high-risk surgery. J Clin Med 2021; 10: 5884. [PubMed: 34945177]
20. Biester T, Nir J, Remus K, et al. DREAM5: an open-label, randomized, cross-over study to evaluate the safety and efficacy of day and night closed-loop control by comparing the MD-Logic automated insulin delivery system to sensor augmented pump therapy in patients with type 1 diabetes at home. Diabetes Obes Metab 2019; 21: 822–28. [PubMed: 30478937]
21. Nicolae A, Semple M, Lu L, et al. Conventional vs machine learning-based treatment planning in prostate brachytherapy: results of a phase I randomized controlled trial. Brachytherapy 2020; 19: 470–76. [PubMed: 32317241]
22. Hong JC, Eclov NCW, Dalal NH, et al. System for high-intensity evaluation during radiation therapy (SHIELD-RT): a prospective randomized study of machine learning-directed clinical evaluations during radiation and chemoradiation. J Clin Oncol 2020; 38: 3652–61. [PubMed: 32886536]
23. Mathenge W, Whitestone N, Nkurikiye J, et al. Impact of artificial intelligence assessment of diabetic retinopathy on referral service uptake in a low-resource setting: the RAIDERS randomized trial. Ophthalmol Sci 2022; 2: 100168. [PubMed: 36531575]
24. Meijer F, Honing M, Roor T, et al. Reduced postoperative pain using nociception level-guided fentanyl dosing during sevoflurane anaesthesia: a randomised controlled trial. Br J Anaesth 2020;125: 1070–78. [PubMed: 32950246]
25. Manz CR, Parikh RB, Small DS, et al. Effect of integrating machine learning mortality estimates with behavioral nudges to clinicians on serious illness conversations among patients with cancer: a stepped wedge cluster randomized clinical trial. JAMA Oncol 2020;6: e204759. [PubMed: 33057696]
26. Wang SV, Rogers JR, Jin Y, et al. Stepped-wedge randomised trial to evaluate population health intervention designed to increase appropriate anticoagulation in patients with atrial fibrillation. BMJ Qual Saf 2019; 28: 835–42.
27. Piette JD, Newman S, Krein SL, et al. Patient-centered pain care using artificial intelligence and mobile health tools: a randomized comparative effectiveness trial. JAMA Intern Med 2022; 182: 975–83. [PubMed: 35939288]
28. Repici A, Spadaccini M, Antonelli G, et al. Artificial intelligence and colonoscopy experience: lessons from two randomised trials. Gut 2022; 71: 757–65. [PubMed: 34187845]
29. Al-Hilli Z, Noss R, Dickard J, et al. A randomized trial comparing the effectiveness of pre-test genetic counseling using an artificial intelligence automated chatbot and traditional in-person genetic counseling in women newly diagnosed with breast cancer. Ann Surg Oncol 2023; 30: 5990–96. [PubMed: 37567976]
30. Lin H, Li R, Liu Z, et al. Diagnostic efficacy and therapeutic decision-making capacity of an artificial intelligence platform for childhood cataracts in eye clinics: a multicentre randomized controlled trial. EClinicalMedicine 2019; 9: 52–59. [PubMed: 31143882]
31. Pavel AM, Rennie JM, de Vries LS, et al. A machine-learning algorithm for neonatal seizure recognition: a multicentre, randomised, controlled trial. Lancet Child Adolesc Health 2020; 4: 740–49. [PubMed: 32861271]
32. Lui TKL, Hang DV, Tsao SKK, et al. Computer-assisted detection versus conventional colonoscopy for proximal colonic lesions: a multicenter, randomized, tandem-colonoscopy study. Gastrointest Endosc 2023; 97: 325–34. [PubMed: 36208795]
33. Xu L, He X, Zhou J, et al. Artificial intelligence-assisted colonoscopy: a prospective, multicenter, randomized controlled trial of polyp detection. Cancer Med 2021; 10: 7184–93. [PubMed: 34477306]

34. Seol HY, Shrestha P, Muth JF, et al. Artificial intelligence-assisted clinical decision support for childhood asthma management:a randomized clinical trial. PLoS One 2021; 16: e0255261. [PubMed: 34339438]
35. Mangas-Sanjuan C, de-Castro L, Cubiella J, et al. Role of artificial intelligence in colonoscopy detection of advanced neoplasias:a randomized trial. Ann Intern Med 2023; 176: 1145–52. [PubMed: 37639723]
36. Wei MT, Shankar U, Parvin R, et al. Evaluation of computer-aided detection during colonoscopy in the community (AI-SEE): a multicenter randomized clinical trial. Am J Gastroenterol 2023; 118: 1841–47. [PubMed: 36892545]
37. Yamaguchi D, Shimoda R, Miyahara K, et al. Impact of an artificial intelligence-aided endoscopic diagnosis system on improving endoscopy quality for trainees in colonoscopy: prospective, randomized, multicenter study. Dig Endosc 2024; 36: 40–48. [PubMed: 37079002]
38. Miquel S-B, Luca L, Vokinger KN. Development pipeline and geographic representation of trials for artificial intelligence/machine learning–enabled medical devices (2010 to 2023). NEJM AI 2023; 1: AIp2300038.
39. Mayo RC, Leung J. Artificial intelligence and deep learning - radiology's next frontier? Clin Imaging 2018; 49: 87–88. [PubMed: 29161580]
40. Pakdemirli E, Wegner U. Artificial intelligence in various medical fields with emphasis on radiology: statistical evaluation of the literature. Cureus 2020; 12: e10961. [PubMed: 33083162]
41. Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ Digit Med 2020; 3: 118. [PubMed: 32984550]
42. Stewart JE, Rybicki FJ, Dwivedi G. Medical specialties involved in artificial intelligence research: is there a leader? Tasman Medical Journal 2020; 2: 20–27.
43. Park SH, Han K, Jang HY, et al. Methods for clinical evaluation of artificial intelligence algorithms for medical diagnosis. Radiology 2023; 306: 20–31. [PubMed: 36346314]
44. Park SH, Choi JI, Fournier L, Vasey B. Randomized clinical trials of artificial intelligence in medicine: why, when, and how? Korean J Radiol 2022; 23: 1119–25. [PubMed: 36447410]
45. Acosta JN, Falcone GJ, Rajpurkar P. The need for medical artificial intelligence that incorporates prior images. Radiology 2022; 304: 283–88. [PubMed: 35438563]
46. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. Nat Med 2022; 28: 1773–84. [PubMed: 36109635]
47. Hennessy EA, Johnson BT. Examining overlap of included studies in meta-reviews: guidance for using the corrected covered area index. Res Synth Methods 2020; 11: 134–45. [PubMed: 31823513]
48. Pearce FJ, Cruz Rivera S, Liu X, Manna E, Denniston AK, Calvert MJ. The role of patient-reported outcome measures in trials of artificial intelligence health technologies: a systematic evaluation of ClinicalTrials.gov records (1997–2022). Lancet Digit Health 2023; 5: e160–67. [PubMed: 36828608]
49. Hassan C, Spadaccini M, Mori Y, et al. Real-time computer-aided detection of colorectal neoplasia during colonoscopy : a systematic review and meta-analysis. Ann Intern Med 2023; 176: 1209–20. [PubMed: 37639719]
50. Cruz Rivera S, Liu X, Hughes SE, et al. Embedding patient-reported outcomes at the heart of artificial intelligence health-care technologies. Lancet Digit Health 2023; 5: e168–73. [PubMed: 36828609]
51. Poblete SA. Standing up together. Clin J Oncol Nurs 2018; 22: 371. [PubMed: 30035780]

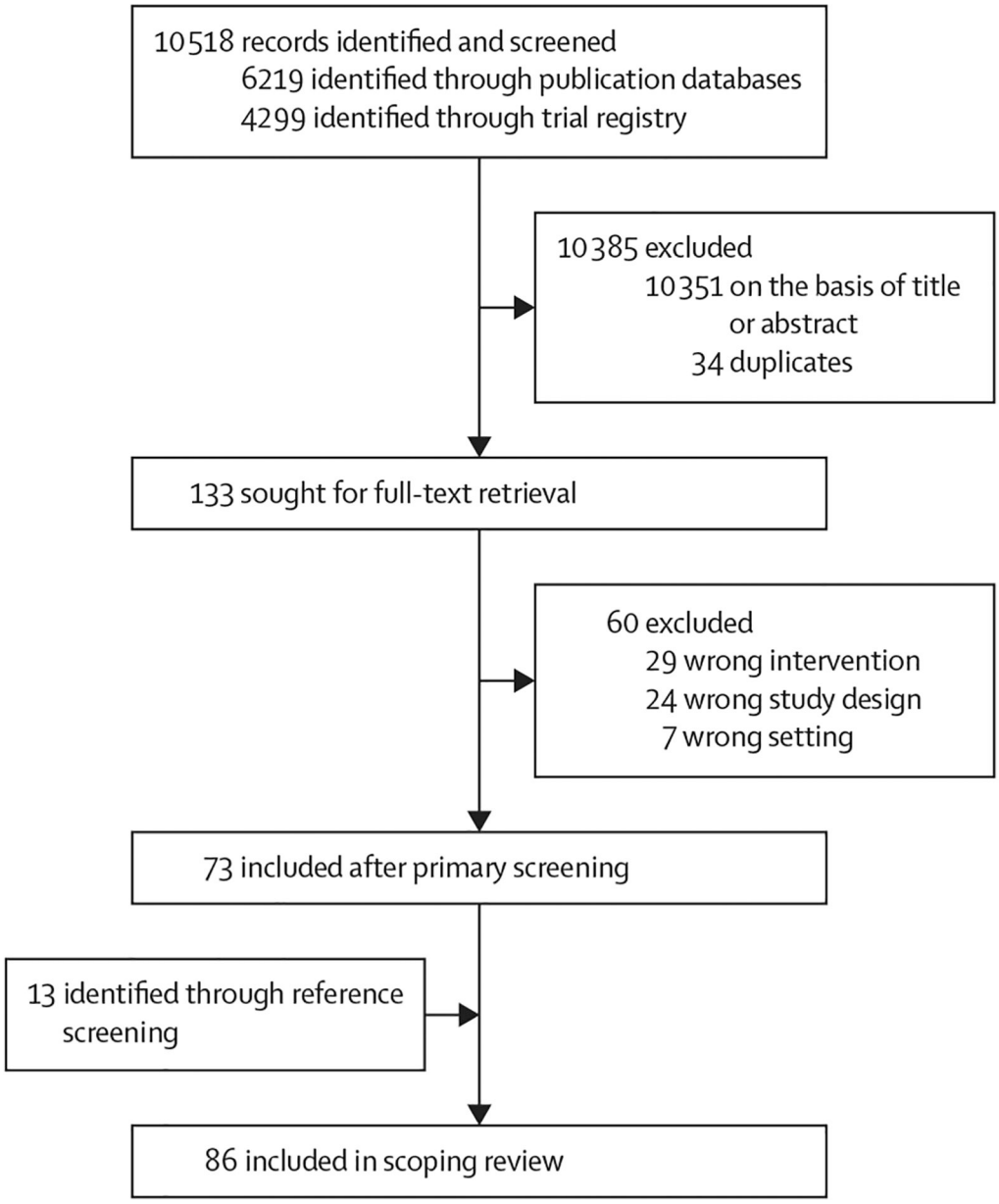


**Figure 1:**
Study selection

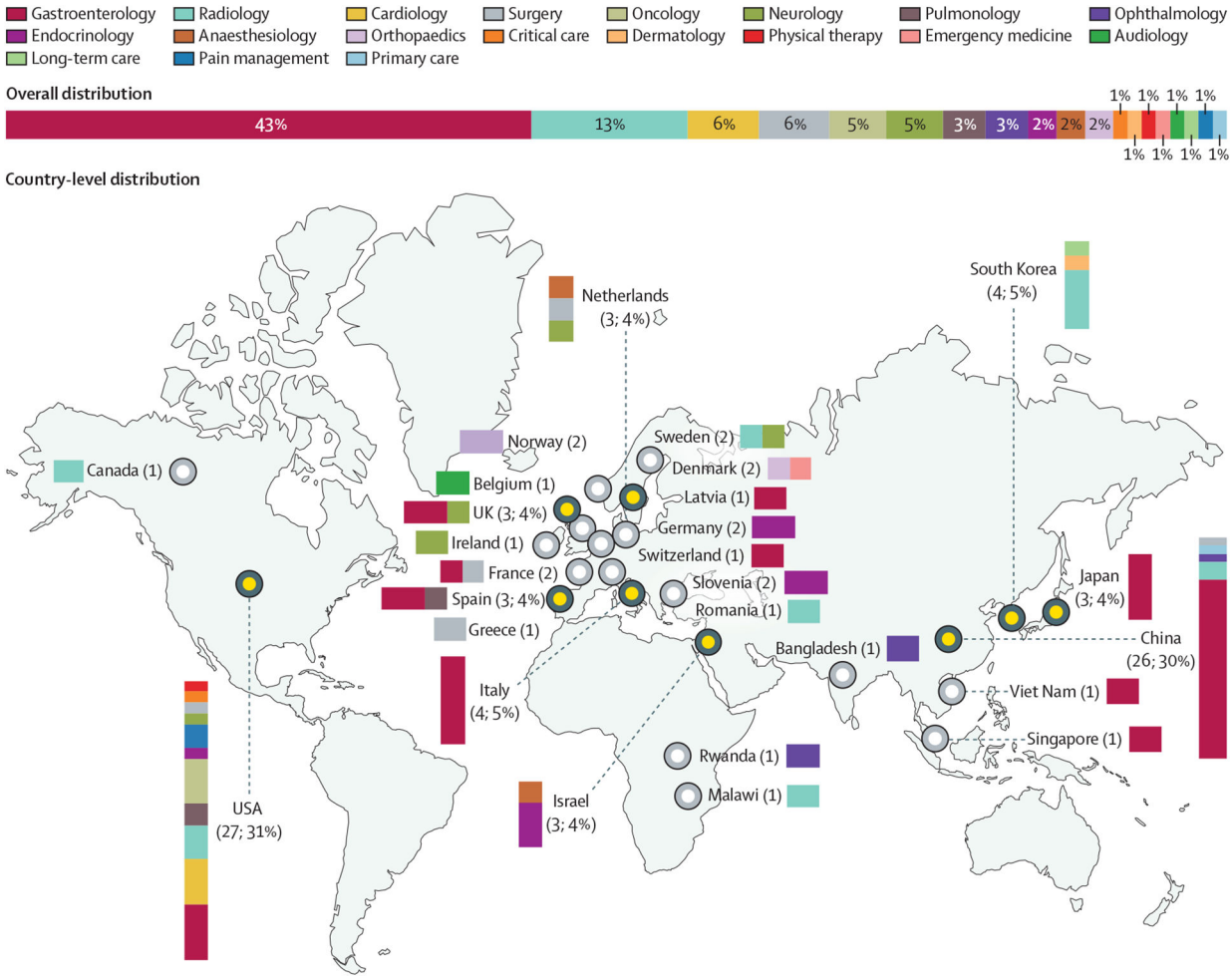


**Figure 2: Randomised controlled trials of artificial intelligence in clinical practice across countries and specialties**

Norway, France, Sweden, Denmark, Germany, and Slovenia each comprise 2% of the distribution. Canada, Belgium, Ireland, Greece, Latvia, Switzerland, Romania, Bangladesh, Rwanda, Malawi, Viet Nam, and Singapore each comprise 1% of the distribution.

**Table 1:**

Primary endpoints and types for randomised controlled trials of artificial intelligence in clinical practice

| | Statistically significant improvement | No statistically significant effect | Showed non-inferiority | Statistically significant deterioration | Total |
|---|---|---|---|---|---|
| Care management | 15 | 1 | 2 | .. | 18 |
| Clinical decision making | 6 | 1 | .. | .. | 7 |
| Diagnostic yield or performance | 34 | 10 | 1 | 1 | 46 |
| Patient behaviour and symptoms | 10 | 3 | 2 | .. | 15 |
| Total | 65 | 15 | 5 | 1 | 86 |

Data are n.

**Table 2:**

Primary endpoint results and group comparisons for randomised controlled trials of AI in clinical practice

| | Statistically significant improvement | No statistically significant effect | Showed non-inferiority | Statistically significant deterioration | Total |
|---|---|---|---|---|---|
| AI *vs* clinician | 3 | 1 | 3 | 1 | 8 |
| AI *vs* routine care | 16 | 4 | ·· | ·· | 20 |
| AI-assisted clinician *vs* unassisted clinician | 46 | 10 | 2 | ·· | 58 |
| Total | 65 | 15 | 5 | 1 | 86 |

Data are n. AI=artificial intelligence.