Randomized Clinical Trials of Machine Learning Interventions in Health Care
A Systematic Review

Deborah Plana, BS; Dennis L. Shung, MD, PhD; Alyssa A. Grimshaw, MSLIS; Anurag Saraf, MD; Joseph J. Y. Sung, MBBS, PhD; Benjamin H. Kann, MD

Abstract

IMPORTANCE Despite the potential of machine learning to improve multiple aspects of patient care, barriers to clinical adoption remain. Randomized clinical trials (RCTs) are often a prerequisite to large-scale clinical adoption of an intervention, and important questions remain regarding how machine learning interventions are being incorporated into clinical trials in health care.

OBJECTIVE To systematically examine the design, reporting standards, risk of bias, and inclusivity of RCTs for medical machine learning interventions.

EVIDENCE REVIEW In this systematic review, the Cochrane Library, Google Scholar, Ovid Embase, Ovid MEDLINE, PubMed, Scopus, and Web of Science Core Collection online databases were searched and citation chaining was done to find relevant articles published from the inception of each database to October 15, 2021. Search terms for machine learning, clinical decision-making, and RCTs were used. Exclusion criteria included implementation of a non-RCT design, absence of original data, and evaluation of nonclinical interventions. Data were extracted from published articles. Trial characteristics, including primary intervention, demographics, adherence to the CONSORT-AI reporting guideline, and Cochrane risk of bias were analyzed.

FINDINGS Literature search yielded 19,737 articles, of which 41 RCTs involved a median of 294 participants (range, 17-2488 participants). A total of 16 RCTs (39%) were published in 2021, 21 (51%) were conducted at single sites, and 15 (37%) involved endoscopy. No trials adhered to all CONSORT-AI standards. Common reasons for nonadherence were not assessing poor-quality or unavailable input data (38 trials [93%]), not analyzing performance errors (38 [93%]), and not including a statement regarding code or algorithm availability (37 [90%]). Overall risk of bias was high in 7 trials (17%). Of 11 trials (27%) that reported race and ethnicity data, the median proportion of participants from underrepresented minority groups was 21% (range, 0%-51%).

CONCLUSIONS AND RELEVANCE This systematic review found that despite the large number of medical machine learning–based algorithms in development, few RCTs for these technologies have been conducted. Among published RCTs, there was high variability in adherence to reporting standards and risk of bias and a lack of participants from underrepresented minority groups. These findings merit attention and should be considered in future RCT design and reporting.

Key Points

Question How are machine learning interventions being incorporated into randomized clinical trials (RCTs) in health care?

Findings In this systematic review of 41 RCTs of machine learning interventions, despite the large number of medical machine learning–based algorithms in development, few RCTs for these technologies have been conducted. Of published RCTs, most did not fully adhere to accepted reporting guidelines and had limited inclusion of participants from underrepresented minority groups.

Meaning These findings highlight areas of concern regarding the quality of medical machine learning RCTs and suggest opportunities to improve reporting transparency and inclusivity, which should be considered in the design and publication of future trials.

Introduction

Machine learning has the potential to improve the diagnosis and prognosis of disease to enhance clinical care. Given the increasing amount of digital data generated from routine medical care, available computational processing power, and research advances, such as deep learning, there has been substantial interest in applying machine learning techniques to improve patient care across...
Models have been investigated for tasks such as improved cancer diagnosis, emergency department triage, and intensive care unit decision support. However, the recent failures to successfully implement machine learning systems in clinical settings have highlighted the limitations of this technology, generating disillusionment and distrust in their potential to impact medicine. These machine learning system failures are often attributable to a lack of generalizability, an inability to adapt a system trained with data from 1 context to perform well in a new one, or an inability to demonstrate a clinically meaningful benefit. Mitigation strategies have been proposed to ensure their applicability, such as the use of larger and more diverse data sets and direct collaborations with clinical experts in model development. We investigated a different and complementary area of study of machine learning model-testing procedures, randomized clinical trials (RCTs), which may affect their ultimate use in heterogeneous clinical settings.

Randomized clinical trials are considered the gold standard for assessing an intervention's impact in clinical care, and the current landscape of RCTs for machine learning in health care continues to evolve. Randomized clinical trials, particularly those with transparent and reproducible methods, are important for demonstrating the clinical utility of machine learning interventions given the inherent opacity and black box nature of these models and the difficulty in deciphering the mechanistic basis for model predictions. Furthermore, machine learning model performance in the clinical setting is dependent on the training data that was used during model development and may not generalize well to patient populations that are out of the training data's distribution.

Factors such as geographic location and racial, ethnic, and sex characteristics of model training data are often overlooked; thus, RCTs that are inclusive of a range of demographic backgrounds are crucial to avoiding known biases that can be propagated and deepened based on flawed training data. Therefore, we performed a systematic review to better understand the landscape of machine learning RCTs and trial qualities that affect reproducibility, inclusivity, generalizability, and successful implementation of artificial intelligence (AI) or machine learning interventions in clinical care. We focused the review on trials that used AI or machine learning as a clinical intervention, with patients allocated randomly to either a treatment arm with a therapeutic intervention based on machine learning or a standard of care arm.

Methods

This systematic review used the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) and Synthesis Without Meta-analysis (SWiM) reporting guidelines. The protocol was registered a priori (CRD42021230810).

Search Strategy and Selection Criteria

A systematic search of the literature was conducted by a medical librarian (A.A.G.) in Cochrane Library, Google Scholar, Ovid Embase, Ovid MEDLINE, PubMed, Scopus, and Web of Science Core Collection databases to find relevant articles published from the inception of each database to October 15, 2021, and final searches were performed in all databases on this date. The search was peer reviewed by a second medical librarian using the Peer Review of Electronic Search Strategies (PRESS) guideline. Databases were searched using a combination of controlled vocabulary and free text terms for AI, clinical decision-making, and RCTs. The search was not limited by language or year. Details of the full search strategies are given in eAppendix 1 in the Supplement. The citation chaser package for R software, version 4.0.3 (R Foundation for Statistical Computing) was used to search the reference lists of included studies and to retrieve articles that cited the included studies to find additional relevant studies not retrieved by the database search.

Citations from all databases were imported into an Endnote 20 library (Clarivate Analytics), in which duplicates were removed. The deduplicated results were imported into the Covidence systematic review management program for screening and data extraction. Two independent screeners performed a title and abstract review, with a third screener to resolve disagreements. This
The phase of screening was performed by 5 of us (D.P., D.L.S., A.A.G., A.S., and B.H.K.). The full texts of the resulting articles were then independently reviewed for inclusion by 2 screeners (D.P., D.L.S., and A.S.) and a third screener (B.H.K.) to resolve disagreements. Articles were included if they were deemed by consensus to be RCTs in which AI or machine learning was used in at least 1 randomization arm in a medical setting. Search strategies are available in the eAppendix 1 in the Supplement, and reasons for exclusion are found in Figure 1.

**Statistical Analysis**

Two of us (D.P. and D.L.S.) independently extracted data and assessed the risk of bias for each study using standardized data extraction forms. The forms completed by each were compared; disagreement was resolved by review and discussion, with another one of us (B.H.K.) serving as final arbiter. Authors were not contacted for additional unpublished data. Risk of bias was assessed using the Cochrane Risk of Bias, version 2 tool for RCTs. This tool was developed to assess risk of bias in RCTs and has 5 domains, including risk of bias owing to the randomization process, deviations from the intended interventions (effect of assignment to intervention), missing outcome data, measurement of the outcome, and selection of the reported result.

To assess reproducibility and reporting transparency, we assessed article adherence to the recently published Consolidated Standards of Reporting Trials–Artificial Intelligence (CONSORT-AI) reporting guideline, which is an extension of the CONSORT guideline developed through an international multi-stakeholders group via staged consensus. Machine learning–based RCTs are recommended to routinely report the extended criteria in addition to the core CONSORT items. Two of us (D.P. and D.L.S.) independently extracted data and assessed each of the 11 CONSORT-AI extension criteria for each article. Disagreement was resolved by review and discussion, with another of us (B.H.K.) serving as final arbiter.

### Figure 1. Screening and Selection of Randomized Clinical Trials

28159 Records identified from databases and registers
25410 From databases
2749 From registers

8422 Records removed before screening (duplicates)

19737 Article titles and abstracts screened

116 Articles sought for retrieval

116 Full-text articles assessed for eligibility

75 Excluded
19 Conference abstracts
16 Wrong study design
14 Wrong intervention
11 Duplicate study data
6 No clinical decision-making
3 No AI or machine learning
2 Preliminary results only
2 Wrong setting
1 No clinical outcome
1 Not in English language

2802 Records identified from citation searches

743 Records removed before screening (duplicates or retrieved in database search)

2059 Article titles and abstracts screened

11 Articles sought for retrieval

11 Full-text articles assessed for eligibility

11 Excluded

AI indicates artificial intelligence.

* Databases and registers included Cochrane Library, Google Scholar, Ovid Embase, Ovid MEDLINE, PubMed, Scopus, and Web of Science Core Collection.
To evaluate inclusivity, we assessed reporting of sex, race, and ethnicity. We calculated the proportion of participants from underrepresented minority groups within each study using the National Institutes of Health definition of groups underrepresented in biomedical research; the definition designates American Indian or Alaska Native, Black or African American, Hispanic or Latino, and Native Hawaiian or other Pacific Islander as underrepresented minority groups. To assess other qualities pertaining to generalizability and clinical adoption, we assessed the use of clinical vs nonclinical end points, whether the trial was conducted at a single site or multiple sites, and geographic location. Other qualities assessed were the use of measures with vs without performance thresholds, the disease setting of the trial, the model training data type, and the machine learning model type. The data for all aforementioned items were independently extracted by 2 of us (D.P. and D.L.S.) for each article, with disagreement resolved by review and discussion, with another of us (B.H.K.) serving as final arbiter. All summary statistics were computed using R software, version 4.0.3.

Results

The search resulted in 28,159 records; after duplicates were removed, 19,737 remained for title and abstract screening, and 19,621 of these were excluded (Figure 1). No additional articles were located from citation chasing. Full-text review was conducted for 116 articles; of those, 75 studies were excluded because they were conference abstracts (n = 19), had the wrong study design (n = 16), performed the wrong intervention (n = 14), contained duplicate study data (n = 11), did not involve clinical decision-making (n = 6), did not use AI or machine learning (n = 3), provided preliminary results only (n = 2), were not conducted in a medical setting (n = 2), did not assess any outcomes that impacted clinical decision-making (n = 1), or were not written in the English language (n = 1) (eAppendix 1 and eAppendix 2 in the Supplement). Overall, 41 RCTs involving a median of 294 participants (range, 17–2488 participants) met inclusion criteria.

Trials were most often conducted in the US (15 [37%]) and China (13 [32%]), and 6 studies (15%) were conducted across multiple countries. In terms of qualities associated with generalizability, 20 RCTs (49%) were conducted at multiple sites, and 21 RCTs (51%) were conducted at a single site. Only 11 trials (27%) reported race and ethnicity (Figure 2); among those trials, a median of 21% (range, 0%-51%) of participants were from underrepresented minority groups.

To assess RCT transparency and reproducibility, we assessed trial adherence to CONSORT-AI standards (Figure 3). We found that no RCT met all the criteria. A total of 13 RCTs (32%) met at least 8 of 11 criteria (eAppendix 3 in the Supplement). The most common reasons for lack of guideline adherence were not assessing poor-quality or unavailable input data (38 trials [93%]), not analyzing performance errors (38 [93%]), and lack of a statement regarding code or algorithm availability (37 [90%]).

The risk of bias for the RCTs is summarized in Figure 4. Overall risk of bias was high in 7 trials (17%), bias from measurement of outcomes was the type most often observed, with at least some concern for bias in 19 trials (46%).

Regarding clinical use cases in RCTs, the most common clinical specialty represented was gastroenterology (16 [39%]), most of these RCTs involved endoscopic imaging, and most studies involving clinical use cases enrolled adult patients (36 [88%]). Four trials (10%) were performed in a...
<table>
<thead>
<tr>
<th>Source</th>
<th>Study location</th>
<th>Study aim</th>
<th>Female, %</th>
<th>Race and ethnicity (%)</th>
<th>Medical specialty</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pavel et al, 26 2020</td>
<td>Ireland, the Netherlands, Sweden, UK</td>
<td>Detect neonatal seizures</td>
<td>40</td>
<td>NR</td>
<td>Neonatology</td>
<td>Neonatal seizures</td>
</tr>
<tr>
<td>Wang et al, 27 2020</td>
<td>China</td>
<td>Detect colorectal adenomas</td>
<td>52</td>
<td>NR</td>
<td>Gastroenterology</td>
<td>Colon polyp and adenoma</td>
</tr>
<tr>
<td>Caparros-Gonzalez et al, 28 2018</td>
<td>Spain</td>
<td>Assess premature infant physiological response</td>
<td>41</td>
<td>NR</td>
<td>Neonatology</td>
<td>None</td>
</tr>
<tr>
<td>Nimri et al, 29 2020</td>
<td>Europe (multiple countries), Israel, US</td>
<td>Optimize insulin dose</td>
<td>52</td>
<td>NR</td>
<td>Endocrinology</td>
<td>Type 1 diabetes</td>
</tr>
<tr>
<td>Vennalaganti et al, 30 2018</td>
<td>US</td>
<td>Detect Barrett esophagus-associated neoplasia</td>
<td>24</td>
<td>African American (2), White (95), other (3)</td>
<td>Gastroenterology</td>
<td>Barrett esophagus-associated neoplasia</td>
</tr>
<tr>
<td>Voss et al, 31 2019</td>
<td>US</td>
<td>Improve socialization in children with autism spectrum disorder</td>
<td>11</td>
<td>Black (3), East Asian (24), Hispanic (23), Middle Eastern (1), South Asian (8), White/European American (35), unknown (17)</td>
<td>Pediatrics</td>
<td>Autism spectrum disorders</td>
</tr>
<tr>
<td>Manz et al, 32 2020</td>
<td>US</td>
<td>Increase serious illness conversations among patients with cancer</td>
<td>54</td>
<td>Non-Hispanic Black (20), non-Hispanic White (70), other (10)</td>
<td>Oncology</td>
<td>End of life</td>
</tr>
<tr>
<td>Persell et al, 33 2020</td>
<td>US</td>
<td>Improve blood pressure control in outpatients with hypertension</td>
<td>61</td>
<td>Black (35), Hispanic (8), White (52), other (9)</td>
<td>Primary care</td>
<td>Hypertension</td>
</tr>
<tr>
<td>Repici et al, 34 2020</td>
<td>Italy</td>
<td>Detect colorectal adenomas</td>
<td>51</td>
<td>NR</td>
<td>Gastroenterology</td>
<td>Colorectal neoplasia</td>
</tr>
<tr>
<td>Wijnberge et al, 35 2020</td>
<td>Netherlands</td>
<td>Detect intraoperative hypotension</td>
<td>43</td>
<td>NR</td>
<td>Cardiac surgery</td>
<td>Intraoperative hypotension</td>
</tr>
<tr>
<td>Shimabukuro et al, 36 2017</td>
<td>US</td>
<td>Predict outcomes in patients with sepsis</td>
<td>54</td>
<td>African American (11), Asian (16), Hispanic (21), White (47), other (5)</td>
<td>Intensive care</td>
<td>Sepsis</td>
</tr>
<tr>
<td>Wang et al, 37 2020</td>
<td>China</td>
<td>Detect colorectal adenomas</td>
<td>49</td>
<td>NR</td>
<td>Gastroenterology</td>
<td>Colon polyp and adenoma</td>
</tr>
<tr>
<td>Gong et al, 38 2020</td>
<td>China</td>
<td>Detect colorectal adenomas</td>
<td>51</td>
<td>NR</td>
<td>Gastroenterology</td>
<td>Colorectal adenomas</td>
</tr>
<tr>
<td>Lin et al, 39 2019</td>
<td>China</td>
<td>Diagnose childhood cataracts</td>
<td>55</td>
<td>NR</td>
<td>Ophthalmology</td>
<td>Childhood cataracts</td>
</tr>
<tr>
<td>Rabbi et al, 40 2015</td>
<td>US</td>
<td>Facilitate weight loss through automated personalized feedback for physical activity and diet</td>
<td>47</td>
<td>NR</td>
<td>Primary care</td>
<td>None</td>
</tr>
<tr>
<td>Auloge et al, 41 2020</td>
<td>France</td>
<td>Assess feasibility of AI/AR tool for vertebroplasty</td>
<td>65</td>
<td>NR</td>
<td>Orthopedics</td>
<td>Vertebral fracture</td>
</tr>
<tr>
<td>Avari et al, 42 2021</td>
<td>Spain, UK</td>
<td>Decrease hypoglycemia episodes with personalized bolus advice for people with type 1 diabetes</td>
<td>52</td>
<td>NR</td>
<td>Endocrinology</td>
<td>Type 1 diabetes</td>
</tr>
<tr>
<td>Wang et al, 43 2019</td>
<td>China</td>
<td>Detect colorectal adenomas</td>
<td>52</td>
<td>NR</td>
<td>Gastroenterology</td>
<td>Colon polyp and adenoma</td>
</tr>
<tr>
<td>Forman et al, 44 2019</td>
<td>US</td>
<td>Facilitate weight loss by predicting and preventing dietary lapses</td>
<td>85</td>
<td>Hispanic or non-White (30), non-Hispanic White (70)</td>
<td>Primary care</td>
<td>Obesity</td>
</tr>
<tr>
<td>Wu et al, 45 2019</td>
<td>China</td>
<td>Reduce rate of blind spots during EGD</td>
<td>52</td>
<td>NR</td>
<td>Gastroenterology</td>
<td>No specific disease</td>
</tr>
<tr>
<td>El Solh et al, 46 2009</td>
<td>US</td>
<td>Predict optimal CPAP using neural network to reduce titration failure</td>
<td>57</td>
<td>NR</td>
<td>Pulmonology</td>
<td>Obstructive sleep apnea</td>
</tr>
<tr>
<td>Luštrek et al, 47 2021</td>
<td>Belgium, Italy</td>
<td>Assess self-management of congestive heart failure using app and wristband</td>
<td>NR</td>
<td>NR</td>
<td>Cardiology</td>
<td>Congestive heart failure</td>
</tr>
<tr>
<td>Chen and Gao, 48 2021</td>
<td>China</td>
<td>Assess AI-based echocardiography for diagnosis of acute heart failure</td>
<td>36</td>
<td>NR</td>
<td>Cardiology</td>
<td>Acute left heart failure</td>
</tr>
<tr>
<td>Seol et al, 49 2021</td>
<td>US</td>
<td>Assess management of childhood asthma</td>
<td>43</td>
<td>White (72)</td>
<td>Pediatrics</td>
<td>Asthma</td>
</tr>
<tr>
<td>Repici et al, 50 2022</td>
<td>Italy, Switzerland</td>
<td>Investigate colon adenoma detection of nonexpert endoscopists</td>
<td>50</td>
<td>NR</td>
<td>Gastroenterology</td>
<td>Colorectal cancer screening</td>
</tr>
<tr>
<td>Kamba et al, 51 2021</td>
<td>Japan</td>
<td>Decrease colon adenoma miss rate</td>
<td>23</td>
<td>NR</td>
<td>Gastroenterology</td>
<td>Colorectal cancer screening</td>
</tr>
<tr>
<td>Liu et al, 52 2020</td>
<td>China</td>
<td>Increase polyp and adenoma detection with CADe</td>
<td>53</td>
<td>NR</td>
<td>Gastroenterology</td>
<td>Colorectal cancer screening</td>
</tr>
</tbody>
</table>
primary care setting, and all of these trials involved user-inputted data.33,40,44,57, 4 other trials (10%) were in cardiology or cardiac surgery and involved electrocardiographic, wearable device, echocardiographic, or arterial waveform data. Two trials (5%) were performed in the neonatal setting,26,28 evaluating seizures and physiological distress, and 3 studies (7%) were performed primarily among pediatric populations more broadly,31,39,49 evaluating asthma, autism spectrum disorder, and childhood cataracts. Most RCTs involved clinical outcome measures (34 [83%])26,27, 29-39,42-46,58,60-65 and outcome measures without performance thresholds (32 [78%]).28,31,36,38-40,42,44,46-58,60-66 In terms of data sources, 15 trials (37%) mostly used endoscopic imaging–based interventions,27,34,37,38,43,45,50-52,54,56,60,63,65,66 5 (12%) used patient-inputted data,33,40,44,55,57 2 (5%) used primary electronic health record data,32,61 2 (5%) used electrocardiogram data,48,59 and 2 (5%) used blood-based data (glucose and insulin levels).29,42 A total of 20 articles (49%) used deep learning neural networks.27,30,34,37,38,43,45,46,48,50-52,54,56,59,60,62,63,65,66

Discussion

This systematic review found a lack of RCTs for medical machine learning interventions and highlighted the need for additional well-designed, transparent, and inclusive RCTs for machine learning interventions to promote use in the clinic. There is growing concern that new machine learning models are being released after preliminary validation studies without follow-through on their ability to formally show superiority in a gold standard RCT.67,68 Of note, there are currently 343 US Food and Drug Administration (FDA)–approved medical AI or machine learning interventions.69
Our finding of 41 medical machine learning RCTs suggests that most FDA-approved machine learning-enabled medical devices are approved without efficacy demonstrated in an RCT. This finding is likely explained, in part, by the lower burden of evidence required for AI or machine learning algorithm clearance (often classified by the FDA as software as a medical device) compared with pharmaceutical drugs. To our knowledge, this review is the first rigorous attempt at quantifying this gap.

Prior work has shown a lack of prospective testing in this field but has not rigorously assessed the quantity of RCTs using a PROSPERO-registered method or tie-breaking arbitration or has analyzed the testing of technologies only related to imaging data. In addition, these studies did not explore study adherence to CONSORT-AI standards or assess the inclusivity of underrepresented minority groups and women in the study populations. Finally, the scope of our review compared with prior work differed; our work specifically focused on the use of clinical AI or machine learning interventions that were used as investigational arms in RCTs. We excluded RCTs that used traditional statistical models and RCTs in which AI or machine learning was included in the study protocol but was not part of the randomized intervention. In this way, we highlighted RCTs that directly compared AI or machine learning with standard of care and RCTs that were designed to demonstrate a high level of evidence for clinical utility. A comparison of the trials included in this study with prior work is available in eAppendix 4 in the Supplement.

A total of 41 randomized clinical trials were included in the analysis. Individuals from underrepresented minority groups were participants in 11 clinical trials in which information on participant race and/or ethnicity was reported. Data for 2021 are from January through October 15. The other medical specialty category includes anesthesiology, cardiac surgery, emergency medicine, general surgery, gynecology, intensive care, ophthalmology, pulmonology, and radiology.
Our initial search of 28,159 records and subsequent yield of only 41 RCTs indicates a translational gap between development and clinical impact. Most of the RCTs included in this review were conducted between January 2019 and October 2021 (36 [88%]), and 16 studies (39%) were conducted between January and October 2021, indicating that the rate of new RCTs for machine learning interventions increased over time. Clinical use cases for these technologies most often involved endoscopic imaging in gastroenterology (15 [37%]) and enrolled adult patients (36 [88%]).

Regarding trial reporting, no RCT included in this review adhered to all the machine learning-specific reporting standards (ie, the CONSORT-AI extension guidelines). Specifically, 37 trials (90%) did not share code and data along with study results, 38 (93%) did not analyze poor-quality or unavailable input data, and 38 (93%) did not assess performance errors, all of which may contribute to issues in reproducibility. These results suggest that machine learning RCT reporting quality needs improvement. The CONSORT-AI guideline helps ensure transparency and reproducibility of RCT methods, and the lack of guideline adherence observed among RCTs in this review may be another barrier to clinical adoption. Of note, the CONSORT-AI standards were published in September 2020, when most of the trials analyzed in this review were conducted.
review would have been either published or under peer review. Future work should reassess the percentage of guideline-adherent RCTs published after 2021 to assess the impact of the CONSORT-AI guideline in RCT design.

Regarding RCT inclusivity, among the trials selected based on our search criteria, we found that only 20 (49%) were conducted at more than one site. In addition, we found a lack of reporting of demographic information across studies, with only 11 RCTs (27%) reporting participant race or ethnicity. Within this subset, studies had a median of 21% of enrolled participants belonging to underrepresented minority groups, a number concordant or slightly higher than proportions reported in prior systematic reviews analyzing medical RCTs. Trials were most often conducted in the US (15 [37%]) or China (13 [32%]), and only 6 studies (15%) were conducted across multiple countries. Taken together, this lack of diversity in patient populations involved in RCTs indicates that the generalizability of their results across clinical sites is unknown—a growing concern for the federal regulation of machine learning interventions as medical devices.

Regarding risk of bias, a high risk was found in 7 trials (17%); although substantial, this proportion was lower than the proportion of high-risk studies found in a cross-sectional study of non–machine learning interventions which found that a median of 50% of studies had a high risk of bias. This difference suggests that deficiencies in design, execution, and reporting of RCTs are not more widespread than those in other trials of medical interventions.

This systematic review found a low but increasing number of RCTs of machine learning interventions in health care. This low number is in contrast to the large number of preliminary validation studies of medical machine learning interventions and the increasing number of FDA approvals in this research area; many of these technologies have reached the clinical implementation phase without a gold standard assessment of efficacy through an RCT. It is not practical to formally assess every potential iteration of a new technology through an RCT (eg, a machine learning algorithm used in a hospital system and then used for the same clinical scenario in another geographic location). In particular, when an algorithm only indirectly affects patient care (eg, risk stratification, enhanced diagnosis), local, independent validation studies may provide an adequate level of evidence to encourage early adoption, although this is an area of ongoing debate. A baseline RCT of an intervention’s efficacy would help to establish whether a new tool provides clinical utility and value. This baseline assessment could be followed by retrospective or prospective external validation studies to demonstrate how an intervention’s efficacy generalizes over time and across clinical settings.

Limitations
This study has several limitations. Of note, this analysis only selected RCTs that assessed a machine learning intervention directly impacting clinical decision-making. Additional work could be done to quantify the use of machine learning in alternative settings (eg, improving clinician-facing tools for workflow efficiency or assessing patient stratification, including biomarker discovery and validation efforts). Future work should aim to incorporate these wider definitions of a clinical tool for assessing the impact of machine learning across diverse steps in the clinical care pipeline. Nonetheless, we hypothesize that such literature contains a similar abundance of preliminary results and a dearth of RCTs assessing the relevance of machine learning in a controlled, clinical setting. An additional limitation is that this area of research is rapidly evolving, and our work is only current through October 2021. Future systematic reviews of machine learning interventions in health care will require more frequent updating and study of available results.

Conclusions
This systematic review found a low but increasing number of RCTs for machine learning interventions in health care. These results highlight the need for medical machine learning RCTs to promote safe
and effective clinical implementation. The findings also highlight areas of concern in terms of the quality of medical machine learning RCTs and opportunities to improve reporting transparency and inclusivity, which should be considered in the design and publication of future trials.


**SUPPLEMENT.**

- **eAppendix 1.** Search Strategies
- **eAppendix 2.** Excluded Studies
- **eAppendix 3.** Included Trial Characteristics
- **eAppendix 4.** Comparison of Study Articles With Prior Literature
- **eReferences**