Artificial intelligence (AI)-based applications have the potential to improve the quality and efficiency of patient care in dermatology. Unique challenges in the development and validation of these technologies may limit their generalizability and real-world applicability. Before the widespread adoption of AI interventions, randomized trials should be conducted to evaluate their efficacy, safety, and cost effectiveness in clinical settings. The recent Standard Protocol Items: Recommendations for Interventional Trials—AI extension and Consolidated Standards of Reporting Trials—AI extension guidelines provide recommendations for reporting the methods and results of trials involving AI interventions. High-quality trials will provide gold standard evidence to support the adoption of AI for the benefit of patient care.

Is artificial intelligence (AI) ready to make a significant contribution to dermatology practice? The application of imaging-based disease recognition is well-suited to dermatology, and impressive examples of AI-driven skin lesion classification have been reported (Esteva et al., 2017; Tschandl et al., 2017). In recent years, several studies have shown that AI algorithms can match or exceed the accuracy of experienced clinicians in correctly classifying controlled clinical or dermoscopic images of different skin lesions in isolation (Esteva et al., 2017; Haenssle et al., 2018; Tschandl et al., 2019). More recently, an algorithm was published that could discriminate between 26 different skin diseases using clinical images and clinical metadata such as past medical history (Liu et al., 2020b). Other applications of AI in dermatology range from providing population-based estimates for the frequencies of different melanocytic lesions through analysis of electronic medical records, generating new datasets on which to train classifiers for other skin diseases, and the identification of malignant lesions using smartphone-based dermoscopic images (Han et al., 2018; Lott et al., 2018; Phillips et al., 2019).

As the technologies become more advanced, a number of AI-enabled imaging devices and smartphone applications have been approved by regulators for clinical decision support, including MelaFind, SkinVision, DERM AI, and MoleAnalyzer Pro (Chuchu et al., 2018; Freeman et al., 2020; Young et al., 2020). Although none of these devices has current United States Food and Drug Administration approval (Benjamens et al., 2020), Breakthrough Device Designations have been granted to 3Derm (Boston, MA) and Skin Analytics (London, United Kingdom) (Freeman et al., 2020). With rigorous validation and consideration of practical issues such as variability in technique or technology, these AI applications have the potential to support clinicians and patients to augment the delivery of dermatology care across triage, diagnosis, and monitoring (Liu et al., 2019; Nagendran et al., 2020).

However, AI interventions have unique characteristics that make them vulnerable to distinct biases relating to the training data, which in turn may...
lead to a failure to translate into improved patient outcomes in clinical practice. Broadly speaking, AI algorithms use the advanced computational analysis of large quantities of training data to define a mathematical function capable of mapping further, similar examples to prespecified outputs. These training data might consist of passages of text or, as is frequently the case in dermatologic applications of AI, images labeled with a corresponding diagnosis. Because these training data constitute the reference frame for which the AI intervention will make decisions in the real world, it is imperative that the training data reflect the setting in which the algorithm will be deployed. As a minimum, studies presenting diagnostic tools utilizing AI should present values for the sensitivity, specificity, and positive and negative predictive values achieved by the device being put forward.

Large datasets collected for the development of AI algorithms may incompletely report clinical metadata such as the indication for an investigation. This lack of information makes it difficult to extrapolate the characteristics of the patient population to determine population characteristics and spectrum of disease. Furthermore, there is a tendency for developers of AI interventions to dichotomize diagnostic tools, often at an unclear threshold, for example, by developing tools that classify skin lesions into melanoma or not melanoma. This does not accurately reflect the full spectrum of disease in the real world and may result in an unacceptably high proportion of over or underdiagnosis.

Concerns regarding clinical generalizability have been raised because most published AI algorithms have been developed and tested using datasets and experimental designs that do not recapitulate actual clinical workflow (Nagendran et al., 2020). Poor generalizability can arise from a lack of skin color data in dermatology machine learning studies and a lack of datasets that come only from a single clinical site or region (Adamson and Smith, 2018; Kinyanjui et al., 2019). A recent study observed that machine learning datasets in the United States are predominantly derived from California, New York, and Massachusetts and may not be representative of the broader population (Kaushal et al., 2020). Consequently, the performance of AI algorithms may be less predictable when applied to patients outside of these groups, creating the potential for harm owing to inaccurate predictions (Adamson and Smith, 2018; Lashbrook, 2018).

There are also issues concerning the accuracy with which images in the dataset are labeled with clinical diagnoses because most datasets have diagnosis labels from dermatologists’ assessment of an image, whereas the gold standard is often clinical-pathological correlation (Esteva et al., 2019). Ideally, all lesions included in the training dataset should have a pathologically confirmed diagnosis and not just those that are most suspicious of malignant disease. Furthermore, there is a worry that the pressure of AI hype may lead to overinflation of claims around the potential for AI to improve patient care without robust evidence supporting these outcomes. This evidence could feasibly influence regulators when making licensing decisions, and the awareness of impressive headline results may cause unrealistic expectations in the clinical and patient communities of the readiness of the technology (Liu et al., 2019).

Given these potential biases, it is essential that AI algorithms are assessed by randomized controlled trials (RCTs)—the gold standard for assessing the effectiveness of health interventions—including new drugs and medical devices. Many AI studies to date describe the superior diagnostic performance of AI systems to that of humans but focus exclusively on algorithm development and its early performance in silico. Few studies evaluate multiple algorithms to establish which is the most reliable when placed in the hands of users in a clinical setting. To date, we are not aware of any RCTs in dermatology comparing the effectiveness of AI systems in real-world scenarios, where the AI system is integrated into a clinical pathway and patient outcomes are evaluated, with that of the current gold standard.

To better define the role of diagnostic AI interventions in clinical practice, we call for increased use of RCTs that evaluate their safety and efficacy. For these RCTs to have the greatest impact, it is critical that they be fully reported, so that their conduct and findings can be understood and interpreted. The Standard Protocol Items: Recommendations for Interventional Trials (SPIRIT) and the Consolidated Standards of Reporting Trials (CONSORT) guidelines set out minimum requirements for reporting new trial protocols and the results of trials, respectively (Chan et al., 2013; Schulz et al., 2011). These standards provide structured guidance for reporting on the design, conduct, and interpretation of the trial across key domains, including many areas where the completeness and quality of reporting will affect the reader’s ability to understand and interpret the study and the risk of bias. By encouraging this full and transparent reporting, the standards can help to promote high-quality studies and study reports. Ultimately, these reporting standards improve the quality of the evidence base from which we make scientific, clinical, and policy decisions regarding the use of therapeutic interventions on patients.

Because there are technology-specific considerations to implementing and evaluating AI systems that warrant evaluation, these unique factors led to the development of the SPIRIT-AI and CONSORT-AI guidelines for RCTs evaluating AI systems (Cruz Rivera et al., 2020; Liu et al., 2020a). The new guidelines, developed by the SPIRIT-AI and CONSORT-AI Working Group, add 15 and 14 new items to the existing SPIRIT 2013 and CONSORT 2010 checklist items, respectively. These new recommendations were made through a process of international multistakeholder consensus, including 103 stakeholders with expertise in clinical trials and machine learning.

The guidelines particularly draw attention to AI-specific concerns. The importance of carefully describing input data for AI interventions has been discussed. The new guidelines call for the inclusion and exclusion criteria to be stipulated at the level of the input

---

data as well as at the patient level to distinguish the selection criteria related to the characteristics of the patient from those chosen for ensuring input data quality. Furthermore, the guidelines call for fully describing input data acquisition methods and accounting for missing or unsuitable input data. The SPIRIT-AI and CONSORT-AI guidelines also stress the importance of clearly describing the role that the AI intervention will play within the context of the trial and clinical pathway. This includes describing any human–AI interactions and the level of training or expertise humans will need to appropriately use the AI device. Similarly, it is important to describe the output of the AI device and how this will contribute to decision making, again including the relevant expertise that users will need in the trial.

Finally, the guidelines advocate that specific technical aspects unique to AI interventions are accurately reported. Examples include the actions taken in response to performance errors made by the AI device, the version of the AI device, and whether the AI device or its code can be accessed by users or researchers. Although created with RCTs in mind, the checklist items may serve as useful points for consideration for other types of study design involving AI research.

The SPIRIT-AI and CONSORT-AI guidelines set the foundation on which the next stages of AI research will stand: informing RCT design for AI systems so that their impact on tangible patient outcomes can be robustly assessed. The time has come to improve the evidence base for these promising technologies in dermatology.

ORCID
Matthew Taylor: http://orcid.org/0000-0001-9480-4832
Xiaoxuan Liu: http://orcid.org/0000-0002-1286-0038
Alastair Denniston: http://orcid.org/0000-0001-7849-0087
Andres Esteva: http://orcid.org/0000-0001-1937-9682
Justin Ko: http://orcid.org/0000-0003-4232-5906
Roxana Daneshjou: http://orcid.org/0000-0001-7988-9356
Ahn-Wen Chan: http://orcid.org/0000-0002-4498-3382

AUTHOR CONTRIBUTIONS
Conceptualization: XL, AD, AWC; Writing — Original Draft Preparation: MT; Writing — Review and Editing: MT, XL, AD, AE, JK, RD, AWC

CONFLICT OF INTEREST
AE is an employee of Salesforce.com and owns CRM stock. JK serves as a clinical advisor to Skin Analytics, Leo Innovation Labs, and Enspecra Health and chairs the American Academy of Dermatology Task Force on Augmented Intelligence. RD is on the advisory board for VisualDx. The remaining authors state no conflict of interest.

REFERENCES