Development of High-Quality Artificial Intelligence in Dermatology: Guidelines, Pitfalls, and Potential

Carrie Kovarik[1]

JID Innovations (2022) **2,** 100157; doi:10.1016/j.xjidi.2022.100157

Artificial intelligence (AI) or the development of computer systems and algorithms to perform tasks that normally require human intelligence is an active area of investigation, with speculation over future innovations as well as integration of current technology seamlessly into everyday life with applications such as face identification on smartphones, recommendations of shows on streaming applications, use of voice to text, and Global Positioning System optimization by ridesharing services. The continuous development of technology and tools within AI also offers great potential for clinical applications within medicine. To be most successful, human intelligence and AI must work together synergistically. The concept of augmented intelligence (AuI) focuses on the role of AI in assisting clinicians to enhance human intelligence rather than replace it. When developed appropriately, AuI has the potential to positively transform the practice of physicians.

AI policies and positions

In the past, medical technology often has been developed without significant input from physicians or the end user, such as in the case of the electronic medical record (Dietsche, 2018). The American Medical Association (AMA) proactively provided policies to advance the role of AuI in enhancing patient care, with the goals of increasing the value for physicians and identifying opportunities to integrate practicing physicians' perspectives into the development and implementation of healthcare AI (American Medical Association, 2018). The AMA also specifically highlighted the need to promote the development of high-quality, clinically validated AI that is transparent, focuses on user-centered design, conforms to leading standards for reproducibility, addresses bias, and avoids introducing or exacerbating healthcare disparities. In 2019, the American Academy of Dermatology published a Position Statement on AuI, reinforcing many of these issues, including the need for technology to be collaboratively developed, while at the same time minimizing the risk of potentially disruptive effects and unintended consequences (Kovarik et al., 2019).

Guidelines and pitfalls when developing high-quality AI/AuI

In clinical dermatology, there are many potential applications of AI/AuI. Some examples include skin lesion diagnosis, monitoring/tracking of skin lesions, diagnostic decision support, histopathology diagnostic support, improving workflows, and uncovering associations between clinical and molecular data (Young et al., 2020). The development and successful implementation of AI/AuI requires high-quality research involving multiple steps, attention to detail, persistence, and a dedication to high-quality clinical care. A deficiency in any critical step can lead to false predictions by the algorithm and poor outcomes for the patients on whom it is used. In the case of any rapidly growing technology, policies, regulations, and ethics may not keep up with development and implementation. Although AI holds promise for dermatology, considering the implications of poorly developed AI and encouraging the creation of quality applications are critical.

Although the development of a useful AI algorithm may be complex, a practical application will be difficult to implement successfully in a clinical setting without basic criteria being met. First, there has to be a clinically relevant question that the AI can answer. The clinical question will guide the type of images required, and it is essential to have ethical access to sufficient, diverse, quality, and appropriately labeled data sets, which are often images or clinical photos in dermatology applications. The algorithm should be trained on an initial set of images, then fine tuned, and validated on a separate set of images. An independent test dataset should be used to evaluate the accuracy of the algorithm. The algorithm should then be validated in a relevant clinical setting, with continued monitoring and evaluation (Kovarik et al, 2019).

This process can be prone to many pitfalls. Image datasets often lack diverse skin types (Guo et al., 2022a); may be incomplete, preventing the algorithm from diagnosing rare diseases or rare presentations of common diseases (Tschandl, 2021); contain images of poor quality; or have improper labels. The method used for image labeling and the classification of the data on which the algorithms are trained (ground truth) should be clearly described and justified. Histopathological diagnosis should be considered the gold standard for labeling neoplasms (Daneshjou et al., 2022b). Many algorithms receive insufficient dataset

[1]*Department of Dermatology, University of Pennsylvania, Philadelphia, Pennsylvania, USA*

Correspondence: Carrie Kovarik, Department of Dermatology, University of Pennsylvania, 2 Maloney Building, 3600 Spruce Street, Philadelphia, Pennsylvania 19104, USA. E-mail: carrie.kovarik@pennmedicine.upenn.edu

Cite this article as: JID Innovations 2022;2:100157

training through inadequate sample sizes, lack of external test sets, homogeneous datasets, and lack of clinical validation. Some algorithms are not developed appropriately for a particular end user or clinical scenario or without clinical collaborators.

Optimally, both images and algorithms would be publically available to allow for assessment of dataset characteristics, potential bias, and robustness (Daneshjou et al., 2021). Dermatology image sets are not always available owing to patient privacy or consent issues; however, patient demographic data, including skin tone, as well as other metrics related to image quality (cameras used, image processing steps) should also be made available because they may affect AI performance. In addition, as noted by Daneshjou et al. (2021), if images were obtained by public sources, they should be clearly delineated and labeled for reference. Making an AI algorithm publically available also presents challenges because there is intellectual property involved, and the complexities around interpreting the code can be complicated. Online platforms have now been created that will host AI models and allow users to interactively test them with their own images or data (Abid et al., 2020).

To develop high-quality clinically relevant AI/AuI, standards need to be met. Earlier this year, Daneshjou et al. (2022b) published the "Checklist for Evaluation of Image-Based Artificial Intelligence Reports in Dermatology." In this consensus statement, the authors highlight key recommendations for developers of AI and reviewers of AI-based reports, focusing on data, technique, technical assessment, and application. Rigorous clinical trials are much needed to evaluate AI algorithms to understand whether they translate to the clinical settings in which they will be used (Daneshjou, 2022a; Han et al., 2022). In addition to evaluating algorithms that aid in diagnostic support, researchers are trying to study whether AI is able to delineate body surface area and localize lesions of several skin diseases, including vitiligo, psoriasis, and atopic dermatitis (Guo et al., 2022b; Han et al., 2022; Hurault et al., 2022; Medela et al., 2022; Schaap et al., 2022).

Using AI for lesion segmentation and severity scoring

Tracking atopic dermatitis over time using a quantitative measurement is important to follow patient progress and assess therapeutic efficacy. This can be a time-consuming task with variable inter- and intra-rater reliability. Both the Eczema Area and Severity Index (EASI) and the Severity Scoring of Atopic Dermatitis Index (SCORAD) have been developed to assess the clinical signs of atopic dermatitis, and they have been validated in the in-person setting (Schmitt et al., 2013). Given the rise of teledermatology, investigators have compared remote atopic dermatitis severity assessments on the basis of photographs with in-person assessments (Ali et al., 2021; Hughes et al., 2021). Results showed strong correlations between assessments performed in person and those done using photographs; however, photographs were taken by study coordinators or physicians in a controlled environment. In addition, data on skin tone were not available for study participants. One study on atopic dermatitis showed that in patients with highly pigmented skin evaluated using photographs, all EASI and SCORAD scores had poor inter-rater reliability regardless of disease severity (Zhao et al., 2015).

Recently, *JID Innovations* published a report by Medela et al. (2022) describing an AI algorithm that was created to automate the SCORAD assessment using photographs. The question of inter- and intra-rater reliability using these assessments makes the question of ground truth and data labeling difficult. High-quality AI begins with accurately labeled training datasets that are derived using best practices and on the basis of evidence. If the AI algorithm is trained using a component of inaccurate eczema severity data, the outcomes will be erroneous. Hurault et al. (2022), also recently published in *JID Innovations*, examined whether high-quality eczema segmentation data can be obtained from dermatologists using images reliably. They found that inter-rater reliability of eczema segmentation varied from image to image, with a poor agreement between the raters on average. These results reinforce the difficulty of reliably and consistently detecting atopic dermatitis from photos. Hurault et al. (2022) offered suggestions for improving poor inter-rater reliability in segmentation data for machine learning models, including letting the algorithm identify eczema regions by itself, using algorithms that can be trained on noisy segmentation labels, improving the training of the raters, and averaging the segmentation from multiple raters. Alternatively, photo assessments that are performed in person may assist in reliability, given that this is the setting in which the tools were validated, including in patients with highly pigmented skin.

The study by Medela et al. (2022) utilized three image datasets, two comprised patients with Fitzpatrick skin types I–III and one from patients with Fitzpatrick skin types IV–VI. Images were collected from online public sources, and demographic data were only available for one dataset. Three experts annotated each image without any context except the image alone. The ground truth labels for lesion segmentation and visual sign classification were obtained by averaging the results of the three annotators. As mentioned by Hurault et al. (2022), this is one method that may reduce inter-rater reliability. Although the results by Medela et al. (2022) show potential, with lesion segmentation annotation generally consistent across datasets, some visual signs such as edema and dryness were difficult to assess in photos, particularly in patients with darker skin tones. The original Consensus Paper on the SCORAD (Stalder et al., 1993) mentions the difficulty in accurately assessing edema from clinical photos as well as the need to assess dryness with palpation. In addition, the majority of images in the study by Medela et al. (2022) were of mild atopic dermatitis, and the algorithm will need to recognize all severities of the disease as well as be able to distinguish what is not atopic dermatitis within the field of view.

Although quantitative measurement of atopic skin disease seems to be a task in which AI could assist, the technology is only in the preliminary stages. The challenge of identifying the actual area of involvement of atopic dermatitis in the skin is a product of poor inter- and intra-rater reliability in disease labeling

as well as the lack of a definitive diagnostic procedure for mapping out the disease. Given the suboptimal agreement by experts, consideration could be given to an unsupervised learning approach or one that trains the algorithm to diagnose atopic dermatitis to create its own segmentation data. Further advancement of AI algorithms will also require an appropriate representation of all skin tones, patient demographics, severity of atopic dermatitis, and complications of disease during the training phase to move this process forward.

AI/AuI innovation in dermatology

There are significant opportunities for future innovations in AI/AuI to improve the practices for physicians and outcomes for patients, including streamlining time-consuming tasks and providing diagnostic decision support. AI also has the potential to assist with tasks where there is limited or poor inter-rater reliability; however, the algorithm is only as good as the data labels from which it is trained. (Schlessinger et al., 2019). All clinical AI applications must ensure the highest data quality and validation standards. Development of high-quality AI is time consuming and costly; however, in the end, there is great potential for improving patient outcomes and satisfaction, maintaining patient safety, reducing cost, and transforming physician practice.

ORCID

Carrie Kovarik: http://orcid.org/0000-0002-3258-3605

CONFLICTS OF INTEREST

The author states no conflict of interest.

REFERENCES

Abid A, Abdalla A, Abid A, Khan D, Alfozan A, Zou J. An online platform for interactive feedback in biomedical machine learning. Nat Mach Intell 2020;2:86–8.

Ali Z, Joergensen KM, Andersen AD, Chiriac A, Bjerre-Christensen T, Manole I, et al. Remote rating of atopic dermatitis severity using photo-based assessments: proof-of-concept and reliability evaluation. JMIR Form Res 2021;5:e24766.

American Medical Association. Artificial intelligence in medicine. 5 July 2022, https://www.ama-assn.org/amaone/augmented-intelligence-ai; 2018.

Daneshjou R. Toward augmented intelligence: the first prospective, randomized clinical trial assessing clinician and artificial intelligence collaboration in dermatology [e-pub ahead of print] J Invest Dermatol 2022a;142:2301–2.

Daneshjou R, Barata C, Betz-Stablein B, Celebi ME, Codella N, Combalia M, et al. Checklist for evaluation of image-based artificial intelligence reports in dermatology: CLEAR Derm consensus guidelines from the international skin imaging collaboration artificial intelligence working group. JAMA Dermatol 2022b;158:90–6.

Daneshjou R, Smith MP, Sun MD, Rotemberg V, Zou J. Lack of transparency and potential bias in artificial intelligence data sets and algorithms: a scoping review. JAMA Dermatol 2021;157:1362–9.

Dietsche E. Why physician input is essential to EMR optimization [internet]. 5 July 2022, https://medcitynews.com/2018/10/physician-input-emr-optimization/; 2018.

Guo L, Yang Y, Ding H, Zheng H, Yang H, Xie J, et al. A deep learning-based hybrid artificial intelligence model for the detection and severity assessment of vitiligo lesions. Ann Transl Med 2022b;10:590.

Guo LN, Lee MS, Kassamali B, Mita C, Nambudiri VE. Bias in, bias out: underreporting and underrepresentation of diverse skin types in machine learning research for skin cancer detection-A scoping review. J Am Acad Dermatol 2022a;87:157–9.

Han SS, Kim YJ, Moon IJ, Jung JM, Lee MY, Lee WJ, et al. Evaluation of artificial intelligence-assisted diagnosis of skin neoplasms. a single-center, paralleled, unmasked, randomized controlled trial. J Invest Dermatol 2022;142:2353–2362.e2.

Hughes ME, Aralis H, Bruhn KW, Cotliar J, Craft N, DeLuca IJ, et al. A reliability study using Network-Oriented Research Assistant to evaluate the use of digital photographs in the assessment of atopic dermatitis. J Am Acad Dermatol 2021;85:725–6.

Hurault G, Pan K, Mokhtari R, Olabi B, Earp E, Steele L, et al. Detecting eczema areas in digital images: an impossible task? JID Innov 2022;2:100133.

Kovarik C, Lee I, Ko J. Ad Hoc Task Force on Augmented Intelligence. Commentary: position statement on augmented intelligence (AuI). J Am Acad Dermatol 2019;81:998–1000.

Medela A, Mac Carthy T, Aguilar Robles SA, Chiesa-Estomba CM, Grimalt R. Automatic SCOring of atopic dermatitis using deep learning: a pilot study. JID Innov 2022;2:100107.

Schaap MJ, Cardozo NJ, Patel A, de Jong EMGJ, van Ginneken B, Seyger MMB. Image-based automated Psoriasis Area Severity Index scoring by Convolutional Neural Networks. J Eur Acad Dermatol Venereol 2022;36:68–75.

Schlessinger DI, Chhor G, Gevaert O, Swetter SM, Ko J, Novoa RA. Artificial intelligence and dermatology: opportunities, challenges, and future directions. Semin Cutan Med Surg 2019;38:E31–7.

Schmitt J, Langan S, Deckert S, Svensson A, von Kobyletzki L, Thomas K, et al. Assessment of clinical signs of atopic dermatitis: a systematic review and recommendation. J Allergy Clin Immunol 2013;132:1337–47.

Stalder JF, Taïeb A, Atherton DJ, Bieber P, Bonifazi E, Broberg A, et al. Severity scoring of atopic dermatitis: the SCORAD index. Consensus Report of the European Task Force on Atopic Dermatitis. Dermatology 1993;186:23–31.

Tschandl P. Risk of bias and error from data sets used for dermatologic artificial intelligence. JAMA Dermatol 2021;157:1271–3.

Young AT, Xiong M, Pfau J, Keiser MJ, Wei ML. Artificial intelligence in dermatology: A primer. J Invest Dermatol 2020;140:1504–12.

Zhao CY, Wijayanti A, Doria MC, Harris AG, Jain SV, Legaspi KN, et al. The reliability and validity of outcome measures for atopic dermatitis in patients with pigmented skin: A grey area. Int J Womens Dermatol 2015;1:150–4.